Collaborative Integration, Publishing and Analysis of Distributed Scholarly Metadata

Dissertation zur
Erlangung des Doktorgrades (Dr. rer. nat.) der
Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Sahar Vahdati
aus dem Tabriz, Iran

Bonn
2019

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Sören Auer
2. Gutachter:	Prof. Dr. Rainer Manthey
Tag der Promotion:	17.01.2019
Erscheinungsjahr:	2019

Abstract

Research is becoming increasingly digital, interdisciplinary, and data-driven and affects different environments in addition to academia, such as industry, and government. Research output representation, publication, mining, analysis, and visualization are taken to a new level, driven by the increased use of Web standards and digital scholarly communication initiatives. The number of scientific publications produced by new players and the increasing digital availability of scholarly artifacts, and associated metadata are other drivers of the substantial growth in scholarly communication. The heterogeneity of scholarly artifacts and their metadata spread over different Web data sources poses a major challenge for researchers with regard to search, retrieval and exploration. For example, it has become difficult to keep track of relevant scientific results, to stay up-to-date with new scientific events and running projects, as well as to find potential future collaborators. Thus, assisting researchers with a broader integration, management, and analysis of scholarly metadata can lead to new opportunities in research and to new ways of conducting research. The data integration problem has been extensively addressed by communities in the Database, Artificial Intelligence and Semantic Web fields. However, a share of the interoperability issues are domain specific and new challenges with regard to schema, structure, or domain, arise in the context of scholarly metadata integration. Thus, a method is needed to support scientific communities to integrate and manage heterogeneous scholarly metadata in order to derive insightful analysis (e.g., quality assessment of scholarly artifacts).

This thesis tackles the problem of scholarly metadata integration and develops a life cycle methodology to facilitate the integrated use of different methods, analysis techniques, and tools for improving scholarly communication. Some key steps of the metadata life cycle are implemented using a collaborative platform, which allows to keep the research communities in the loop. In particular, the use of collaborative methods is beneficial for the acquisition, integration, curation and utilization of scholarly metadata. We conducted empirical evaluations to assess the effectiveness and efficiency of the proposed approach. Our metadata transformation from legacy resources achieves reasonable performance and results in better metadata maintainability. The interlinking of metadata enhances the coherence of scholarly information spaces both qualitatively and quantitatively. Our metadata analysis techniques provide a precise quality assessment of scholarly artifacts, taking into account the perspectives of multiple stakeholders, while maintaining compatibility with existing ranking systems. These empirical evaluations and the concrete applications with a particular focus on collaborative aspects demonstrate the benefits of integrating distributed scholarly metadata.

Kurzfassung

Die Forschung wird zunehmend digital, interdisziplinär und datengetrieben und beeinflusst neben der akademischen Welt auch unterschiedliche Umgebungen wie Industrie und Verwaltung. Die Drastellung, Veröffentlichung, Gewinnung, Analyse und Visualisierung von Forschungsergebnissen werden auf eine neue Ebene gehoben, angetrieben durch den verstärkten Einsatz von Webstandards und digitalen Initiativen zur wissenschaftlichen Kommunikation. Die Anzahl der wissenschaftlichen Publikationen neuer Akteure und die zunehmende digitale Verfügbarkeit wissenschaftlicher Artefakte und der damit verbundenen Metadaten sind weitere treibende Kräfte für das starke Anwachsen der wissenschaftlichen Kommunikation. Die Heterogenität wissenschaftlicher Artefakte und ihrer Metadaten, die über verschiedene Webdatenquellen verteilt sind, stellt für Forscher eine große Herausforderung in Bezug auf Suche, Ausfinden und Erkunden der Metadaten dar. So ist es beispielsweise schwierig geworden, den Überblick über relevante wissenschaftliche Ergebnisse zu behalten, über neue wissenschaftliche Veranstaltungen und laufende Projekte auf dem Laufenden zu bleiben und potenzielle zukünftige Mitarbeiter zu finden. Die Unterstützung von Forschern bei der breiteren Integration, Verwaltung und Analyse wissenschaftlicher Metadaten kann daher zu neuen Möglichkeiten und Formen der Forschung führen. Das Problem der Datenintegration wurde in den Bereichen Datenbanken, Künstliche Intelligenz und Semantic Web ausführlich behandelt. Ein Teil der Interoperabilitätsprobleme ist jedoch domänenspezifisch und neue Herausforderungen in Bezug auf Schema, Struktur oder Domäne ergeben sich im Rahmen der wissenschaftlichen Metadatenintegration. Daher ist eine Methode erforderlich, um Wissenschaftsgruppen bei der Integration und Verwaltung heterogener wissenschaftlicher Metadaten zu unterstützen, um aussagekräftige Analysen (z.B. Qualitätsbewertungen wissenschaftlicher Artefakte) abzuleiten.

Diese Arbeit beschäftigt sich mit dem Problem der Integration von wissenschaftlichen Metadaten und entwickelt eine “Lebenszyklusmethode”, um den integrierten Einsatz verschiedener Methoden, Analysetechniken und Werkzeuge zur Verbesserung der wissenschaftlichen Kommunikation zu erleichtern. Einige wichtige Schritte des Metadaten-Lebenszyklus werden über eine kollaborative Plattform umgesetzt, die es ermöglicht, die Forschungsgemeinschaften auf dem Laufenden zu halten. Insbesondere der Einsatz kollaborativer Methoden ist für den Erwerb, die Integration, die Kurierung und die Nutzung wissenschaftlicher Metadaten von Vorteil. Wir haben empirische Evaluationen durchgeführt, um die Effektivität und Effizienz des vorgeschlagenen Ansatzes zu beurteilen. Unsere Metadatentransformation aus Legacy-Ressourcen erreicht eine angemessene Leistung und führt zu einer besseren Wartbarkeit der Metadaten. Die Verknüpfung von Metadaten erhöht die Kohärenz der wissenschaftlichen Informationsräume qualitativ und quantitativ. Unsere Metadatenanalyseverfahren ermöglichen eine präzise Qualitätsbewertung wissenschaftlicher Artefakte unter Berücksichtigung der Perspektiven mehrerer Interessengruppen bei gleichzeitiger Kompatibilität mit bestehenden Rankingsystemen. Diese empirischen Auswertungen und die konkreten Anwendungen mit besonderem Fokus auf kollaborative Aspekte zeigen die Vorteile der Integration von verteilten wissenschaftlichen Metadaten.

Contents

CHAPTER 1

Introduction

Initially, the Web was proposed [22] as an infrastructure interconnecting scientific documents at CERN, the largest physics laboratories[1]. The aim was to assist researchers in browsing through scientific information such as scholarly concepts, documents, project reports, also retrieving citation information between documents. The disconnected, heterogeneous and inflexible structure of the data caused the need for such a system. In addition, a local keyword search was the only available information retrieval mechanism, which was limited to a smaller community of the users being aware of such predefined terms. The identified problems in this local environment have shown a miniature model of the rest of the world. Thus, the proposed solution had to be globally applicable. Therefore, Tim Berners-Lee's proposal, the "World Wide Web" with a global vision, on developing a network of documents using a Hypertext Markup Language (HTML)[2] made through a successful development. In later years, the so called "Web of Documents" merged with the Internet in public use with primary focus on human consumption of the published information. The ubiquitous availability of computers and their connection via networks, and the Internet gave rise to the Web as a global, distributed information system. It sparked a global wave of creativity, collaboration and innovation and became the most quickly adopted communication platform. The World Wide Web has became the main publishing space of information for almost every real world domain. Enormous amounts of content have been made available by diversity of individuals, stakeholders and organizations through online repositories, web pages, digital libraries etc.

As the nature and the scale of the data being created or plugged into the Web changed, the classic paradigm of data management and integration approaches became in need of new proposals. The big giants of the Web such as Google and Microsoft reported about the characteristics of the vast amount of data and the deep web sources and their corresponding problems [173]. Primarily, the data integration and management approaches have been developed to support information systems with a reasonable size and unified schema of the underlying data [305]. The diversity of data schema on the Web of Documents has also changed the assumption of having structured data sources [96]. It was not possible to see the Web as the classical databases with elements that can be organized, stored, and indexed in a certain manner [145]. In addition, diverse and independent data providers cause the data quality and consistency issues on the Web. In order to get reasonable exploration results over the Web, search engines needed to understand semantics and interrelationship of different and disparate datasets. With the appearance of social networks, electronic commerce, audio and video portals, the Web have become increasingly interactive, Web 2.0 (user-generated content) [201]. Thus added up to the heterogeneity and diversity of the published data.

[1] https://home.cern/

[2] https://www.w3.org/html/

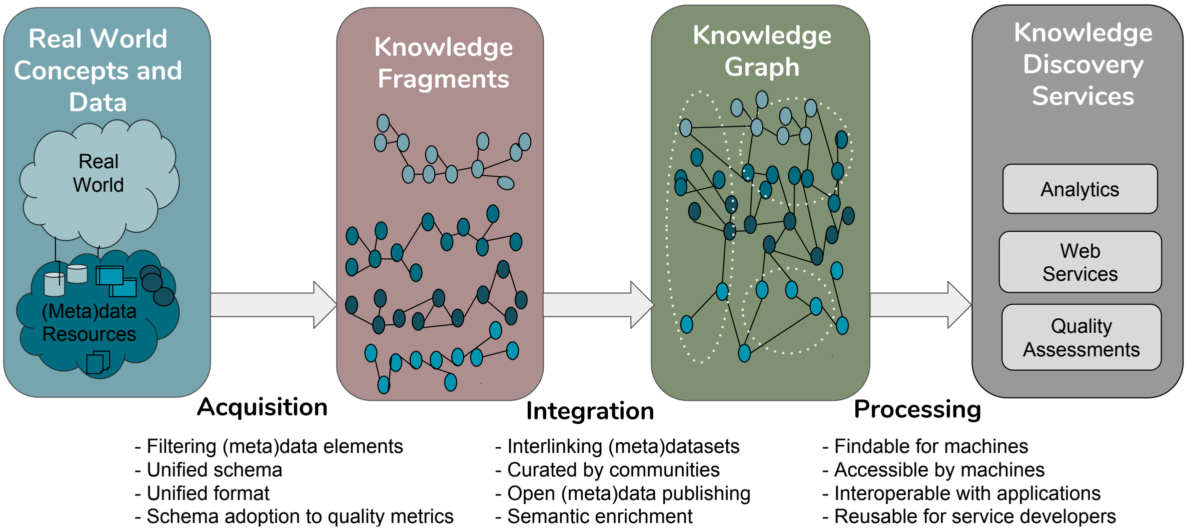


Figure 1.1: **A Pipeline for Metadata Integration**. Heterogeneous (meta)datasets are integrated for creating a knowledge graph. Curation methods are used to provide high quality assessment and representations, metadata management, web services and applications and analytics.

In order to boost the search engines on the Web, semantic representation of the concepts and relationships of the data and metadata became a mandatory requirement. The “Web of Documents” had to change to the “Web of Data” where it represents information in a machine-readable way and interweaves abstract concepts as well as descriptions of real-world entities in a giant graph-like structure. Considering the information already represented in various web pages as uniform structured data, the term Linked data refers to a set of best practices for publishing such information on the Web. Automatic extraction, transformation and integration of information following the linked data principles by using Uniform Resource Identifiers (URIs) allows to identify separate objects on web pages or databases. Linking URIs enables exploration of other data sources and retrieve of associated data rather than querying an individual database of information. The Web of Data employs Linked Data standards, i.e., RDF data model (Resource Description Framework) as a *lingua franca* for knowledge representation, SPARQL as a query language for RDF, and the Web Ontology Language (OWL) as a logical formalism to encode ontologies. Ontologies are used to create a basic, logical, machine-readable description of concepts and their relations in a chosen domain of discourse. The Web is presently evolving into a semantic “Web of Data” [23] which means instead of linking documents of web pages, the intention moves towards linking individual objects. Data elements contained in a document are identified and made universally accessible and useful. Such level of connected Big Data [63] changed the concepts from information spaces to *knowledge graphs* [248].

Figure 1.1 shows the the pipeline of metadata integration starting from real world objects as metadata resources, and extracting knowledge fragments of specific domains, creating knowledge graphs, finally exploitation knowledge. The conceptualization of the real world and representation of the Knowledge graphs are means of storing and using data, which allows people and machines to better tap into the connections in datasets. Knowledge graphs enable not only the description of the meaning of data, but the integration of data from heterogeneous sources and the discovery of previously unknown patterns. Knowledge semantically represented in knowledge graphs can be exploited to solve a broad range of problems in the respective domain. This opens up new technical possibilities as it allows data from across the Web to become comprehensible for machines first, and humans later, to be examined and compared automatically. Nevertheless, to exploit the semantics encoded in such knowledge graphs, a deep analysis of the graph structure as well as the semantics of the represented relations, is required. By applying semantic web and *Linked Data* technologies and creating a big scholarly knowledge graph, the aim is to facilitate management of metadata by extracting, organizing, and processing viable knowledge out of the integrated, interlinked or crowd-sourced input metadata. In the context of scholarly communication, scientific results mainly publications, have been made available on the Web with low marginal costs and easily accessible with regard to legal permissions and licenses. However, the characteristics of the scholarly metadata are influenced by the data integration and management challenges on the Web. In order to enable machines to integrate and exchange such sources of data and have the meaning of that information automatically interpreted, semantic interoperability levels need to be identified. Although, standards and formats have addressed this issue, the search and retrieval of such information on the Web still remains challenging. Because the data that are maintained in the documents of the web pages still need to be examined manually. This thesis developed strategies for exploiting the possibilities of recent technology advances for scholarly communication.

1.1 Motivation

Life of scientists involve continuous exploration and knowledge acquisition about related artifacts and their corresponding metadata [7] from diverse resources. Along with all the other domains, technology has changed the way scholarly data and metadata have been created and shared. With the advent of the Internet and the web, vast amount of research work rapidly published in recent years increased the amount of information and scholarly metadata. Despite certain improvements for example increasing accessibility of certain artifacts and decreasing efforts and costs in creating them, discovery of relevant metadata remains an ongoing challenge for scholarly communities. By reason of the sheer amount of information in unstructured formats makes data on the web heterogeneous and hard to be exchanged and used. Due to this problem, keeping track of relevant information and inferring analytics becomes a hard task for stakeholders involved in document-based scholarly communication. We motivate the problem of difficulty in finding particular information with the following four examples.

Example 1: Overview of the scholarly artifacts and their metadata. Every Junior researcher involving into new research topics needs to go through a learning curve and get overviews of relevant information about research artifacts, events, people etc. For senior researchers, staying up to date with all the developments happening at the relevant communities is a vital and continuous task. Let us assume two groups of researchers, one preparing a survey study about *Link Discovery* and the other group is seeking information to have an overview of that research domain. Consider *Alice*, a researcher from the *Data Integration* community, who has little knowledge about *link discovery* and is in need of getting an overview about the relevant tools, developments, active research groups and overall status of this domain. In contrast, *Axel*, a senior researcher, created a survey paper on this topic entitled *A Survey of Current Link Discovery Frameworks* [194]. It took Axel and three members of his group a considerable effort and time to conduct this survey and develop a reasonable comparison framework. The survey paper covers 10 different linked discovery tools and compares their functionality based on a common set of criteria. At the time of writing this dissertation, by using the keyword "Link discovery survey" on Google Scholar as one of the most used search engines for scholars, Axel's survey paper is the second hit with 71 citations; thus, this is one of the relevant survey papers that Alice would analyze and compare. However, there are at least 10 more survey papers that look relevant, and Alice would face the challenge of studying them in detail or making an informed selection. Despite of all efforts in making a comprehensive survey, Alice might need a different set of comparisons that requires herself tracing some of the original descriptions of those 10 frameworks or maybe more. An approach that is able to generate overviews of the most relevant related work automatically would allow for the identification of the *must read* related work and *must know* frameworks and developments. There are many such use cases that require structured representation of promotion and developments of the community. Community members are the best source for such information and making the metadata available for the rest of the community. A collaborative content creation by the whole community could minimize the effort and time of scholars in providing such surveys of the topic. In addition, it can maximize the comprehensiveness of such knowledge for researchers in need of gaining it.

Example 2: List of potential scholars from the community to collaborate. Different kinds of scholarly metadata are distributed and published by individuals and organizations. Researchers often query about information that needs to be explored from such discrete and distributed resources. We present an example of cooperation recommendation for researchers based on possible but not discovered co-authorship relation. The example starts with the discovery of such co-authorship relation between researchers working on data-centric problems in the Semantic Web area. Generally, researchers get to know each other either during scientific events or projects, or based on recommendations of other community members or by discovery of a related work. In order to discover possible cooperation with other people from the community, researchers need to find and explore profile of relevant community members. Profile of researchers and their co-authorship information is present on services for example DBLP [3] as a bibliographic database for computer science. There are many cooperation and authorship possibilities that never happen because of the lack of awareness about the existence of another party or procrastination of the collaboration. More concretely, the profiles of two selected researchers one from Semantic Web and the other from Data Management and Integration communities are checked for the time between 2015 and 2017. Their networks of co-authorship are being compared within other metadata repositories. While till 2016 there has been not a single collaboration or co-authorship, after 2016, these two researchers started to work in the same research lab, and a large number of scientific results, e.g., papers and projects were produced. Scholarly communities need automatic recommendation about similar use cases in order to increase the impact and value of research results. An approach being able to discover such potential collaborations automatically by metadata analytics would allow for the identification of the best collaborators and, thus, for maximizing the success chances of scholars and researchers working on similar scientific problems.

Example 3: List of scholarly venues ranked by quality metrics. One of the main challenges for researchers is to find a *right* venue to publish their research results. The selection criteria for venues ranges from venue location, deadline, topics to the acceptance format, registration fees etc. We motivate the problem of filtering and extracting metadata about scientific events from call for paper (CfP) emails of mailing lists with the following scenario. Besides having a different portfolio of services to support researchers, every research community has its own way of distributing such information. Our focus is on mailing lists, i.e., a communication medium often used by research communities as a specific channel for distributing, e.g., announcements of releases of software packages or datasets, CfPs of upcoming scientific events, and research related opinions and questions. Active Researchers receive a vast amount of emails about conferences and scientific progress every day. Subscribing to such mailing lists increases the enormous number of announcements every day. Suppose a researcher who has subscribed to such a mailing list needs to identify upcoming related scientific events. A researcher in our scenario has to trace the emails on a list and to decide which ones to have a closer look into. Although this process looks straightforward and is one of the favorite communication channels for researchers, a lot of relevant information might either be overlooked or overwhelm recipients.

[3] https://dblp.uni-trier.de/

1.2 Problem Statement and Main Challenges

Researchers in different fields have different needs on metadata analytics. In addition to scholarly articles, there are other types of artifacts such as Open Educational Resources (OER), events that are generated as digital products provided by different stakeholders in scholarly communication. Efficient research thus requires awareness of such additional related information and the overall status of artifacts [148]. Different stakeholders communicating in scholarly ecosystem dealing with all types of scholarly artifacts face a major obstacle in the preparation of complete and accurate metadata. They struggle with collecting metadata from the community, with the need to minimize the burden on researchers. The technology already made great leaps forward in terms of discoverability and accessibility. It is now possible but limited to integrate metadata about affiliations, grants, and research outputs between systems that use persistent identifiers for people, places, etc. However, the entire scholarly communication has the potential to shift to a new paradigm by comprehensive, accurate, up-to-date metadata. The examples in section 1.1 shows the issues aroused by the current paradigm of scientific communication for researchers. They need to explore, evaluate and decide on many things that are based on metadata of different scholarly artifacts and stakeholders etc. The information that researchers are seeking depends on discovery, access, integration, analysis, and reproducibility of metadata about all possible kinds of produced and shared artifacts. Due to the limited machine-interpretability of these documents, innovative assistance services for researchers to explore and retrieve required information are lacking. In order to facilitate knowledge discovery by assisting humans and machines, FAIR principles[4] have been introduced as a set of guiding principles to make (meta)data Findable, Accessible, Interoperable, and Reusable. Despite the attempts in developing services for supporting scholarly communication (more details in section 2.2), incomplete metadata (Example 1), missing *semantic* links between repositories (Example 2) of all kinds of artifacts and data heterogeneity (Example 3), keeps the challenge still remaining [47]. The status of current scholarly metadata distributed in repositories inherit the characteristics of the of big data [130] specified for scholarly metadata [297] as *6 Vs of big scholarly data*: v1) high *volume* of scholarly metadata about scholarly artifacts being made available, v2) *variety* of entities and relationships among these different types of artifacts, v3) *velocity* representing the growth rate of scholarly data and metadata, v4) *value* and quality of scholarly metadata and impact evaluation of artifacts, people or events and v5) *veracity* of metadata such as author disambiguation and de-duplication. A sixth characteristic is added in [297] for scholarly metadata, v6) *variability* of the meanings of the metadata. In addition, the current information retrieval approaches for most of these repositories are based on keyword search. Keyword search is increasingly inadequate for retrieving information from the enormous and ever growing amount of metadata. Therefore, such characteristics add challenges towards providing a comprehensive approach for the current paradigm of scholarly communication:

Challenge 1: Collecting and Curating Metadata from Multiple Distributed Sources including Databases and Members from the Research Community. The origin of metadata is the scholarly communities and individual members and sources such as researchers, publishers, libraries and data repositories. In the past decades, scholarly communication has witnessed a rapidly growing number of published artifacts and their metadata. Thus, a large and widely spread amount of unstructured data about scholarly artifacts have been made available via communication channels not specifically designed for that purpose e.g., survey papers, emails, homepages. However, these metadata are often duplicated, disconnected and not readily reusable by other systems [96]. In addition, most of the other fundamental information remain as the community information and disconnected from artifacts. There have been attempts to collect structured metadata from research communities. For example, manuscript submission

[4] `https://www.force11.org/group/fairgroup/fairprinciples`

systems aimed at collecting metadata from researchers directly at the time of submission. However, the collected metadata needed pre-processing and curation to become reusable with the purposed of the underlying system. In addition, this approach often needed duplicate entry of metadata and viewed as too complex and time consuming by some authors. On the other hand, such metadata about authors, title, abstract of the manuscripts are limited and do not support the needs of researchers in seeking certain information. Example 1 in section 1.1 shows one use case of such queries that requires analytics on metadata created from content of scholarly artifacts. In addition, browsing through these metadata to identify significant characteristics of a certain artifact requires lots of effort and is a time consuming task. The enrichment and interlinking of such metadata collections advances scholarly pursuits for the benefit of scholarly communication. Synchronization and automation are the key steps in this challenge.

Challenge 2: Integration of Heterogeneous Metadata Resources. In recent years, the challenges of data integration have changed dramatically [180]. The previously proposed approaches for data integration has has scale to Web data. The domain of scholarly communication and the corresponding metadata created and published by researchers and other stakeholders are not exceptional from this fact. Therefore, the heterogeneity of *big scholarly metadata*, a term coined by Xia et al. [297] creates obstacles for services which are based on metadata integration. Example 2 in section 1.1 shows one use case that required integration of metadata from different resources. Scholarly metadata are published in big quantities (volume) and about different types of artifacts (variety). Publishing of the scholarly artifacts and their associated metadata are increasingly growing (velocity). There are structural differences (veracity) across representation of information related to scholarly artifacts of the same type. Integration and evaluation of scholarly metadata play important role in the life of scholars (value).

Challenge 3: Systematic Quality Assessment of Scholarly Artifacts. Currently, the space of information around scholarly artifacts is organized in a cumbersome way, thus preventing stakeholders from making informed decisions. Scholarly data analysis involves various applications in better understanding *science of science* using quality indicators [86]. Most of the currently available measurement services about quality (*fitness for use* [127, 144]) of scholarly artifacts are limited to certain indicators. For example, the number of citations for publications are often used for success measurement of a research work of a researcher which does not relate directly to the quality of the work in the meaning of fitness for use. Because of the diversity and wide range of possible indicators, it is not an easy task to define a centralized service for quality assessment of scholarly metadata and derive meaningful insights [168]. The problem of current services not being able to offer quality based recommendations arises from the current metadata representation and management. In addition, there is hardly any comprehensive formalization or implementation of ontologies about other criteria for quality of scholarly artifacts on which the communities are agreed up. Example 3 in section 1.1 shows a use case about the needs of scholars on venue recommendation. That motivates a comprehensive conceptualization of the scholarly communication with regards to the quality, *fitness for use*, of the scholarly entities.

Challenge 4: Providing Services Addressing the Information Needs of Many Different Kinds of Stakeholders. Scientific communication is composed of a variety of stakeholders with different interactions in scientific communities [40]. Thus, scholarly metadata have been published and expected to be consumed by individual researchers, scholarly organizations, institutes and research centers. Therefore, services for scholarly communication require to support a broad range of users [168]. Apart from search engines that are designed for general information exploration purposes, most of the current scholarly services are focused on limited use cases and research domains. Often, researchers of different disciplines need to get particular information from other communities, see the examples in section 1.1. This requires awareness of the information exchange channels and services of the target community. Taking into account the roles of researchers, e.g., reviewer, organizers in the scholarly communication, gaining access to the right information based on what they search and where there search is a time consuming and challenging task. A comprehensive system with rich and connected metadata can support different stakeholders of scholarly communication.

1.3 Research Questions

The ultimate purpose of this thesis is to facilitate scholarly commutation with semantifying scholarly metadata. In order to do so, corresponding to each of the challenges explained in section 1.2, four research questions have been defined to be addressed in this thesis:

Research Question 1: How can we leverage semantic representation techniques to facilitate the acquisition and the collaborative curation of scholarly metadata?

With the help of Semantic Web technologies, building more explicit and interoperable, machine-readable representations of information has become possible. Considering this question, the aim is to explore the possible improvements on the current paradigm of scholarly communication with regard to the FAIR principles. A collaborative acquisition of scholarly metadata facilitates creation and curation of knowledge bases for scholarly communication. Community involvement in the curation synthesizes complex information and increases their comprehensiveness and visibility.

Research Question 2: To what extent can we increase the coherence of scholarly communication artifacts by semi-automatic linking?

Scholarly communication artifacts, such as bibliographic metadata about scientific publications, research datasets, citations, description of projects, profile information of researchers, are often published independently and isolated. With the help of Linked Data technologies, interlinking of semantically represented metadata have been made possible. We investigate on discovering and providing links between the metadata of scholarly artifacts. The links are generated retrospectively by devising similarity metrics over sets of attributes of the artifact descriptions. Interlinking of such metadata makes it sharable, extensible, and easily re-usable.

Research Question 3: How can the quality of scholarly artifacts be assessed systematically?

Discovering high quality and relevant research-related information have certain influence on the life of researchers and other stakeholders of the communication system. For examples, scholars search for quality in the meaning of fitness for use in questions such as "the venues should a researcher participate" or "the papers should be cited". In this regard, the impact and usability of scholarly artifacts, events and researcher profiles are directly affected by their quality. Assisting researchers with a deeper quality assessments of scholarly metadata and providing recommendations can lead to new opportunities in research.

Research Question 4: What analytic services can fulfill the information needs of the stakeholders in scholarly communication?

There are already attempts to assist researchers in this task, however, resulting recommendations are often rather superficial and the underlying process neglects the different aspects that are important for authors. Providing recommendation services to researchers and a comprehensive list of criteria while they are searching for relevant information. Furthermore, having access to the networks of a paper's authors and their organizations, and taking into account the events in which people participate enables new indicators for measuring the quality and relevance of research that are not just based on counting citations. The proposed approach will provide a crowd-sourcing platform to support recommendation services about scientific venues, projects, results, etc. based on quality assessment.

1.4 Publications Associated with this Dissertation and Contributions

The following articles were produced during the preparation of this dissertation. The following chapters are based on the contributions presented in these articles:

- *Journal Articles*:

  1. Behnam Ghavimi, Philipp Mayr, **Sahar Vahdati**, Christoph Lange, Sören Auer, *Semi-Automatic Approach for Detecting Dataset References in Social Science Texts*, IS&U 2016;
  2. Anastasia Dimou, **Sahar Vahdati**, Angelo Di Iorio, Christoph Lange, Ruben Verborgh, and Erik Mannens, *Challenges as Enablers for High Quality Linked Data: Insights from the Semantic Publishing Challenge*, PeerJ 2017.

- *Conference and Workshop Papers*:

  3. **Sahar Vahdati**, Sören Auer, Christoph Lange, *OpenCourseWare Observatory – Does the Quality of OpenCourseWare Live up to its Promise?*, LAK 2015;
  4. **Sahar Vahdati**, Farah Karim, Jyun-Yao Huang, Christoph Lange, *Mapping Large Scale Research Metadata to Linked Data: A Performance Comparison of HBase, CSV and XML*, MTSR 2015;
  5. **Sahar Vahdati**, Natanael Arndt, Sören Auer, Christoph Lange, *OpenResearch: Collaborative Management of Scholarly Communication Metadata*, EKAW 2016;
  6. Giorgos Alexiou, **Sahar Vahdati**, Christoph Lange, George Papastefanatos, Steffen Lohmann, *OpenAIRE LOD Services: Scholarly Communication Data as Linked Data*, SAVE-SD 2016;
  7. **Sahar Vahdati**, Anastasia Dimou, Christoph Lange, Angelo Di Iorio, *Semantic Publishing Challenge: Bootstrapping a Value Chain for Scientific Data*, SAVE-SD 2016.
  8. Behnam Ghavimi, Philipp Mayr, **Sahar Vahdati**, Christoph Lange, *Identifying and Improving Dataset References in Social Sciences Full Texts*, ElPub 2016;
  9. Shirin Ameri, **Sahar Vahdati**, Christoph Lange, *Exploiting Interlinked Research Metadata*, TPDL 2017, **Second best paper award–honorary mention**;
  10. Said Fathalla, **Sahar Vahdati**, Christoph Lange, Sören Auer, *Analysing Scholarly Communication Metadata of Computer Science Events*, TPDL 2017;
  11. Said Fathalla, **Sahar Vahdati**, Sören Auer, Christoph Lange, *Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles*, TPDL 2017;
  12. Rebaz Omar, **Sahar Vahdati**, Christoph Lange, Maria-Esther Vidal and, Andreas Behrend, *SAANSET: Semi-Automated Acquisition of Scholarly Metadata using OpenResearch.org Platform*, ICSC 2018;
  13. **Sahar Vahdati**, Rahul Jyoti Nath, Guillermo Palma, Maria-Esther Vidal, Christoph Lange, Sören Auer, *Unveiling Scholarly Communities of Researchers using Knowledge Graph Partitioning*, TPDL 2018.

- *Working Draft*:

  14. **Sahar Vahdati**, Christoph Lange, Sören Auer, Andreas Behrend, *Towards a Comprehensive Quality Assessment Model for Scientific Events*, Scientometrics Journal.

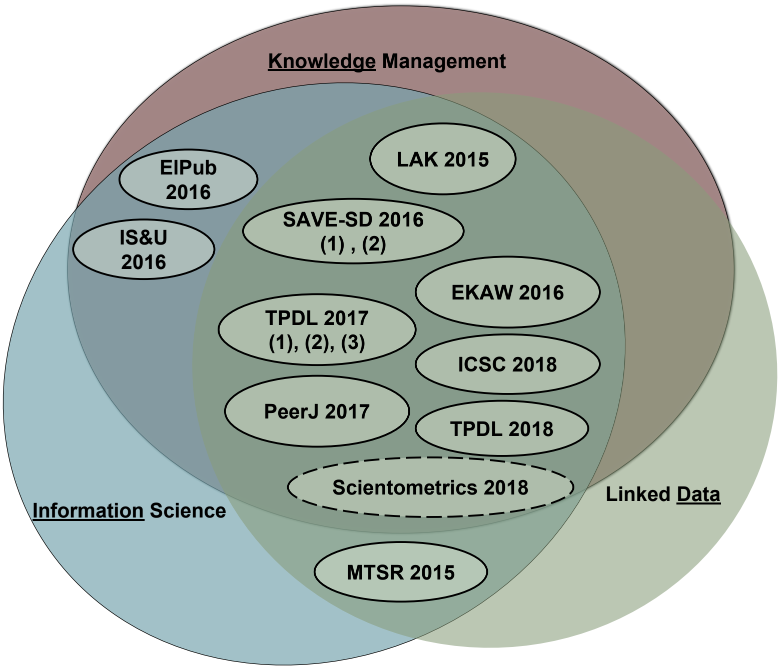


Figure 1.2: **Overview of the main research areas covered by this thesis**. The publications associated to this thesis have been distributed through the following research domains: Knowledge Management, Linked Data, Information Science.

This research has an impact on three main research communities: Information Science as the domain of focus for the identified gap in current needs and available services, and Knowledge Management and Linked Data as the technical support for the proposed approach. The distribution of the research results of this thesis through the related research domains is shown in figure 1.2. The contributions of this research are as follows:

- A scholarly knowledge graph integrating data from several external datasets;
- A knowledge-driven framework for data acquisition and curation platform following a crowd sourcing approach;
- A set of possible recommendations and analytics; and
- A systematic and comprehensive quality assessment of scholarly artifacts.

Parts of the contributions of this dissertation which is mentioned earlier were achieved as the result of effective teamwork. The papers co-authored by the following people are the result of theses (master and bachelor) closely supervised by the author of this dissertation: Behnam Ghavimi, Shirin Ameri, Rebaz Omar, and Rahul Jyoti Nath. Apart from leading of the theses projects, the author of this dissertation (Sahar Vahdati) has significant contributions in the process of writing and publishing of the research results. The contributions of Sahar Vahdati in the papers co-authored with Said Fathalla are mainly related to the implementation and analysis of his ontology on the OpenResearch.org platform. The author (Sahar Vahdati) will use the "we" pronoun throughout this dissertation, but all of the contributions and materials presented in this work, except the previously mentioned collaborative works, originated from the work of the author solely by herself.

1.5 Thesis Structure

In this thesis, we focus on analysing the problems of the scholarly communication and providing approaches for their implementation. A systematic and comprehensive management scholarly metadata is proposed based on Linked Data technologes. The steps of metadata management are introduced in the form of a life cycle. Some of the steps of the life cycle are implemented in a platform for automating and crowd-sourcing the collection and integration of semantically structured metadata (knowledge graph) about scholarly communication in order to reduce the effort for researchers to find "suitable" and "related" (according to different metrics) artifacts. Therefore, this research aimed at contributing towards a research knowledge graphs with the following research goals: (i) defining a comprehensive quality based measurement for scientific artifacts [269, 270, 272, 275], (ii) developing a platform for collaborative and semantic scholarly metadata management [271]; (iii) providing services for semantically enriching and interlinking of scholarly communication metadata [2]. The proposed platform establishes possibilities for the evaluation and assessment of scholarly artifacts considering a set quality metrics defined by community and provides a cross-domain service for managing metadata of artifacts. This supports easy and flexible data exploration using Linked Data technology based on structured scholarly metadata. To prepare the reader for the upcoming chapters, an overview of the thesis investigated is presented.

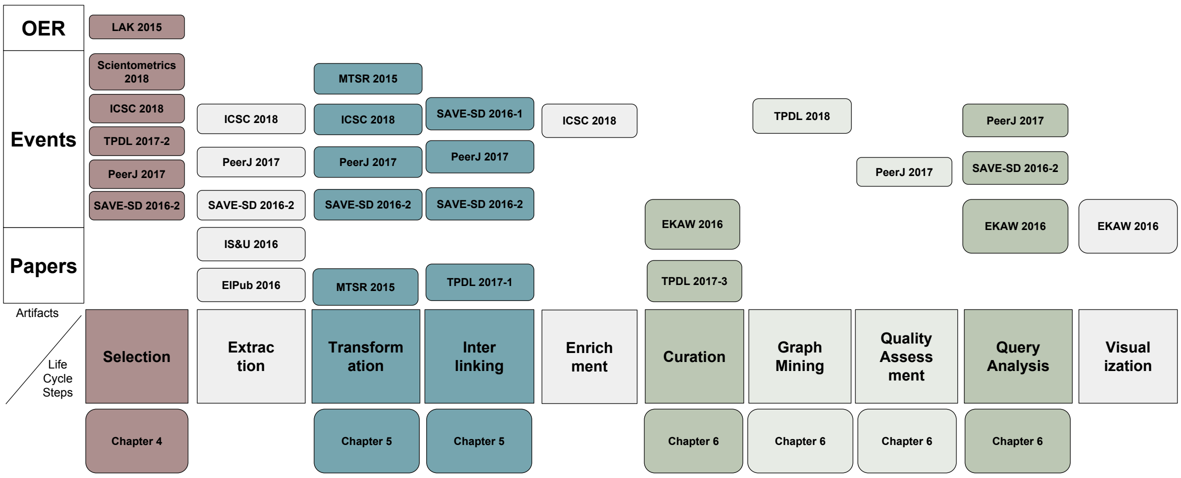


Figure 1.3: **Distribution of contributions from publications through chapters of this dissertation.** The X axis represents metadata management stages from the proposed life cycle, the Y axis represents three example scholarly artifacts that was the use case of this research.

Chapter 1. is the introduction of the thesis and **Chapter 2.** provides information about the development of scholarly communication and the current services. Figure 1.3 shows the design of the chapters based on the main contributions from the published papers. The proposed metadata management life cycle will be presented in **chapter 3**. Contributions to each of the steps described in the **metadata life cycle** are presented in the corresponding publication related to this thesis. Figure 1.3 shows the relevance of the publications on stages and the addressed artifact. Same colored stages and their publications have been described in the same chapter. **Chapter 4.** (purple) represents contributions related quality assessment of artifacts and events. **Chapter 5.** (blue) describes the research work related to transformation and extraction of metadata. **Chapter 6.** (green) is about the curation process and utilization of the created and curated metadata. Since contributions to the other stages (gray) have been relevantly limited, they are skipped to appear in chapters. **Chapter 7.** provides a conclusion and possible future directions.

CHAPTER 2

Scholarly Communication Then and Now

Science being the enterprise of discovering knowledge, scientific communication is intended as a knowledge exchange ecosystem. Scholars disseminate their research results by publishing written documents. This way of communication has developed over time and consists of certain steps and corresponding stakeholders such as publishers, authors, reviewers, and organizers. In recent years, scholarly communication has faced rapid changes in terms of producing a large volume of scholarly artifacts and their accessibility [216]. The need to retrieve information from such a complex and heterogeneous system increased the number of investigations in providing support for individual scholars or research communities.

This chapter reviews the history of knowledge exchange among scholars from its origin to the present status. Section 2.1 looks back to the development of scholarly communication through time. We observe the evolution of the required steps for disseminating research results. The section summarizes the impact of publishing in the life of scholars and the importance of being involved in scholarly communication. We also overview the development of scholarly artifacts through time starting from ancient time till the digital era. The second part of this chapter, Section 2.2, provides the state-of-the-art of services developed to facilitate the involved stakeholders in scholarly communication. The early physical systems supporting publishing and dissemination are out of the scope of this section. However, we focus on summarizing digital services developed for the online assistance of scholars through different stages of scholarly communication. This chapter aims at providing a comprehensive overview of the area and support in justifying the gap of facilitating scholarly communication by already existing services.

2.1 Development of Scholarly Communication

Scholarly communication is the process of propagating scientific knowledge and research results to make them publicly available. For scientific activities, a certain communication system has been established over time. Apart from the quality of facilities in science forced by geographical or political conditions, there are two main sides of activities in academia namely education and research. Considering education, the population is expected to pass through a certain educational system and gain academic knowledge and their corresponding degrees. Academic lectures are held by advanced scholars who present an exposition of the given subject with the purpose of training the target audience. For the research side, after certain achievements, individuals involve themselves in knowledge discovery activities called ”research“. Eventually, the groups of scientists with common research interest have built research communities.

Researchers produce essays as written documents in order to exchange results within scientific communities. Such scientific literature is a textual representation of a research work which has been accomplished in a research institution. For many decades, scientific publishing has been the main communication channel for scholars. The whole scholarly communication system is established gradually and was affected by technological development. Through time, a lot of incremental changes have happened in terms of the roles of people, organizations, artifacts as well as their impact on the reputation of people and communities. Based on a systematic analysis, we overview the development of scholarly communication from the viewpoint of four main aspects:

- **Publishing and artifacts**: Disseminating research results has been the main communication form for scholars. The type of scholarly artifacts has changed over time depending on the technological development. Moving from physical artifacts to digital artifacts brought a lot of new facilities. Nowadays, with the help of digitization, there are digital monographs, books, micropublication, blog posts, videos, datasets etc. Subsection 2.1.1 overviews the development of publishing and their corresponding artifact types over time.
- **Collaboration**: The Internet has brought people together virtually and increased interactions and collaborations. Scholarly stakeholders are using a combination of the World Wide Web, email system, and discussion groups, etc. to share knowledge and support each other, and organize events etc. As a consequence, scholarly collaborations are made broadly possible across institutional and geographical boundaries. In science and academia, collaboration ranges from commenting on results of each other to actually conducting research and producing results together. Collaboration plays a significant role in scholarly communication and scientific results because of the interdisciplinary nature of the science. Subsection 2.1.2 analysis of the development of scholarly communication in terms of changes in collaborations.
- **Quality control**: With the expanding growth of the publications, the methods for approving the innovation, quality, and soundness of the claims about scientific results are also changing. From ancient times, the value of research results has been controlled by senior and qualified researchers of the corresponding community. Nowadays, various methods and quality control systems have been developed for this purpose such as peer-review of publications. Subsection 2.1.3 summarizes the attempts on creating such quality control systems and investigates on advantages and disadvantages of the proposed and in-use methods.
- **Success measures**: The level of productivity and the impact of scholars in their field of interest determines their success rate. It has always been measured with several metrics related to their achievements and research results. With the rapid changes of digital publishing, the metrics for measuring success and reputation of individual scientists, groups, and organizations have become increasingly changing. In early times, unique innovations and extraordinary findings by individual scientists have been the only way of such measurements. By emergence of scientific publishing, a lot of performance-driven metrics have also developed such as bibliometric metadata and citation counts etc. Subsection 2.1.4 provides an overview of the developed metrics for measuring the success rate of scholars and communities.

2.1.1 Publishing and Artifacts

Creating written documents has been the predominant knowledge exchange paradigm until recently. Some of the earliest communication in the writing form recorded to be symbols scratched on stones of caves that date back to the 65th millennium BC[115]. Early written symbols were based on pictographs (pictures which resemble what they signify) and ideograms (symbols which represent ideas). Ancient Sumerian, Egyptian, and Chinese civilizations began to adopt such symbols to represent concepts. One of the earliest representations of systematic writing goes back to the seventh millennium BC at Jiahu [166] where 16 symbols were used to represent natural elements. Through time, such symbols have been developed into the sophisticated alphabets of today [60] and end up with long texts transferring knowledge.

Later in medieval Europe, book and manuscript production was confined largely, however only in wooden frames or clay tablets [268]. The documents and the information written in this form were only findable by its main authors or maintainers and accessibility was only defined for certain people. Providing transcribes and re-using of knowledge was a major challenge because of the certain restrictions in creating them. Therefore, a collection of such documents used to be stored in one place and accessed by people. To have a centralized storage of such documents that were initially collected in temples, libraries started to emerge. One of the most famous libraries of early times with a huge collection of written documents was the library of Alexandria that functioned as a major center of scholarly artifacts from its construction in the 3rd century BC. Mainly it was none-serial documents written in one volume or in a limited number of volumes that were stored in such repositories. Scholarly metadata management has already started in such libraries by using catalogs and such documents became well-known as *monographs* [263]. In the beginning, the catalogs were subject-only e.g., philosophy, mathematics and the classification of the corresponding artifacts have been mainly done by language or material. Through time, library catalogs turned to manuscript lists, arranged by format or author names.

After the use of paper as the main writing medium (starting in Egypt and China), the printing and publishing industry thrived. Printed catalogs of libraries have been published as dictionary catalogs in the early modern period and enabled scholars outside a library to gain an idea of its contents. More individual publishers also started to distribute manuscripts by the change Johannes Gutenberg brought to the printing industry in Germany and Europe. He established a new profession as *publisher* in 1450 which becomes the favorite activity of some scholars who could get the printing and publishing license from rulers. However, libraries remained as the main data and metadata storage.

One can relate the history of scholarly events to the history of the libraries where libraries operated as important venues for scholars to gather in one place and share ideas, knowledge, and their original work. Until the 1600s, apart from library meetings, research results were communicated privately in letters, lectures or in books. The French Journal Des Sçavans and the English Philosophical Transactions of the Royal Society in 1665 were the two first scientific journals to systematically publish research results as manuscripts [169]. Journals made the chaos of science accretive by bringing the possibility for announcing advance inventions as well as short-term and steady reports of experiments. All these started to build the scientific communication through publishing research results in scientific documents which are often called papers.

Through the establishment and development of this communication model, several stakeholders emerged based on the available dissemination technology and requirements of the research community. Publishing houses are one of the early emerging stakeholders of scholarly communication. One of the pioneers in natural science is Springer that is founded in 1842 by Julius Springer who had a publishing house in Berlin. After 175 years, the name Springer stands for one of the globally active publishers.

With the increasing amount of published manuscripts and journals, the need for more systematic metadata management inside data repositories increased. Librarians started to propose and use new classification models. Although indexing has been designed earlier, the first card catalogs appeared in the late 19th century after the standardization [78]. Until the digitization of library catalogs, which began in the 1980s, card catalog was the primary tool to locate documents, books, and manuscripts in the libraries. Card indexing enabled more flexibility in the management of such metadata and made exploration bibliographic items and related enquirers easier. It was also the basis for the development of the online public access catalogs in the 20th century.

An evolutionary period started for communication channels through which news, education, data, and messages were disseminated with media such as radio, TV and in later times the Internet. This was the time moving from physical artifacts to digital artifacts. Till now, recordings have been used for educational lesson broadcasting, oral history and storytelling, frequent question answering and research finding transferring. With the invention of video tapping, lecture recording in both audio and video became active scholarly artifacts especially for educational resources or broadcasting event till today. After the emergence of early personal computers in the 1960s and invention of the web, physical libraries have been transformed into digital libraries. They have been facilitated to online manuscript cataloging to enhance the usability of digital libraries and scholarly manuscript repositories by providing a dynamic search facility over the stored metadata e.g., author, title, keyword.

Most of the online catalogs allow searching for any word in a title or other field, increasing the ways to find a record. Digital libraries made the information more accessible to many people with disabilities. Digitization and online catalogs reduced the space of physical storage considerably. Metadata versions and updates on each version have been made significantly more efficient. Although there has been always a historical revolution of content, the development of scholarly communication has been mainly focused on artifacts and reduction of the marginal costs in preparation of the communicated objects. Especially, digitization reduced a lot of marginal costs in preparing of such materials, the effort of exploring and accessing such scholarly artifacts had been a challenge. One of the initial movements towards this direction started with proposals about Open Access material as the underlying policy of publishing. These policies aimed to make the content of scientific works available for everyone, anywhere in the world to read and access and build upon the work of others.

The Open Access movement dates back more than thirty years where the Gutenberg project started with the aim of making most consulted books digitally available to the public as eBooks [109]. The first free journals were published on the Internet in the late 1980s and early 1990s. By having early web pages [22], online archives of scientific documents started to be disseminated by individual researchers or organizations. *ArXiv*[1] [93](launched in 1994) is one of the early online repositories of electronic preprints (before peer review) of scholarly publications. This repository which is still in function is one of the few repositories providing free of cost access to scientific publications. It contains basic metadata of publications such as title, author names, abstract etc.

Through the existence of such services following Open Access movement, free availability of huge volumes of monographs, peer-reviewed articles, and reports have been made possible that has enormously increased the impact and quality of research works. In order to be able to use them effectively, researchers and others need help to navigate their way around, organize, analyze and explore the content and metadata relevant to their work. To handle the growing volume of electronic publications, new tools and technologies such as digital libraries have to be designed to allow effectively automated and semantically classified search facilities. The concept of digital libraries(DL) became the trend where it was emerged in 1892 by Paul Outlet with the vision of building a search system and interlink documents and image files together [294]. One of the early examples was created by the Education Resources Information Center (ERIC) as a digitized version of the scholarly resources of that institute. In 1994, after the existence of early web pages, the Digital Libraries initiative was launched with the purpose of providing more online facilities to access the libraries online through the Web [239].

Digital libraries have been defined as a virtual organization with the purpose of collecting, managing and preserving of digital content, and offers specialized functionality on that content with regards to quality [156]. Although digital libraries have made a huge change in the availability of resources, the accessibility remained limited. The dissemination of digital resources on the web by libraries often requires special permissions or subscriptions in the organizational level. In the early 2000s, the Open

[1] `https://arxiv.org/`

Access(Archives) Initiative Protocol (OAI-PMH[2] was proposed to harvest (or collect) the metadata descriptions of the records in an archive so that services can be built using metadata from many archives. It develops and promotes interoperability standards that aim to facilitate the efficient dissemination of scholarly artifacts to increase their availability in scholarly communication.

The fundamental technological framework and standards that are developing to support this work are, however, independent of both the type of content offered and the economic mechanisms surrounding that content. As a result, the Open Archives Initiative is currently an organization and an effort explicitly in transition and is committed to exploring and enabling this new and broader range of applications. As we gain greater knowledge of the scope of applicability of the underlying technology and standards being developed and begin to understand the structure and culture of the various adopter communities, we expect that we will have to make continued evolutionary changes to both the mission and organization of the Open Archives Initiative.

FAIR principles [3]have been made in order to bring guidelines for artifact and metadata dissemination [293]. It introduces four main criteria for data and metadata to be findable, accessible, interoperable and reusable. The assumptions of findability are that each element represented by metadata should be assigned a globally unique and eternally persistent identifier. In addition, both data and data are required to be registered or indexed in a searchable resource. In terms of accessibility data and metadata are considered to be disseminated in a format that is retrievable by their identifier using a standardized communications open, free policies. Metadata authentication is highly respected under the FAIR principles and metadata are accessible, even when the data are no longer available. With regard to interoperability, both data and metadata should be presented in a formal and broadly applicable language (using vocabularies that follow FAIR principles). Metadata is considered re-usable with respect to free licenses which is associated with their provenance.

2.1.2 Collaboration

Most of the early scientific publications have been recorded with solo authors [164]. In the current scholarly communication, scientific collaboration is more prevalent than it was decades ago [296]. Co-authorship is one of the valid criteria for measuring the collaboration of scientists and communities. Technology revolution also brought multidisciplinary researchers with diverse scientific backgrounds and perspectives in close collaboration. If researchers with complementary skills join a research project, it can reduce the effort by half in contrast to a solo scholar.

A report is published by Thomson Reuters for each year between 1998 and 2011, showing the number of papers with more than 50, 100, 200, 500 and 1000 co-authors [140]. The statistics of papers and number of co-authors show collaborative authoring in science increasingly outperforms individual authorship. The trend of papers with 50 to 100 authors goes upward from the late 90s to the mid-00s. In the study by Thomson Reuters, the highest number of authors in 1981 is recorded as 118 which was multiplied by 5 only 8 years later. Scholarly communication is currently done in very large scopes in terms of co-authorship and collaborations where there exist scientific articles with 2000 co-authors.

Another study [287] reports the group authorship increased from virtually zero to over 15 percent. The changes in the way research used to be done, methods and facilities have made collaborations necessary. However, sharing of authorship does not directly reflect a tangible engagement. Nevertheless, collaborative papers tend to get cited more often. For example, between continents and countries such as those published jointly by UK and US authors are cited on average more often than either nation domestically. It also works at the institutional level.

[2] `https://www.openarchives.org/`

[3] `https://www.force11.org`

In some countries, the collaboration between the research and industrial sectors has become more apparent. In addition, there is also a correlation between collaboration and higher impact in science [302]. Some of the publishing systems established a contribution recognition approach where authors need to clearly state their responsibility. More collaboration in science is visible because of the changes the Web and Internet brought to the private and professional lives of people. Special social network for scholars connects researchers to each other in a virtual space that can easily lead to scientific contributions. With more travel funding, scientific events and projects, the overall scholarly communication have been facilitated with a more interactive research methodology. However, none of the currently available services are able to predict or recommend effective candidates for collaboration.

2.1.3 Quality Control

Due to cumulative nature of scientific knowledge, quality and trust are particularly important. As reported in [120] currently, many published research findings are false or exaggerated, and an estimated 85 percent of research resources are wasted. Researchers need to be supported by automated systems to ensure that they have effective and high-quality channels through which they can publish and disseminate their findings and that they perform to the best standards by subjecting their published findings to rigorous peer review. In order to build such systems, quality assessment frameworks for each type of scholarly artifacts need to be established. Such assessments ensure that papers published in scientific venues or journals answer meaningful research questions and draw accurate conclusions based on professionally executed experimentation.

Although *Peer review* is now a fundamental quality control measure implemented during the publishing process, the practice as we know it today is quite different from how it was envisioned almost two centuries ago. From the very early times, there had been discussions about reviewing written work of scientists. One of the pioneer review process ideas was first described around 854 AC by a physician named Ishaq bin Ali al-Rahwi from Syria, in his book Ethics of the Physician [254]. However, development of a systematic evaluating process with the purpose of publishing started with the invention of printing for public and publishing of the first scholarly journals. That was mainly editing proposals by peers to regulate the quality of the written material that became publicly available and less about the validity of the research. A first global method for generating and assessing new science is proposed by Francis Bacon in 1620. Later in 1669, experts elected by the French Academy of Science to write reports about ideas and inventions of other scientists for the King.

The first rejection of a scientific work is recorded for the same time by Oldenburg, the Royal Society's first secretary [190]. Shortly after the publishing of first research journals, the peer review process was added in addition to the editing process. The Royal Society of Edinburgh described their peer review in 1731 as follow: "Memoirs sent by correspondence are distributed according to the subject matter to those members who are most versed in these matters. The report of their identity is not known to the author."[284]. Later in 1752, the Royal Society of London adopted this review procedure and developed the "Committee on Papers" to review manuscripts before they were published. For the first time, papers were distributed to reviewers with the intent of authenticating the integrity of the research study before publication. In 1831, William Whewel of the Royal Society of London suggested that reports are commissioned for the incoming papers, to be included in the new version of journal *proceedings* [9].

Peer review in a more systematized form has developed immensely since the Second World War, at least partly due to the large increase in scientific research during this period. A trusted form of scientific communication is provided through peer review, however, critics argue that the peer review process delays publication and stifles innovation in experimentation, and acts as a poor screen against plagiarism. Nowadays, it is used not only to ensure that a scientific manuscript is experimentally and ethically sound, but also to determine which papers sufficiently meet the required standards of quality and originality before publication. Peer review is now standard practice by most credible scientific events and journals. It is an essential part of determining the credibility and quality of work submitted. The *Research Excellence Framework* (REF) [225] for assessing the quality of research in UK higher education institutions, classifies publications by the venues they are published in. This facilitates assessing every researcher's impact based on the number of publications in conferences and journals. Providing such information to researchers supports them with a broader range of options and a comprehensive list of criteria while they are searching for events to submit their research contributions. *Overlay journal* An overlay journal or overlay journal [191] is a term for a specific type of open access academic journal, almost always an online electronic journal (journal). Such a journal does not produce its own content but selects from texts that are already freely available online. While many overlay journals derive their content from pre-print servers, others, such as the Lund Medical Faculty Monthly, contain mainly papers published by commercial publishers but with links to self-archived pre- or post-prints when possible.

Automated benchmarking platforms are the other evaluation methods for more practical research results are automated benchmarking platforms for scientific competitions There have not yet been a foolproof system developed to take the place of peer review, however, researchers have been looking into electronic means of improving the peer review process. Unfortunately, the recent explosion in online only/electronic journals has led to the mass publication of a large number of scientific articles with little or no peer review. This poses the significant risk to advances in scientific knowledge and its future potential. For scholarly events, the *Google Scholar Metrics (GSM)*[4] provides ranked lists of conferences and journals by scientific field based on a 5-year impact analysis over the Google Scholar citation data. 20 top-ranked conferences and journals are shown for each (sub-)field. The ranking is based on the two metrics h5-index[5] and h5-median[6]. GSM's ranking method only considers the number of citations, whereas we intend to offer a multidisciplinary service with a flexible search mechanism based on several quality metrics.

2.1.4 Success Measures

The research communities of past times could recognize the scientific excellence oby peers [199]. Based on a report by UNESCO, already in the period 2007 to 2015 the global population of researchers increased by 20 percentage [7]. In today's big scholarly communication, the career of scholars generally depends on the extent to which their success is recognized by the community. This fact has forced the need for implementing success measurement frameworks by scientific communities. To be able to deal with increasing competition, the metrics for defining success rate of scholars have changed over time. In the past, pioneers and innovators were considered as reputed and successful scientists by the contributions they have been having for humanity and societies. Those not accepted or recognized during their own life would still be acknowledged at some point with the evolution of societies, science, and technology. With the establishment of scholarly communication through publishing scientific articles, success measures also mainly considered around the publishing rates and several metrics related to that. Consequently, different assessment frameworks have been defined with the purpose of identifying scientific success and impact of research communities, organizations and individual researchers.

Research publications have been the key elements of scholarly communication and considered in most scientific communities as main research outputs. The bibliometric parameters have been used as

[4] `https://scholar.google.com/intl/en/scholar/metrics.html`

[5] h5-index is the h-index for articles published in the last 5 complete years.

[6] 5-median is the median number of citations for those articles in the h5-index.

[7] UNESCO Science Report Towards 2030 `http://en.unesco.org/` science report

proxies for excellence in assessment by most funding agencies and universities/research organizations. For example, the number of publications has been often considered as the key indicator of science productivity [160]. With the established habit of referencing other works inside the publications, citation counts became crucial for evaluating the academic achievements of researchers. In many of the research communities, scholars are frequently evaluated on the perceived significance of their work with the citation count. Thus, most methods for evaluating research and scholars are now based on bibliometric indicators, such as various publication-based and citation-based metrics. This has pushed researchers to publish as many articles as possible and crucially follow the number of citations gained from the community. Therefore, the number of publications has substantially increased over the last few decades. Thus, most of the excellence evaluation services established around the citation count as the indicator of a researcher's scientific performance. In the current era, many institutions and universities have to attribute credit scores to their academic publications. In [111], a list of criteria through which researchers get credits are mentioned: Articles, Arguments, Data, Staff, Equipment, Funds, Recognition. Mainly bibliometric information is used as the most commonly used metric for most such frameworks, for example, h-index, citation counts etc. However, it is proved in a recent survey [36] that the prediction of citation counts, as well as the h-index of the corresponding author, do not necessarily correlate to the significance of the work from the community point of view.

The authors of this work concluded that peer judgments of importance and significance differ from metrics-based measurements. The same fact is applicable for the Journal Impact Factor (JIF) which is used for evaluation of research works and authors. Originally, JIF emerges in 1972 as a tool for librarians in making decisions on the purchase of journal subscriptions [159]. Later, it became a common success measure while widely acknowledged to be a poor indicator of the quality of individual papers. Some of these methods have been used in order to evaluate organizations and institute. The damages or advantages such success measurement are bringing in the scholarly communication, ranking systems and the career of researchers is explained in [161]. Lawrence has clearly stated the example cases and the impact of such measurement of science. For example, he stated the fact of having one paper in a journal with high JIF and receiving the high number of citations can change the prospects of a postdoc from nonexistent to substantial. Two or three such papers can make the difference between unemployment and tenure. The fact is, it is not only the incomplete measurement of success by these approaches, it is also the effect they have on the whole scholarly communication. The growth of open access is also being held back by success measuring factors related to publishing. The need to maximize publications and citations makes the large research groups benefit from the group size in gaining more citations or number of papers However, other factors such as number of supervision, recruitment promotion, research prizes could also be considered.

In general, the success rate of researchers or a scholarly organization cannot be evaluated with a single number. The problem arises from stakeholders which favor numerical evaluation of performance and reward compliance inside the scholarly communication. Increasingly complex grant applications requirements in research excellence result at the expense of research effort. Institutions, research groups, and researchers find themselves in a competitive scholarly communication system Scholars have complex merits and achievements that involve different variables. This makes the evaluation of their success and judgment about their excellence impossible and unfair to only be summarized by a single figure. Publish or perish culture where quality and relevance are subordinate to quantity forces science to follow a close system. Under such constrains, initiatives towards open science, FAIR principles, Open access etc. are not powerful enough. The Leiden Manifesto [113] attempts in proposing basic policies for metrics of research evaluation. One of the main points mentioned in the manifesto is to consider quantitative evaluation of excellence with a support on qualitative, expert assessment. Since scientists have diverse research missions, no single evaluation model applies to all contexts. Success measures should consider metrics relevant to policy, industry or the public aspects of science. There are a criticism about the limitations of such metrics built on English only literature. Therefore, the manifesto suggests consideration of local excellence metrics. A better approach is through multidimensional criteria evaluation, taking into consideration what is expected from a researcher and what is relevant for the career of any researcher. A multidimensional and comprehensive assessment of researchers by their employers and funders in a broader scope is required due to the mobility of researchers across borders, in all scientific domains and at all career stages. Changing practice from the traditional paradigm in most disciplines will require a fundamental shift.

2.2 State-of-the Art of Services Supporting Scholarly Communication

In order to position the proposed approach, it is crucial to explore the already existing systems facilitating scholarly communication. A short overview of highlighted attempts is discussed in the remaining of this section. Most of the currently available services are custom implementations with a focus on covering certain problems. Looking deeply into the present systems, it is clear that supporting interoperability and services based on the quality assessment of artifacts have not yet been comprehensively realized; for example, a cross-disciplinary publication venue recommendation system is missing.

- **Domain Modeling**: Sharing a common understanding of the structure of information is one of the more common goals in developing ontologies. Ontologies have become common on the World-Wide Web. Ontology-based languages have been developed for encoding knowledge on Web pages to make it understandable for human and machines while exploring knowledge. Domain modeling and development of ontologies are often used as milestones in providing better knowledge management and exploration. Scholarly Publishing, as well as other domains, witnessed the development of specific ontologies. One of the main research areas in semantic publishing is the development of semantic models of scholarly communication (more details in subsection 2.2.1).
- **Scholarly Metadata Extractors** A lot of information is already carried inside the scholarly artifacts. Especially publications as the main means of the scholarly communication contain a lot of metadata. The research contributions are introduced in using different sections and representation types such as text, tables, figures, bullet points etc. Despite various formats, almost all scientific publications include basic segments such as title, author, affiliation, abstract, list of keywords, publisher, year, number of pages, and list of references. Metadata extraction from scholarly artifacts especially from publications is a crucial task in building a scholarly knowledge graph. Subsequently, some of the selected metadata extraction tools are introduced (subsection 2.2.2).
- **Datasets and Repositories**: Along with the development of the Web, a huge amount of datasets have been published. In the domain of scholarly communication, different artifacts have been made available by individuals and organizations. Online repositories have been created in order to have a centralized management of such artifacts and their metadata. Different research communities have developed their own repositories and the culture of publishing datasets. With a focus on Computer Science field, a set of relevant related work will be discussed in details (subsection 2.2.3).
- **Services**: Diverse type of services have been developed in order to make the life of researchers in making use of the artifacts and the metadata disseminated over the Web. Such services have a wide range of types such as digital libraries, search engines or statistical and analytical web pages. Such services mostly have a focus on supporting researchers with regard to a particular artifact. For example, there are search engines for publications and different search engines for events etc. An overview of most-used and related services are respectively introduced (subsection 2.2.4).

2.2.1 Domain Modelings

The preliminary requirement of building services with the help of the Semantic Web technologies is to have the domain conceptualization and vocabularies at hand. It enables building a knowledge graph for representing the research findings in a structured and semantic format. There are several ontologies developed for describing the domain of scholarly communication and scholarly artifacts mainly publications. One of the early attempts in modeling the scholarly communication domain is *Functional Requirements for Bibliographic Records (FRBR)* which was developed using the entity-relationship model to conceptualize online library catalogs and bibliographic databases from a user's perspective. FRBR conceptualizes three groups of entities. The first group considered the scholarly artifacts as research results (e.g., publications). Group two focuses on those entities responsible for the content of scholarly artifacts (a person or corporate body).

The third group includes the entities that serve any research effort (concept, object, event, and place). FRBR represent a series of structured ideas about bibliographic records and can be used for basic assumptions entities involved in default publishing activities. However, FRBR has limited focus on some special aspects of the scholarly communication and further developments require extensions of the model. *Metadata Encoding and Transmission Standard* (METS) is one of the initial schema modeling attempts as a metadata standard for encoding descriptive, administrative, and structural metadata regarding objects within digital libraries [179]. A comprehensive version of METS was published as a standard for descriptive cataloging labeled as *Resource Description and Access* (RDA). METS and RDA are still considered as the pioneer guidelines and instructions for creating a library and cultural heritage resource metadata. They have been updated and the vocabularies have been used in developing metadata registries using Linked Data technologies.

The need for more concrete ontologies in different aspects of scholarly communication lead to the development of specific vocabularies. As one of the early attempts, the Dublin Core (DC for short) schema developed to describe metadata terms related to digital and physical resources. It was initially proposed as a simple set of fifteen metadata elements with the focus on scholarly artifacts such as books or CDs (contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type). *Dublin Core Abstract Model* (DCAM) is a reference model independent of any particular encoding syntax [213]. The metadata vocabularies of DC have been widely used for applications of the linked data cloud and Semantic Web implementations. However, it is not sufficiently comprehensive to describe specific properties of objects of the scholarly domain for example properties of online courses.

Semantic Web for Research Communities (SWRC)[8] is an ontology for modeling entities of research communities such as persons, organizations, publications (bibliographic metadata) and their relationships [259]. The *Semantic Web Conference Ontology* (SWC)[9] models knowledge about conferences. It covers the sub-domains of *describing papers* and *modelling the roles*. The defined terms include the authors and their affiliations, the role of the researchers in different venues. The above-mentioned ontologies have focused on metadata about scholarly artifacts, however, there are also content modelings. The ontology of Rhetorical Blocks (ORB)[10] is a formalization capturing the coarse-grained rhetorical structure of scientific publications [52]. The *Scholarly Article* (SA)[11] ontology comprises a set of concepts related to published articles such as *article*, *keywords*, *contributor* and *citation*. Moreover, it comprises a set of properties such as *isStyleOf* and *dateRejected*.

[8] SWRC: http://ontoware.org/swrc

[9] SWC: http://www.scholarlydata.org/ontology/doc/

[10] ORB: https://www.w3.org/2001/sw/hcls/notes/orb/

[11] SA: http://ns.science.ai/

The *scientific EXPeriments Ontology* (EXPO)[12] is a core ontology to provide a structured framework for scientific experiments by formalizing the generic concepts of experimental design, methodology, and results representation [251]. Discourse Elements Ontology (DEO) is an ontology for describing articles in terms of its main components such as Abstract, Introduction, Reference List and Figures [55]. Linked Science Core is an ontology for describing scholarly communication resources involving Publication, Researcher, Method, Hypothesis, and Conclusion [13]. The Semantic Survey Ontology (SemSur) is an ontology for capturing the content of survey articles involving research approaches, problems, implementations, publications and evaluations [82]. Due to the important role of bibliographic citations, there has been a number of ontologies developed only with this focus.

One of the widely used ontologies is *SPAR*[13], family of Semantic Publishing and Referencing ontologies. Two of the ontologies from SPAR family ontologies are focusing on citation information. The *Citation Typing Ontology* (CiTO)[14] provides a set of object properties related to citing published articles, such as “is cited by” and “cites” [209]. The *Bibliographic Ontology* (BIBO)[15] covers the main concepts and properties for describing citations and bibliographic references [59]. Scholarly artifacts other than publication witnessed a less amount of attention with regard to the ontologies developed for describing them. For example, there are few ontologies for scientific events still in a very preliminary status.

The *Scholarly Event Description Ontology* (SEDE)[16] is a comprehensive ontology for describing scholarly events in terms of agents (e.g., persons, committees), places (e.g., cities, venues) and time (e.g., start/end date). SEDE provides a basis to represent, collect, and share from scholarly event data [125]. The Semantic Web Dog Food (SWDF) dataset[17] and its successor *ScholarlyData* are among the pioneers of datasets of comprehensive scholarly communication metadata. There was hardly any ontology found specifically developed for modeling of the online courses. The Teaching Core Vocabulary [18] providing terms about course materials and documents. It is based on practical requirements set by providing seminar and course descriptions as Linked Data. Due to the broad spectrum of the scholarly domain, the addressed ontologies in this section are limited to the artifacts focused in this thesis (Publications, OCW, events).

2.2.2 Scholarly Metadata Extractors

Rule-based metadata extraction is of the widely used methods in extracting the fine-grained metadata. Most of the developed tools are using the upper part of the first pages of the publications as the actual source of the metadata. Constructing author profiles by extracting author names, affiliations, contact addresses, the research grant is one of the main use cases for article search services. These approaches perform promising with certain format and style of documents. The performance of the tools also highly affected by the name ambiguity. Machine-learning approaches are used with the purpose of extracting information from the complex and diverse type of documents. Different tools have been developed to deal with metadata extraction from different types d scholarly documents such as books, theses [89, 295]. However, the range of the metadata being extracted still stays limited to a particular set of properties. Citations play an important role in evaluating the research work of scientists and the impact of their work, extracting citation information. As a common style of publishing scientific documents, usually, the last page of articles are considered as the section pointing to the citations. A survey is done about the

[12] EXPO: `http://expo.sourceforge.net/`

[13] `http://www.sparontologies.net/`

[14] CiTO: `http://purl.org/spar/cito`

[15] BIBO: `http://bibliontology.com/`

[16] SEDE: `http://eventography.org/sede/`

[17] `http://data.semanticweb.org`

[18] `http://linkedscience.org/teach/ns#`

development of the citation extraction tools over the period of 2006-2010 [304]. They also analyze the coupling task of citation and impact indicators in the field of information science. The section is often called as “bibliography”, “references” or “resources”. In order to provide services evaluating research impact, co-citation networks have been built using citation extraction techniques. *SemEval 2018* is a challenge that has been held on semantic relation extraction and classification in scientific papers [117]. Neural Networks [292] have been used as the main method in the approaches proposed by the participants of the challenge. The reported results show the extraction of such instances is a challenging task. The challenges in obtaining high-quality metadata require working with a smaller corpus, dealing with specialized vocabularies. Moreover, the scarcity of annotated data and available domain-specific resources influence the quality of extracted metadata. In [152], the authors are automatically extracting 5178 terms from Wikipedia. They have collected the titles of all the mentioned scientific articles in these wiki pages and terms redirecting to them.

The extracted terms are used to categorize the publication under the fields of“physics”, “applied and interdisciplinary physics”, “theoretical physics”, “emerging technologies” and their direct subcategories. Citation networks of the pages and the articles are built to reveal scientific memes. In order to evaluate the performance of this approach, human annotations are used for categorization of the same publications. This work shows the wide spectrum of the possible analytics with scholarly metadata and citation networks. *FLUX-CIM* is a tool for extraction of citation metadata using flexible unsupervised techniques [56]. This approach does not rely on patterns encoding specific citation style and as claimed by the authors, it suffers from expensive training phase for the learning. In order to do so, a number of citation parsing tools have been developed. *ParsCit* is a similar tool that uses machine learning approaches for reference string extension. It was used in CiteSeer, however, it was limited to a certain training data with low scalability. In [202], researchers from CiteSeer mention the issues of the previously developed algorithms. It is explained that new methods are using Web crawling approaches and metadata integration in order to benefit from the already available scholarly metadata. *TeamBeam* is a scholarly metadata extractor that performs by classification of the text blocks [136]. Depending on the layout of the input article, the quality and diversity of extracted metadata vary. The TeamBeam algorithm exploits layout information and contains dictionaries for names. The algorithm used for extraction outperforms ParsCit. However, its performance depends on the number of metadata and only performs for a specific format of the article. Recently developed service namely *EXCITE*[19] is a generic tool for extracting reference information from scholarly documents in PF format. The still ongoing activities in this direction show the open challenges in metadata extraction. Due to the diversity of scholarly artifacts and the need for various properties and metadata being extracted, the approaches are also offered in wide ranges yet not comprehensive.

2.2.3 Datasets and Repositories

Several efforts on publishing reusable, machine-readable metadata (i.e. *linked open data*) related to scholarly data such as publication, scientific events, authors and etc, have been motivated quality considerations [244]. Our own ongoing work on extracting linked data from the CEUR-WS.org open access computer science workshop proceedings volumes is also motivated by quality assessment. We run a few dozen of quality-related queries such as “What workshops have changed their parent conference?” against the linked dataset in order to assess the quality of the workshops published at CEUR-WS.org and to validate different information extraction implementations [121, 170]. Both the work of Bryl, Birukou, Eckert and Kessler and ours have in common that they lack a systematic, comprehensive definition of quality dimensions. Currently, many RDF data are made available, the Semantic Web Dog Food (SWDF)

[19] https://github.com/exciteproject/

dataset[20] as one of the pioneers and *ScholarlyData*[21] that provides RDF dumps for scientific events. The Springer LOD dataset[22] about their conference proceedings (Lecture Notes in Computer Science) serves trust-related questions of stakeholders. Bryl, Birukou, Eckert and Kessler mention questions such as “Shall I submit a paper to this conference?”, and point out that the data that is required for answering such questions is not easily available but, e.g., hidden in conference management systems [40].

DBLP[23] is one of the most widely known bibliographic databases in computer science. It provides information mainly about publications but also considers related entities such as authors, editors, conference proceedings and journals. The metadata of events, deadlines, and subjects are out of the scope of the DBLP database. However, it allows event organizers to upload XML data with bibliographic data for ingestion. The dataset of DBLP is available in multiple formats as well as an RDF dump [24].

OpenAIRE[25], Open Access Infrastructure for Research in Europe, is an aggregator of scholarly metadata collected from thousands of repositories, libraries, institutes, publishers and individual data providers. OpenAIRE collects metadata about open access publication, projects and research datasets [175]. In addition, the schema of the OpenAIRE database management system covers metadata about the scholarly organization, people, software. *CORE*[26] is a gate to access research papers under FAIR principles. Metadata enrichment is done by text-mining approaches. Similar to OpenAIRE, CORE aggregates metadata about scientific papers from data providers from all over the world including institutional repositories, subject-repositories, and journal publishers. They call the process of collecting metadata, harvesting which allows CORE to offer search, text-mining and analytical services. CORE also collects the full-text of the research papers and applies text-mining in order to extract metadata. *SciGraph* similar to OpenAIRE aggregates data sources from Springer Nature and key partners from the scholarly domain.

Zenodo created by OpenAIRE and CERN, acts as a repository for research datasets from different disciplines. It enables any individual, scientific community or research institution to load their datasets freely. The users keep the ownership over their unique community collections. Upload allowance per each piece of data is 1GB. Metadata management and enrichment of the entries inside Zenodo is directly done by OpenAIRE. This makes Zenodo be able to offer a strong search facility. However, updates of the already existing data are not automatic and versioning requires specific steps to be done by the data provides together with the managers of Zenodo. Similarly, there are other online repositories in order to store and share scholarly artifacts. *FigShare* is an online digital repository of different kinds of scholarly artifacts including figures, datasets, images, and videos [247].

Different categories of data, publications, preprints and manuscripts can be uploaded by individuals before the process of peer review. In order to authors retain ownership, the Figshare repository makes data available under the Creative Commons Attribution License (CCAL). Sharing research results on such online repositories allows the authors to receive early feedback and may be helpful in revising and refining the article for final submission. Another example of this category is *DataHub*[27]. It is an free online tool to share and discover high quality datasets. Gradually, every community have built their own data repository such as PANGAEA[28] for earth and environment science, NCBI-PubMed[29] for medical science. More details on such repositories are out of the scope of this thesis. Therefore, more details in

[20] http://data.semanticweb.org/

[21] http://www.scholarlydata.org/dumps/

[22] http://www.lod.springer.com/

[23] http://www.dblp.uni-trier.de/

[24] http://www.dblp.l3s.de/d2r/

[25] https://www.openaire.eu/

[26] https://www.core.ac.uk/

[27] https://www.datahub.io/

[28] https://www.pangaea.de/

[29] https://www.ncbi.nlm.nih.gov/

this regard is skipped. Due to diversity and heterogeneity of such repositories, there are directories to support both repository administrators and service providers in sharing best practice and improving the quality of the repository infrastructure. *OpenDOAR* is one of the academic metadata aggregating tools of open access repositories. The main focus of OpenDOAR is to provide a quality assured list of artifacts which are openly accessible. Along with digitization, as ownership and credits to research works play an important role in the entire scholarly communication, having unique and persistent identifies became very important.

Crossref[30] is a cooperative effort to enable persistent cross-publisher citation linking. Citation data are not usually freely available to access, however, *OpenCitations* is a scholarly infrastructure organization that provides a Data Model for citation information and uses the SPAR (Semantic Publishing and Referencing) Ontologies for encoding scholarly bibliographic and citation data in RDF [210]. It has developed the OpenCitations Corpus (OCC) of open downloadable bibliographic and citation data recorded in RDF. OCC is a database that harvests citation data from Crossref and other sources. *Initiative for Open Citations* (I4OC)[31] makes citations available through a REST API.

2.2.4 Tools and Systems

In this section, the current services that are provided by different communities and organizations in order to serve the needs of stakeholders through the scholarly communication are captured. Due to the complexity of the scholarly communication pipeline and the broad range of stakeholders and their needs, we only focus on the relevant state-of-the-art for this thesis. The sections follow a discussion about the systems that provide services around the three artifacts focused on this thesis namely publications, online courses, and scientific events. Three types of services are selected to be reviewed in this section:

- **Social Networks** such as Facebook and Twitter have changed the way people and communities have been interacting with each other. These social networks are a new environment for communication and information sharing. Along with all the other domains, academic social networks emerged for the target group of researcher, students and all the stakeholders involved in scholarly communication. Social networks enhance the possibility of managing and disseminating scholarly ideas, results, events, and discoveries. Furthermore, they are influencing collaborations, education, and research. In this section, we focus on covering some of the most used and famous social networks in the context of scholarly communication.
- **Digital Libraries** are containers of digital collections which constitute of a significant number of documents in it combined with metadata for each. Such collections are organized by a group of people or organizations and classified according to a set of certain criteria. Usually, a centralized metadata repository is needed to closely couple with the collection of documents to store information about them. In the narrow sense, scholarly metadata repositories are in the similar usage of catalog cards of physical libraries. A digital library is a digital library system if and only if it contains a digital library management system and at least one collection of documents and at least one catalog of its content with an (optional) interface for offering search facilities.
- **Search Engines** are built of sets of programs which are used to search and collect for information within a specific realm. Typically, web search engines retrieve answer of queries by sending out a spider to fetch as many matches as possible. Another program is called an indexer which reads these matches and creates an index based on the words contained in each match. Each search engine works based on a specific algorithm to and the deal reaction is to return for meaningful

[30] `www.crossref.org`

[31] `www.i4oc.org`

> results for each query. A Search Engine is a set of programs in the shape of an integrated system that takes an entry query and browses indexed catalogs without offering any content and displays the matched results with the search keyword.

A huge amount of scholarly data is published after the appearance of the Web 2.0 as scholars are using social media in communicating with other community members about their research results and activities. One of the platforms is *ResearchGate* [32], a social network designed for researchers in order to create their scientific profiles, to list their publications among others and to interact with each other by sharing research results out of the official publishing limits. It also provides researchers with a functionality to create discussion groups, share updates, results, and resources with their networks, and internal search engine that allows users to search through major databases. In addition, researchers can upload their published articles onto their personal profile pages and access events such as scientific conferences. *Mendeley* [33] is also a social a desktop and web program produced by Elsevier for managing and sharing research papers. Researchers use Mendeley in discovering research data and collaborating online [85]. Although it was developed as a social network service, it also has several other features such as metadata extraction, RDF viewers, search facilities and citation management.

Academic.edu and *VIVO* [34] are other examples of social networks for scholars. Academia.edu is the platform for scholars to share their research, monitor deep analytics around the impact of their research, and track the research of academics they follow in specific fields. Since its inception in September 2008, over 22 million users signed up and added about 6 million papers and 1.5 million research interests. It also attracts over 36 million unique visitors per month. VIVO is developed for recording, editing, searching, browsing, and visualizing.

There are some digital libraries (DL) which are hosted by organizations with different initial goals. Most of DL(s) in this category are constructed to manage the documents of various specific subjects such as ACM[35] and IEEE Xplore [36] digital library. The ACM digital library is a comprehensive collection of full-text articles and their bibliographic information. It provides search facilities on top of the library. Special research institutes are granted with free access to the articles, otherwise, access is under close licenses. IEEE Xplore is the online indexing facility for material published by IEEE. Access to the material requires a subscription and is under close and payment-based permission. Some type of digital libraries are collections of digitized documents of libraries to survive the old versions and make the content of the libraries remotely available. Elsevier as one of the active multidisciplinary publishers provides several services as well as digital libraries such as ScienceDirect [37] and Scopus [38]. ScienceDirect is a large collection of scientific and medical research and provides access to full-text articles. Both services require subscription and metadata of authors and citations are the only information provided is search features. Both have subscription-based access to the material and pay-per-view purchase. Web Of Science [39] has a similar access policies for the collection of articles provided by Thomson Reuters publisher[40]. It can also be considered as a citation indexing system with a search facility on top of the underlying collection. Another special digital library in this category which is built in the purpose of a

[32] `https://www.researchgate.net/`
[33] `https://www.mendeley.com/`
[34] `https://www.vivoweb.org/`
[35] `http://dl.acm.org/`
[36] `https://ieeexplore.ieee.org/Xplore/home.jsp`
[37] `https://sciencedirect.com/`
[38] `https://www.scopus.com`
[39] `https://clarivate.com/products/web-of-science/`
[40] `https://thomsonreuters.com/en.html`

digital library and directed by an informal steering committee is NDLTD [41](Networked Digital Library of Theses and Dissertations).

Other digital libraries investigated in this thesis are namely the digital library of Congress [42], JeromeDL[43], BUILD-ER[44]. Google Scholar [45] is an online, freely accessible search engine that realized in 2004 and allows users to look for both physical and digital copies of articles. It searches a wide variety of sources, including academic publishers, universities, and preprint depositories looking for Peer-reviewed articles, Theses, Books, Technical reports, Abstracts, Reprints. The regular version of Google crawls over web pages but Google Scholar gets the data from three sources. Google Scholar is in cooperation with most of the scientific communities such as publishers, institutes, societies. They are the first source which provides scholarly materials, abstract, and citation data which is not available via regular Google search. Second, Google Scholar uses an algorithm which runs over the Internet to identify web documents that look scholarly and are publicly available. The third source is the reference part of the content of the scholarly documents.

More recently, Google Scholar has added a feature that allows authors to take control of their own publications enabling these to be presented as a corpus of work. The facility includes the ability to include keywords that allows grouping of authors, although there is no control of these keywords or linking of similar terms. The other search engine is CiteSeerX [46] which is developed as a specific service for the computer science domain in order to explore the scholarly artifacts. Although it might be considered as a digital library as it makes the access to the PDF format of the searched document, it retrieves from metadata and data from cached pages. Therefore, we categorize it under the currently available search engines for scholarly documents. Aminer [47] is a mining and search engine service for researchers. They can create profiles and track their publishing records. Aminer provides a graphical view of statistics about individual researchers. It provides advanced search facilities in order to explore metadata about authors, publications, events, citations and research topics. Inside Aminer, a ranking system is developed that collects information and calculates h-index for all the considered artifacts. The whole metadata set of Aminer is freely available for developers and service providers of scholarly communication. The Bielefeld Academic Search Engine (*BASE* [48]) runs over massive academic web resources. It has a faceted browser and uses an indexing technique for retrieving the metadata. BASE provides access to the full texts of about 60 percent of the indexed documents for free (Open Access).

There is a number of search engines developed for exploring metadata about scientific events. *CFP Manager* [123] is an information extraction tool specific to the domain of computer science; it extracts metadata of events from an unstructured text representation of CfPs. Because of the different representations and terminologies of CfPs across research communities, this approach requires domain-specific implementations. The extracted data is limited to the keywords used in the content of CfPs. In addition, the CFP Manager does not support data curation workflows involving multiple stakeholders. Hurtado Martin et al. proposed an approach based on user profiles, which takes a scholar's recent publication list and recommends related CfPs using content analysis [118]. Xia et al. presented a classification method to filter CfPs by social tagging [299]. Wang et al. proposed another approach to classify CfPs by implementing three different methods but focus on comparing the classification methods rather than

[41] `http://www.ndltd.org/`

[42] `http://loc.gov/library/libarch-digital.html`

[43] `http://jeromedl.org/`

[44] `http://builder.bham.ac.uk`

[45] `https://scholar.google.de/`

[46] `http://www.citeseerx.ist.psu.edu/`

[47] `https://www.aminer.org/`

[48] `https://www.base-search.net/`

services to improve scientific communication [281]. *DBWorld*[49] collects data about upcoming events and other announcements in the field of databases and information systems. Each record comprises event title, deadline, event homepage and the full-text description. *WikiCFP*[50] is a popular service for publishing CfPs. Similar to DBWorld, WikiCFP only supports a limited set of structured event metadata (title, dates, deadlines), which results in limited search and exploration functionality. WikiCFP employs crawlers to track high-profile conferences. Although WikiCFP claims to be a semantic wiki, there is no collaborative authoring, versioning, minimal structure and the data is not downloadable as RDF or accessible via a SPARQL endpoint.

Cfplist[51] works similar to WikiCFP but focuses on social science related subjects. Data is contributed by the community using an online form. *SemanticScholar*[52] offers a keyword-based search facility that shows metadata about publications and authors. It uses artificial intelligence methods in the back-end and retrieves results based on highly relevant hits with the possibility of filtering. *Conference.city*[53] is a new service initialized in 2016 that lists upcoming conferences by location. For each conference, title, date, deadline, location, and number of views of its conference.city page is shown. Based on the location of the conference, Google plug-ins are used to recommend flights, accommodation, and restaurants. The service collects data mainly from event homepages and from mailing lists. In addition, it allows users to add a conference using a form.

PapersInvited[54] focuses on collecting CfPs from event organizers and attracting potential participants who already have access to the ProQuest service[55]. ProQuest acts as a hub between repositories holding rich and diverse scholarly data. The collected data is not made available to the public. The ISO 20121 international standard supports organizers of events of all types – sports, business, culture, politics – in integrating sustainability with their activities.[56] The standard provides general guidelines but also mentions some of the metrics of our model, such as event sponsoring registration methods and other types of financial support of events.

The currently available services for open educational resources and OCW are not developed as much as the services for other scholarly artifacts. A thorough search of the literature indicates that work related to OCW quality assessment is still rather scarce. Most of the previous works consider repositories and their impact on education rather than quality of courses. In [279], a set of quality assurance criteria is introduced considering four aspects of OCW: 1. content, 2. instructional design, 3. technology and 4. courseware evaluation. About half of the dimensions that we consider in this work (such as availability, multilinguality) are also introduced in Vlădoiu's work. However, some of them are not considered in this work because either they are subjective (e.g., self-containedness) or difficult to measure (e.g., relevance of the content for self-learning) or out of the scope of assessing course quality (e.g., interoperability of the interface). In [187], a machine learning approach has been devised to support automatic OCW quality assessments. A problem here, however, is the availability of suitable training data, which could be provided by expert sample assessments obtained using the methodology presented in this paper.

The University of Berlin and the university of MIT also invested in OCW services. The OCW services names OpenHPI [57] from Berlin university provides online free course with multimedia materials. The courses are designed along the semester period and include exercise and exam material. High-quality

[49] https://www.research.cs.wisc.edu/dbworld/

[50] http://www.wikicfp.com/

[51] https://www.cfplist.com/

[52] https://www.semanticscholar.org

[53] http://www.conference.city/

[54] http://www.papersinvited.com/

[55] http://www.proquest.com/

[56] http://www.iso.org/iso/news.htm?refid=Ref1598

[57] https://www.open.hpi.de/

Service Name	Addressed Entities	Accessibility of Artifacts	Quality Criteria	Community Contribution	Advanced Search	Publishing Metadata
ACM DL	Publications	✗	✗	✗	✗	✗
IEEEx DL	Publications	✗	✗	✗	✗	✗
Arnetminer	Publications	✗	✗	✗	✗	✗
arXive	Publications	✓	✗	✗	✗	✗
DBLP	Publications	✗	✗	✗	✗	✓
Google Scholar	Publications	✗	✗	✗	✗	✗
Mendeley	Publications	✓	✗	✓	✗	✗
ScienceDirect	Publications	✓	✗	✗	✗	✗
Scopus	Publications	✗	✗	✗	✗	✗
WOS	Publications	✗	✗	✗	✗	✗
OpanAIRE	Publications, Datasets	✓	✗	✗	✗	✓
ResearchGate	Publications	✓	✗	✓	✗	✗
SciGraph	Publications	✗	✗	✗	✗	✗
CFP Manager	Events	-	✗	✗	✗	✗
Zenodo	Publications, Datasets	✗	✗	✗	✗	✗
VIVO	Publications	✓	✗	✗	✗	✗
CEUR-WS	Publications	✓	✗	✗	✗	✗
SemanticScholar	Publications	✓	✗	✗	✗	✗
Springer LOD	Events	-	✗	✗	✗	✓
ProQuest	Publications	✓	✗	✗	✗	✗
PubMed	Publications	✓	✗	✗	✗	✗
OpanDOAR	Publications	✓	✗	✗	✗	✗
SPAR	Publications	✓	✗	✗	✗	✗
OpenCitations	Publications	✓	✗	✗	✗	✓
CrossRef	Publications	✓	✗	✗	✗	✗
I4OC	Publications	✓	✗	✗	✗	✓
EasyChair	Publications, Events	-	✗	✗	✗	✗
DBWorld	Events	-	✗	✗	✗	✗
CORE	Events	-	✗	✓	✗	✗
CFP Manager	Events	-	✗	✗	✗	✗
BASE	Publications	✓	✗	✗	✗	✗
Academic Search	Publications	✓	✗	✗	✗	✗
SlideWiki	OpenCourseWare	✓	✗	✗	✗	✗
OpenHPI	OpenCourseWare	✓	✗	✗	✗	✗
GSM	Events	-	✓	✗	✗	✗
MIT OER	OpenCourseWare	✓	✗	✗	✗	✗
Jorum	OpenCourseWare	✓	✗	✗	✗	✗
OER Commons	OpenCourseWare	✓	✗	✗	✗	✗
Temoa	OpenCourseWare	✓	✓	✗	✗	✗
WikiCFP	Events	-	✗	✓	✗	✗
Cfplist	Events	-	✗	✗	✓	✗
CiteSeer	Publications, Events	✗	✗	✗	✗	✓
ScholarlyData	Events	-	✗	✗	✗	✓
Springer LOD	Events	-	✗	✗	✗	✓
Conference.city	Events	-	✗	✗	✗	✗
PapersInvited	Events	-	✗	✗	✗	✗
SemanticScholar	Publication, Persons	✗	✗	✗	✗	✗

Table 2.1: **Comparison of existing services**. A selective group of services has been analysed based on their support in accessibility, quality, community involvement, search options and availability of metadata.

videos are recorded from the lecturers. However, the metadata availability and search functions are not fully semantified. MIT OpenCourseWare [58] only provides text material for courses as well as syllabus exam material. The students can train themselves by using such material. However, grading system and value of courses remain out of the system to the physical world. These courses are all made for practice. OER Commons [59] quoted “he worldwide OER movement is rooted in the human right to access

[58] http://www.ocw.mit.edu/index.html

[59] http://www.oercommons.org/

high-quality education". Quality issues are critical for scholarly artifacts. It provides the subscribers a platform to create, share, and discuss resources with others in your network. More than 250 institutions worldwide are openly publishing courses today. Further OCW repositories have been made available by organizations and will be discussed in section 4.3.

The Open Education Consortium[60] as the central community of institutions and organizations aggregates open education lists 26,611 courses from 80 providers. Many of the repositories mentioned above are members of the Open Education Consortium. MIT OpenCourseWare as one of the popular OCW repositories reports that they have made 2,150 courses available so far. Since its launch, 137 million people have been visiting MIT OpenCourseWare annually. OpenCourseWare Consortium [61] is an organization to provide policies in support and advance openness in education around the world. Many of the OER and OCW services work under the framework of the OpenCourseWare Consortium. The current visible problem of OER services is that accessibility has been considered as the main challenge. Quality of OCW has not been seriously taken into account. Therefore, similar to other scholarly artifacts OCW users suffer from quality issues. In addition, quality related metadata of courses are also not specifically made available for the users and the developers. Temoa [62] is a knowledge hub that eases a public and multilingual catalog of OCW. Temoa supports the users to find resources and materials based in their needs for teaching and learning (fitness for use).

It is possible to search the OCW based on the material type such as video, text, audio. It provides a faceted browser in order to filter results based on several metrics. However, Temoa does not provide information about multilingualism, license, and content of the OCW. Xpert [63] is an integrating system that contains metadata and resources from data providers. Bulk import of material is made available for registered users. It has a single search button that can take several keywords and retrieves the results based on query terms. Learning material in Xpert is tagged by the corresponding science domain and topics. However, quality metrics are not considered in this service also. *SlideWiki* [64] is a collaborative platform for creating educational material in slide presentation. It is based on semantic technologies and provides features to ease multilinguality, license assurance, linking information. It also provides an interactive environment for students and teachers.

Table 2.1 shows a selected list of most relevant services for this thesis categorized by the artifacts they focus. So far the development of services was discussed which has happened independently of the criticism of the overall process of current scientific communication. The current paradigm of scholarly communication comprises of specific steps such as: preparation of manuscripts, organization of publishing channels, peer-review process and publishing [30]. However, analysis of the state-of-the-art services shows the main focus of service providers and the scientific communities has been on *how* research articles are evaluated and disseminated. The services are marked with regard to the extent they support accessibility, quality, collaborative creation of artifacts and community involvement, search options and availability of metadata. This gap shows the inadequacy of current services in providing comprehensive support for researchers with regard to the life-cycle stages of discovering, integrating, sharing, evaluating and re-using metadata about scholarly artifacts. In order to realize modern scientific communication exploiting the potential of digitization, stakeholders involved in scientific communication, from researchers to policymakers, publishers, etc., require a joint and comprehensive reference model for scholarly metadata management.

[60] http://www.oeconsortium.org/

[61] http://www.ocwconsortium.org/

[62] http://www.temoa.info/

[63] http://www.xpert.nottingham.ac.uk/

[64] https://www.slidewiki.org/

CHAPTER 3

Metadata Integration and Management

Metadata-driven technologies and smart infrastructures for metadata management and analytics are increasingly improving web-based services for a broad range of application domains. The enormous amount of data generated day by day, demands the development of integration, management approaches to provide high quality and accurate analytics. The focus of this research is on using a metadata life cycle as a methodology for metadata management steps in order to transform such information into the actionable knowledge that enables such useful analytics. Artifacts as the subjects of their metadata are engaging in the scope of this research with regards to specific criteria, e.g., quality or how their metadata are "FAIRly" manageable.

In order to follow the discussions, section 3.1 discusses the concept of "metadata" and the different perspectives of research communities (mainly the Semantic Web community and the Database Community). In this chapter, a methodology in the form of a life cycle is introduced to structure the required steps for metadata management. There are already several attempts using a life cycle structure in order to define (meta)data management steps for different purposes. The section 3.3 discusses the life cycle structure as a suitable methodology for the objective of this research. section 3.2 presents the required technical foundations before introducing the proposed life cycle for scholarly metadata management in section 3.4.

3.1 Data and Metadata

The two key concepts of data and metadata have been used by scholars mostly without specific borderlines. In some cases, the two terms utilized interchangeably. From an ontological point of view, the term "data" is the plural [1] form of the Latin word "datum" [70]. The term "datum" means "a piece of information" or "something to give" [50]. The term "metadata" is a modified version of the term "data" with a Greek prefix that means "after", "behind", or "higher" [100] and used for emphasizing on transcending a concept such as metamathematics and metatheories, which are mathematical theories about other mathematical theories. A common interpretation of "metadata" is data that provides information about other data [215]. However, this definition neither encapsulates the full scope of the term metadata nor differentiates it from the concept of data. This section includes discussions to clarify the notation, origin and perspectives of research communities on the meaning of the two terms,"data and metadata". The already existing definitions about the two concepts of "data" and "metadata" have been explored and collocated from scientific literature and standards. A set of characteristics have been derived from the highlighted definitions to clarify the meaning of the term "metadata" within the scope of this thesis.

[1] It is widely recommended to use the term in a plural form. also followed in this thesis (same for the term metadata).

Data is a broadly used term by almost all science disciplines. In a general sense, the term "data" refers to the representation of the real world objects. In the digital era, such representations are considered to be in the form of numbers, characters, symbols or signals of binary codes etc., to be given to a system or a computer. In the pyramid of *Data*, *Information*, *Knowledge*, and *Wisdom* (DIKW) [230], "informaton" is defined as inferred descriptions with meaning from data. The concept of "knowledge' is considered as information having been processed or structured. Knowledge can provide a framework for inference of new information from a mix of experimental results, contextual information, etc. "Wisdom" is articulated as integrated knowledge.

Research communities are widely using the term "data" to express the application domain they deal with e.g., medical data, bibliographic data. Greenberg in the Encyclopedia of Library and Information Sciences (ELIS) [98] considers "data objects are ranging from information resources, such as a monograph, newspaper, or photograph – to activities, events, persons, places, structures, transactions, relationships, execution directions, and programmatic applications etc." Instance of individual objects are represented in smaller granularity as data elements [84]. A data element is the smallest unit of a class of objects captured and represented by specific attributes [157].

Exploration of the literature shows the meaning of the term "data" have been influenced by the development and usage of information systems over time. As a trend in the computer science field, the term "data" is combined with certain adjectives to emphasize the specific characteristics about the structure of the data e.g., relational model of data, linked data, big data, and smart data. In terms of structure, "Data" can be divided into three categories:

- Structured data are modelled and organized either in the form of tables or in some other way. The searching and accessing information from such type of data are easily facilitated e.g., data stored in the relational databases [165].
- Unstructured data are unorganized and require advance tools and software to access information e.g., web contents, wikis and emails [33].
- Semi–Structured data are basically structured data that are unorganized. For example, JSON (JavaScript Object Notation) files, BibTex files, CSV files, XML and other markup languages are the examples of semi–structured data found on the Web [43].

Relational model data is structured representation of data in terms of tuples (rows) and attributes (columns), grouped into relations (tables) [53]. From the classical databases and data management point of view, data have been mostly considered as structured representation of entities. Relational database management systems (RDBMSs) are originally designed for a single server in order to maintain the integrity of the table mappings and avoid the problems of distributed computing [3]. RDBMSs are designed for low–latency retrieval or update of data. With the massive data present on the Web and the new application domains with requiring immediate processing, traditional RDBMS do not scale to be used for large analysis.

Linked data refers to a specific representation of structured data including semantics that enable interlinking in the scale of the Web data [28]. The Semantic Web is a Web of Data. Linked data provides freely available data on the Web that are identified by the Uniform Resource Identifier. Such data can be retrieved directly via HTTP. The exponential growth of the information exchange through innovative platforms on the Web over the past years brought a surge in creation of diverse data types in huge quantities. The Web of Data employs Linked Data standards see section 3.2, i.e., RDF data model (Resource Description Framework) as a *lingua franca* for knowledge representation, SPARQL as a query language for RDF, and the Web Ontology Language (OWL) as a logical formalism to encode ontologies.

Big data is about huge quantities of information represented in heterogeneous data structures. The five dimensions of heterogeneity of big data (Volume, Velocity, Variety, Value, and Validity) are discussed in chapter 1. Nowadays, the concept of "data" covers text documents, multimedia content, audio and video, as well as log files and recorded web activities which are mostly in unstructured formats. Big data statistics of the year 2017 shows 80% of the available data on the web is unstructured while only 20% is structured [233]. Data with such characteristics are too complex for traditional data management systems. Therefore, management steps for big data starting from capturing data, sharing, querying to visualization follow a different process. The MapReduce programming model has been introduced to handle such data with parallel processing of data [155].

Smart data is focused on representation of valuable and actionable information [291] often with immediate analysis on dynamic data. Big and smart data are designed for zero-latency in data processing and information retrieval. The data on the Web is contently changing. For the applications with requirements on continuous decision making, meaningful insights from data becomes crucial. Such smart analysis heavily depend on format, characteristics and temporality of data and requires specific management than traditional system.

Data (in every format) are quantified, counted, collected or measured through experimental activities by either human or a series of automated processes by machines [178]. The output of computation and processing activities are also "data" as derived information from the input "data". A comprehensive data management model can be defined based on the characteristics of the underlying data see section 3.4

Stated By	Quotation
ISO [122]	*"Re-interpretable **representation** of information in a formalized manner suitable for communication, interpretation or processing".*
UNECE [54]	*"The physical **representation** of information in a manner suitable for communication, interpretation, or processing by human beings or by automatic means."*
IFIP [286]	*"A **representation** of facts or ideas in a formalized manner capable of being communicated or manipulated by some process".*
OECD [76]	*"A **representation** of facts and concepts in a formalized manner suitable for communication, interpretation or processing by human beings or by automatic means".*
Landry [157]	*"Facts that are the **result of observation or measurement**".*
Guptill [101]	*"Generally viewed as elements that model or **represent** real-world phenomena".*
Bequai [17]	*"Any **representation** of fact or idea in a form that is capable of being communicated or manipulated by some process".*

Table 3.1: **Data.** Collected statements and quotations about the term "data" from literature and standards.

Table 3.1 shows a selective list of quotations about the term data given by researchers of different communities such as librarians and computer scientists. Most of these attempts goes back to the classic interpretation of the term in the field of computer science in the efficiency of the retrieving, compressing and storing information. One of the common characteristics of data given by the definitions is to represent facts or elements.

As a general conclusion of the presented quotations, the concept of "data" is considered as representation of prime objects of focus from real world. Thus, the working assumption of this thesis is that any representation of real world objects including digitized versions of artifacts are considered as data e.g., digitized documents, videos, figures, evaluation datasets.

Metadata – Notion and Origin The term "metadata" is known to have been in use since the late 1960s by scholars of data management and statistics communities [98]. One of the early usages of the term was documented in 1968 in Philip Bagley's book on *Extension of programming language concepts* [116].

Howe describes meta level data from data container and storage point of view. The first appearance of the term in a scientific publication is in the dissertation of Sundgren [258], where several meta level concepts for data management such as “metadata” and “metainformation” are introduced [252].

In the early 1990s, NASA’s Global Change Master Directory (GCMD) [34] released Earth and space as a data collection to describe data where users can understand what the data is about. NASA started to report the huge investigations on managing “data about data” and the challenges encountered with managemnet of such data [34]. At the same time, DARPA was investing on the languages for representing data and early proposals of “infradata” [2] had been suggested [83]. A combination of all the challenges encountered with the previous definitions of the term “metadata” and needs emerging from these challenges led to the adoption of the concept of “metadata” in the metadata registry standard. Since then, the term “metadata” has been generally (not exclusively) associated with digital information systems and related topics.

To address all the issues defined around this concept, a metadata workshop was held by the National Center for Supercomputing Applications (NCSA) [288]: As a result, Dublin Core (DC), a set of vocabulary terms about metadata elements, was described in a publication by the participants. The initial version of Dublin Core was to facilitate the discovery of objects in a networked environment. This was in parallel with a proposal of a generic metadata model for the World Wide Web [23] , the starting point of the Semantic Wed. The Dublin Core Metadata Initiative (DCMI) combined both ideas and made use of DC vocabularies and defined more domain specific terms in a broader scope [119]. By embedding DC metadata into web pages, the precision of information discovery improved. For example, Web–based search services adopted such vocabularies and and increased information retrieval.

Nowadays, metadata is inevitable to be mentioned when it comes to digital infrastructures and systems preserving and supporting discovery, access, and use of information. Today, it plays a vital role in information communication and discovery, especially on the Web as it became the main information dissemination channel. This is evidenced by the wide range of developments and implementation of data integration tools and digital asset management systems and increasing need of enterprise applications [99].

Metadata – Description and Quotations The most common definition of the term “metadata” is “data about data” [215]. As the two terms of “data” and “information” have been also used interchangeably (despite of their distinct meanings), it is not excluded to define “metadata” as “information about information” or “data about information” or “information about data”. It means metadata represent informative and relevant details of the underlying data.

Codd proposed a systematic use of a relational model for organizing data that became the foundation of relational databases. In this classical view of representing data by means of the theory of relations , the database’s schema (data about data) is stored in “data dictionary”, and is disconnected from the database as the information storage (data) [53]. However, the old paradigm of information dissemination has changed with the Web. As stated by Greenberg, considering the Web and Internet, metadata management and related activities are far beyond the simple information cataloging [99]. Metadata are data about data where “data” refers to any resources as a prime interest of the observer(s) [98]. Such resources cover data or digital and physical objects. This is in contradiction with the classical view on the two concepts of “data” and “metadata” and creates disagreements. From a data manager perspective, “metadata” is “data about data” which only means the schematic information. Whereas, from a data scientist point of view, “metadata” is “data about data” which means both schematic alongside instance level information. Table 3.3 shows a comparison of data and metadata with regards to the defined list of defined characteristics.

[2] Infradata is a special kind of metadata in a networked infrastructure.

Stated by	Quotation
ISO [122]	*“Metadata is data that **defines and describes other data**.”*
Scott [105]	*“Metadata is a love note to the people and machines after you.”*
Liu [222]	*“Metadata is a **semiotics framework for analyzing data provenance** research.”*
NISO [215]	*“Metadata is **structured information** that describes, explains, locates, and makes it easier to retrieve, use, or manage an information resource.”*
McCarthy [178]	*“Metadata is **descriptive information about data** contents and organization.”*
Bergman [20]	*“Metadata is data providing information about one or more aspects of the source data, thus **data about data**.”*

Table 3.2: **Metadata.** Collected statements and quotations about the term “metadata” from literature and standards.

A list of criteria is deducted from the proposed and accepted definitions framework of characteristics [101] and are adopted to the FAIR principles [293]. The following explanation of the FAIR principles is taken from the guidelines.

To be findable:

- “F1. (meta)data are assigned a globally unique and eternally persistent identifier.”
- “F2. data are described with rich metadata.”
- “F3. (meta)data are registered or indexed in a searchable resource.”
- “F4. metadata specify the data identifier.”

To be accessible:

- “A1 (meta)data are retrievable by their identifier using a standardized communications protocol.”
  - “A1.1 the protocol is open, free, and universally implementable.”
  - “A1.2 the protocol allows for an authentication and authorization procedure, where necessary.”
- “A2 metadata are accessible, even when the data are no longer available.”

To be interoperable:

- “I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.”
- “I2. (meta)data use vocabularies that follow FAIR principles.”
- “I3. (meta)data include qualified references to other (meta)data.”

To be re-usable:

- “R1. meta(data) have a plurality of accurate and relevant attributes.”
  - “R1.1. (meta)data are released with a clear and accessible data usage license.”
  - “R1.2. (meta)data are associated with their provenance.”
  - “R1.3. (meta)data meet domain-relevant community standards.”

Criteria	FAIR	Data	Metadata
Discovery	Findable	Data being openly available can not influence discovery of its own or underlying facts.	Metadata makes underlying data more discoverable [172].
Accessibility	Accessible	Data needs to acquire an identified set of data.	Metadata is an identified set of data about another data.
Quality	Interoperable	Fitness of data for use from representation, completeness maintenance point of views.	Metadata determines if a set of data meets a specified need [101].
Representation	Interoperable	Data refers to physical elements in the realty.	Metadata refers to digital or physical objects.
Structure	Reusable	Completely unstructured digitized artifacts can be considered as data.	Metadata has at least a basic level of structured representation of the referred elements.

Table 3.3: **Metadata Characteristics.** Characteristics of the term metadata is summarized and classified with FAIR princples.

As stated in the guidelines:"The Principles define characteristics that contemporary data resources, tools, vocabularies and infrastructures should exhibit to assist discovery and reuse by third-parties. FAIRness is a prerequisite for proper data management." The Table 3.3 shows the classification of the metadata and data characteristics based on the FAIR principles.

As depicted in Figure 3.1, "metadata" are generated in two possible ways: (a) Sequential Order of Data and Metadata. Data and metadata can be distinguished depending on the status of an observer. Real world objects are the artifacts under the focus of observation from which raw information have been collected. Data elements are representation of such real world objects. Considering the data already in the abstract level, metadata represent information about data i the previous level. (b) Metadata collected directly from data resources either from real world by human using forms, or through automated application directly from data resources.

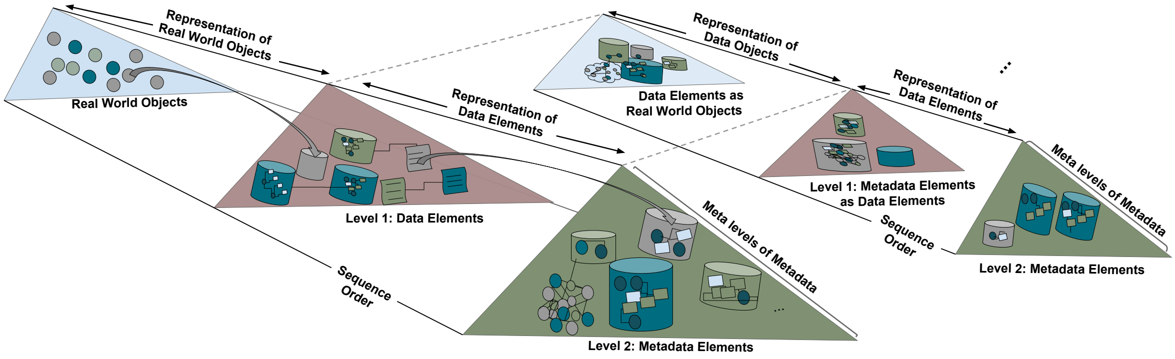

Representation of Real World Objects
Real World Objects
Representation of Data Elements
Level 1: Data Elements
Sequence Order
Meta levels of Metadata
Level 2: Metadata Elements
...
Representation of Data Objects
Data Elements as Real World Objects
Representation of Data Elements
...
Level 1: Metadata Elements as Data Elements
Sequence Order
Meta levels of Metadata
Level 2: Metadata Elements

Figure 3.1: **Data and Metadata Generation**. Metadata are generated either in sequential order after having data or they can be directly collected from data resources.

In a basic level of observation, data represent elements as raw values collected from real world domains. Metadata represent information about the underlying data in a second abstract level. State of the real world can change depending on the view of the observant. The already defined or collected metadata can be considered as real world in the next level. Thus, meta level of already created metadata defines the concepts of "data" and "metadata". Therefore, "metadata are data about data" where:

1. the real world objects have already been observed at least once and represented as data,
2. previous representation of such data is captured as metadata and
3. observant defines a meta level on top of the abstractions of the objects of prime interest.

Metadata describe a wide variety of information about the underlying resource. Thus, metadata increase consistency and maintenance of represented data elements. Search engines of the Web are mainly keyword-based. Metadata increases visibility of digital artifacts (documents, videos, images etc.) by providing identifiers. Therefore, discovery of such resources by the right consumers (machines or human) increases. Furthermore, the embedded semantics and knowledge in textual documents are impossible to be discovered without metadata. Without metadata, resources are isolated information stored in separated silos on the Web. Metadata allows a resource to be understood by both humans and machines. For example, the metadata elements are crucial for machines in order to automatically discovering and connecting with suitable Web application programming interfaces (APIs). Explicit knowledge about the structure and datatypes of such APIs are required to be made available under certain criteria such as the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Analytics are outcome of a series of human–machine metadata life cycle.

In this thesis, metadata are are considered as data about data where the second "data" refers to resources of a prime interest. Describing a resource with metadata increases its discovery, promotes interoperability across systems and facilitates integration with other relevant information [215]. The discussed characteristics of the concept "metadata" represents "Big Data" characteristics which is a virtual representation of each real–world entity captured and stored in data sources. The complexity of such data is known as the *6Vs* as characteristics of Big Data [249]. The term Big Scholarly Data (BSD) [297] is coined to represent the vast amount of information about scholarly networks including stakeholders and artifacts such as authors, organizers, papers, citations, figures. The heterogeneity and complexity of data and their associated metadata distributed on the Web bring new issues and challenges with respect to general semantic interoperability issues. In order to further proceed with the vision of this research in management of scholarly metadata, it is required to understand the main characteristics of the available metadata sources and integration issues.

The 6Vs of Scholarly Metadata The following is the detailed explanation of the 6Vs for big scholarly metadata which was already discussed.

- Volume refers to ability to ingest and store very large datasets; in the context of scholarly metadata, at least over 114 million scholarly documents [137] are recorded in 2014 to be available in PDF formats. In computer science, the total number of publications of the different types is reaching 4 million [224]. Different types of publication in different formats is being published every day in other scientific disciplines, more details have been discussed in subsection 2.1.1.
- Velocity denotes the growth rate generating such data; the average growth rate of scientific publishing is measured as 8 to 9 percentage [37].
- Variety indicates multiple data formats and models, the domain of scholarly communication is a complex domain [15] including many different types of entities with complexity interrelationship among them.
- Value concerns the impact of high quality analytics over data; as discussed in subsection 2.1.4, certain facts play enormously important role in the reputation and basically life of research stakeholders. Providing precise and comprehensive statistics supports researchers with already

existing success measurement tools such as number of citations. In additions, deep and mined knowledge with flexible analytics can provide new insights about artifacts and people involved in the scholarly communication domain.

- Veracity refers to the biases, ambiguities, and noise in data; this characteristic is especially applicable in the context of scholarly communication domain due to deduplication problems [176] and ambiguity problem for various scholarly artifacts as well as person names.
- Variability of the meaning of the metadata [297].

Semantic Interoperability Conflicts Transformation of scholarly metadata into knowledge will enable domain understanding and providing better services for the scholarly communities. Knowledge graphs enables integration of such metadata sources, which evolve over time and can reach large dimensions [87]. However, in order to integrate such resources in a unified way, semantic interoerability challenges need to be studied [96]. The aforementioned heterogeneity and complexity characteristics of the scholarly metadata leads integration and interoperability challenges. That affects ability of an underlying metadata management system or infrastructure to be engaged in the ongoing activity process of other system. Such issues originates from difference in modeling of the same real world entities, representation of different or same entities in various formats.

A systematic categorization of interoperability issues have been introduced in [19]. The following is the identified and adopted categories of interoperability issues in the context of scholarly metadata:

- Structure: scholarly metadata among the already existing resources are described in different formats e.g., structured, unstructured or semi-structured data.
- Schema: metadata resources of the scholarly communication domain are using different schemas for modeling of the entities, attribution and relationships. Due to complexity of the domain, schematic issues also include modeling conflict in attributes and classes e.g., properties of one modeling are entities of the other model.
- Domain: various interpretations of the same domain can occur. For example the same term can be used for different meanings or different terms can be used for the same concept. As a common practice, different acronyms are given to the same concept.
- Representation: while collecting or representation scholarly metadata, various granularity levels of details for the same concepts can be captured by different resources. Scholarly metadata are represented in different scales and units and languages.

The heterogeneous scholarly metadata published on the Web are disconnected. As a preliminary step towards serving information needs of the target users (scholars), scholarly metadata on the Web need to follow the FAIR principles. As a summary, metadata is created to improve resource discovery, resource management and content rating. It is also recorded for other reasons including administrative control, security and preservation. In the context of this research, metadata management is aimed with the purpose of improving quality, discovery and interoperability of resources.

3.2 Technical Foundations

The Semantic Web technologies are employed as technical foundations of this thesis. Therefore, we introduce the Resource Description Framework (RDF), the data model used as the underlying representation of Semantic Web data. The discussion follows by the introduction of SPARQL, the querying language of RDF data. This research aims at using Semantic Web technologies to collaboratively create and curate a scholarly knowledge graph, the definition of knowledge graphs in the context of information management and retrieval. Furthermore, we describe the Wikibase software that is used for community involvement in metadata collection and curation.

Resource Description Framework (RDF) The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications[3], originally designed as a metadata data model. It is a semantic graph–based data model tailored for representing semi–structured data on the Web. The main building block of RDF is a triple. RDF triple consists of three elements shown in Figure 3.2: subject, predicate, object. Subject denotes a resource to which predicate and object belong; only URIs or blank nodes can be subjects in RDF. Predicate denotes a trait of the subject, i.e., a relationship between the subject and object; only URIs can be predicates in RDF. Object specifies predicate with a particular value; URIs, blank nodes or string literals can represent a value of the predicate. As anything can be the subject of multiple triples at the same time, and often also the object of other triples, things are becoming connected with each other in a network structure in graphs.

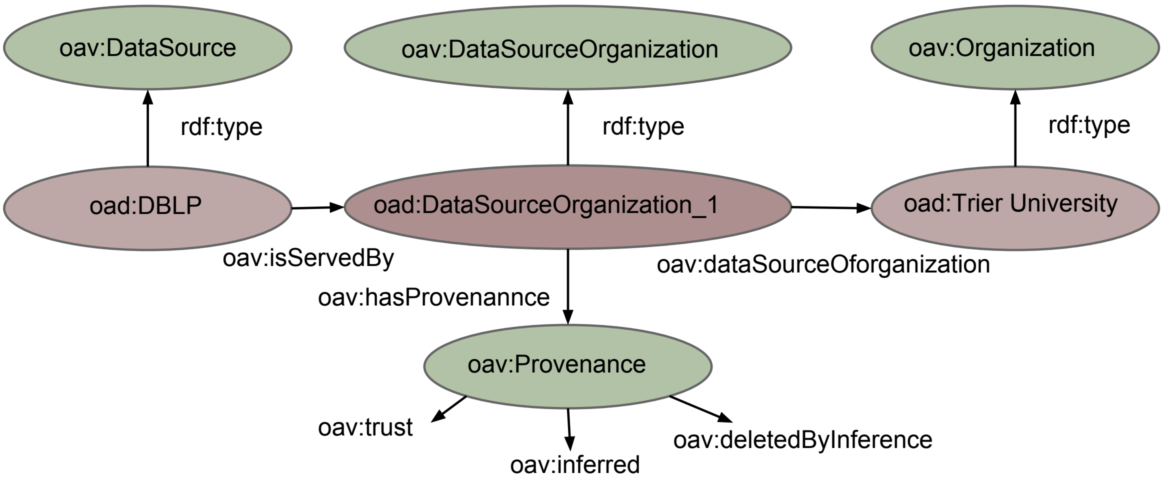


Figure 3.2: **An example of RDF graph.** Representation of an RDF graph and instance level.

Best practices for publishing such graphs on the Web in a way that is as reusable as possible are subsumed under the term “Linked Data”. Linked Data technology involves standards such as the Uniform Resource Identifier (URIs), HyperText Transfer Protocol (HTTP) to encode dereferenceable URIs and RDF to represent the objects. It primarily enables the machines to explore the Web of Data, but in a second step also humans who use machine services, such as search engines. RDF has come to be used as a general method for conceptual description or modeling of information that is implemented in Web resources, using a variety of syntax notations Turtle [4], N-Triples [5]. It is also used in knowledge management applications. The Resource Description Framework (RDF) was adopted as a W3C recommendation in 1999 and today is a standard for exchanging data on the Web.

SPARQL as Querying Languages SPARQL is the W3C recommend language [108] to query RDF datasets and stands for SPARQL Protocol and RDF Query Language. It is able to retrieve and manipulate data stored in RDF format. Therefore, SPARQL queries are executed against RDF datasets, consisting of

[3] https://www.w3.org/standards/semanticweb/

[4] https://www.w3.org/TR/turtle/

[5] https://www.w3.org/TR/n-triples/

RDF graphs. It is a W3C standard, and it is recognized as one of the critical technologies of the semantic web. A SPARQL query consists of triple patterns, conjunctions, disjunctions, and optional patterns. A SPARQL endpoint accepts queries and returns results via HTTP. Triple patterns are similar to RDF triples where the subject, predicate, or object may be variables. In a query, variables act like placeholders which are bound with RDF terms to build the solutions of the query. The expressive power of SPARQL comes in the ability to combine data properties as well with the schema of the data. A SPARQL query consists of up to five parts:

- **Prefix Declaration:** A list of URI prefixes to avoid writing complete URIs in the query.
- **Dataset Clause:** Similarly to SQL databases, where the user specifies the schema to be used, in the dataset clause is specified which graph is going to be queried.
- **Result Clause:** In this clause the type of query (SELECT, ASK, CONSTRUCT or DESCRIBE) and the variables to return are specified.
- **Query Clause:** The query clause contains the patterns that have to be matched in the graph. Resources fulfilling the specified patterns will be associated with the corresponding variables in the result clause.
- **Solution Modifiers:** The results of the queries can be paginated, ordered or sliced.

The results of SPARQL queries can be returned and/or rendered in a variety of formats such as XML, JSON, RDF. SPARQL variables start with a ? and can match any node (resource or literal) in the RDF dataset. Triple patterns are just like triples, except that any of the parts of a triple can be replaced with a variable (pattern matching). Variables named after the SELECT keyword are the variables that will be returned as results. Listing 3.1 shows an example of SELECT query. The *rdf:type* predicate links individual instances to *rdfs:Class* types.

```
PREFIX ex: <http://example.org/2017/03/schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?name
WHERE {?s rdf:type ex:Person .
  ?s ex:affiliation ex:BonnUniversity .
  ?s ex:name ?name. }
```

Listing 3.1: **SPARQL Example 1.** An example of SELECT clause in a simplest for is shown in SPARQL query. It retrieves a list of people who are affiliated at the University of Bonn.

Listing 3.2 shows another example of a SPARQL query using CONSTRUCT query. According to the SPARQL Query Language for RDF W3C Recommendation, CONSTRUCT returns a graph; a set of triples. It is useful to fetch a set of triples out of a triplestore, especially a remote triplestore and more importantly to create new triples and import them into the graph.

```
CONSTRUCT
  ?s rdf:label ?name .
  ?s rdf:type ?Researcher .
WHERE {
  ?s rdf:type ex:Person .
  ?s ex:affiliation ex:BonnUniversity .}
```

Listing 3.2: **SPARQL Example 2.** An example of CONSTRUCT clause in a simplest for is shown in SPARQL query. The query fetches all persons who are researchers and affiliated at the University of Bonn.

The access interfaces to query raw RDF dumps are either through SPARQL Endpoints or Linked Data Fragments. SPARQL endpoints are HTTP–based means that are easy to use, as they allow highly specific fragment selection. However, public endpoints have low availability as each unique query requires time on the server. Linked Data Fragments provide Web scale querying by offering datasets as fragments that requires little time on the server for query processing.

Knowledge Graphs A graph (G) is an ordered pair $G = (V, E)$ where V is the set of vertices and E is the set of edges. The vertices are the entities of the graph, and the edges denote the connections or associations between pairs of vertices. A Knowledge Graph (KG) a representation of knowledge in graphs in such a way that entities are represented by nodes and the relationships between entities are represented by edges of the graph[283]. More formally, let $\mathcal{E} = \{e_1, \cdots, e_{N_e}\}$ be the set of entities, $\mathcal{R} = \{r_1, \cdots, r_{N_r}\}$ be the set of relations connecting two entities, $\mathcal{D} = \{d_1, \cdots, d_{N_d}\}$ be the set of relations connecting an entity and a literal, i.e., the data relations, and $\mathcal{L}$ be the set of all literal values. A knowledge graph $\mathcal{G}$ is a subset of $(\mathcal{E} \times \mathcal{R} \times \mathcal{E}) \cup (\mathcal{E} \times \mathcal{D} \times \mathcal{L})$ representing the facts that are assumed to hold.

Knowledge graphs expand our understanding of metadata management using more flexible schemes such as Linked Data [28]. In 2012, the concept of knowledge graphs was used by Google to refer to their graph–based collections of knowledge [248]. There are several global knowledge graphs such as WikiData[6] and DBpedia[7]. However, under certain characteristics, any data can be represented as knowledge graphs. In this thesis, we consider the knowledge graphs as RDF graphs with explicit schema provided by ontologies by following the definition given in [207] which shows a knowledge graph:

- describes real world entities and their interrelations,
- defines possible classes and relations of entities in a schema,
- allows for potentially interrelating arbitrary entities with each other
- covers various topical domains.

Wikibase Software A wiki is a website on which users collaboratively create and curate content and structure directly from the web browser. A wiki is a system using wiki software or engine which is a type of content management system. However, the content is not owned by any specific agent neither a person nor an organization. The online encyclopedia project Wikipedia[8] is the most popular wiki–based website with hundreds of wikis. MediaWiki (MW)[9] is a free and open-source wiki software. It has large number of configuration settings and extensions for enabling various features to be added or changed. MediaWiki is used as a knowledge management system for groups of people who collaboratively create and modify content. It uses an extensible lightweight wiki markup. A form-based interface appears for the registered user. The users can use the editing environment or directly from the forms.

Semantic MediaWiki (SMW)[10] is an extension to MediaWiki that allows for annotating semantic data within wiki pages. Every semantic annotation within SMW is a *property* that represents a metadata entry. The instances of such metadata are created in the form of RDF. Markup language using brackets is used for the representation of the properties e.g., [[is Conference::ISWC 2018]]. Every Wiki page is a subject and the metadata is the property. The Object is the value to which the semantic link is created. Similar to MW, specific *Templates* can be designed to store metadata. Semantic forms enables user–created forms for adding and editing pages that use semantic data, see chapter 6.

[6] https://www.wikidata.org/
[7] https://wiki.dbpedia.org
[8] https://en.wikipedia.org/
[9] www.mediawiki.org
[10] www.semantic-mediawiki.org

3.3 Metadata Management in the Form of Life Cycle

Data and metadata management have been regarded as a series of activities for the administration of data for decades [97]. However, given the growth rate of data collections and increasing heterogeneity of their associated metadata in our era (as discussed in the previous section), new challenges have been raised for the tasks of integrating, managing, and analysing such metadata. The ultimate objective of data management activities is delivering efficient, interoperable and extensible services. Such services have been designed to describe, share and access data to which underlying metadata refers. Data management processes consist of actions perform under the control of certain rules as guidelines [6]. Several series of management activity steps have been introduced by leading research domains of computer science such as information science, databases and the web communities [51, 97].

In the context of digital libraries, metadata management is defined [154] as a set of design decisions that coordinates required activities to create, transform, preserve and maintain metadata about physical resources. As of the earliest comprehensive works is a survey [193] from 1970s that lists all the required contemporary data processing methods that are used in a wide range of applications. The need for metadata management of distributed and heterogeneous resources becomes more and more critical. In order to facilitate scholarly metadata management, a collaborative and partially decentralized environment is required to enable domain specific metadata capturing, transformation, community-based curation including reuse of already existing datasets. Such an approach would need to be able to represent metadata semantically to provide comprehensive interlinking from different resources of other relevant artifact types and datasets. Since the automated data acquisition methods alone do not achieve the required coverage and accuracy, a semi-automated method including community contributions is required. Several obstacles have been reported in metadata management process [285] such as being expensive and difficult to implement etc. Therefore, metadata creation and management need to be done as efficiently as possible through an already examined trusted process.

Types of the topology of the action steps in data management models depend on the type of data under consideration, objectives of the activities of each stage and their interrelations and the type of data manager. Based on a study of proposed models (Table 3.4), the possible patterns of modeling for such data/metadata management activities can be categorized into four classes:

- Sequenced modeling is used for a process with a start step and an end point. Each action of the sequenced model is following the other;
- Tree model is the type of modeling used for depicting parallel and dependent stages of a series of data management activities;
- Centralized Cycle model composes of stages controlled by an action manager (human or machine);
- Decentralized Cycle is the model used for representing a series of stages that are repeatedly happening independent of a central control.

These four models are depicted in Figure 3.3. An overview of the seven most used models for scientific data is discussed in [14]. In addition, with a survey study on 51 data and metadata models in [62], the decentralized and cyclic model seems to be the best practice for representing metadata management activities. While the concept of **life cycle** mostly appears in biology-related literature, it has been used in a metaphorical way by different communities in particular economy and business [103], energy science [234] and data management [14]. Multiple versions of a data life cycle exist with differences attributable to variation in practices across domains or communities. In computer science, the life cycle idea is applied in several but related different variations. For example, the sequence of changes that data undergoes applied by specific systems are considered as data management life cycles [62].

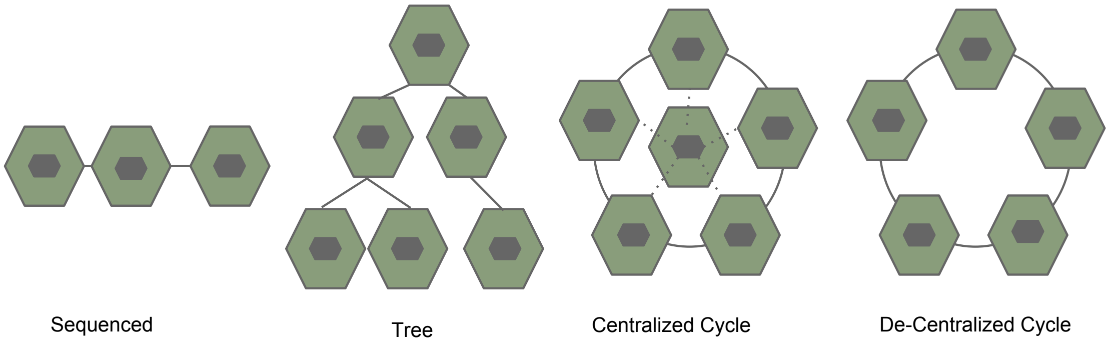


Figure 3.3: **Possible positioning of the data management stages in data management models.** Each node is representing an action step and their relation and order with regards to the other steps define the type of the model.

One of the active projects in data management, DataOne [11], describes the data life cycles from their point of view as a process that provides a high level overview of the stages involved in successful management and One of the highlighted models is the DCC Curation life cycle [114]. The core part of the model is consist of the digital data inside databases. The proposed steps for curation of metadata for such objects are: information representation, curation preservation planning, and community involvement. However, the model lacks to of required stages before and after curation and only focuses on limited systems and possible metadata curation activities for them. In another work, a more specific model is presented for research data management [205] in order to support researchers as data managers.

Similarly, the Research360 Project has developed a data management life cycle model with six stages [126]. A specific work is done for research data management that analysis literature to determines four key process areas in the form of a life cycle [58]. These four proposed phases cover acquisition, representation, dissemination and storage. In the context of heterogeneous data management on the Web, these phases remain incomplete. More precise and extended stages are required to be considered in order to provide approaches for challenges of metadata interoperability and semantics. In addition, dissemination is a best practice in many discussions of open science and data. Table 3.4 is a summary of the proposed life cycles for data management purposes. While there are several variations of this spelling e.g., life cycle, lifecycle, life-cycle, in this thesis we adopt to LOD life cycle and will use the concept as "life cycle". A specific life cycle is proposed for linked open data [10]. This model focuses on required stages to cover aligned tools which support the whole life cycle of Linked Data. In a different work a life cycle is proposed for big data [261] which focuses on stages for management of business data.

It has nine steps starting from business case evaluation, data identification and acquisition. Although, they give the name of life cycle for the proposed process, it is a sequence of steps specifically designed for business purposes. Considering 6Vs of big data, the proposed stages remains incomplete. Big data is about heterogeneous data that is created in different formats and requires transformation, curation, and mining which are not considered in the big data life cycle. The proposed life cycle in this thesis is fundamentally adapted to the LOD life cycle towards a big scholarly metadata life cycle. The required stages for metadata management in the context of scholarly communication are significantly different from pure linked data. For example, selecting of the eligible elements for the activity cycle is the primary stage for metadata management which is not considered in the LOD life cycle. The focus of big data life cycle is mainly business projects whereas the focus of this thesis is metadata management. Therefore, a new version of the processes for scholarly metadata management is created which will be discussed in more details in section 3.4.

[11] www.dataone.org/

Name	Proposed Stages	FAIR	5Vs
The Digital Curation Centre (DCC) model [114]	Curation: Create or Receive, Appraise, Select, Access, Use and Reuse, Transform; Preserve: Ingest, Preserve	Accessible, Reusable	Veracity, Volume
DataONE: Data Life Cycle Management [182]	Analyze, Collect, Assure, Describe, Deposit, Preserve, Discover, Integrate	Findable	Volume, Variety
Linked Open Data Life-cycle [10]	Extraction, Storage and Querying, Authoring, Interlinking, Classification, Quality Analysis, Evaluation, Search and Browsing	Findable, Interoperable, Reusable	Volume, Variety, Veracity
Big Data Life-cycle [212]	Evaluation, Identification, Acquisition and Filtering, Extraction, Validation and Cleansing, Aggregation and Representation, Analysis, Visualization, Utilisation of Analysis Results	Findable, Reusable	Volume, Variety, Velocity, Veracity, Value
Research Data Life-cycle [105]	Discovery and planning, Collection, Processing and Analysis, Publishing and Sharing, Long-term management, Reusing data	Reuse	Volume
JISC Research Data Life-cycle [171]	Plan, Create, Use, Appraise, Publish, Discover, Reuse	Findable, Reuse	Volume
Scientific Metadata Management Framework (Prabhune) [214]	Acquisition, Verification Assignment, Registration, Deposition, Extraction and Transformation, Discovery and Access, Analysis and Visualization	Findable, Accessible, Interoperable	Volume, Variety

Table 3.4: **An overview of already proposed Data Management Life-cycles.** Several other life cycles and sequence of data management processes have already been proposed. The collected list of life cycles has been classified considering the stages an coverage of FAIR principles and 6Vs of big scholarly metadata.

3.4 MEDAL: A Metadata Life Cycle

Adopted to the Linked Data life cycle [10], we propose a management cycle for scholarly metadata. The aim of this cycle is to provide a comprehensive structure of steps required for metadata management. MEDAL (MEtaDAta Life cycle) is an integrated distribution of aligned steps which covers the whole management processes from eligibility checking of the associated metadata of artifacts to interlinking, providing smart analytics. The steps of the life cycle look discrete, however they do not exist in isolation from each other. In fact, step n is triggered by the output of its predecessor step and provides input for its successor step, see Figure 3.5.

The life cycle starts with eligibility checking of target metadata and resources to be considered for the entire management process (**Selection**). Such heterogeneous metadata in different formats are embedded in the artifacts, resource objects, and datasets. Some parts of the target metadata need to be projected out from the reference resources (**Extraction**). A step is required to converts information represented in unstructured, structured or semi-structured forms to a unified format (**Transformation**). Integration of generated or collected metadata with other already existing datasets expands the information space (**Interlinking**). Semantic enrichment of interlinked data by relation discovery adds missing or overlooked metadata and supports creation of a rich knowledge graph **Enrichment**. Data acquisition from heterogeneous resources have the potential of being incomplete. In order to provide a comprehensive knowledge graph of gathered metadata, data cleaning, annotation and manual correcting is needed (**Curation**). The information gathered and curates as a knowledge graph could be inferred by applying graph mining and machine learning techniques (**Mining**). Metadata improves the quality of underlying resources. Based on the present metadata, one can define quality metrics and provide complex assessment that would not be possible without metadata. Such assessments can give insights about quality artifacts (**Quality Assessment**). At this stage, having a rich metadata at hand would enable comprehensive

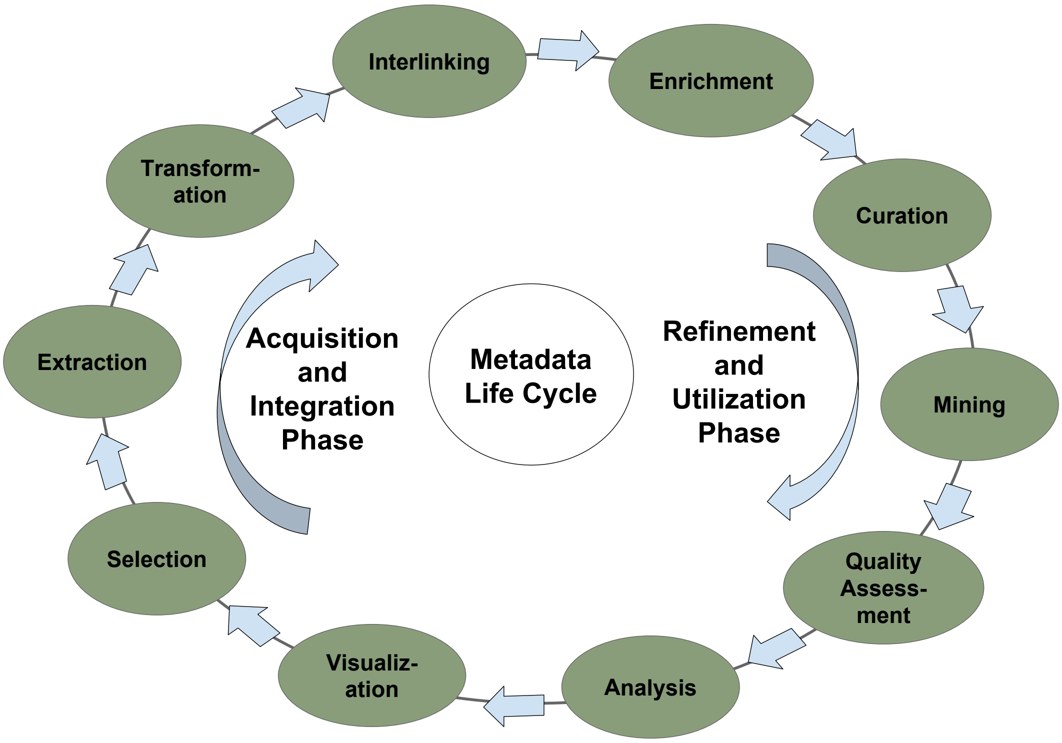


Figure 3.4: **Metadata management cycle**. A life cycle is used as a methodology to order the required steps for management of scholarly metadata.

querying of data with the purpose of providing analytics (**Analysis**). In order to make the results easily understandable for human, a better representation is necessary (**Visualization**).

Figure 3.4 depicts the overall process required for such facilitation of scholarly metadata management. The cycle is not limited to scholarly metadata, and it can be applied for any type of metadata with the specific characteristics introduced in this thesis. For example the same life cycle could be applied for metadata related to music collections.

There are multiple implicit information around the life cycle.

The proposed action steps have been characterized into three main phases that are required for metadata management:

- **Acquisition and Integration Phase**: includes steps to select metadata and gather from different types of sources, and follow with metadata aggregation.
- **Refinement and Utilization Phase**: includes steps regarding the enrichment and mining of the collected metadata as well as curation. The result of this phase is a knowledge graph and analysis and visualization of the results are also covered in this phase.

Inspired from the success of software engineering community by re-modeling software development process as a series of spirals [4, 35, 135], a complete or partial application of the life cycle is envisioned as spirals. All the three proposed phases and their corresponding stages are present in individual spiral. Each spiral builds on previous work, and requirements are addressed through multiple application of life cycle. The idea of life cycle spirals has been already presented in a recent work in [102]. However the proposed life cycle for data in that work focuses on required process for project acquisition in the

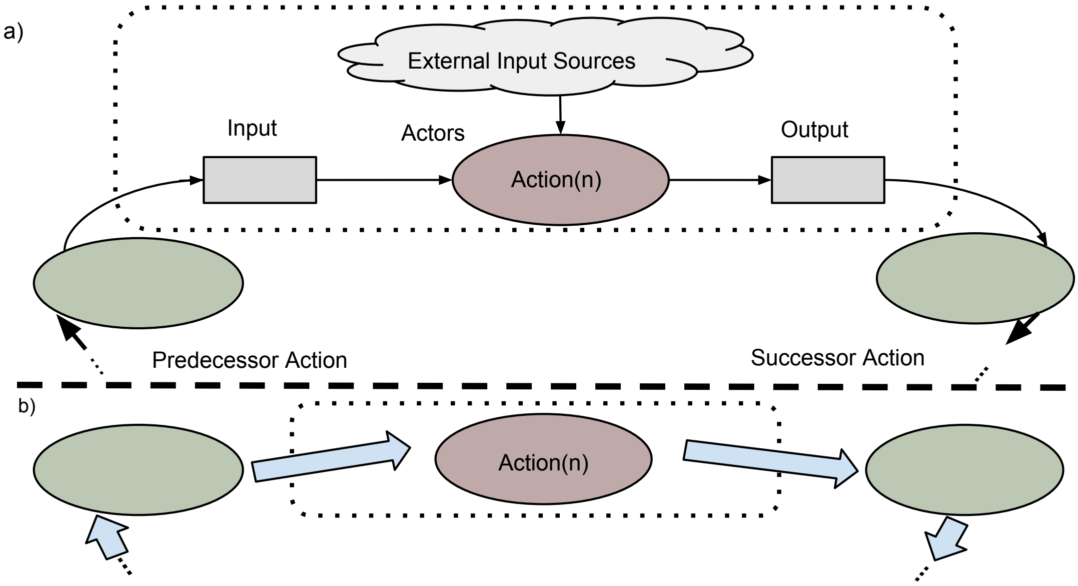


Figure 3.5: **Implicit information of the life cycle**. Each of the action steps of the life cycle requires input resources. The input of action steps can be the output of previous action step. Each action step of the life cycle is done by an actor that can be either human or machines.

institutional level and describes stages required for scientific data generated alongside. The detailed description of the phases and their associated action steps f the life cycle is presented in the upcoming subsections.

3.4.1 Acquisition and Integration Phase

The initial phase in the metadata management is acquisition and integration of required datasets that includes the following steps: **Selection**, **Extraction**, **Transformation**, and **Interlinking**. The detained description of each step will be discussed in this section.

Selection Metadata integration and management actions are applicable over a selection of resources with to be considered as an input. The confirmation whether a resource is an eligible input to be taken through the further management steps is made based on the requirements of the individual use case and specific objectives of that action. At this stage of the life cycle, metadata is not directly involved. However, criteria for eligibility checking of resources are expected to be defined based on the available metadata of the target resources at the time of selection. Resources are selected per their matching with the least required characteristics. A resource in this level can be an individual artifact or a repository of metadata or artifacts as well as collection of content.

Selection of the resources to be carried to the other steps is the starting point of the life cycle. In relation to the first step which is a required starting point, any of the other steps have the potential to become the second step. More precisely, the eligibility checking is the step that always comes in front of the next selected step that is required to trigger the life cycle. The other steps of the life cycle can be skipped according to the status of the input resources and the required action. However, the order of the steps have to be followed as it is proposed. As a consequence of this assumption, the checking criteria can change in relation to the next upcoming step of the life cycle. Therefore, for extraction as the second proposed step, the checking criteria would be considered different than the selection process of resources for interlinking etc. The selection step acts as an observatory of the underlying resources of the main focus. It is composed of three main sub-action steps:

- **Conceptual modeling** is the process of understanding the logical structure of an application domain for which an information management system is aimed to be designed. In other words, the conceptual schema designing starts with documenting all detailed requirements in the application landscape to represent "concepts" and relationship types between them. Since many decades ago, data modeling in invented to assist in the design of databases in particular relational databases. This type of modeling is aiming at the exploration of the real world and meaning of concepts. The conceptual model of a domain is to express the meaning of terms and concepts used by domain experts to discuss the problem, and to find the relationship types between different concepts and their attributes. New models are continually being developed and varieties of existing models are extended over years. There are different methodologies to use in the purpose of conceptual modeling such as object oriented modeling and entity-relationship types modeling and ontology languages etc. A decade ago, from a completely different direction than database community, the artificial intelligence world, the concept of semantics has arisen as a subject of focus for data modeling and conceptualization of domains. Modeling of a domain provides a basis for collecting data according to the defined categories, and its corresponding database design. The purpose is to classify them so that computers can make inferences from them.

- **Criteria defining** is another step after conceptualization. In this step, domain experts together with data managers capture fundamental criteria to be considered for checking the eligibility of metadata resources and artifacts. This step requires a deep understanding of the final use cases of the metadata management and utilization of the results. Based on the conceptual modeling of the domain and the envisioned endpoint, the criteria are defined. Such criteria are the bases for the quality assessment stage.

- **Eligibility checking** is the final step of selection where eligible artifacts, repositories, datasets or the target metadata will be filtered. Those resources passing the eligibility checking test are passed through the next stage of the life cycle.

The Selection stage and the corresponding sub-stages have a close connection with the quality assessment stage. The checking criteria defined in the selection stage are considered as the quality assessment metrics.

Example 1. Publication-Related Finding and selecting the list of the scientific publications to be read or cited is currently done through the trusted ranking provided by the available search engines. Researchers often use keywords combined with filtering of metadata about publication year, author name or event name.

Example 2.Event-Related Researchers with different incentives and needs have interest on a group of events in their domain to submit research results of participate. Events with low acceptance rate is usually considered as the most successful one. However, this is not the only criteria to target an event for publishing research results. There are characteristics such as location, fees, reputation of the organizers and speaker also play an important role. Senior researchers often collect a subjective opinion over years about events and other venues and possibilities for publishing their research results within their domain. Exploration and accessibility of such metadata is often challenging for researchers out of the exact research domain.

Example 3. OCW-Related The selection of repositories to take online courses from are done through a systematic approach. Certain criteria should be defined based on the needs of the users. List of the repositories and the courses with most fitness for use can be selected.

Extraction Most of the information required for providing sophisticated querying, comprehensive services and analytics are embedded in unstructured form of artifacts or adhered to the other structured or semi-structured formats. In order to use the selected information (pre-specified in the selection stage) within the metadata life cycle, the second default stage is extraction of such metadata. Metadata extraction from unstructured resources of information such as text documents has been initiated to provide unified services on top of heterogeneous datasets [106].

Information extraction was introduced as one of the main steps towards implementation of the Semantic Web where the knowledge contained within these documents are extracted and made more accessible for machine processing. Considering the amount of documents plugged in to the web. a constant extraction of metadata from unstructured data is required in the process of metadata management for each application domain. Metadata extraction enables automation of knowledge-driven activities such as content classification, integrated information exploration and uncovering hidden relationships, etc.

The extraction process involves transforming an unstructured text or a collection of texts into sets of facts (structured, machine-readable statements). Depending on the type of the artifacts (text documents, multimedia documents such as video, audio, etc.) as the input for this stage, the extraction process can be performed either automatically or in a semi-automatic way. Information extraction from text documents is one of the core application domains of the Natural Language Processing (NLP) and Artificial Intelligent (AI) approaches. In the case of scientific text documents, the paradigm of publishing follows certain templates but varies for different publishers. Typically, the following main subtasks are involved for metadata extraction from text documents:

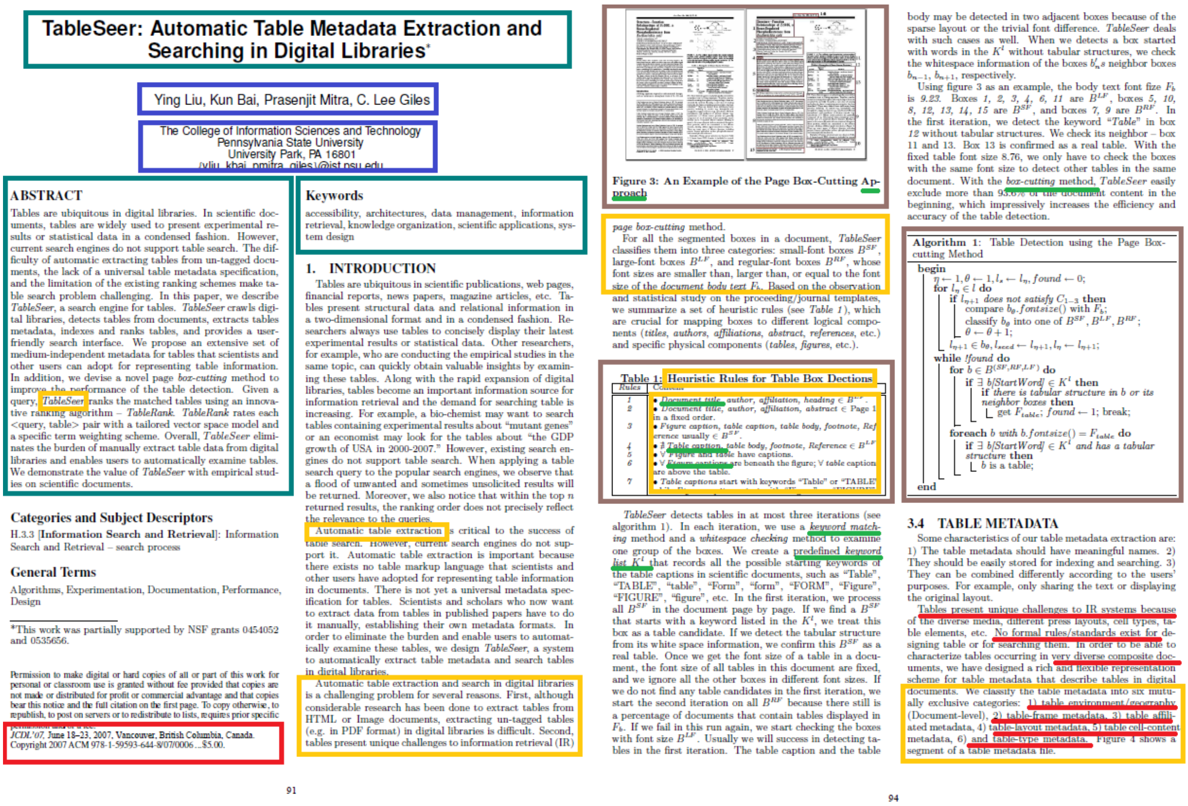
- **Pre-processing** of the text is required in order to identify the method of extraction. Templates of the target documents are being processed by computational linguistics tools Tokenization and sentence splitting are the mostly used processes of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining.

- **Execution** of the extraction approaches usually starts with fact extraction either by keyword extraction or identification of the proper names or named entities mentioned in a text. The task of named entity recognition (NER) is to find each mention of a named entity in the text and label its type (pre-specified in selection step). Several approaches such as hidden markov models have been applies for NER purposes. The results of NER are to link, or cluster, these mentions into sets that correspond to the entities from real world Metadata extraction can be done with the purpose of relation extraction to find and classify semantic relations among the text entities. Relation extraction methods are highly coalesced with supervised machine learning approaches with large training datasets and pattern recognition approaches with best practices on use cases lacking the existence of large training datasets. More complex approaches are required for text documents including other type of information such as figures, tables and multimedia elements.

- **Post-processing** is the task of identifying relationships between the extracted concepts. Furthermore, a unification of the extracted metadata into a standard form is needed to be applied. This step is done in order to normalize the extracted information and is different that the transformation stage. The post-processing continuous with a step of getting rid of the noise that involves eliminating duplicate data. Finally, the extracted knowledge are ingested and stored into a data management system for further use and application of other stages.

Regarding the multimedia documents, although improvements in the corresponding technology and standards, provide important functionalities, extraction and management of information embedded in the content of such artifacts is left to the content and metadata manager [153]. Therefore, semi-automated metadata extraction is mostly done together with semantic annotation of multimedia artifacts [32]. There are many other activities in this direction that are involved with speech recognition and image processing, sentiment analysis approaches which are beyond the scope of this research.

Example 1. Publication-Related Scientific documents often called papers are the main means of scholarly communication. Researchers represent and publish their results with a certain writing style and structure [92]. Scholars represent the research problem and the domain of interest as well as the proposed approaches, developed tools implementation, evaluation and results inside scientific papers. In addition to the text, certain elements are expected to be appear within scientific content such as tables, figures, drawings (charts, graphs and diagrams), footnotes and captions. However, because of the current scholarly communication paradigm, such important knowledge is locked within the text of scientific papers. Figure 3.7 depicts a set of sample metadata embedded in a scientific document (paper). Blocks of information are highlighted to show an example of information to be extracted in order to provide a semantic representation of such scientific documents in a structured way.



TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries*

Ying Liu, Kun Bai, Prasenjit Mitra, C. Lee Giles

The College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16801
{yliu, kbai, pmitra, giles}@ist.psu.edu

ABSTRACT

Tables are ubiquitous in digital libraries. In scientific documents, tables are widely used to present experimental results or statistical data in a condensed fashion. However, current search engines do not support table search. The difficulty of automatic extracting tables from un-tagged documents, the lack of a universal table metadata specification, and the limitation of the existing ranking schemes make table search problem challenging. In this paper, we describe *TableSeer*, a search engine for tables. *TableSeer* crawls digital libraries, detects tables from documents, extracts tables metadata, indexes and ranks tables, and provides a user-friendly search interface. We propose an extensive set of medium-independent metadata for tables that scientists and other users can adopt for representing table information. In addition, we devise a novel page *box-cutting* method to improve the performance of the table detection. Given a query, *TableSeer* ranks the matched tables using an innovative ranking algorithm – *TableRank*. *TableRank* rates each <query, table> pair with a tailored vector space model and a specific term weighting scheme. Overall, *TableSeer* eliminates the burden of manually extract table data from digital libraries and enables users to automatically examine tables. We demonstrate the value of *TableSeer* with empirical studies on scientific documents.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – search process

General Terms

Algorithms, Experimentation, Documentation, Performance, Design

*This work was partially supported by NSF grants 0454052 and 0535656.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'07, June 18–23, 2007, Vancouver, British Columbia, Canada.
Copyright 2007 ACM 978-1-59593-644-8/07/0006 ...$5.00.

Keywords

accessibility, architectures, data management, information retrieval, knowledge organization, scientific applications, system design

1. INTRODUCTION

Tables are ubiquitous in scientific publications, web pages, financial reports, news papers, magazine articles, etc. Tables present structural data and relational information in a two-dimensional format and in a condensed fashion. Researchers always use tables to concisely display their latest experimental results or statistical data. Other researchers, for example, who are conducting the empirical studies in the same topic, can quickly obtain valuable insights by examining these tables. Along with the rapid expansion of digital libraries, tables become an important information source for information retrieval and the demand for searching table is increasing. For example, a bio-chemist may want to search tables containing experimental results about "mutant genes" or an economist may look for the tables about "the GDP growth of USA in 2000-2007." However, existing search engines do not support table search. When applying a table search query to the popular search engines, we observe that a flood of unwanted and sometimes unsolicited results will be returned. Moreover, we also notice that within the top *n* returned results, the ranking order does not precisely reflect the relevance to the queries.

Automatic table extraction is critical to the success of table search. However, current search engines do not support it. Automatic table extraction is important because there exists no table markup language that scientists and other users have adopted for representing table information in documents. There is not yet a universal metadata specification for tables. Scientists and scholars who now want to extract data from tables in published papers have to do it manually, establishing their own metadata formats. In order to eliminate the burden and enable users to automatically examine these tables, we design *TableSeer*, a system to automatically extract table metadata and search tables in digital libraries.

Automatic table extraction and search in digital libraries is a challenging problem for several reasons. First, although considerable research has been done to extract tables from HTML or Image documents, extracting un-tagged tables (e.g. in PDF format) in digital libraries is difficult. Second, tables present unique challenges to information retrieval (IR)

91

Figure 3: An Example of the Page Box-Cutting Approach

page box-cutting method.

For all the segmented boxes in a document, *TableSeer* classifies them into three categories: small-font boxes B^{SF}, large-font boxes B^{LF}, and regular-font boxes B^{RF}, whose font sizes are smaller than, larger than, or equal to the font size of the *document body text* F_b. Based on the observation and statistical study on the proceeding/journal templates, we summarize a set of heuristic rules (see *Table 1*), which are crucial for mapping boxes to different logical components (*titles, authors, affiliations, abstract, references,* etc.) and specific physical components (*tables, figures,* etc.).

Table 1: Heuristic Rules for Table Box Dections

Rules	Content
1	• *Document title, author, affiliation, heading* $\in B^{LF}$.
2	• *Document title, author, affiliation, abstract* $\in$ Page 1 in a fixed order.
3	• *Figure caption, table caption, table body, footnote, Reference* usually $\in B^{SF}$.
4	• $\nexists$ *Table caption, table body, footnote, Reference* $\in B^{LF}$
5	• $\forall$ *Figure and table* have captions.
6	• $\forall$ *Figure captions* are beneath the figure; $\forall$ *table caption* are above the table.
7	• *Table captions* start with keywords "Table" or "TABLE"

TableSeer detects tables in at most three iterations (see algorithm 1). In each iteration, we use a *keyword matching* method and a *whitespace checking* method to examine one group of the boxes. We create a predefined *keyword list* K^l that records all the possible starting keywords of the table captions in scientific documents, such as "Table", "TABLE", "table", "Form", "form", "FORM", "Figure", "FIGURE", "figure", etc. In the first iteration, we process all B^{SF} in the document page by page. If we find a B^{SF} that starts with a keyword listed in the K^l, we treat this box as a table candidate. If we detect the tabular structure from its white space information, we confirm this B^{SF} as a real table. Once we get the font size of a table in a document, the font size of all tables in this document are fixed, and we ignore all the other boxes in different font sizes. If we do not find any table candidates in the first iteration, we start the second iteration on all B^{RF} because there still is a percentage of documents that contain tables displayed in F_b. If we fail in this run again, we start checking the boxes with font size B^{LF}. Usually we will success in detecting tables in the first iteration. The table caption and the table body may be detected in two adjacent boxes because of the sparse layout or the trivial font difference. *TableSeer* deals with such cases as well. When we detects a box started with words in the K^l without tabular structures, we check the whitespace information of the boxes $b'_n s$ neighbor boxes b_{n-1}, b_{n+1}, respectively.

Using figure 3 as an example, the body text font fize F_b is *9.23*. Boxes *1, 2, 3, 4, 6, 11* are B^{LF}, boxes *5, 10, 8, 12, 13, 14, 15* are B^{SF}, and boxes *7, 9* are B^{RF}. In the first iteration, we detect the keyword *"Table"* in box *12* without tabular structures. We check its neighbor – box 11 and 13. Box 13 is confirmed as a real table. With the fixed table font size 8.76, we only have to check the boxes with the same font size to detect other tables in the same document. With the *box-cutting* method, *TableSeer* easily exclude more than 93.6% of the document content in the beginning, which impressively increases the efficiency and accuracy of the table detection.

Algorithm 1: Table Detection using the Page Box-cutting Method

```
begin
    $\eta \leftarrow 1, \theta \leftarrow 1, l_s \leftarrow l_\eta, found \leftarrow 0$;
    for $l_\eta \in l$ do
        if $l_{\eta+1}$ does not satisfy $C_{1-3}$ then
            compare $b_\theta.fontsize()$ with $F_b$;
            classify $b_\theta$ into one of $B^{SF}, B^{LF}, B^{RF}$;
            $\theta \leftarrow \theta + 1$;
        $l_{\eta+1} \in b_\theta, l_{seed} \leftarrow l_{\eta+1}, l_\eta \leftarrow l_{\eta+1}$;
    while !found do
        for $b \in B^{(SF,RF,LF)}$ do
            if $\exists$ b[StartWord] $\in K^l$ then
                if there is tabular structure in b or its neighbor boxes then
                    get $F_{table}$; $found \leftarrow 1$; break;
        foreach b with $b.fontsize() = F_{table}$ do
            if $\exists$ b[StartWord] $\in K^l$ and has a tabular structure then
                b is a table;
end
```

3.4 TABLE METADATA

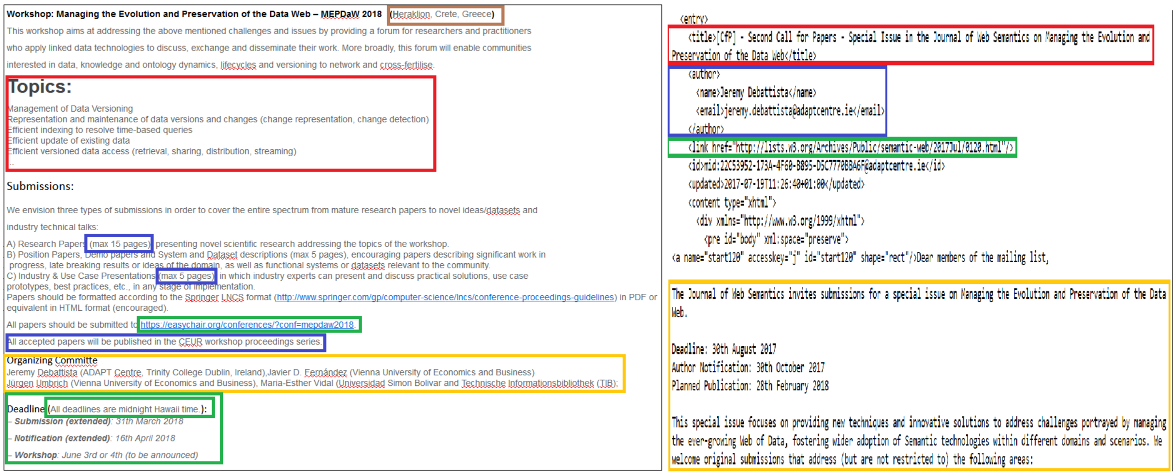
Some characteristics of our table metadata extraction are: 1) The table metadata should have meaningful names. 2) They should be easily stored for indexing and searching. 3) They can be combined differently according to the users' purposes. For example, only sharing the text or displaying the original layout.

Tables present unique challenges to IR systems because of the diverse media, different press layouts, cell types, table elements, etc. No formal rules/standards exist for designing table or for searching them. In order to be able to characterize tables occurring in very diverse composite documents, we have designed a rich and flexible representation scheme for table metadata that describe tables in digital documents. We classify the table metadata into six mutually exclusive categories: 1) table environment/geography (Document-level), 2) table-frame metadata, 3) table affiliated metadata, 4) table-layout metadata, 5) table cell-content metadata, 6) and table-type metadata. Figure 4 shows a segment of a table metadata file.

94

Figure 3.6: **Embedded Metadata within a Scientific Paper**. Embedded metadata inside scientific papers includes information about the authors, venue, publisher, title, affiliation and email, ORCID id of authors. Furthermore, information about the domain of focus, the research approach, references, achievements, analysis etc.

Example 2. Event-Related Mailing lists are used as a popular way [250] of exchanging announcements or spreading discussions easily among researchers. They form one of the most reliable sources of information about upcoming events because of the large coverage of events by Calls for Papers (CfPs) disseminated in those mailing lists. The principal reasons for using email as a scientific communication channel are the known target group, speed and immediacy it offers. However, the sheer amount of emails sent through those mailing lists makes it difficult for one individual to keep track of them. Although data from mailing lists is a reliable source of information about upcoming events, it is hard for one individual to extract specific information from them. To obtain the information they are interested in, subscribers are required to first filter a huge amount of emails by relevance, and then, in the worst case, read the full text of the relevant ones. Researchers have to trace the emails on a list and to decide which ones to have a closer look into. Although this process looks straightforward and is one of the favorite communication channels for researchers, a lot of relevant information might either be overlooked or overwhelm recipients.



Workshop: Managing the Evolution and Preservation of the Data Web – MEPDaW 2018 (Heraklion, Crete, Greece)

This workshop aims at addressing the above mentioned challenges and issues by providing a forum for researchers and practitioners who apply linked data technologies to discuss, exchange and disseminate their work. More broadly, this forum will enable communities interested in data, knowledge and ontology dynamics, lifecycles and versioning to network and cross-fertilise.

Topics:

Management of Data Versioning
Representation and maintenance of data versions and changes (change representation, change detection)
Efficient indexing to resolve time-based queries
Efficient update of existing data
Efficient versioned data access (retrieval, sharing, distribution, streaming)
..

Submissions:

We envision three types of submissions in order to cover the entire spectrum from mature research papers to novel ideas/datasets and industry technical talks:

A) Research Papers (max 15 pages) presenting novel scientific research addressing the topics of the workshop.
B) Position Papers, Demo papers and System and Dataset descriptions (max 5 pages), encouraging papers describing significant work in progress, late breaking results or ideas of the domain, as well as functional systems or datasets relevant to the community.
C) Industry & Use Case Presentations (max 5 pages) in which industry experts can present and discuss practical solutions, use case prototypes, best practices, etc., in any stage of implementation.
Papers should be formatted according to the Springer LNCS format (http://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines) in PDF or equivalent in HTML format (encouraged).

All papers should be submitted to https://easychair.org/conferences/?conf=mepdaw2018.

All accepted papers will be published in the CEUR workshop proceedings series.

Organizing Committe
Jeremy Debattista (ADAPT Centre, Trinity College Dublin, Ireland),Javier D. Fernández (Vienna University of Economics and Business)
Jürgen Umbrich (Vienna University of Economics and Business), Maria-Esther Vidal (Universidad Simon Bolivar and Technische Informationsbibliothek (TIB);

Deadline (All deadlines are midnight Hawaii time.):
- ***Submission (extended)***: *31th March 2018*
- ***Notification (extended)***: *16th April 2018*
- ***Workshop***: *June 3rd or 4th (to be announced)*

```
  <entry>
    <title>[CfP] - Second Call for Papers - Special Issue in the Journal of Web Semantics on Managing the Evolution and
Preservation of the Data Web</title>
    <author>
      <name>Jeremy Debattista</name>
      <email>jeremy.debattista@adaptcentre.ie</email>
    </author>
    <link href="http://lists.w3.org/Archives/Public/semantic-web/2017Jul/0120.html"/>
    <id>mid:22C53952-173A-4F60-B895-D5C7770BBA6F@adaptcentre.ie</id>
    <updated>2017-07-19T11:26:40+01:00</updated>
    <content type="xhtml">
      <div xmlns="http://www.w3.org/1999/xhtml">
        <pre id="body" xml:space="preserve">
<a name="start120" accesskey="j" id="start120" shape="rect"/>Dear members of the mailing list,

The Journal of Web Semantics invites submissions for a special issue on Managing the Evolution and Preservation of the Data
Web.

Deadline: 30th August 2017
Author Notification: 30th October 2017
Planned Publication: 28th February 2018

This special issue focuses on providing new techniques and innovative solutions to address challenges portrayed by managing
the ever-growing Web of Data, fostering wider adoption of Semantic technologies within different domains and scenarios. We
welcome original submissions that address (but are not restricted to) the following areas:
```

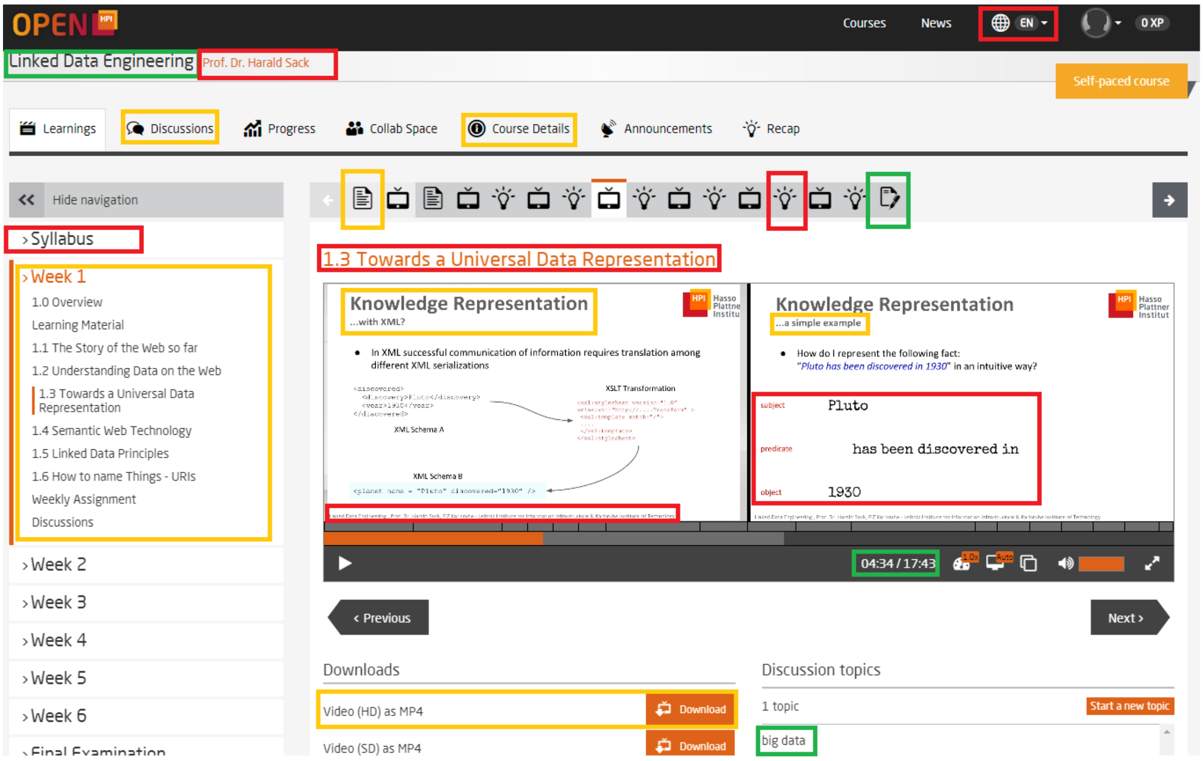
Figure 3.7: **Embedded Metadata within a call for paper of an event**. Metadata of scientific events, including their research community, event name, topics covered, the date and location of the event, deadlines and list of organizers are knowledge encoded in CfP emails.

Example 3. OCW-Related Online courses are often created by using existing materials. However, those materials are usually proposed as unstructured form of information on their aims and the typology of a limited group of users which they are targeting.

Moreover, often the content is not clearly synthesized so that a complete analysis of the whole materials requires further efforts by human only. Therefore, a researcher wanting to get a quick training about a topic using online courses is never certain whether she has found all or most of the truly relevant courses. Furthermore, interesting relationships, similarities, and other indirect connections between courses and other material are missed. Thus, a researcher may find a few relevant courses and videos in her search, but may miss a significant number of courses directly “on point” and will likely not see many videos that have some interesting relationship to the subject being searched. The reasons for this are varied, but they essentially suffer from the lack of a common way to tag or provide metadata for such courses when they are created and stored.

Figure 3.8[12] shows part of the metadata embedded in an online course with videos and slides. All this information requires human checks and machines are not unable to read and process such unstructured

[12] The screen shot of the course is from OpenHPI (the educational Internet platform of the German Hasso Plattner Institute, Potsdam), the linked data engineering course. https://open.hpi.de/courses/semanticweb2016



In XML successful communication of information requires translation among different XML serializations

```
<discovered>
  <discovery>Pluto</discovery>
  <year>1930</year>
</discovered>
```

```
<planet name = "Pluto" discovered="1930" />
```

How do I represent the following fact: *"Pluto has been discovered in 1930"* in an intuitive way?

Figure 3.8: **Embedded Metadata within an Online Course**. Embedded metadata inside video or audio courses, the slides and their content shown in the course carries important metadata. In addition, researchers also make their choices of reusing or learning from a course by knowing about syllabus, language, duration, assigned material etc. All these information are represented in current services and systems as unstructured or semi-structured data.

and heterogeneous data. Whereas, the users of interactive platform can ask their questions and discuss points of interest with each other in the course discussion forum which is actively moderated by the teaching team.

Transformation Metadata originate from distributed and heterogeneous sources which makes it massive and complex for traditional systems to handle. In order to better management and utilization within the scope of proposed life cycle, such datasets are required to be represented in semantic format. A unified, linked, and less complex data representation is needed in order to achieve speedy, simple and effective applications. As discussed before, one of the widely accepted yet simple representation of data is possible by using Resource Description Framework (RDF). The resources represented by RDF are aimed to be interpreted by machines with the use of reasoning and logical rules. Different transformation tools are being introduced for different data models such as Triplify and RML to transfer relational data to RDF.

Example 1. Publication-Related Datasets including metadata about several important scholarly artifacts are published by different platforms and services. For different technical reasons, the OpenAIRE infrastructure has three different representation of metadata: XML, HBase, CSV. For example some of the highlighted metadata about publications in the CSV format are the following properties: id, title, publisher, year, license, DOI (Digital Object Identifier), access url, type, harvested date. Due to presence of the characters that are usually used as delimiters, a special one, #!#, is used in CSV representation of metadata by OpenAIRE. Listing 3.3 is a CSV representation of metadata from OpenAIRE project about a publication from the OpenAIRE infrastructure is as follows:

CSV

```
#dedup_wf_001::39b91277f9a2c25b1655436ab996a76b#!#The Data Model of the OpenAIRE
Scientific Communication e−Infrastructure#!#null#!#null#!#Springer#!#null#!#null
#!#null#!#null#!#2012#!#2012−01−01#!#Open Access#!#Open Access#!#Access#!#null#!#
0#!#null#!#nulloai:http://helios−eie.ekt.gr:!#publication#10442/13187oai:pumaoai.
isti.cnr.it:cnr.isti/cnr.isti/2012−A2−040#!#1#!
```

Listing 3.3: **CSV Example.** An example representation of metadata in CSV format is shown.

The following Listing 3.4 is the RDF representation of metadata about the same instance:

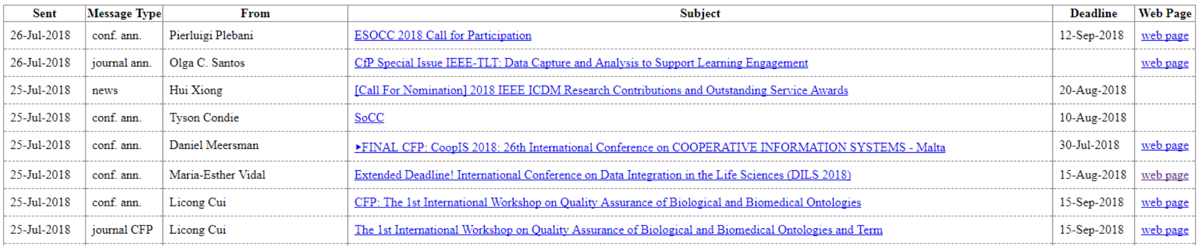
RDF

```
@prefix oad: <http://...oa.eu/data/> .
@prefix oav: <http://...oa.eu/vocab#> .
# further prefixes omitted; see \url{http://prefix.cc}.

oad:result/...001::39b9... rdf:type oav:Result, bibo:Publication;
    dcterms:title "The Data Model of the OpenAIRE Scientific Communication
    e−Infrastructure"@en ;   dcterms:dateAccepted "2012−01−01"^^xsd:date ;
    dcterms:language "en";   oav:publicationYear 2012 ;
    dcterms:publisher "Springer";   foaf:firstName "Paolo"; foaf:lastName "Manghi".
```

Listing 3.4: **RDF Example.** An example representation of metadata in RDF format is shown.

Example 2. Event-Related Most of the communities use their own way of announcing call for papers. One of the most used formats is representation of upcoming scientific events in relational format. In this way, the metadata of events are fragmented in several fixed properties and represented inside a Figure 3.9. In order to gain more information, researchers need to trace the list and follow the link manually to the homepage of the event.



Sent	Message Type	From	Subject	Deadline	Web Page
26-Jul-2018	conf. ann.	Pierluigi Plebani	ESOCC 2018 Call for Participation	12-Sep-2018	web page
26-Jul-2018	journal ann.	Olga C. Santos	CfP Special Issue IEEE-TLT: Data Capture and Analysis to Support Learning Engagement		web page
25-Jul-2018	news	Hui Xiong	[Call For Nomination] 2018 IEEE ICDM Research Contributions and Outstanding Service Awards	20-Aug-2018	
25-Jul-2018	conf. ann.	Tyson Condie	SoCC	10-Aug-2018	
25-Jul-2018	conf. ann.	Daniel Meersman	▸FINAL CFP: CoopIS 2018: 26th International Conference on COOPERATIVE INFORMATION SYSTEMS - Malta	30-Jul-2018	web page
25-Jul-2018	conf. ann.	Maria-Esther Vidal	Extended Deadline! International Conference on Data Integration in the Life Sciences (DILS 2018)	15-Aug-2018	web page
25-Jul-2018	conf. ann.	Licong Cui	CFP: The 1st International Workshop on Quality Assurance of Biological and Biomedical Ontologies	15-Sep-2018	web page
25-Jul-2018	journal CFP	Licong Cui	The 1st International Workshop on Quality Assurance of Biological and Biomedical Ontologies and Term	15-Sep-2018	web page

Figure 3.9: **Call for papers of the Database community**. A list of call for papers announced on DBWorld in a relational data format.

Example 3. OCW-Related

The following Listing 3.5 shows a representation of metadata corresponding to an online course in XML format. Harvested metadata from repositories providing such metadata can be transformed to a unified format.

XML

```
<or:ocw>
  <title schemename="dnet:cource_title" classname="main title"
   schemeid="dnet:cource_title" classid="main title">Knowledge Engineering</title>
  <dateofstart>2012−01−01</dateofstart>
  <organizer>University of Bonn</organizer>
</or:ocw>
```

Listing 3.5: **XML Example.** An example representation of metadata in XML format is shown.

Interlinking as an integration approach has been fused into the core vision of the Linked Data principles [28]. While interlinking approaches for other data models has been technically possible and in-use, a lightweight solution for connecting the isolated and disconnected datasets was a recurrent problem. Along with emergence of Linked Open Data, a series of best practices have been established on the usage of RDF data model. Consequently, the LOD cloud[13] has emerged as one of the largest collections of interlinked datasets on the web. In the proposed *5-star* labeling of data [21], interlinking is the final required step that makes the dataset connected to the LOD cloud. In this context, interlinking is mainly done in order to increase the comprehensiveness, quality and usage of the datasets. It is the discovery of (ideally) all instances that represent the same real-world objects located in different datasets [139]. To this aim, certain steps are required that are listed as below:

- **Selecting candidate datasets** to be considered as the target datasets for interlinking is the initial step in this regard. It requires a basic understanding of the domain and promising strategy in finding the relevant resources. Usually domain experts already know about the published datasets (This sub-step has a close connection to the main selection step (3.4.1)).
- **Bi-directional modeling** of the two datasets, both the source and the target, is required in order to define the candidate entities and their properties inside the interlinking rules. Since the terms used in different ontologies for datasets can be different but with the same meaning, mapping the entities and properties of two datasets can be a challenging task. Usually a documentation is attached to the datasets with explanation of the reused or defined vocabularies. The linking administrator needs to gain a satisfactory understanding of the meaning for both sides in order to define linking candidates.
- **Selecting interlinking tools** is an important step as the performance and fitness for use of the tools defines the number of identified links between the two datasets. A number of software tools has been developed for this purpose. Among the existing tools, we selected Silk[14] [280] and LIMES[15] [196] because of their results outperforming other tools. However, depending on the characteristics of the underlying source and target datasets, an evaluation of their performance is required. Interlinking is not an stand alone task. In order to plug the tools and execution of the linking process into a scalable and automated infrastructure, both LIMES and SILK are adoptable by technical changes.

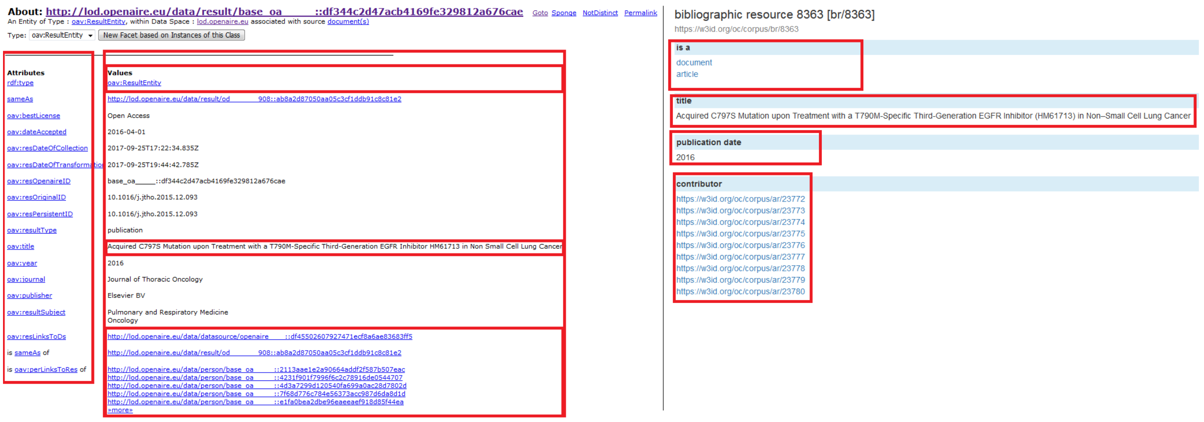
Example 1. Publication-Related There are a lot of repositories and digital libraries that are publishing metadata of scientific publications. *OpenAIRE.eu* is a metadata aggregator mainly for Open Access research results. Publications are one of the core entities of OA data model. *OpenCitations* contains information about publications and their references. Figure 3.10 shows the interlinking results of OpenAIRE LOD and OpenCitations and the RDF description of one single publication in these two datasets. The results of interlinking on the *title* property for publications identifies these entries as the same instances from the real world.

Example 2. Event-Related The following example is the metadata offered by Springer LOD about a specific conference series. The information about the same instance is present in OpenResearch.org platform. However, the metadata coverage of the two datasets have a different focus. By interlinking the two datasets on the shared properties such as the name of the conference series, the missing properties in either datasets can be unveiled through the *sameAs* links. The selected example in Listing 3.6 shows the

[13] http://lod-cloud.net/

[14] http://aksw.org/Projects/LIMES.html

[15] http://silkframework.org/



About: http://lod.openaire.eu/data/result/base_oa_____::df344c2d47acb4169fe329812a676cae Goto Sponge NotDistinct Permalink
An Entity of Type : oav:ResultEntity, within Data Space : lod.openaire.eu associated with source document(s)
Type: oav:ResultEntity ▾ New Facet based on Instances of this Class

Attributes	Values
rdf:type	oav:ResultEntity
sameAs	http://lod.openaire.eu/data/result/od_______908::ab8a2d87050aa05c3cf1ddb91c8c81e2
oav:bestLicense	Open Access
oav:dateAccepted	2016-04-01
oav:resDateOfCollection	2017-09-25T17:22:34.835Z
oav:resDateOfTransformation	2017-09-25T19:44:42.785Z
oav:resOpenaireID	base_oa_____::df344c2d47acb4169fe329812a676cae
oav:resOriginalID	10.1016/j.jtho.2015.12.093
oav:resPersistentID	10.1016/j.jtho.2015.12.093
oav:resultType	publication
oav:title	Acquired C797S Mutation upon Treatment with a T790M-Specific Third-Generation EGFR Inhibitor HM61713 in Non Small Cell Lung Cancer
oav:year	2016
oav:journal	Journal of Thoracic Oncology
oav:publisher	Elsevier BV
oav:resultSubject	Pulmonary and Respiratory Medicine Oncology
oav:resLinksToDs	http://lod.openaire.eu/data/datasource/openaire____::df45502607927471ecf8a6ae83683ff5
is sameAs of	http://lod.openaire.eu/data/result/od_______908::ab8a2d87050aa05c3cf1ddb91c8c81e2
is oav:perLinksToRes of	http://lod.openaire.eu/data/person/base_oa_____::2113aae1e2a90664addf2f587b507eac http://lod.openaire.eu/data/person/base_oa_____::4231f901f7996f6c2c78916de0544707 http://lod.openaire.eu/data/person/base_oa_____::4d3a7299d120540fa699a0ac28d7802d http://lod.openaire.eu/data/person/base_oa_____::7f68d776c784e56373acc987d6da8d1d http://lod.openaire.eu/data/person/base_oa_____::e1fa0bea2dbe96eaeeaef918d85f44ea »more»

bibliographic resource 8363 [br/8363]
https://w3id.org/oc/corpus/br/8363

is a
document
article

title
Acquired C797S Mutation upon Treatment with a T790M-Specific Third-Generation EGFR Inhibitor (HM61713) in Non–Small Cell Lung Cancer

publication date
2016

contributor
https://w3id.org/oc/corpus/ar/23772
https://w3id.org/oc/corpus/ar/23773
https://w3id.org/oc/corpus/ar/23774
https://w3id.org/oc/corpus/ar/23775
https://w3id.org/oc/corpus/ar/23776
https://w3id.org/oc/corpus/ar/23777
https://w3id.org/oc/corpus/ar/23778
https://w3id.org/oc/corpus/ar/23779
https://w3id.org/oc/corpus/ar/23780

Figure 3.10: **Interlinking Example.** The two figures show one publication instance with different properties in two different datasets, OpenAIRE LOD (a) and OpenCitations (b). Interlinking of the two datasets on the *title* and *year* properties identifies these two publications as *sameAs* links.

two entries about the *European Semantic Web Symposium* in the OR and Springer datasets. The metadata of this particular instance on OpenResearch.org lack the information about the former name of the event series. However it includes the Twitter account of the event series. Interlinking these metadata would increase the completeness of the two datasets.

NT–Same Event from OR and LOD Springer

```
<http://...confId> <...#label> "ConferenceSeries: European Semantic Web Symposium".
<http://.../core/ConferenceSeries> <...#label> "Class: Conferenceseries"@en.
<http://...confId> <http://.../core/name> "Extended Semantic Web Conference".
<http://...confId> <http://.../core/scigraphId> "...confId".
...
<http://...wikiPage/ESWC> <...#title> "Extended Semantic Web Conference".
<http://...wikiPage/ESWC> <...#hasTwitter> "@eswc_conf".
...
<http://openresearch.org/ESWC> \textbf{<.../owl#sameAs>} <http://lod.springer.com/
    esws>.
```

Listing 3.6: **XML Example.** An example representation of metadata in NT format is shown.

Example 3. OCW-Related Each OCW is created for a particular research topic. Researchers and research centers offering OCW have an ongoing research in parallel. However, the content alignment of the offered courses with an up to date ongoing research by the Creator or the community is only achievable through manual observations mainly by domain experts. However, interlinking on the topics covered in the OCW with the topics called for by publishing venues of that community can give insights in this regard. A researcher attending a conference can access a list of OCW available for each of the topics covered by that event. In contrast, students studying research topics using OCW can easily find relevant venue to submit research results or participate in.

3.4.2 Refinement and Utilization Phase

Once the required metadata is gathered and integrated, a phase of refinement is required to achieve a clean and high quality dataset. The required steps are **Enrichment**, **Curation**, **Mining**, and **Quality Assessment**. The utilization of the achieved knowledge graph ends with the two steps of **Analysis** and **Visualization** The following part of this section discusses these steps in detail.

Enrichment is a step of refinement in the metadata management cycle. This process is required when the ultimate purpose is to simultaneously improve qualitative and qualitative aspects of the underlying metadata. Enrichment creates smart data pieces containing highly-structured and informative notes for machines to refer to. The metadata extracted from resources only includes metadata from the original content. Additional enrichment of metadata is required towards increasing the completeness of the underlying dataset. Application of the enrichment methods can be divided into three use cases:

- **Materialization of the linked set** is an enrichment process directly applied on the interlinking results. Although interlinking includes unveiling links, not all of the *sameAs* links are understandable by the tools consuming them. In order to materialize links into the regional dataset, additional properties are needed to be generated and added.

- **General enrichment** steps can be performed for different case studies. In this process quality and completeness problems are solved other than the ones covered or affected by interlinking. Along the whole management process, new entities and properties can be discovered to be added to the already existing metadata. In addition,

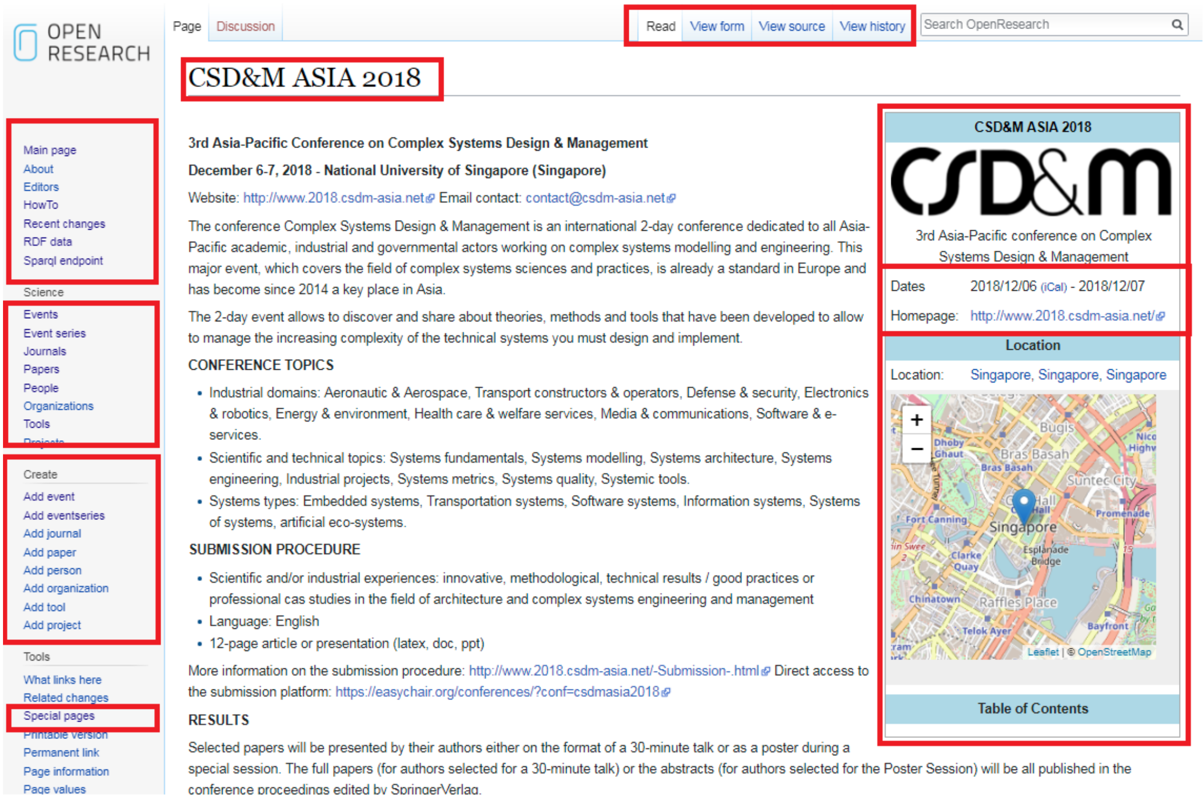
One type of enrichment is attaching additional information to various concepts (e.g., people, things, places, organizations, etc.) in a given text or any other content, so called annotation. Unlike classic text annotations, which are for the reader's reference, semantic annotations are used by machines. Any resource of data e.g., document, video enriched with semantic tags becomes a source of information that is easy to interpret, combine and reuse by machines. Semi-automatic semantic enrichment of metadata in knowledge graphs is a fundamental step for information discovery and recommendation services to explore and suggest information about items of interest. It can be used to discover related patterns and missing relationships between semantically similar or related items. In consequence, knowledge discovery and ranking services can be provided on top of the graph concepts. The set of identified metadata is semantically enriched by linking and integrating with upper level ontologies.

Example 1. Publication-Related Let us assume a list of researchers and their publications to which the desired enrichment is to connect the link of their homepage. Interlinking is not helpful as the *sameAs* links are only possible to be provided when there is a dataset containing the whole information at once. This level of enrichment is required to be executed through several individual resources. In this particular case, the information can only be found by crawling the Web by using search engines and discovering the required information and retrieving it in a reusable format. From the retrieved information, only parts of it can be added to the underlying knowledge graph.

Example 2. Event-Related Generally event metadata in any format include the information about the address of the event venue. However, with more facilitation such solid addresses can be enriched for a better usage. For example, by enriching the raw addresses and connecting to the OpenStreetMaps[16], one can provide services related to locations for event participants shown in Figure 3.11.

Example 3. OCW-Related Education material are usually provided for a local group of audience in a particular language and local sentiments. The access boundaries to scientific and educational material have been increasingly reduced with the existence of Internet and OCW. However, the majority of the metadata and the content present on the Web is limited to a number of languages, English and Chinese on top [256]. Assuming a dataset containing metadata for OCW of a certain topic where everything has an English label, it is a hard and manual task to obtain such labels in different other language. Other integration approaches lack a comprehensive support for such cases. For example, interlinking can not be helpful in this refinement phase. However, achieving multilingual labels for such metadata is possible

[16] `https://www.openstreetmap.de/`



CSD&M ASIA 2018

3rd Asia-Pacific Conference on Complex Systems Design & Management

December 6-7, 2018 - National University of Singapore (Singapore)

Website: http://www.2018.csdm-asia.net Email contact: contact@csdm-asia.net

The conference Complex Systems Design & Management is an international 2-day conference dedicated to all Asia-Pacific academic, industrial and governmental actors working on complex systems modelling and engineering. This major event, which covers the field of complex systems sciences and practices, is already a standard in Europe and has become since 2014 a key place in Asia.

The 2-day event allows to discover and share about theories, methods and tools that have been developed to allow to manage the increasing complexity of the technical systems you must design and implement.

CONFERENCE TOPICS

- Industrial domains: Aeronautic & Aerospace, Transport constructors & operators, Defense & security, Electronics & robotics, Energy & environment, Health care & welfare services, Media & communications, Software & e-services.
- Scientific and technical topics: Systems fundamentals, Systems modelling, Systems architecture, Systems engineering, Industrial projects, Systems metrics, Systems quality, Systemic tools.
- Systems types: Embedded systems, Transportation systems, Software systems, Information systems, Systems of systems, artificial eco-systems.

SUBMISSION PROCEDURE

- Scientific and/or industrial experiences: innovative, methodological, technical results / good practices or professional cas studies in the field of architecture and complex systems engineering and management
- Language: English
- 12-page article or presentation (latex, doc, ppt)

More information on the submission procedure: http://www.2018.csdm-asia.net/-Submission-.html Direct access to the submission platform: https://easychair.org/conferences/?conf=csdmasia2018

RESULTS

Selected papers will be presented by their authors either on the format of a 30-minute talk or as a poster during a special session. The full papers (for authors selected for a 30-minute talk) or the abstracts (for authors selected for the Poster Session) will be all published in the conference proceedings edited by SpringerVerlag.

CSD&M ASIA 2018

3rd Asia-Pacific conference on Complex Systems Design & Management

Dates	2018/12/06 (iCal) - 2018/12/07
Homepage:	http://www.2018.csdm-asia.net/

Location

Location:	Singapore, Singapore, Singapore

Table of Contents

Figure 3.11: **Enriched metadata**. The text address of conference venues are enriched by open street maps.

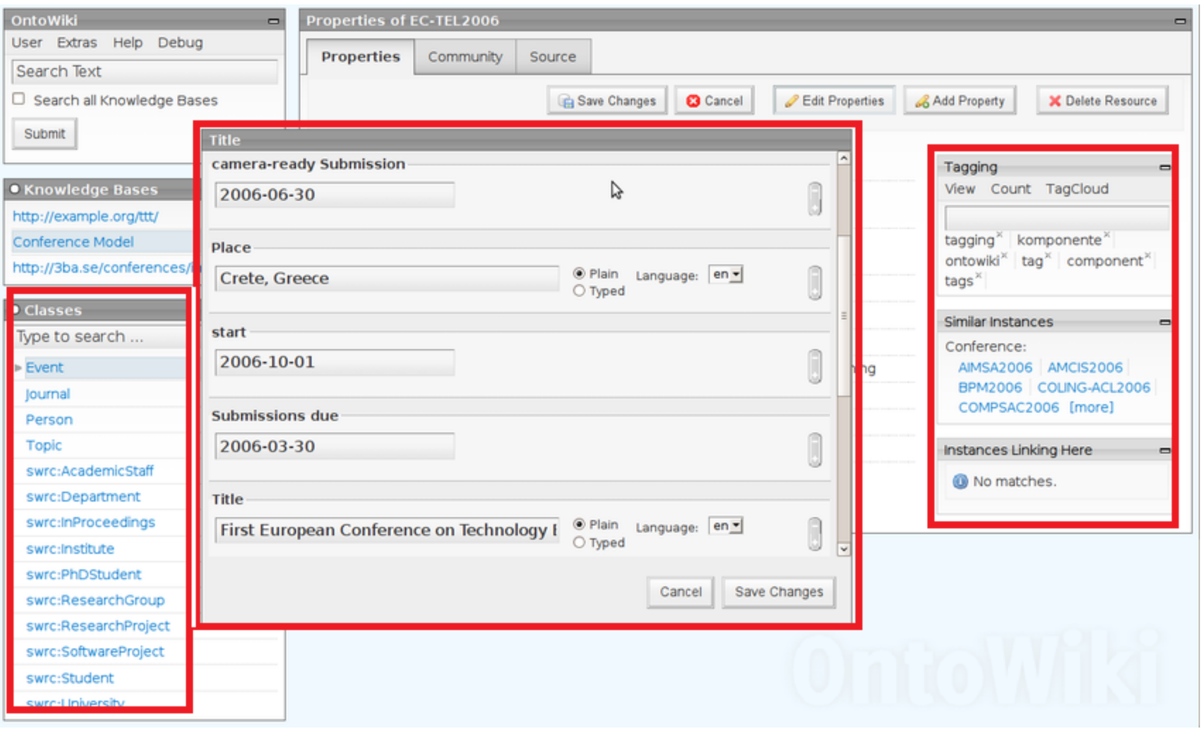
through individual enrichment approaches e.g., defining a dictionary and mapping the labels. Such approaches provide search facilities for a broader range of audience.

Curation enables availability of comprehensive and clean data. In a broad range, curation means a range of activities done to clean, manage and validate the metadata components. One of the main aspects in curation is fixing the mistakes in the data properties. Occurrence of such errors in the data originates from previous steps. However, none of the previous steps covers fixing and cleaning the collected metadata. Curation influences all the principles related to acquisition, maintenance and management of underlying metadata. In the era of big data, the curation of metadata has become more prominent, particularly for high volume and complex metadata.

The exact curation process depends on the volume of the data, the amount of noise in the data and the expected correctness level to have the data ready to use. It increases the high probability of correct data retrieval and advances maintenance of services [57]. It is typically user initiated and expected to be done by domain experts. Wiki pages and collaborative authoring systems developed for creation and curation of knowledge by community of experts. Therefore, the wiki pages and the related collaborative authoring tools have been the initial candidate for such usages. There have been several authoring tools developed for this purpose such as Semantic MediaWiki (SMW) [151] as an extension of Media Wiki, OntoWiki [11], RDF Editor [131] etc. Similar to SMW, another attempts on extending MediaWiki have been done, for example IkeWiki [237], RDFa Authoring [240].

Curation using SMW supports the user with a flexible environment enabling easy knowledge creation. It provides an interface to markup with minimal knowledge of RDF terms and syntax. A set of references to existing ontologies on the Internet for use in the markup support users with an easy way of understanding the semantics involved in wiki pages. An accurate and complete RDF dump with the ability to make modifications easily can be made available easily for the external users.

With authoring tools for RDF format, users are enabled to create precise, unambiguous encoding of information in a machine readable form.



Properties of EC-TEL2006

Title

camera-ready Submission: 2006-06-30

Place: Crete, Greece ● Plain ○ Typed Language: en

start: 2006-10-01

Submissions due: 2006-03-30

Title: First European Conference on Technology E ● Plain ○ Typed Language: en

Figure 3.12: **Metadata curation.** Semantic curation forms are used for a scientific document on OntoWiki

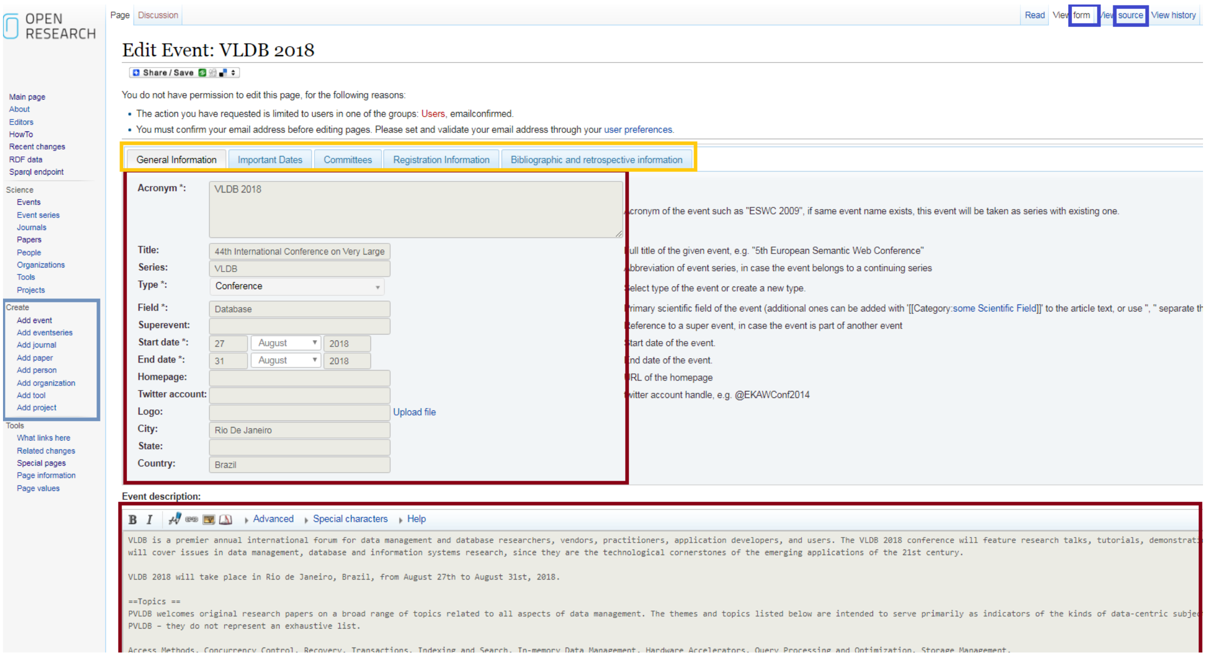
.

Example 1. Publication-Related Although, there has been a lot of progress in representing bibliographic metadata of scholarly publications e.g., Bibtex [17], curation of such data from community members and experts is not comprehensively served by any of the state of the art services. While using OntoWiki as the knowledge management system, domain experts are able to easily curate the underlying metadata. Figure 3.12 shows the form-based editor of ontowiki for a list of existing publication metadata.

Example 2. Event-Related As explained before, Semantic MediaWiki is used as one of the collaborative authoring platforms for generating knowledge by communities and domain experts. The following Figure 3.13 shows a semantic form to collect information about scholarly events. Any researcher who is either an organizer of an event or simply aware of a relevant event is able to fill the form and create a wiki page. The data filled by the creator is automatically represented semantically.

The already existing data can be curated by the same creator or a different collaborator of the platform. SMW provides two ways of curation for users: 1) semantic forms, 2) editing the source code. Using the markup language of wiki pages, one can edit the content and metadata of created articles. Wiki markup or Wikicode, consists of the syntax and keywords used by the MediaWiki software to format a page. Similar to any other platform using MW, editing can be done either through the classic editing through wiki markup (wikitext) or through a new VisualEditor (VE). In this way, everyone authorized in the system

[17] `http://www.bibtex.org`



Edit Event: VLDB 2018

You do not have permission to edit this page, for the following reasons:

- The action you have requested is limited to users in one of the groups: Users, emailconfirmed.
- You must confirm your email address before editing pages. Please set and validate your email address through your user preferences.

Acronym *:	VLDB 2018	cronym of the event such as "ESWC 2009", if same event name exists, this event will be taken as series with existing one.
Title:	44th International Conference on Very Large	ull title of the given event, e.g. "5th European Semantic Web Conference"
Series:	VLDB	bbreviation of event series, in case the event belongs to a continuing series
Type *:	Conference	elect type of the event or create a new type.
Field *:	Database	rimary scientific field of the event (additional ones can be added with '[[Category:some Scientific Field]]' to the article text, or use ", " separate th
Superevent:		eference to a super event, in case the event is part of another event
Start date *:	27 August 2018	tart date of the event.
End date *:	31 August 2018	nd date of the event.
Homepage:		RL of the homepage
Twitter account:		witter account handle, e.g. @EKAWConf2014
Logo:		
City:	Rio De Janeiro	
State:		
Country:	Brazil	

Event description:

```
VLDB is a premier annual international forum for data management and database researchers, vendors, practitioners, application developers, and users. The VLDB 2018 conference will feature research talks, tutorials, demonstrati
will cover issues in data management, database and information systems research, since they are the technological cornerstones of the emerging applications of the 21st century.

VLDB 2018 will take place in Rio de Janeiro, Brazil, from August 27th to August 31st, 2018.

==Topics ==
PVLDB welcomes original research papers on a broad range of topics related to all aspects of data management. The themes and topics listed below are intended to serve primarily as indicators of the kinds of data-centric subje
PVLDB - they do not represent an exhaustive list.

Access Methods, Concurrency Control, Recovery, Transactions, Indexing and Search, In-memory Data Management, Hardware Accelerators, Query Processing and Optimization, Storage Management,
```

Figure 3.13: **Wiki page of a scientific event**. Editing event metadata using semantic forms of SMW.

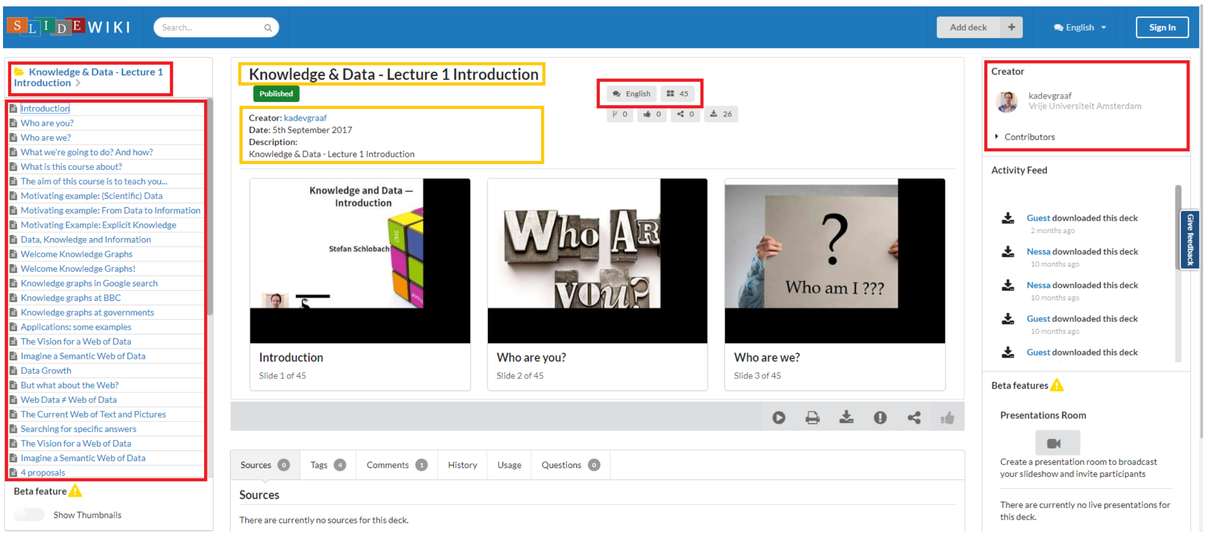
can improve articles immediately for all readers. For some special cases, the wiki pages can be protected from editing.

Example 3. OCW-Related Online courses and their material are currently following certain formats such as representation in slides, explanation in audios or videos, etc. Therefore, collaborative authoring and curation of material and metadata about educational courses requires a special platform. SlideWiki[18] is a collaborative platform that was developed for creating material for courses and their metadata. The metadata of OCW in semantic representation is shown in Figure 3.14.

Mining/Prediction Due to the dynamic nature of science, the knowledge graphs related to scientific communication can be considered incomplete by default because relations among graph entities might be unestablished, unknown or broken at the time of graph creation. In addition, many of the ranking and quality criteria in the context of scholarly communication are about the *impact* of that particular object on research. Basically, assessing the impact of something recent will require looking into the future. The knowledge graph of such a metadata management system can be used to offer predictions about what status will any object obtain in the future such as citation impact or topic movement. The approaches for knowledge extraction from huge networks by uncovering patterns and predicting emergent properties of the network can facilitate link prediction activities. Link prediction using *knowledge graph embbedings* (KGEs) received strong interest in the last years. The idea behind KGEs is to represent entities and relations of a knowledge graph (KG) into a low dimensional vector space.

Using mining and AI approaches, different types of recommendations for scholarly community (co-author recommendation for future collaboration, event recommendation for future attending, etc.) can be done by generating a scholarly knowledge graph, enriched by textual descriptions for entities, and using knowledge graph embedding models that can take advantages of textual descriptions of entities. Recommendation can be done by the entity ranking obtained from score function of embedding models.

[18] https://stable.slidewiki.org/



Knowledge & Data - Lecture 1 Introduction

- Introduction
- Who are you?
- Who are we?
- What we're going to do? And how?
- What is this course about?
- The aim of this course is to teach you...
- Motivating example: (Scientific) Data
- Motivating example: From Data to Information
- Motivating Example: Explicit Knowledge
- Data, Knowledge and Information
- Welcome Knowledge Graphs
- Welcome Knowledge Graphs!
- Knowledge graphs in Google search
- Knowledge graphs at BBC
- Knowledge graphs at governments
- Applications: some examples
- The Vision for a Web of Data
- Imagine a Semantic Web of Data
- Data Growth
- But what about the Web?
- Web Data ≠ Web of Data
- The Current Web of Text and Pictures
- Searching for specific answers
- The Vision for a Web of Data
- Imagine a Semantic Web of Data
- 4 proposals

Knowledge & Data - Lecture 1 Introduction

Creator: kadevgraaf
Date: 5th September 2017
Description:
Knowledge & Data - Lecture 1 Introduction

Knowledge and Data — Introduction
Stefan Schlobach

Introduction
Slide 1 of 45

Who are you?
Slide 2 of 45

Who am I ???

Who are we?
Slide 3 of 45

Sources

There are currently no sources for this deck.

Creator

kadevgraaf
Vrije Universiteit Amsterdam

Contributors

Activity Feed

Guest downloaded this deck
2 months ago

Nessa downloaded this deck
10 months ago

Nessa downloaded this deck
10 months ago

Guest downloaded this deck
10 months ago

Guest downloaded this deck

Create a presentation room to broadcast your slideshow and invite participants

There are currently no live presentations for this deck.

Figure 3.14: **Collaborative Creation of content for educational courses**. A course is shown in slidewiki which contains semantic representation of corresponding metadata.

Example 1. Publication-Related Hidden metadata in scholarly knowldge graphs or textual and visual representation of artifacts can help providing recommendations about relevant scientific results as well as to find potential future collaborations. Metadata about affiliation of people is embedded in scientific papers. Not only the keywords mentioned by the authors but the most used keywords in description of the work is needed to be analysed and added to the knowledge graph. Figures and other illustrations of scientific papers can be used in mining approaches.

Example 2. Event-Related Similar to publications, event metadata together with other scholarly metadata can be taken to a next level of analysis using mining approaches. An example would be event recommendation based on deep analytics.

Example 3. OCW-Related By using mining approaches the metadata about OCW and in combination of other scholarly metadata about research topics and the active people can lead to specific predictions and recommendations.

Quality Assessment A lot of data is being published on the Web with variety in quality of information covering various domains since data is merged together from autonomous sources. Datasets often contain inconsistencies as well as miss-represented and incomplete information. The quality of data is one of the important topics that affects the other steps of the life cycle. Certain criteria can be considered to define the quality of a dataset. For examples, completeness, accuracy, consistency and validity are often used for quality assessment of data.

Example 1. Publication-Related Assuming a dataset containing metadata about publications, the quality of the data cab be defined under certain criteria for example completeness. The amount of missing values for the defined metadata and properties disqualify the completeness of the dataset. Therefore, it affects the quality of the whole dataset.

Example 2. Event-Related A dataset about scholarly event can be considered as high quality data considering criteria of accuracy and timeliness. Recency of such data makes their fitness for use.

Example 3. OCW-Related In terms of OCW, consistency of the dataset can be an important quality metadata. Publishing OCW metadata and the recency of such dataset supports the users in better utilization with different purposes.

Analysis Quality metrics are “procedure(s) for measuring a quality dimension”, which “rely on quality indicators and calculate an assessment score from these indicators using a scoring function” [26]. The quality of assessed objects is analysed during this stage based on the defined metrics. This stage is about implementations of certain quality metrics specific to the domain of Scholarly Communication which can be used for ranking, filtering and recommending different component such as events in a flexible and user-defined way. The resulting framework supports the definition of quality aspects which are relevant for different stakeholders including authors/researchers, event participants, event organizers, publishers, reviewers, sponsors and organizations.

While the step considered in the life cycle is labled with the term quality of “data”, the main line of this research is about quality assessment of scholarly artifacts which is considered as a specific part of the analysis. As an essential fact for researchers, they need to keep themselves up to date about the developments of approaches, tools, and achievements related to their topics of interest. Junior researchers need to have an overview of the already existing related work and senior ones mostly need to be aware of ongoing research activities of other parties. Such information are mostly embedded inside the scientific papers, technical reports or web pages introducing any particular development. For example, the content of a scientific paper contains information about the problem addressed by that work.

The types of metadata that should be considered for extraction and evaluation varies for every domain. The proper identification of metadata is particularly challenging and important when the metadata is planned to be exploited for determining as quality criteria for the domain objects. An expert or knowledge engineer identifies a set of metadata items and related quality metrics. Any metric has a precise definition by which its exact value can be computed from metadata. We propose a framework for identification and classification of quality indicators that follows the standard terminology of *data quality* research, with the key terms of *category*, *dimension* and *metric*.

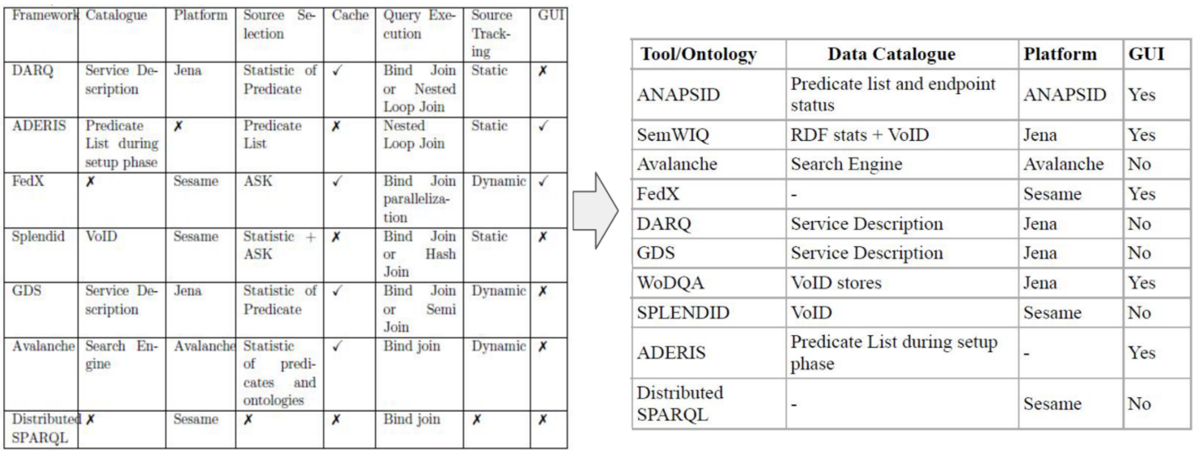
Example 1. Publication-Related Survey papers include information about different but high qualified papers, tools, approaches of an specific research topic. It is often a challenge to provide a comprehensive survey paper. It requires a very high level understanding and broad view of the domain. Due to the vast amount of information published everyday, many recent and important research results can be overlooked o be considered while preparing a survey paper. On the other hand, for readers of the survey papers, information is often transferred with a lot of pointers to the original work. Semantic representation of information about tools, frameworks, and any research results helps in generating systematic and comprehensive overviews and analytics. The Figure 3.15 shows a result of analysis that is done over semantic representation of several tools and approaches developed for “Federated Query Languages”. A survey table is generated out of the metadata semantically represented for relevant information. The table is closely identical and more comprehensive than the summary provided in relevant survey papers for this topic.

Example 2.Event-Related Using the *ASK* queries of the mediawiki on OpenResearch.org platform, a table of alalytics is created for scientific events with certain criteria about their CORE ran and acceptance rate. Listing 3.7 shows this example generated on OR.

ASK query

```
{{#ask:[[Category:Event series]] [[Category:{{#urlget:field}}]]
| ?Title
| ?Homepage
| ?has CORE2017 Rank
| ?has Average Acceptance Rate
}}
```

Listing 3.7: **Ask query.** An example of a query about events with CORE rank and certain acceptece rate is shown.

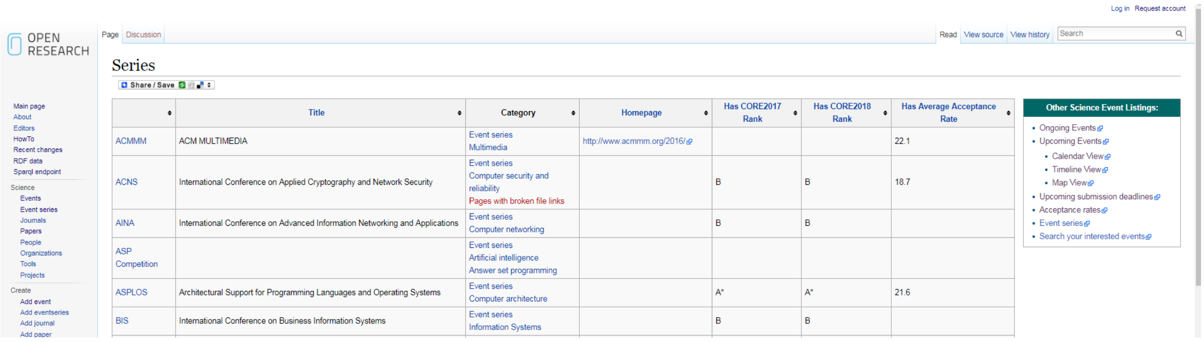


Framework	Catalogue	Platform	Source Selection	Cache	Query Execution	Source Tracking	GUI
DARQ	Service Description	Jena	Statistic of Predicate	✓	Bind Join or Nested Loop Join	Static	✗
ADERIS	Predicate List during setup phase	✗	Predicate List	✗	Nested Loop Join	Static	✓
FedX	✗	Sesame	ASK	✓	Bind Join parallelization	Dynamic	✓
Splendid	VoID	Sesame	Statistic + ASK	✗	Bind Join or Hash Join	Static	✗
GDS	Service Description	Jena	Statistic of Predicate	✓	Bind Join or Semi Join	Dynamic	✗
Avalanche	Search Engine	Avalanche	Statistic of predicates and ontologies	✓	Bind join	Dynamic	✗
Distributed SPARQL	✗	Sesame	✗	✗	Bind join	✗	✗

Tool/Ontology	**Data Catalogue**	**Platform**	**GUI**
ANAPSID	Predicate list and endpoint status	ANAPSID	Yes
SemWIQ	RDF stats + VoID	Jena	Yes
Avalanche	Search Engine	Avalanche	No
FedX	-	Sesame	Yes
DARQ	Service Description	Jena	No
GDS	Service Description	Jena	No
WoDQA	VoID stores	Jena	Yes
SPLENDID	VoID	Sesame	No
ADERIS	Predicate List during setup phase	-	Yes
Distributed SPARQL	-	Sesame	No

Figure 3.15: **A framework comparison table generated manually by researchers and the same table by a query on OR**. A table included in a survey paper (left-side) is compared to the table generated (right-side) by our approach as the results of querying the Aurora knowledge graph. A more detailed and fine-grained description of the surveyed approaches can be generated.

The result of such queries can be shown in a table view by default. Every query on OR or any SMW platform can be saved for general usage. Side bars are used in special pages for saving such queries as shown in Figure 3.16.



Series

	Title	Category	Homepage	Has CORE2017 Rank	Has CORE2018 Rank	Has Average Acceptance Rate
ACMMM	ACM MULTIMEDIA	Event series Multimedia	http://www.acmmm.org/2016/			22.1
ACNS	International Conference on Applied Cryptography and Network Security	Event series Computer security and reliability Pages with broken file links		B	B	18.7
AINA	International Conference on Advanced Information Networking and Applications	Event series Computer networking		B	B	
ASP Competition		Event series Artificial intelligence Answer set programming				
ASPLOS	Architectural Support for Programming Languages and Operating Systems	Event series Computer architecture		A*	A*	21.6
BIS	International Conference on Business Information Systems	Event series Information Systems		B	B	

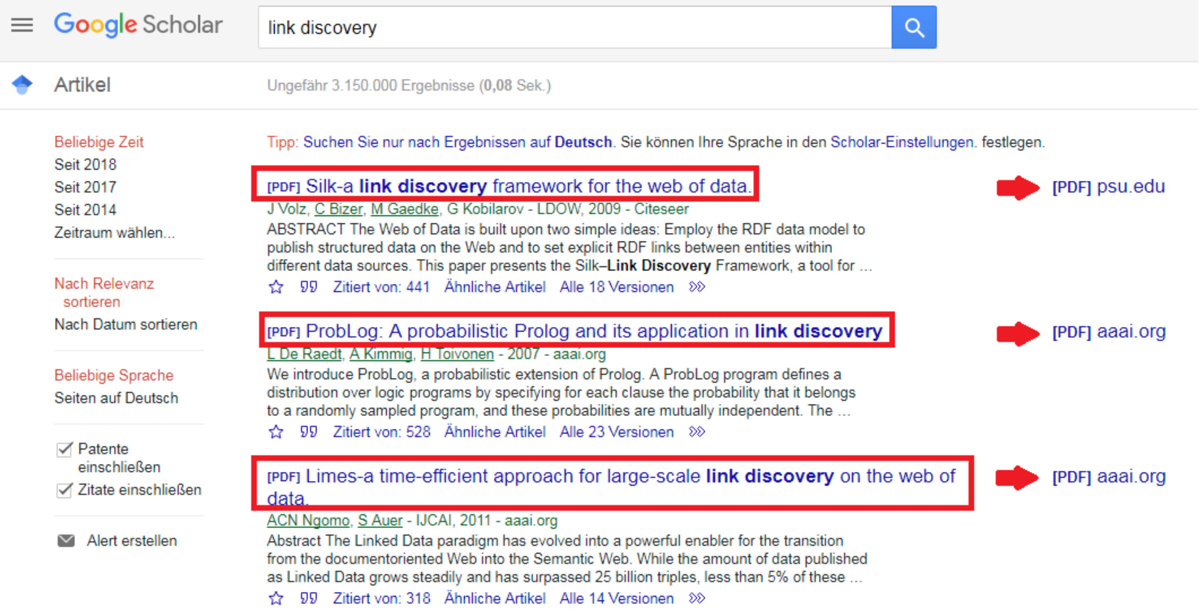
Other Science Event Listings:
- Ongoing Events
- Upcoming Events
  - Calendar View
  - Timeline View
  - Map View
- Upcoming submission deadlines
- Acceptance rates
- Event series
- Search your interested events

Figure 3.16: **Predefined queries**. Fixed queries can be saved as a side bar in a special page.

Example 3. OCW-Related A faceted browser can be used in order to provide analytics about OCW. It is often a need to find OCW inside an online platform under certain criteria. For example, analysis over the number of creation and contributors of the course material with regards to the diversity in language, illustrations etc.

Visualization Metadata of scholarly communication is heterogeneous with a large variety of entities and many types of interrelationships. Each entity is different from others in number and type of attributes. Visualization of such a heterogeneous metadata graph using different views and models such as timelines, calendars, and etc can help users to have a better understanding. As a final stage of the metadata management cycle, metadata visualization is required. Visualization helps in gaining a better understanding of the underlying knowledge.

Example 1. Publication-Related Currently the result of keyword based queries over the popular search engines for scholars are shown as a list, see Figure 3.17.



Ungefähr 3.150.000 Ergebnisse (**0,08** Sek.)

Tipp: Suchen Sie nur nach Ergebnissen auf **Deutsch**. Sie können Ihre Sprache in den Scholar-Einstellungen. festlegen.

[PDF] Silk-a **link discovery** framework for the web of data.
J Volz, C Bizer, M Gaedke, G Kobilarov - LDOW, 2009 - Citeseer
ABSTRACT The Web of Data is built upon two simple ideas: Employ the RDF data model to publish structured data on the Web and to set explicit RDF links between entities within different data sources. This paper presents the Silk–**Link Discovery** Framework, a tool for ...
Zitiert von: 441 Ähnliche Artikel Alle 18 Versionen
[PDF] psu.edu

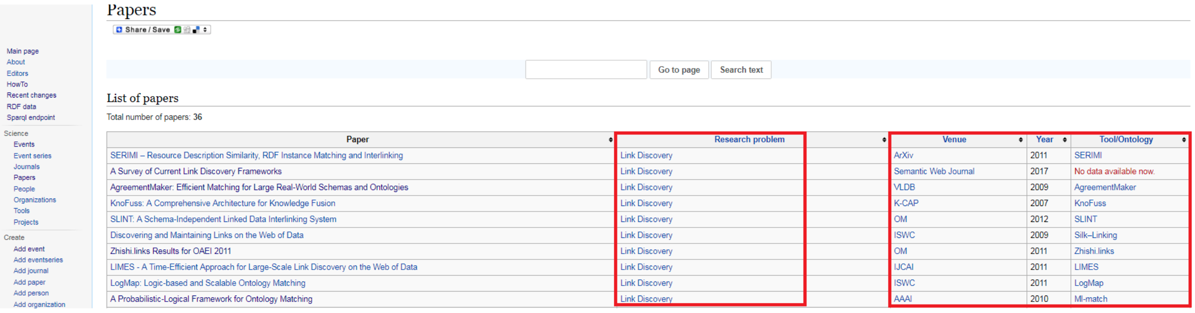
[PDF] ProbLog: A probabilistic Prolog and its application in **link discovery**
L De Raedt, A Kimmig, H Toivonen - 2007 - aaai.org
We introduce ProbLog, a probabilistic extension of Prolog. A ProbLog program defines a distribution over logic programs by specifying for each clause the probability that it belongs to a randomly sampled program, and these probabilities are mutually independent. The ...
Zitiert von: 528 Ähnliche Artikel Alle 23 Versionen
[PDF] aaai.org

[PDF] Limes-a time-efficient approach for large-scale **link discovery** on the web of data.
ACN Ngomo, S Auer - IJCAI, 2011 - aaai.org
Abstract The Linked Data paradigm has evolved into a powerful enabler for the transition from the documentoriented Web into the Semantic Web. While the amount of data published as Linked Data grows steadily and has surpassed 25 billion triples, less than 5% of these ...
Zitiert von: 318 Ähnliche Artikel Alle 14 Versionen
[PDF] aaai.org

Figure 3.17: **Publication metadata on Google Scholar**. The list of publications and their metadata is retrieved from keyword search on Google Scholar.

In order to access the full article, researchers need to proceed several links provided aside of the each entry in the list. The list is ranked based on the Google page ran as well as the number of citation. Full title of the scientific papers are retrieved with their matching part on the search keyword.

However, the semantic representation of metadata enables simple but comprehensive visualization. The Figure 3.18 shows a different representation of the metadata about scientific papers on OR.



Papers

List of papers

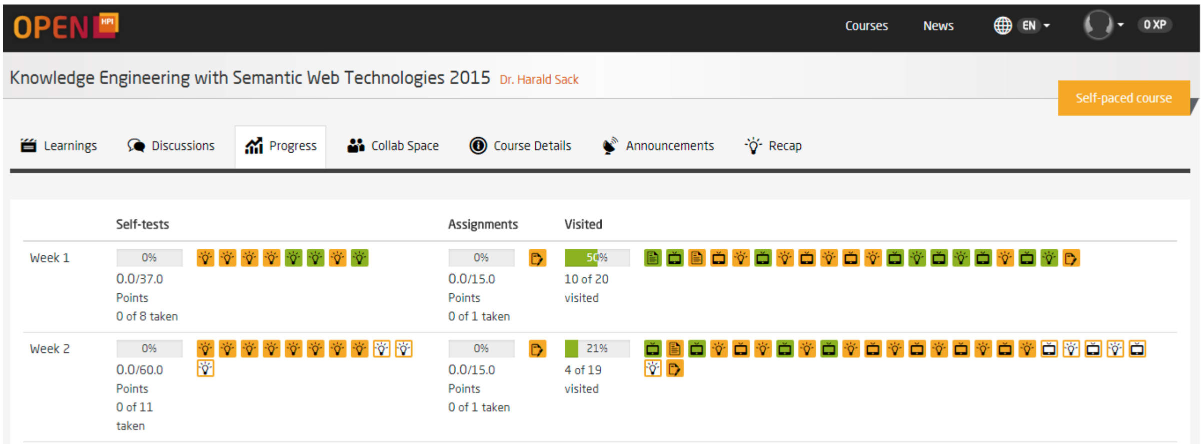
Total number of papers: **36**

Paper	Research problem		Venue	Year	Tool/Ontology
SERIMI – Resource Description Similarity, RDF Instance Matching and Interlinking	Link Discovery		ArXiv	2011	SERIMI
A Survey of Current Link Discovery Frameworks	Link Discovery		Semantic Web Journal	2017	No data available now.
AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies	Link Discovery		VLDB	2009	AgreementMaker
KnoFuss: A Comprehensive Architecture for Knowledge Fusion	Link Discovery		K-CAP	2007	KnoFuss
SLINT: A Schema-Independent Linked Data Interlinking System	Link Discovery		OM	2012	SLINT
Discovering and Maintaining Links on the Web of Data	Link Discovery		ISWC	2009	Silk–Linking
Zhishi.links Results for OAEI 2011	Link Discovery		OM	2011	Zhishi.links
LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data	Link Discovery		IJCAI	2011	LIMES
LogMap: Logic-based and Scalable Ontology Matching	Link Discovery		ISWC	2011	LogMap
A Probabilistic-Logical Framework for Ontology Matching	Link Discovery		AAAI	2010	MI-match

Figure 3.18: **Table visualization of publication metadata**. Currently the already existing services retrieve a list.

Example 2.Event-Related Similar to publication, different visualizations are possible to ease the understanding of the metadata about events. In this way, inference of the information can be easily and effectively done. For example, the topic movement in research depending on emergence of scientific events can be visualized.

Example 3. OCW-Related Figure 3.19 shows the visualization of metadata bout an online course over OpenHPI. Such visualization makes the tracking of the progress easier for lecturers and the students.



Knowledge Engineering with Semantic Web Technologies 2015 Dr. Harald Sack

	Self-tests	Assignments	Visited
Week 1	0% 0.0/37.0 Points 0 of 8 taken	0% 0.0/15.0 Points 0 of 1 taken	50% 10 of 20 visited
Week 2	0% 0.0/60.0 Points 0 of 11 taken	0% 0.0/15.0 Points 0 of 1 taken	21% 4 of 19 visited

Figure 3.19: **Progress Metadata depicted on OpenHPI**. Visualization of progress shows how much of the course is studied by the student as well as the remaining chapters and exercises.

The proposed life cycle explicitly aims to provide a detailed and comprehensive list of the processes and practices necessary for metadata management and can be used a reference to data science, metadata curation and metadata management activities. The cycle is a comprehensive checklist of data management practices that merit attention in a data management plan.

CHAPTER 4

Quality Assessment of Scholarly Artifacts

Due to the often subjective nature of the concept of *quality* in research, there exist several definitions by different researchers. Quality is defined as excellence, value, conformance to specifications, or meeting user expectations [129]. More generally, it is widely accepted as *fitness for use* [127, 144]. Application of this meaning to the scholarly communication reflects the extent to which the totality of features and characteristics of an artefact led to a successful fulfillment of scholar's needs. The quality of scholarly artifacts and other elements of scholarly communication such as events have multiple characteristics. Researchers combine assessments of these characteristics in different ways depending on their view or task. For researchers, upcoming events on a specific topic can be interesting with regard to the closeness of the location, the validity of the publisher and the reputation of speakers and organizers. Another researcher can only focus on the reputation of the event with regard to the acceptance rate. Therefore, depending on incentive and objectives of the individual researchers or communities, there are wide range of requirements and needs in the context of scholarly communication domain. As pointed out in Table 2.1, the current services lack a comprehensive support of quality-related aspects of scholarly communication. To automate systematic quality assessment at a large scale, it is, therefore, crucial to supply such characteristics with corresponding rich metadata for assessing the characteristics.

In order to facilitate the scholarly metadata management, modeling of the domain is the starting point. Within the scope of this thesis, the domain modeling and further implementations for scholarly communication has been done for three types of artifacts in the context of three different projects: OCW (SlideWiki [1]), Publication and Datasets (OpenAIRE [2]) and Events (OpenResearch [3]).

OpenCourseWare as the one of the addressed artifact in this thesis has been the main focus of SlideWiki project. In this regard, the domain modeling for leaning material has been done theoretically and the implementation of the defined concepts left on further versions of the platform.

Publication and Datasets have been addressed in OpenAIRE project which aggregates metadata about research (projects, publications, people, organizations, etc.) into a central Information Space. OpenAIRE LOD aims at increasing interoperability and reusability of scholarly metadata and open access publications by exposing it as Linked Open Data (LOD). By following the LOD principles, it is now possible to further increase interoperability and reusability by connecting the OpenAIRE LOD to other datasets about projects, publications, people and organizations. Doing so required us to represent the OpenAIRE data model using linked open vocabularies.

Events are the focused entities of the OpenResearch project (will be discussed in details in chapter 6),

[1] http://www.slidewiki.org
[2] http://www.openaire.eu
[3] http://www.openresearch.org

the objective is to provide a community based platform in order to manage scholarly metadata of events e.g., conferences, workshops. In section 4.1, a comprehensive description of the scholarly metadata domains on the example of Events, OCW and Publications is provided with regard to quality aspects. The methodology that is used in this thesis in order to design the quality assessment framework for scholarly artifacts and events is explained in section 4.2. The quality assessment of scholarly artifacts with this methodology is done on the example of OCW section 4.3 and scholarly event metadata section 4.4. In addition, an alternative approach is used for quality assessment of publication metadata which will be discussed in section 4.5. Two different steps of the life cyclesection 3.4 is addressed in this chapter:*Eligibility Checking* and *Quality Assessment*. These two stages are complementary from the following perspectives:

Pre-extraction perspective which includes *Eligibility checking* of scholarly artifacts that is generally done for the purpose of identifying candidate metadata to be extracted. This process has been applied over a selective dataset of OCW in order to obtain a better understanding of their quality. Another attempt was to bootstrap a value chain for scientific data to enable services, such as assessing the quality of scientific output with respect to novel metrics. Description of this research is presented in the remainder of this chapter based on the following publications[4].

> **Sahar Vahdati**, Christoph Lange, Sören Auer. *OpenCourseWare observatory: does the quality of OpenCourseWare live up to its promise?* In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge 2015;
> **Sahar Vahdati**, Anastasia Dimou, Christoph Lange, Angelo Di Iorio. *Semantic Publishing Challenge: Bootstrapping a Value Chain for Scientific Data* In Proceedings of Semantics, Analytics, Visualisation: Enhancing Scholarly Data Workshop 2016.

Post-extraction perspective *Quality assessment* of scholarly artifacts by analysing extensively extracted, curated or crowdsourced metadata. Semantic representation of scholarly event metadata has been considered in this process. A comprehensive framework of assessment metrics for evaluating scientific events and their series is developed. The resulting quality metrics are specified based on a conceptual model of events, their stakeholders, and the publications that result from them. Description of this research is presented in the remainder of this chapter based on the following publication.

> **Sahar Vahdati**, Christoph Lange, Sören Auer, Andreas Behrend. *Towards a Comprehensive Quality Assessment Model for Scientific Events* In Scientomerics Journal 2018;
> Anastasia Dimou, **Sahar Vahdati**, Angelo Di Iorio, Christoph Lange, Ruben Verborgh, and Erik Mannens, *Challenges as Enablers for High Quality Linked Data: Insights from the Semantic Publishing Challenge*, PeerJ 2017.

Where open scholarship encompasses all forms of openness such as open data, open educational resources, still, a large number of scholarly communication processes and artifacts (other than publications) are not currently well supported. In addition, publication-related evaluations have been used for almost any measurement in research-related rankings without any serious emphasis on the assessments of the other scholarly artifacts.

Providing such information to researchers supports them with a broader range of options and a comprehensive list of criteria. Current part of this research describes how to analyze and identify a set

[4] **Own Manuscript Contributions.** The author of this thesis has been the first author of the mentioned publications with main contributions on conception and doing research work with significant supervision of the seniors. The articles co-authored with Dimou et al. are a join work with co-chairs of semantic publishing challenge series. Vahdati with supervision of Lange was mainly responsible for Task 1, definition as well as the design and evaluation of the quality assessment queries

of novel indicators for the quality assessment of scholarly artifacts and correlate them to channels of dissemination. The building blocks of this approach are a) semantic enrichment strategies and b) quality assessment methods. An additional goal of this work is to develop strategies for identifying high-quality artifact recommendations using enriched metadata.

4.1 Metadata Domains of Scholarly Communication

The prerequisite to have an information system facilitating scholarly metadata manage is a deep understanding of the domain and representation of entities and relationships Since the proposed approach in this thesis is based on Link d Data technologies, the domain is modeled to a RDF knowledge graph. Towards this objective, a step of ontology engineering is needed in order to do the conceptualization and representation of the domain data model in a suitable standard LOD vocabularies [5]. The specifically tailored nature of the scholarly communication, its large amount of quantity and the frequent updates, pose high requirements on the technology chosen for representing scholarly metadata in Linked Open Data (LOD). It is very important to have a clear specification of the selected or defined vocabularies. During the modeling process, a data engineer has to decide with which vocabularies to express the data [238]. According to Linked Data best practices, modeling starts by reusing already existing vocabularies [29]. We use Linked Open Vocabularies (LOV) [6] to explore existing vocabularies. The challenging aspect of this step is the conceptualization of the concepts and making decision about the vocabulary to reuse in modeling the scholarly communication domain. The current scholarly communication considering publishing habit, has certain stakeholders, research results as artifacts and events, organizations with

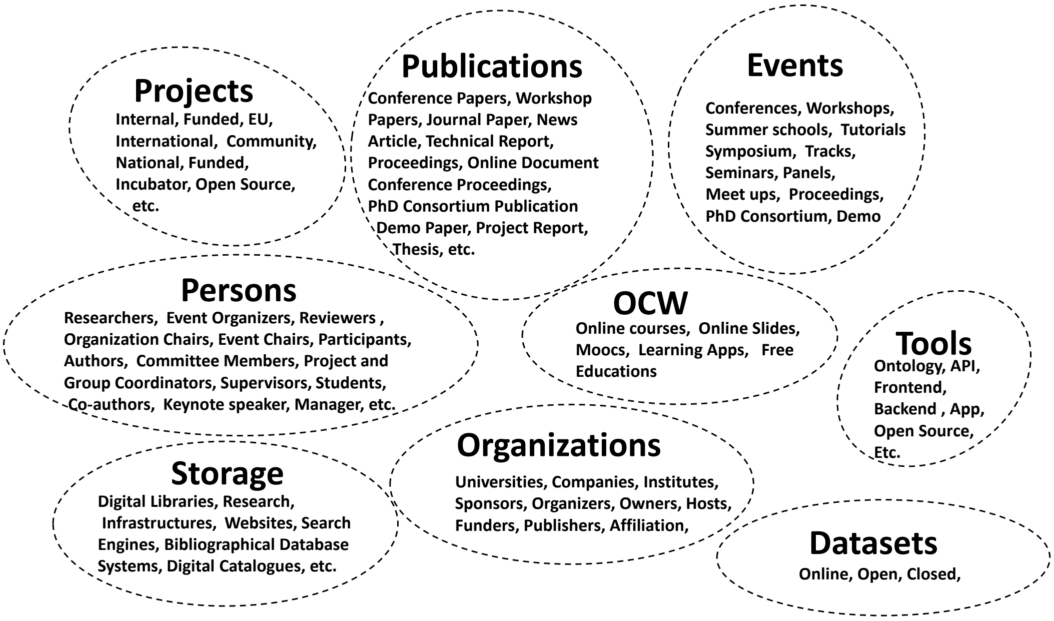

Projects
Internal, Funded, EU, International, Community, National, Funded, Incubator, Open Source, etc.
Publications
Conference Papers, Workshop Papers, Journal Paper, News Article, Technical Report, Proceedings, Online Document Conference Proceedings, PhD Consortium Publication Demo Paper, Project Report, Thesis, etc.
Events
Conferences, Workshops, Summer schools, Tutorials Symposium, Tracks, Seminars, Panels, Meet ups, Proceedings, PhD Consortium, Demo
Persons
Researchers, Event Organizers, Reviewers , Organization Chairs, Event Chairs, Participants, Authors, Committee Members, Project and Group Coordinators, Supervisors, Students, Co-authors, Keynote speaker, Manager, etc.
OCW
Online courses, Online Slides, Moocs, Learning Apps, Free Educations
Tools
Ontology, API, Frontend, Backend , App, Open Source, Etc.
Organizations
Universities, Companies, Institutes, Sponsors, Organizers, Owners, Hosts, Funders, Publishers, Affiliation,
Storage
Digital Libraries, Research, Infrastructures, Websites, Search Engines, Bibliographical Database Systems, Digital Catalogues, etc.
Datasets
Online, Open, Closed,

Figure 4.1: **Domains in Scholarly Communication**. Artifacts and stakeholders and events involved in the scholarly communication is represented.

[5] Throughout this document the terms “ontology” and “vocabulary” will be used indistinctly.

[6] `http://lov.okfn.org/dataset/lov/`

complex relationships. A high-level representation of the concepts in the scholarly communication domain is shown in Figure 4.5. Due to the complexity of the relationships between entities and their properties, a more detailed version of the model is skipped. However, examples of the modeling will be discussed about the focused artifacts in this thesis namely OCW, Publications and Datasets, and Events.

4.1.1 Conceptualization

The selected concepts of the scholarly communication have been conceptually modeled. In modeling the ontology of scientific communication, we followed the best practices of *reusing* terms from existing ontologies [208] and applying ontology *design patterns* [91]. For any further terminology not sufficiently covered by existing ontologies, we defined our own ontology. The methodology used for modeling follows organizing all these information is a directed graph. The modeling is based on a three level categorization: a) Core classes: the entities whose information is continuously and incrementally fed to the information space; namely Events, OCW, Publication, Dataset, Person, Organization, DataSource , Projects. On language level a class is a thing with *rdf:type* property has owl:Class as value. b) Properties: the relationship between classes, used either to connect in a semantic-agnostic way two (or more) core entities for example affiliation of a researcher to an organization or the relation name of the property for example an acronym of an event. A property value can be a data value described by RDF literals (“string”, “decimal”, “date” etc.) or a property value can be a link to another thing. In this section, the core entities and their relations to each other are represented.

Modeling of OpenCourseWare Domain: In order to model the domain of scholarly communication related to online courses, a comprehensive study of the OpenCourseWare repositories has been done [270]. In this study, the main classes and relationships between them have been defined with the perspective of an ontology engineer. As a results, fine main entities have been defined as follows:

- OpenCourseWare is a scientific and educational output designed in the form of a lecture.
- Person is a person involved in the scholarly communication chain in creating the material for the course or can be a participant of the course. In a collaborative environment for creating content, participants can also be treated as content creators.
- Platform represents online service through which the online courses have been made available for the target users (learners, students, lecturers etc.).
- Organization addresses the institute in which the content creator is affiliated, thus the course is taught there or the the organization to which the platform belongs.
- Material represents the content of a course created by a person. A course can be offered in several formats such as video, text, audio etc.

Relationships of the entities can be complex in this mode. In Figure 4.2, a high-level sketch of the OCW domain is presented. Online courses are created by individual researcher or a group. The course and its material need to be offered via an online platform. Generally, institutes and universities have their own platform. Properties of the main entities will be discussed in section 4.3.

Modeling of Artifacts (Publication and Datasets) on the LOD version of OpenAIRE.eu: The preliminary requirement is to understand the overall schema of OpenAIRE data [174]. After a systematic study on the OpenAIRE data model, the main entities and their relationships have been captured. In the second step, conceptualization of the entities and relationships as well as finding suitable already existing vocabularies is important. The mapping should be faithful to the specification of the OpenAIRE data model (no information should be lost). The resulting LOD should be useful, where useful means easy to connect to other linked datasets and easy to consume (for example: easy to query). A set of vocabularies has been used in order to capture the OpenAIRE data model in RDF graph. However, due to technical

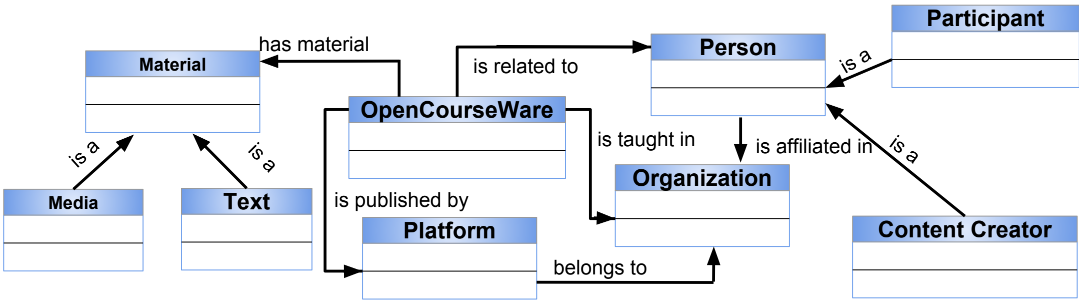


Figure 4.2: **The core concepts of the OCW domain**. Part of the main concepts related to OCW is shown.

issues all the identified vocabularies have been mapped to the two specific ontologies of OpenAIRE data model: OpenAire Vocabulary (abbreviated as OAV)[7], OpenAIRE Data(abbreviated as OAD)[8]. The ontology OAV ontolody is a specification of all metadata terms used or created in providing OpenAIRE LOD services including properties, vocabulary encoding schemes, syntax encoding schemes, and classes. A detailed version of the ontology is described in [274].

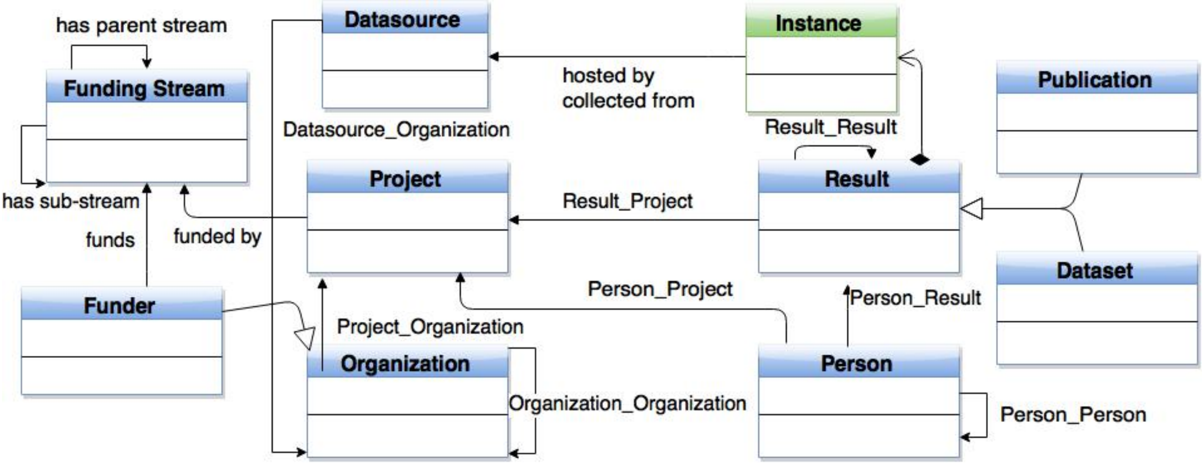


Figure 4.3: **The core concepts of the OpenAIRE ontology** Part of the main concepts related to scholarly publications and datasets is shown.

In the OpenAIRE data model, there are six core entity types.

- Result is a scientific output resulting out of one or more projects. A Result entity can either be a Dataset or a publication.
- Person is a person involved in the scholarly communication chain, such as scientific publications' authors, contributors, data scientists and project coordinators.
- Project is a research project.
- Organization addresses an organization involved in the scholarly communication chain, such as companies, research centres and institutions involved as project partners or as being responsible for operating data sources.

[7] http://lod.openaire.eu/vocab
[8] http://lod.openaire.eu/data

- Datasource represents the metadata of a provider exporting (meta)data about scholarly communication objects.
- Funding Stream identifies the hierarchies of fundings. Funding streams can be nested in a tree of sub-funding streams, including the funder as root and context, program and framework program as trunks. Projects are typically associated to the funding stream “leaves” of such trees.

Based on the definition of each entity, the identical terms have been selected from standard vocabularies shown in the following Listing 4.1.

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix cerif: <http://www.eurocris.org/ontologies/cerif/1.3> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .
@prefix swpo: <http://sw-portal.deri.org/ontologies/swportal> .
@prefix prov: <http://www.w3.org/TR/prov-o/> .
@prefix schema: <http://schema.org/> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix api: <http://purl.org/linked-data/api/> .
@prefix frapo: <http://purl.org/cerif/frapo/> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix dcite: <http://purl.org/spar/datacite/> .
```

Listing 4.1: **Prefixes.** Part of the selected vocabularies for reuse in OpenAIRE LOD ontology.

Certain problems have been identified because of the maintenance issues in the selected ontologies for reuse. Therefore, all the selected terms from the already existing ontologies have been mapped with *owl:sameAs* to the name space of *oav*. Currently all terms and relations in OpenAIRE LOD is defined under *oav* ontology and benefit from hierarchical connection to other reused ontologies. The properties has certain meaning for example *Description* is the abstract of the Result entity and *Subject* the scientific discipline(s) covered by the Result. The same procedure is applied analogously to define all the other concepts in OpenAIRE data model.

Attribute Name	Property	Range
Title	cerif:name	xsd:string
DateOfAcceptance	dcterms:dateAccepted	xsd:date
Publisher (optional)	dcterms:publisher	xsd:string
Description	dcterms:description	xsd:string
Language	dcterms:language	xsd:string
Subject	oav:resultSubject	xsd:string
Country	dbpedia-owl:country	xsd:string

Table 4.1: **Sample Properties of the core entities**. Properties of the publication entity is shown.

Modeling of Events for OpenResearch.org: The central objects of scholarly communication are publications together with datasets by which scientists exchange knowledge. Scientific events and journals are the main channels of this communication. Organizations including companies, research centers, institutions and etc are involved as project partners or as responsible of operating data sources.

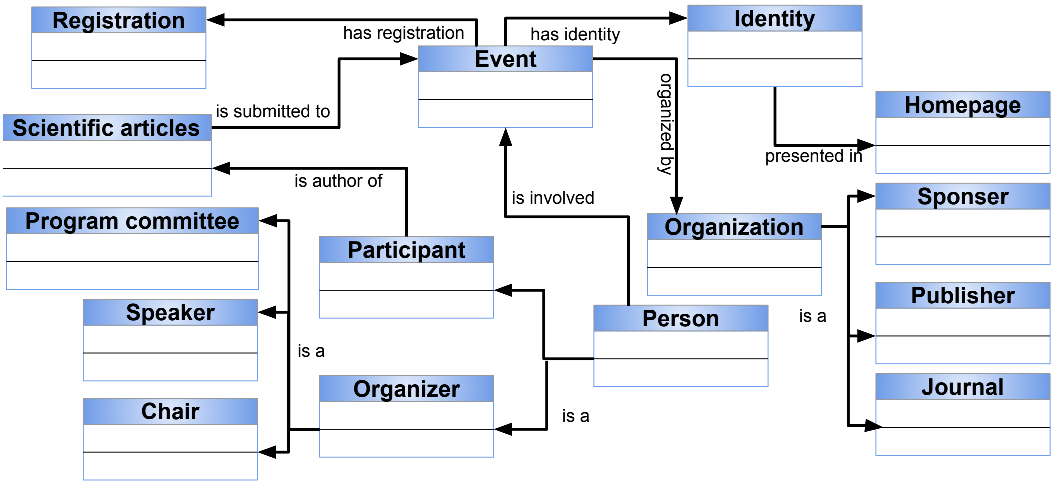


Figure 4.4: **The core concepts of the OpenResearch.org ontology** Part of the main concepts related to scholarly events is shown.

Metadata domains include all the artifacts and stakeholders in the scholarly communication system, such as Publications, Datasets, Persons, Events, Organizations, Projects, and Tools. Focusing on scholarly events, representation of the core classes in or ontology are: *or:ScientifcEvent)*, *or:EventSeries*, *or:Track*, *or:Symposium*. In order to represent sub-classes, let us consider *or:ScientifcEvent* class which will follow a subclass *or:Conference*. Again we would have a triple *or:Conference rdf:subClassOf or:ScientificEvent*. To show this object property we draw an arrow with a white arrowhead from the subclass to the superclass. *orEventSeries* is superclass of *Event (single event)*. Formalization of the model was not deeply explored since it was out of scope for this thesis.

The conceptual model of the events domain has been illustrated in Figure 4.4, there are six core entity types.

- Event is a scientific gathering of scholars who are working on similar topics. Research results as articles are submitted to the events and accepted ones are presented.
- Person is a scholar involved in the scholarly communication chain during the organization and holding phase of the event, such as scientific chairs, other organizers, reviewers, participants, authors, speakers etc.
- Organization addresses the institutes or universities which are holding the event. Usually this points to the affiliation of the main chairs.
- Scientific Articles are the communication means of the scholarly events. Researchers submit their research results and those passing the review phase successfully are presented in the event.
- Registration to the event is one of the main activities. It is not enough to have an accepted work, scholars need to register to the vents and it has its own process.
- Identity shows the ways the abstract concept of event is presented to the scholarly communities. It can point to the event homepage, call for paper emails etc.

4.1.2 Implementation

In this section, the discussion on the domain modelings follows with the implementation of the developed models for the three artifacts, OCW, Publications, Events (as discussed before). Due to the technical needs and priorities of this research with regard to its objectives, the implementation of the data models for these three domains have not been developed in the same level nor in the same platform. The conceptualized domain of the scholarly communication is aimed to be shown as RDF triples by utilizing RDF [9], RDF Schema [10] and OWL [11].

Implementation of the OpenCourseWare Model: The model that was conceptualized for OCW was extended as a quality framework for online courses (will be discussed in section 4.3). The implementation of the defined concepts was left on the shoulders of the SlideWiki.org developers. To the best of our knowledge, most of the defined terms have been turned to a feature inside the platform. The semantic representation of the concepts such as content Creator, language information and material of the OCW are developed [77] in the recent version of SlideWiki.org platform which was based on the work done by the conceptualization of this research.

Implementation of the OpenAIRE LOD Model: The database management system used for OpenAIRE LOD is Virtuoso[12]. It provides an environment to crate the graph of data and based on the injected ontology. The ontology developed for OpenAIRE LOD based on the initial data model has been created as a graph also. At the time of loading data into Virtuoso, the ontology is also imported using the Conductor user interface [13]. All the selected or defined ontologies are imported using the following command: *SPARQL LOAD URL of the Ontology1; Ontology 2,...*. Then the data is connected to the imported ontologies. In this way the OpenAIRE LOD graph has been created. In addition to the concepts defined inside the ontology, one needs to introduce the ontology, its license, online link, the graph namespace prefix, and date or creation at the time of importing the ontology. Parts of the technical challenges that was faced in this step is skipped to be discussed here. A detailed description of the required steps can be found in the main help page by Virtuoso[14].

Implementation of the OpenResearch.org Model: In modeling our ontology of scientific events and their stakeholders (participants, organizers, publishers, sponsors, etc.), we followed the best practices of *reusing* terms from existing ontologies (cf. [208]) and applying ontology *design patterns* [91]. Domain-specific candidates for reuse include generic ontologies about publishing as well as ontologies about scientific conferences. The GND ontology [94] defines authorities established in publishing, including events and persons and their roles. The Semantic Web Conference ontology (SWC) considers academic conferences [188]; it has originally been designed to support the European and International Semantic Web Conferences. The Ontology Alignment Evaluation Initiative (OAEI) provides further conference ontologies that vary in size, language, domain, and modeling style (cf. [301] for the 2015 conference ontologies and [262] for the full series of OAEI evaluation events). For any further terminology not sufficiently covered by existing ontologies, we defined our own ontology called OpenResearch (OR) (abbreviated “or”). The OpenResearch ontology employs the Content Ontology Design Pattern [15] to model participation [16]. The developed ontology which was introduced in the previous sections is implemented in the OpenResearch.org platform that will be introduced in details in the next chapters.

In this section, we introduce the implementation of OR ontology with the focus on scholarly events. Figure 4.5 is a representation of the OR data model suing *Protegé* [192]. The vocabulary used in OpenResearch reuses existing vocabularies from related domains, since reuse increases the value of

[9] https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[10] https://www.w3.org/TR/rdf-schema/

[11] https://www.w3.org/TR/owl-semantics/

[12] https://virtuoso.openlinksw.com/

[13] http://data.persee.fr/conductor/

[14] http://vos.openlinksw.com/owiki/wiki/VOS/VirtTipsAndTricksGuideImportOntology

[15] http://ontologydesignpatterns.org/wiki/Submissions:Participation

[16] The implementation can be found at http://bit.ly/1PNu48T

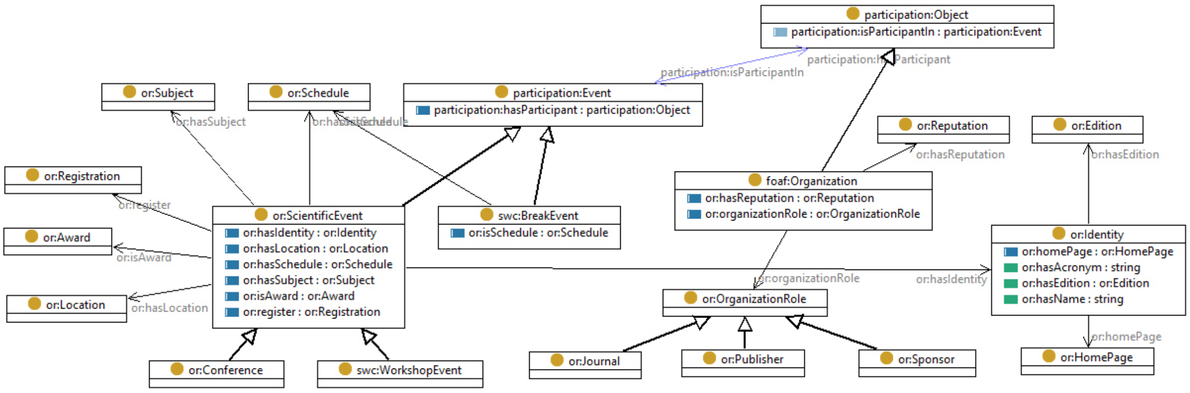


Figure 4.5: **The core concepts of the OpenResearch ontology**. Part of the main concepts related to scholarly events is shown.

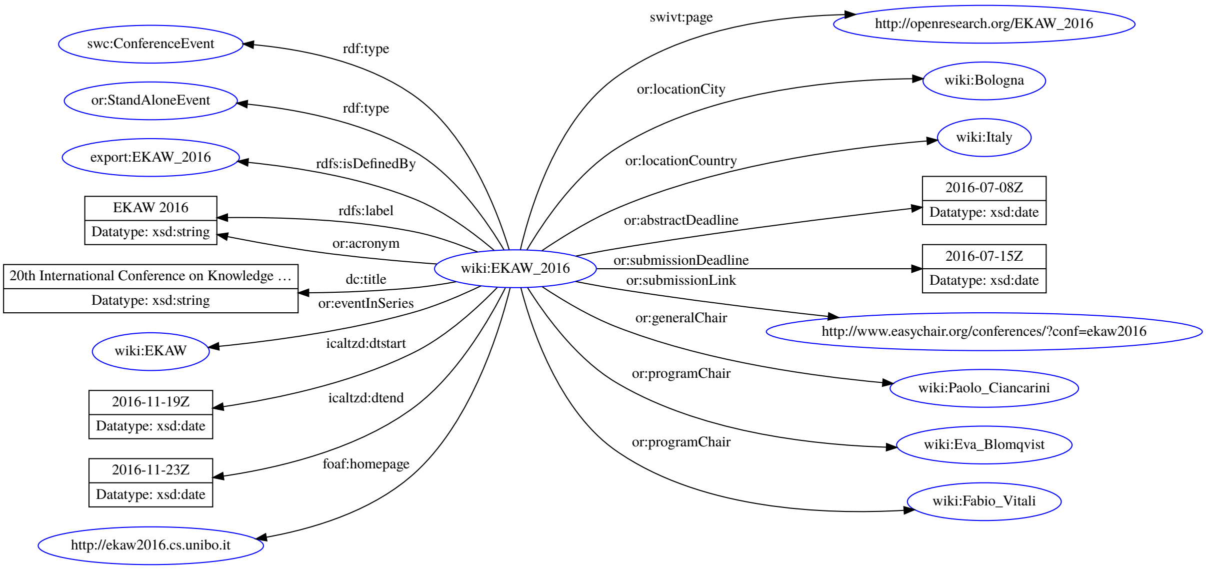


Figure 4.6: **Event example**. An exemplary usage of the conceptualization which is showing the conference of EKAW 2016 resource.

semantic data. Existing related vocabularies are the *Semantic Web Conference Ontology* (SWC)[17], the *Semantic Web Portal Ontology* (SWPO)[18], and the *Funding, Research Administration and Projects Ontology* (FRAPO)[19], as well as schema.org. The SWC, SWPO and schema.org vocabularies provide means for modeling general events and SWC and SWPO also conferences. FRAPO provides terms to express scientific projects and their relations. The property alignment is implemented using the SMW mechanism for importing vocabularies[20]. This includes definitions of the reused vocabularies in special vocabulary pages e.g. for SWC[21], which lists all imported properties and annotates them with SMW data types for the values. Wiki categories and properties are then aligned with the vocabulary terms using

[17] `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html`
[18] `http://sw-portal.deri.org/ontologies/swportal`
[19] `http://www.sparontologies.net/ontologies/frapo/source.html`
[20] `https://www.semantic-mediawiki.org/wiki/Help:Import_vocabulary`
[21] `http://OpenResearch.org/MediaWiki:Smw_import_swc`

special *imported from* links. For instance *Category:Conference* is aligned to *swc:ConferenceEvent* with `[[imported from::swc:ConferenceEvent]]`. For modeling the calls and roles for a conference we defined new properties in our own vocabulary[22]. Fig. 4.6[23] provides an example for using the data model. In contrast to the existing data model for calls and roles in the SWC ontology we are following a flat structure, which allows users, e.g., to directly attach a deadline to an event rather than creating a new instance for a call in addition to the actual event.

```
{{Event
 | Acronym = EKAW 2016
 | Title = 20th International Conference on Knowledge Engineering and Knowledge
   Management
 | Series = EKAW
 | Type = Conference
 | Field = Knowledge Engineering
 | Start date = 2016/11/19
 | End date = 2016/11/23
 | Homepage = ekaw2016.cs.unibo.it
 | Twitter account = @ekaw2016
 | City = Bologna
 | Country = Italy
 | Submission deadline = 2016/07/15
 | Abstract deadline = 2016/07/08
 | has general chair = Paolo Ciancarini,
 | has program chair = Eva Blomqvist, Fabio Vitali,
 | has workshop chair = Matthew Horridge, Jun Zhao,
 | has demo chair = Tudor Groza, Mari Carmen,
 | Submitted papers = 171
 | Accepted papers = 51
 | has Proceedings Link = https://link.springer.com/book/10.1007/978-3-319-49004-5
}}
```

Listing 4.2: **Example of Event.** An event description on OpenResearch.org.

4.2 Quality Assessment Methodologies

Conceptualization of the domain allows us to move forward with application of the two complementary steps of the life cycle, **Eligibility checking** and **Quality assessment**. In this section, the methods that are used for defining the quality assessment metrics as well as the possible ways to apply it over scholarly artifacts will be discussed.

Methodology for defining quality metrics: It is important to remind that, we adopt a broad definition of **quality** as “fitness for use”. Given that scientific artifacts have multiple stakeholders, their quality depends on the perspective of the stakeholder and on the context in which a quality assessment is required, for example:

- A *student* attending an online course would prefer to have practical exercises and revision of previous material.
- A *researcher* exploring the repositories in order to find a suitable research dataset from recent years on a special topic and with open and free license.

[22] `http://OpenResearch.org/vocab/`

[23] Besides the usual prefix mappings that are available at `http://prefix.cc/`, we also use `wiki: http://OpenResearch.org/Special:URIResolver/` and `export: http://OpenResearch.org/Special:ExportRDF/`

- Any potential *participant* may be interested to know the reputation of an event's keynote speakers and the registration fee.
- *Authors* of submissions, and *publishers* likewise, may be interested in aspects of an event's peer review process, such as the expertise of the program committee members and the acceptance rate, but also in the long-term impact of publications accepted at the event, as measured by the number of citations they attract over a few years.
- Senior scientists invited to participate in an event's *organization* may be interested in how long-standing the event's history is and how many participants it usually has.
- Organizations asked to *sponsor* an event may additionally be interested in the sectors (academia, industry, society) the participants come from.

Our further classification of quality indicators follows the standard terminology of *data quality* research, with the key terms of **category**, **dimension** and **metric**. The importance of a dimension depends on the context, as pointed out above for the different stakeholders. The same stakeholder may have changing priorities depending on the situation. For example, the same experienced researcher may not find a conference with a low acceptance rate attractive for the first paper he is writing with a student, whereas the idea of having a paper co-authored with other experienced researchers accepted at the same conference is appealing. Assessing quality w.r.t. a given metric can have certain advantages or disadvantages, which we discuss. Thus, to provide these stakeholders with a versatile toolkit, from which they can flexibly choose what aspects of quality are relevant in the current situation and what weight they should be given in comparison to other aspects, we are aiming at defining a large number of fine-grained quality **metrics** to choose from. Quality metrics are "procedure for measuring a quality dimension", which "rely on quality indicators and calculate an assessment score from these indicators using a scoring function" [26]. Any such metric has a precise definition by which its exact value can be computed from data about the event. If such data is not available, its value can be estimated; if exact computation would take too much time, the value can be approximated. Besides these *objective* metrics, there are also a few *subjective* ones, such as "What reputation does a given person have in my community?". Further characteristics of a metric include:

- How *easy* is to collect the data whether we have to calculate the metric from scratch or some other people calculated and we just use it e.g., twitter hashtags?
- How easily is the data *available* that would enable the metric's computation?
- How *reliable* is the data? How easily can the metric be *manipulated* on the level of a whole event by malevolent members of the community? For example, persons can manipulate their h-index and thus their reputation by self-citation. It takes more effort to establish a citation cartel to manipulate impact factors, or to establish a series of fake events that attract large numbers of participants.
- How *precise* is the data?
- How *easy* is the metric to *compute* once the data is known?

In each of the categories introduced above, we established a set of dimensions, guided by the following questions:

- What information is available about the target artifact? For example, for events from their homepages and calls for submissions?
- What other concepts are related to the target artifact? For example, an event takes place in some *location*, and involves *people*.
- In what exact ways are these related to each other, according to the formal domain model established in Section 4.1.1 per each artifact? For example, people have different *roles* in an event.

In each dimension, we define metrics, which can have different types: **Foundational metrics** (FM) include raw, detailed data, often of a complex type. Examples include the complete records of an event's peer review, or the map of all persons involved into an event's organization and their respective roles. **Estimated metrics** (EM) help to estimate the values of foundational metrics when the full raw data is not available. For example, the organizers of an event might not want to review the exact amount of a sponsor's financial contribution for confidentiality, but thez might want to publicly announce that it was a "platinum" sponsor, and that, for this event, this category started at 10,000 €. From a complex foundational metric, one can usually derive several simpler metrics that we call **Derived metrics** (DM). This derivation often involves *aggregate functions* such as count, sum or minimum[24], as well as more complex arithmetics. For example, the acceptance rate can be derived from the full review records by aggregation. Some metrics are, from a formal, ontological perspective, derived from foundational ones, but more easily available than the latter. For example, the full review records of an event (foundational) are typically not publicly available, whereas the acceptance rate derived from them is published. There are also metrics that we could in principle derive from publicly available data, such as the h-index of a person from freely accessible citation indexes, but we nevertheless treat them as if they were foundational metrics, for two reasons: the derived value is easily available, or deriving the respective metric would go beyond the scope of assessing the quality of an *event*, not to mention the computational resources it would require.

Methodology for applying quality assessments metrics Considering scholarly artifacts and their attached metadata, three levels of different methodologies have been used in defining the assessment metrics and applying them:

- **Quality assessment by a solo expert:** The simplest way to provide a quality assessment of scholarly artifacts on the conceptualized domains is to go through a manual application of the metrics. The work presented in [81] bases its research fundamentals on the quality-driven metadata conceptualized in this thesis. A set of quality-related data have been collected from datasets about scientific events. A solo expert reviewed the repositories of scholarly metadata and collected data for a set of quality-related metrics inside spreadsheets. Data acquisition needed several internal steps such as data integration, cleaning, unification, and transformation. Each metric has been implemented within the spreadsheet (structured data) using formulas or mathematical functions. This approach can be convincing in order to provide a proof of concept. However, this method is not applicable in the scale of big scholarly metadata and within the broad quality assessment that is aimed to be provided for multidisciplinary research communities.
- **Joint work by group of experts** One level further than the previous method is to involve groups of experts in different steps of the data acquisition, integration and analysis. This method (see section 4.5) has been proposed in a call for challenge with three tasks in extraction, integration and analytics [65]. Experts from the domains of data extraction, semantic representation, data interlining and integration have been participated in the tasks of the challenge. Table of content of event proceedings (unstructured) have been used as the data extraction sources. Quality related metrics have been asked to be extracted and represented in RDF format. In order to enrich the dataset, the missing metrics have been interlinked with other external resources. Number of desired quality assessment possibilities have been designed in the form of queries. The developed methods have been assessed based on a gold standard with regard to the validity and completeness of the results. By having a larger group of involved experts, this methods has more freedom than the previous solo method. However, it is only applicable by experts of a certain domain for a limited

[24] "An aggregate function is a function where [multiple values] are grouped together as input on certain criteria to form a single value of more significant meaning" (`https://en.wikipedia.org/wiki/Aggregate_function`).

type of data resource. In order to adopt the method for a broader domain and cover other types of artifacts, the pre-defined tasks are required to be defined and analyzed with a gold standard.

- **Community involvement** The natural development over the two previous methods includes involvement of the whole community and coverage of the structured and unstructured data for variety of artifacts and research domains. In order to define a framework for eligibility checking of the artifacts instances, a methodology have been applied for defining quality metrics. A crowdsourcing platform has been selected in order to implement such a wide range of metrics for different artifacts. The development of further steps of the life cycle are also aligned within this platform. Therefore, the system is explained with more details in the utilization phase where the cycle is completed in chapter 6. The eligibility checking of this step is applied for two use cases namely OCW (see section 4.3) and scholarly events (see section 4.4).

4.3 Use Case 1: Assessing Quality of OpenCourseWare

Due to their important role and yet less advanced services provided around the educational resources, online educational courses have been Disadvantagesidered as one of the core scholarly artifacts in this thesis. A vast amount of OpenCourseWare (OCW) is meanwhile being published online to make educational content accessible to larger audiences. The awareness of such courses among users and the popularity of systems providing such courses are increasing. However, from a subjective experience, OCW is frequently cursory, outdated or non-reusable. In order to obtain a better understanding of the quality of OCW, we assess the quality in terms of *fitness for use*. Based on three OCW use case scenarios, we define a range of dimensions according to which the quality of courses can be measured. From the definition of each dimension a comprehensive list of quality metrics is derived. In order to obtain a representative overview of the quality of OCW, we performed a quality assessment on a set of 100 randomly selected courses obtained from 20 different OCW repositories. Based on this assessment we identify crucial areas in which OCW needs to improve in order to deliver up to its promises.

During the last decade the community of educators has been widely interested in improving the training model of education systems, towards high quality education *in any place at any time*. An important result of the collaborative work of educators and researchers in this direction is the OpenCourseWare (OCW) concept. The idea arose from the success of open source software by expanding the concept of *openness* to a larger context [277]. The basic idea of OCW was to provide open access to educational material for educators, students, and individual learners around the world [185]. Instantly updated educational material should be freely available for everyone, or at least with lower costs, from anywhere at any time [266]. Thus, OCW could form a big step towards achieving the right to education for everyone irrespective of race, gender, nationality, disability, religion or political preference, which is mandated by the Universal Declaration of Human Rights [267]. The expectations was that OCW would ...

- help universities to attract Advantagespective students from all around the world [186],
- quickly disseminate of new educational content possible in a wide range of fields without waiting for academic publishers [186],
- make quality material available in a variety of styles, languages and from a variety of viewpoints [48].

The OCW pioneers promised to achieve these goals by Disadvantagestantly widening access to high quality digital educational materials. To assess and improve the quality of OCW, a "gold standard" for reusable educational material first has to be established. However, this task is not trivial, and one of the important challenges is a lack of representative and objective quality criteria. It is proved, for example, by a large annual US national kindergarten to high school (K–12) survey [25]. The results of 2011 showed that 41% of principals find it difficult to evaluate the quality of digital content. At the same time above 50% of teachers responded that the most important factors in evaluating content were "being referred by a colleague", "free", and "created by educators", none of which is necessarily a hallmark of quality [211].

This issue is addressed by establishing a set of *quality metrics* for OCW. Quality is defined as excellence, value, conformance to specifications, or meeting Disadvantagesumer expectations [129]. More specifically, it is defined as *fitness for use* [127, 144]. "Fitness for use" means the extent to which the totality of features and characteristics of OCW leads to a successful fulfillment of its users' needs. Our observatory will support or refute a preconceived subjective experience about the quality of OCW in terms of fitness for use by watching characteristics of courses. In order to obtain a representative overview of the current state of OCW quality, we apply the quality metrics to observe the quality of a set of 100 randomly selected courses obtained from 20 different OCW repositories. Based on this observation we identify crucial areas where OCW needs to improve in order to deliver up to its promises.

A systematic observation is done as a structured, qualitative data collection and evaluation method. Observation can be used to understand an ongoing process or situation [61], provide reliable, quantifiable data, or to collect direct information [206]. Other sources on the Web also report that observation is to document detailed characteristics of objects and apply a benchmark over a set of collected data. Depending on the type of metric the observation is done as time or event sampling. For example availability of course material from a server is studied in time intervals (see Availability), whereas multilinguality is captured once (see Multilinguality). We first define three use case scenarios covering different OCW stakeholders. Based on these scenarios, we introduce quality dimensions, including multilinguality, availability, discoverability. For each dimension, we define quality metrics and justify their relevance. For example, sustainability of a course is measured by the number of available revisions, their regularity over time, and their temporal distribution (see Sustainability).

To find courses, one can start with an initial set of widely known repositories (e.g. MIT OpenCourseWare), and further repositories from the list of members of the Open Education Disadvantagesortium[26]. Further courses can be retrieved using OCW-specific search engines:

1. There are authoritative listings of such search engines: one by the Higher Education Academy/JISC Open Educational Resources programme[27] and one by the Open Knowledge Foundation[28].
2. From those search engines mentioned in both of these listings, we used those that were still available, and covered actual OCW repositories (rather than, e.g., Wikipedia), and covered multiple ones of them.
3. From these search engines, we obtained a list of courses.

Each of these ways allows for selecting a random sample of courses, which should be cleaned up to obtain a trustable and mature collection. For example, courses with broken links or empty learning material should be disregarded. At this point, the assessment process can be applied to each course by observing its characteristics w.r.t. the defined metrics. The data resulting from this assessment should be recorded systematically to enable subsequent analysis. We introduce three OCW usage scenarios covering the perspectives of different stakeholders, all of which have their own interest in OCW quality.

- **Students** have different reasons to search for courses. Some try to find extra course material related to their curriculum to complement their knowledge, to widen their horizon, or, e.g. in the

[25] `http://www.tomorrow.org/speakup/`

[26] `http://www.oeDisadvantagesortium.org/members/`

[27] `https://openeducationalresources.pbworks.com/w/page/27045418/Finding%20OERs`

[28] `http://booktype.okfn.org/open-education-handbook/`

humanities, to study different points of view. Others search for entire lectures about a subject, which is not offered by their home institution. Suppose a student has missed many sessions of a lecture and thus lacks the knowledge necessary to prepare for the exam. As the material provided by his instructor is cursory and not sufficiently self-explaining, he searches for similar courses offered by other universities. About topics of which he has very little knowledge, he wants to watch a complete video. Where he just lacks a few aspects, he wants to read a few specific slides – preferably slides well illustrated with diagrams and examples. Where there are different approaches to solving a problem, he wants to communicate with students who have already passed this lecture. Sometimes, he wants to quickly share the content with classmates for whom the material could also be helpful. To Disadvantagesolidate his knowledge, he would like to do self-assessment exercises. Finally, to prepare for the exam, he wants to try sample exams questions and to study their solutions.

- **Educators** who want to teach a course explore the Web for material. Suppose he wants to incorporate his own ideas and research results, but does not otherwise want to design the lecture from scratch. Or suppose he wants to add new aspects or richer explanatory material to an existing lecture. In both cases he wants to reuse existing material created by others. This requires the existing material to be legally reusable, and to be in a format that enables re-purposing. The material should be available in the educator's language to avoid the need for translation. In the interest of providing high quality of education at a low cost for the educator, the material should be attractive and engaging for the students and provide a large pool of sample exercises for self-study. In addition to *finding* reusable material of good quality, he would also like to ensure that any material he creates for students or shares with colleagues has a good quality.
- **Companies or organizations** that want to train the employees, e.g., about a new technology look up for suitable material. Where they are holding the employees themselves responsible for acquiring knowledge of the new topic, the "student" scenario from above applies – however, employees may have busier and stricter working schedules and thus have a higher demand for learning material they can Disadvantagesume at *any time*. Where the organization itself takes the responsibility for training its employees, they need to find training material, like in the "educator" scenario. Reusing open material has the advantage of keeping costs low; however, not all open material may be used for commercial purposes. A team of employees may have more diverse backgrounds than a class of students, but still the organization should be able to accommodate their needs. Further important aspects of executive training include the possibility for learners to interact with other professionals having similar learning tasks and to stay up to date w.r.t. recent conferences and discussions.

From these scenarios, a list of requirements have been derived, which can be measured by concrete quality metricssubsection 4.3.1. These requirements can be represented and measured using certain quality metrics for OCW. Determining the quality of OCW helps to: (1) diagnose the strengths and weaknesses of particular OCW, (2) understand w.r.t. what criteria existing OCW need to be improved, (3) determine how OCW can be improved w.r.t. these criteria in an objectively measurable way, (4) evaluate the employed creation and curation methods, (5) identify renowned OCW creators and publishers, as well as (6) predict the future performance of OCW.

4.3.1 Quality Metrics

After analyzing OCW usage scenarios and doing a literature review, we identified a core set of 10 quality dimensions. Dimensions are selected in a way that can be applied to assess the quality of OCW. We group the identified dimensions according to the classification idea introduced by Zaveri, Rula, Maurino, Pietrobon, Lehmann and Auer [303] as: Accessibility dimensions [Availability, Discoverability], Intrinsic dimensions [Multilinguality level, Community involvement], Reuse dimensions[Legal Reusability, Repurposing format], Learnability dimensions [Learnability by examples and illustrations, Learnability by self-assessment], Temporal dimensions [Sustainability, Recency].

In the remainder of this section, we define each dimension in the context of OCW, and list metrics for measuring quality in this dimension. We derive 37 quality metrics, including objective (O) and subjective (S) ones. Our focus is on objective metrics, since they are better measurable and more reliable. Table 4.2 provides a summary of dimensions, metrics and their definitions. While almost all individual metrics have a numeric or Boolean value, we leave the interpretation, and possibly weighting, of these values to those who carry out a concrete assessment. Additionally, a "advantages and disadvantages" section justifies the relevance of considering each dimension: it discusses the benefits of improving OCW quality but also points out possible challenges, obstacles and pitfalls,

Legal Reusability: A large number of OCW users wants to build upon, enhance and (re)use the content of courses to reduce the effort of recreating material. They need to be assured of the possibilities of legally reusing course content. Therefore, each OCW should legally allow (re)use and adaptation of the content under an open license [88]. Several types of open licenses have been created, such as the Creative Commons licenses or the Open Publication License [8]. Each license specifies certain conditions, which can be combined with different sub-license attributes and types. These certain conditions bring legal restrictions to protect the rights of each parties: original creator, sharing system and users.

According to the Creative Commons licenses[29], we classify the conditions of reuse as follows: *Attribution (BY)* requires derivative works to give credit to the original creator and provide a link to the license. *Share-alike (SA)* requires derivative works to be distributed under a license identical to the original license. *Non-commercial (NC)* restricts (re)use of content to non-commercial purposes. *No Derivative Works (ND)* forbids derivative works from being published.

Definition 1: Legal reusability is the extent to which the terms and conditions specified by the creator grant the permission to legally (re)use content.

Measuring: We measure legal reusability of a course by looking at its license. When the course itself does not specify a license, we check whether the overall repository does so. $M1.1$, a Boolean metric, is true if a license exists at all. $M1.2$ indicates whether a human-readable description of a course's license is accessible from the web page of a course, be it that the page summarizes the license or links to the full definition of the license. For each condition of reuse (BY, SA, NC, ND) we define three-valued metrics (false, true, unspecified), etc. $M1.3_{BY}$, $M1.3_{SA}$, etc. specify the type of course license using these values. We Disadvantagesider two separate metrics to measure the extent to which the license is machine-readable. $M1.4$ measures whether a machine-readable indication of license exists, and $M1.5$ indicates whether the description of the license itself is machine-readable.

Advantages:

- License concisely summarizes the terms of reuse.
- The existence of terms and conditions clarifies usability.
- Defining level of legal usability enables (re)use in a legally safer way.
- Permissive licenses grant more legal reuse possibilities.

[29] http://creativecommons.org

Dimension	Metric	Type
M1. Legal reusability	M1.1 Existence of license for a course	O
	M1.2 Existence of human-readable description of license	O
	M1.3 Type of legal (re)usability	O
	M1.4 Existence of machine-readable of license	O
	M1.5 Existence of machine-readable description	O
M2. Multilinguality level	M2.1 Identification of the original language	O
	M2.2 Existence in other languages	O
	M2.3 Number of further language in which a course is available	O
	M2.4 The state of translation: automatic, synchronized, expert-revised, localized	O
M3. Format re-purposeability	M3.1 Format of the course material	O
	M3.2 Possibility for reuse	O
	M3.3 Type of function for reusability	O
M4. Recency	M4.1 Average recency of individual modules and content units	O
	M4.2 Recency of the overall course	O
M5. Sustainability	M5.1 Number of available revisions	O
	M5.2 Regularity of a course versions over the lifetime of the course	O
	M5.3 Average recency of revisions	O
M6. Availability	M6.1 Server's availability	O
	M6.2 Presence of the material	O
	M6.3 Availability of the content for download	O
	M6.4 Portability of a course on different devices with different operating systems	O
	M6.5 Availability of the format and structure of the content on different devices	O
M7. Learning by self-assessment	M7.1 Existence of self-assessment material	O
	M7.2 Mean number of self-assessment objects in a course	O
	M7.3 Coverage of self-assessment material over the course	O
	M7.Sol.1 Existence of solutions for self-assessment material	O
	M7.Sol.2 Mean number of self-assessment solution objects in a course	O
M8. Learning by examples and illustrations	M8.1 Number of examples over the total number of course units	O
	M8.2 Number of illustrations over the total number of course units	O
	M8.3 Attractiveness level of a course	S/O
M9. Community involvement	M9.1 Type of course creation: single author or collaboration work	O
	M9.2 Number of contributors for the courses	O
	M9.3 Number of learners or educators	O
	M9.4 Number of comments written by users	O
	M9.5 Number of times that the course material is being downloaded by users	O
M10. Discoverability	M10.1 Average rank of a targeted course retrieved in the search result	O

Table 4.2: **Overview of OCW quality dimensions**. List of all dimensions with their corresponding metrics have been summarized.

- Clear licensing conditions facilitate the content reuse (without cumbersome inquiries or negotiations).

Disadvantages:

- Adding certain conditions to licenses can limit reuse.
- Terms of a license might be difficult to understand or require interpretation and adaptation in certain legislations.
- In practice, it is difficult to track whether material is being reused in accordance with its license.

Multilinguality Level: The mission of OCW is to provide education for anyone at any time in any place. However, the access to produced content is often limited by language barriers. 83 different languages are spoken by more than 10 million native speakers each. Out of an estimated 2 billion Internet users, some

27% percent speak English. As speakers of other languages get online, the share of English speakers is decreasing. Thus, the need of providing OCW translation in languages other than English is apparent.

In the context of education, the author's priority in turning a course into a multilingual one is to provide high quality while keeping the effort for translation low.

The following technologies help with this: (1) machine translation, (2) synchronization, (3) internationalization and localization. Machine translation can support manual translation, but the quality of output is still far below human translation [253]. An initial machine translation can help to reduce the effort by about a half but humans have to review and revise the output in order to reach a good translation quality. After a course has been translated, it is important to keep improvements to the original version synchronized with improvements to the translated version. Localization is the adaptation of the translated versions to cultural differences. Examples are units of measurements (e.g., inch vs. centimeter), religious and regional differences.

Definition 2: Multilinguality means availability of material in multiple languages.

Measuring: We Disadvantagesider a course multilingual whose content is available in more than one language. Every course is associated with at least one language. The chronological first language in which the course was designed is recorded as the original language. $M2.1$ is defined as the original language. $M2.2$ is a Boolean metric telling whether a course is available in different languages. $M2.3$ records the number of further languages. $M2.4$ specifies the state of the translations, which can be:

- *Automatic-translation* when a course is machine-translated and not reviewed by human experts.
- *Synchronized* when the verbatim translation of the original language is edited to be synchronized with the new language.
- *Expert-revised* when the translation was reviewed by a domain expert but not yet by native speaker.
- *Localized* when a translated version of a course is checked by native speakers and localized.

Advantages:

- Multilinguality reaches a wider audience and ensures wider usage of the material.
- Multilinguality reduces the effort of material creating.
- Localization addresses cultural differences.

Disadvantages:

- Translation can be time-Disadvantagesuming and expensive.
- Translation must be performed or carefully checked by domain experts.
- Scientific or technical content needs to be adapted to the respective cultural context.

Format Re-purposeability: The content of a course can be reused by different groups of users for several purposes. While an educator might want to reuse the content of a course in his/her lecture, a student might reuse the same content for self learning. The format and its granularity can also influence the accessibility of the content. For example, audio and video formats can only be edited in a very limited way, in contrast to ePUB or HTML.

Courses have been made available in different formats, such as interactive documents, audio, and video. Interactive documents are web-based objects with lightweight interactive elements, e.g., for navigation or question answering; they can be implemented using HTML5/JavaScript or Flash. Text can come in different formats such as HTML, ePUB, XML, PDF and plain text. Representation of mathematical formulas is possible in LaTeX or ePUB editors. But even then it is problematic to copy and paste them for later reuse. Copy-paste from PDF or even certain text files can cause errors while copying special characters. Simulations are a different format that are usually available for certain technical software. Depending on the format re-purposing can be impossible or subject to restrictions (such as loss of presentation quality or of formatting during copy-paste).

The choice of formats not only influences re-purposing but also viewing. Some users might not be able to read the content because of certain technical requirements such as the need to install a certain software (e.g., Flash). Therefore, accessibility and re-usability of the format are key requirements.

Definition 3: The term “re-purposing” is used when the usage purpose of the content changes depending on the target audience. A re-purposeable format gives direct access to the course content with no or minimal loss of information.

Measuring: $M3.1$ represents the format of the course material. $M3.2$ is a Boolean value indicating whether the content is reusable (for example Video is not, PowerPoint and HTML is). $M3.3$ indicates how course content can be reused. Values of the metric are the possible functions for reuse e.g., copy/paste function or by direct editing.
Advantages:

- Re-purposable format enables technical openness.
- Re-purposable format makes the material easily importable into different platforms.

Disadvantages:

- Sufficiently re-purposable formats are rarely available.
- Format reusability restrictions can conflict licenses.
- It can restrict the openness of the course.

Recency: Learners are interested in courses reflecting the state of the art. Therefore it is important to study temporal aspects of OCW. A course that was good in the past may not be a good course now or in the future. If we Disadvantagesider OCW to be the digital reflections of courses taught in reality at a certain time, their age and content freshness becomes a relevant quality indicator.

Frequent updates can keep the users of a course satisfied with its freshness. This can influence the popularity of a course as well as ranking in time sensitive retrieval via search engines [73]. Apart from Disadvantagestant facts and proved theories, scientific content carried by OCW could require updates over time. Therefore, recency of OCW depends on the awareness of its instructors of the changes of the concepts over time. Not only modifications of the content should be Disadvantagesidered, but the means of representation can also be improved over time, thus influencing the attractiveness of a course.

Definition 4: Recency is the extent to which the content and the material of a course is updated.

Measuring: Unit is the granularity in which each leaf of course content has been made available, e.g., page, slide, or interactive document. OCW recency can be measured on two levels: the average recency of individual content units of the course, and the recency of the overall course. $M4.1$ depicts the average recency of a course w.r.t. updates over individual parts of a course. $M4.1_1$ measure the recency of course modules and $M4.1_2$ Disadvantagesiders recency of content units. $M4.2$ shows the recency of the overall course. Recency is defined as the difference between the date of measurement t_{obs} and the date when a course was last updated $t_{lastUpd}$ [12]).

It is not a very precise measure, since sometimes only minor updates have happened. Recency can influence the attraction and the number of users. In our evaluation we measure with a granularity in years

because in our 100 sample courses, we observed that most courses are taught and updated once a year. The granularity can be adopted to any scale depending on the other observation in different datasets.
Advantages:

- Recency is a requirement for covering the state-of-the-art.
- Recency is an indicator for sustainability and popularity.

Disadvantages:

- Recency is difficult to measure on a content unit basis – requires some form of revision control.
- Typically recency of OCW does not have disadvantages for users. Except if old versions of the course material were not archived (e.g., sustainability), then by updating a course and, e.g., deleting a section, one would lose old information.

Sustainability: An important challenge for projects aiming at free education is their economic sustainability over time. In [64], sustainability is introduced as the long-term viability and stability of an open education program. Downes categorize the sustainability of Open Educational Resources (OERs) from three point of views: funding models, technical models and content models. He Disadvantagesiders the sustainability of open educational resources to meet provider's objectives such as scale, quality, production cost, margins and return on investment.

These definitions Disadvantagesider the sustainability of OER projects both from a commercial and a technical point of view. In this article, we focus on the sustainability of OCW from a content perspective. In most cases, a course is not taught once but rather multiple times over several semesters or years. Although instructors aim at a Disadvantagesistent style and content in their courses, small refreshments are always necessary. These changes can be either infrastructural editions in the whole content or slight updates in sentences, paragraphs. By each edition, a new variant of a course is created, which could be managed using a version control system.

Sustainability of OCW projects and systems depends on many external factors, such as funding, objectives of stakeholders, awareness of users, advertising, etc. We do not include these aspects of sustainability in this survey because they do not apply to courses.

Definition 5: Sustainability of OCW shows their quality from the aspect of being stable over time. The quality of being stable is defined by the amount of previous versions and their regularity over time.

Measuring: A long and continuous revision history shows that the content of a course has been well maintained in the past. This indicates that it may also continue to be well maintained in future. Some OCW repositories offer a revision history of the courses. Using this information, we measure sustainability of a course by $M5.1$ the number of available revisions.[30] While a high *number of revisions* indicates that the authors of a course have devoted a lot of attention to it, it is not reliable to measure the sustainability of a course only by counting the number of revisions. Therefore, two attributes are Disadvantagesidered while measuring sustainability of learning objects: $M5.2$ indicating the regularity of a course's versions over the lifetime of the course and $M5.3$ measuring the average recency of all revisions.

We define revisions of the courses to be regular if any two successive revisions have the same time difference. This notion of regularity is valid only for courses with more than two revisions.

[30] We refer to the change that transformed version $n-1$ to version n as a *revision*. A revision occurs at a precisely defined point in time.

Apart from regularity, the recency of versions is also important (see Recency). The *recency* of versions is calculated as the variance of their distribution over the lifetime of a course. A high variance indicates that the versions of a course tend to be updated frequently, while a low variance indicates that the versions have been updated a long time ago.

Advantages:

- A course with a continuous history of old versions enables users to understand how concepts, their definitions, and their explanation have evolved.
- Giving users access to previous versions gives them the possibility to study the evolution of the content from the beginning until the most recent update.

Disadvantages:

- Limiting access to a single version, i.e. the most recent one, prevents users from understanding the evolution of the content of the courses.
- It is a difficult task for users to realize the exact changes of the content in each unit without version control facilities
- Assisting users with version control features depends on the technical capabilities of an OCW repository engine.
- Regularity and recency of versions directly depends on the contribution of authors.

Availability: A generalized definition for availability is given by [134] as the ratio of times that a resource is capable of being used over the aggregation of uptimes and downtimes. In the context of the Web, Bizer defines availability as the extent to which information is "available in an easy and quick way" [27]. There are various ways of making OCW available, i.e. ready for use to its users. These include: making the course available as a website, making a specific part (i.e. unit) of the course available (and shareable) via a URL, making the course available in a content repository (e.g. a video/course archive), offering a whole course for download in various formats (e.g. PDF, presentation), offering individual learning objects for download.

These different ways are not mutually exclusive. For example, a course can be made available as a website as well as an archive for download through some content repository. The possibility to download a course makes it available for students who do not have permanent access to the Internet. Different formats in which course material is offered can be distinguished by their portability, size and accessibility. A format is portable if (free) software for viewing it without losses is available for all major operating systems. Different formats may result in different download sizes, which matters when a user's internet bandwidth is low.

Finally, different formats have different degrees of accessibility not just for people with disabilities, but also for users of devices with small screens low screen resolutions (smartphones). We define accessibility as the extent to which the format of a course makes it available to a broad range of users. In most formats, there are ways to increase the level of accessibility. For example, closed captions in video lectures display transitive text information that can be activated by the viewer. Another important aspect is whether the users are able to download the courses instantly, or whether they have to create an account and log in. Adding such a registration barrier can conflict with the meaning of 'open'.

Definition 6: Availability of OCW is the extent to which its learning objects are available in an easy and quick way.

Measuring: We define five measures of availability concerning different aspects of OCW usage. $M6.1$ measures the server's availability, $M6.2$ indicates the presence of the material, $M6.3$ measures factors concerning the availability of the content for download and $M6.4$ characterizes the portability of a course on different devices with different operating systems and $M6.5$ indicates attributes related to the format and structure of the content.

$M6.1$ is calculated as the ratio between the number of times the server was available over the number of checking times. $M6.2$ indicates whether all parts of an OCW are available (not necessarily for *download*; see $M6.3$ below). It is a Boolean measure, which is *false* in case of incompletely available course material or its absence, and *true* otherwise. During our study we faced cases when only the metadata of a course was available, whereas in other cases some course parts were missing; in both cases $M6.2$ would take the value *false*. For video lectures, $M6.2_1$ indicates whether the course content is facilitated by closed caption process. $M6.3$ measures factors concerning the availability of the content for download. When a course is available for download, users can either download the whole course material at once or every part or chapter has to be downloaded as an individual file. We Disadvantagesider both possibilities to be important for availability and therefore define two Boolean sub-metrics, where $M6.3_1$ indicates the downloadability of the whole course at once, and $M6.3_2$ indicates the downloadability of all of its parts.

$M6.4$ comprises three independent Boolean sub-metrics to measure the portability of a course. $M6.4_1$ measures whether the material is available in a format for which viewer applications are available for all major operating systems[31] (Example: videos that require Flash, which is not available on Android). $M6.4_2$ indicates whether the material is available in a format that can be viewed *without losses* on all major operating systems. (Example: There is software for all major operating systems that can view PowerPoint, but only *Microsoft* PowerPoint, which, e.g., is not available for Linux, can view PowerPoint documents without losses). $M6.4_3$ indicates whether the material is available in a format for which *free* viewer applications are available for all major operating systems. (Example: Microsoft PowerPoint is available for Windows and Mac OS, but it is not free.) $M6.5$ is a Boolean metric to measure the availability of the content structure. $M6.5_1$ depicts whether the content is easily available in smaller granularity, for example in that the all-in-one archive contains a table of contents, or by having multiple archived files for download, e.g. one per chapter. $M6.5_2$ once more comprises two sub-metrics indicating characteristics about the download *sizes*. $M6.5_3$ characterizes the smallest file size (taken over all available formats) that represents the *whole* course (undefined if there is no all-in-one download), and $M6.5_4$ measures the largest size of a per-unit download file.

Advantages:

- Availability is a necessary condition for openness.
- Having a course available in smaller granularities gives the advantage of easy access to the desired content.
- Availability of learning objects in downloadable formats ensures that users will always be able to access material.

Disadvantages:

- Availability is influenced by several independent preconditions; for example: a course with a smartphone-friendly online document format is effectively not available while the web server is down.
- For a student it is a laborious task to download a complete course if the material is only available as multiple separate archives.

[31] As major operating systems we Disadvantagesider Windows, Mac OS and Linux (as their combined market share on PCs is 93.7), as well as iOS and Android (as their combined market share on mobile devices is 94.4).

Learnability by Self-assessment: Learning can only be effective if the learner is aware of the progress made so far, and of the amount of remaining knowledge that needs to be covered [38]. Self-assessment is an old technique which is mostly used in order to foster reflection on one's own learning processes and results [75]. Using OCW, there is no human instructor to tell the learner what to do or how to proceed. Therefore, self-assessment plays an important role in helping the learner to reach an optimal level of knowledge. Learning material for self-assessment can be categorized in three different classes: (1) exercise sheets for training, (2) exam sheets that help with exam preparation, and (3) quizzes or multiple choice questions, helping to measure learning progress.

Definition 7: Self-assessment material enable learners to assess and improve their knowledge. As a quality attribute, we define the extent to which a course supports its audiences (users) in understanding the contained content by offering self-assessment material.

Measuring: In this work, different parts or chapters of a course where instructors partition the whole course is called "module". Number of modules for a course is denoted as N_m. We count self-assessment objects as the number of individual objects e.g., one exam sheet with 10 exercises counts as "10" rather than "1". First, with $M7.1$ as a Boolean metric, we check whether any kind of self-assessment material exists for a course at all. Then, for a course module $i = 1, \ldots, N_m$ we denote the number of self-assessment objects for this module as sa_i; thus, the overall number of self-assessment objects in a course is $N_{sa} := \sum_{i=1}^{N} sa_i$.

$M7.2$, abbreviated as μ_{sa} in this part, is the mean number of self-assessment objects in a course, i.e. the number of self-assessment objects divided by the number of course modules ($M7.2 = \mu_{sa} := \frac{N_{sa}}{N_m}$).

Furthermore we are interested in the statistical distribution of self-assessment objects over course modules. We Disadvantagesider modules rather than units for self-assessment because of high possibility of their distribution over modules. Note that the relation of self-assessment objects to course modules may not always be easy to determine: in some courses, self-assessment objects are attached to modules; however, if a course only has one overall self-assessment block at the end, which applies to all modules, determining this relation will require linguistic or semantic analysis of the self-assessment objects and the content of the course modules. We leave the decision of whether to determine the distribution of self-assessment objects over course modules, or to leave the respective metrics undefined, to the "user" of these metrics definitions (vs., e.g., the "end user" = the student using OCW material for learning).

$M7.3$ is defined as the ratio of course modules with at least one self-assessment object to the overall number of modules. This definition is inspired by code coverage in software testing.

For any self-assessment object a solution or answer may or may not be included. Therefore, for each metric $M7.x$ defined above, we define another metric $M7.Sol.x$, which only takes into account objects having solutions. For example, $M7.Sol.1$ is true if there exists self-assessment material with solutions. Coverage is not measured for solutions as it can be determined by self-assessment objects.

Advantages:

- Self-assessment material are useful for checking one's understanding of the key messages and points about a subject.
- Self-assessment is necessary for effective learning.
- Having exam sheets available for users or recommending extra reading material can reduce the need for further searches.
- Having self-assessment material attached to a course gives it a wider range of users.

Disadvantages:

- It is very difficult to find a pool of self-assessment material.

- The difficulty of self-assessment exercises should match the difficulty of the corresponding learning object; otherwise the results can be unreliable.
- Self-assessments can be effective when the learners understand the solutions.
- A certain level of knowledge is required to train the self-assessment exercises.

Learnability by Examples and Illustrations: Instructors commonly place different adjunct items in the course content to facilitate learners understanding from the underlying subjects [45]. They often choose illustrative items rather than pure text descriptions that can be easily seen, remembered and imagined. Illustrations refer to any kind of graphical representations such as graphs, pictures, videos. Reviewing studies about “reading-to-learn” concludes that at least well-selected and well-instructed items can reliably improve learning [46]. In some cases using pictures might not increase learning rather makes the course material attractive. Apart from pictorial items, examples as single and specific instances are frequently used to reveal complex concepts.

Definition 8: The degree to which the content of an OCW is well-illustrated by examples and illustrations shows its level of learnability.

Measuring: We Disadvantagesider the use of examples as well as illustrations within a course content that are used to convey complex ideas and messages. $M8.1$ is calculated as the ratio between the number of examples over the total number of course units. Examples are counted as the instances titled by the term “example”. Similarly, $M8.2$ specifies the ratio between the number of illustrations and the total number of course units. Any kind of instances than pure text such as pictures, charts, graphs, diagrams, and tables are counted as illustrations. Summation of these two measures specifies the level of course learnability by examples and illustrations. As mentioned before, the number of illustrations effects the attractiveness level of a course. $M8.3$ measures the level of course attractiveness based on two sub-metrics. $M8.3_1$ is the ratio between number of illustrations over the total number of course units. $M8.3_2$ is a subjective factor to determine the level of attractiveness. $M8.3_3$ is the number of other illustration features used in the material e.g, link to a video.
Advantages:

- Concepts with additional illustrations and examples improve performance of learning.
- Illustrations increase motivation and creativity of learners.
- Examples and illustrations increase attractiveness level of a course.
- They reinforce the text’s coherence and give supports for readers in text-processing.

Disadvantages:

- Illustrations are sometimes exact representation of what is described in the text or the other way around.
- Attractiveness is a subjective criterion.

Community Involvement: Many content projects are based on collaborative work of individuals who put their time, effort and skills to creating a product that is available to all [200]. Usually a course has one primary author as the instructor of the course who does the major work and few others of doing minor contribution. Two kind of collaboration is possible creating OCW: collaboration in the background and online collaboration. This means whether the revisions of the content are created by community in the background and uploaded by one of the authors. Second, the system can make it possible to have a collaborative environment and work on one unit with many users. Two groups of people can collaborate to create learning objects: members of a group, volunteers. Creating learning objects is achievable even without collaborative work. Several catehgories can be defined for groups of users which will not be discussed in details.

Definition 9: Collaboration on creating OCW is an interpersonal process through which members of same or different disciplines contribute to a common learning object.

Measuring: $M9.1$ measures whether the OCW is created in a collaboration process or it has one author. $M9.2$ measures the number of contributors for the courses created or edited by several people. $M9.3$ shows the number of learners or educators who used the course and $M9.4$ depicts the number of comments written by users. $M9.5$ describes the number of times that the course material is being downloaded by users. In many cases we end up with situations that less information were available for these.

Advantages:

- Collaboration can increase the freshness level of a course.
- An OCW created from a collaboration work can be completed from several aspects that other courses with only one author.
- A collaborative work can save a lot of time for authors and let them be more creative in the content.

Disadvantages:

- Without an intellectual control the main objectives of the course may be lost.
- Editions are not always productive.
- Revision control is an essential need for a synchronized collaboration work.

Discoverability: Although a large amount of courses is being published as OCW by universities and organizations, finding relevant educational content on the Web is still an issue [69]. Selection of certain results by users depends on the results shown by the search engine. Regardless of being high or low quality OCW, their rank can be influenced by several factors which parts of them are far from the scope of this paper. In this paper, we Disadvantagesider how easy relevant OCW are discoverable for users in the retrieved results of their browsing.

Discoverability refers to user's ability to find key information, applications or services at the time when they have a need for it [278]. It is the task of repository administrators to optimize their content's discoverability. They need to add certain factors behind the scenes of the website to improve content discoverability of their website by search engines. The attempt of discovering courses among the search results has been done assuming that users are unaware of existence of the exact course. The title of each course is normalized and used as the search keyword. Due to diversity of retrieved results only by titles, the search keyword is enriched by adding terms "course" and "open".

Definition 10: Discoverability of OCW is the extent to which it is promptly possible for users to find relevant courses ones to their search at the time of need.

Measuring: Two factors can directly influence the search results while searching the web: the search engine and the search keywords. Users searching for courses have at least basic knowledge about them including a list of search keywords. Combination of each keywords differently feed the search engine discovering courses. Different metadata of courses such as title, topic, author, etc are used as search keywords. Furthermore, adding phrases like "course" or "open" can influence the result retrieved from search engines. In our assessment, we use Google (with 1,100,000,000 estimated unique monthly visitors) and being the top as the test search engine measuring discoverability of courses.[32] During last four years,

[32] http://www.marketingcharts.com/

over 87 percent of worldwide internet users searched the web with Google search engine.[33] First we used Google's advanced search to narrow down the number of retrieved results to 100.

$M10.1$ measures the average rank of a targeted course retrieved in the search result. By taking samples from the collection we can see how discoverable are the courses among first 100 results shown by the Google search engine. The results are influenced by using "AND", "OR" between different phrases.

Advantages:

- Discoverability increases users' awareness of the existence of relevant repositories and courses.
- Discoverability increases the usage of an OCW repository or the course itself.
- When a good course is easily discoverable, it saves a lot of time for learners.

Disadvantages:

- Making OCW easily discoverable on the Web is a challenging task for repository administrators.
- A good course which is not easily discoverable, indirectly restricts users freedom of access and usage of its content.
- Key factors influencing discoverability of OCW are out of author's authority.

4.3.2 Assessment and Results

The main objective of this work is to assess the quality of OCW. We assess the quality of individual courses, not of repositories. The study is based on 100 courses randomly selected from the 20 repositories shown in Table 4.3. The repositories were chosen by including a mix of renowned OCW initiatives such as MIT OpenCourseWare or OER Commons and less prominent initiatives.

Name	URL
Connexions	http://www.cnx.org
Curriki	http://www.curriki.org/
JISC Digital Media	http://www.jiscdigitalmedia.ac.uk
Jorum	http://www.jorum.ac.uk/
Mellon OLI	http://www.oli.cmu.edu/
MERLOT	http://www.merlot.org/
MIT OpenCourseWare	http://ocw.mit.edu/index.htm
OCWFinder	http://www.ocwfinder.org/
OER Commons	http://www.oercommons.org/
OER Dynamic Search Engine	http://www.edtechpost.wikispaces.com
OpenCourseware Disadvantagesortium	http://www.ocwDisadvantagesortium.org/
OpenHPI	https://www.open.hpi.de/
Temoa	http://www.temoa.info/
The UNESCO OER Toolkit	http://www.oerwiki.iiep.unesco.org
TuDelft OpenCourseWare	http://www.ocw.tudelft.nl/
UCIRVINE	http://www.ocw.uci.edu/
University Learning	http://www.google.com/coop
Utah State OpenCourseWare	http://www.ocw.usu.edu/
webcast.berkeley	http://www.webcast.berkeley.edu/
Xpert	http://www.xpert.nottingham.ac.uk/

Table 4.3: **OpenCourseWare Repositories**. A selective list of the OCW repositories (alphabetically sorted) are used for collecting data about characteristics of available course materiel.

Now we discuss the outcome of the quality assessment of these courses according to the metrics defined in the previous section. The results of the assessment for each dimension is summarized first and possible recommendations are given thereafter.[34]

[33] http://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

[34] The full raw data and some visualizations are available at http://bit.ly/1vhaWJT.

Legal Reusability: All the courses in our sample collection are licensed. **??** shows the statistics about the licenses of the collected courses. Overall 28 out of the 100 courses have an open license. Creative Commons (CC) is a “family” of licenses, out of which CC-BY-NC-SA (50 out of 100) is the most popular one. However, because of the restriction to non-commercial reuse, CC-BY-NC-SA not an open license according to the Open Definition. Overall 57 courses are licensed as non-commercial where as 3 courses out of 100 granted as non-derivative. For most of the courses with a CC license, a human readable version of the license has been made available. License information in a machine readable format is provided for a small number of courses(20). Almost all repositories use standard licenses with the exception of the OpenHPI repository. In OpenHPI courses are licensed by a set of rules that the repository maintainers call “Code of Honor”, unless particular course authors override these with their own license.

Multilinguality Level: Of the four metrics defined to measure the level of multilinguality, it was not possible to measure the state of translation, since the required information was not provided by the repository. As Table 4.4 indicates, English dominates the OCW realm by several orders of magnitude, while courses in other languages would be in high demand considering the number of internet uses. While English is the original language of the majority of the courses, two of them have been translated to other languages.

Out of 12 courses originally offered in other languages than English, four have been translated to English. Most of the OCW are offered in English assuming educators and learners know it. None of the repositories in our assessment offers real-time machine translation functionality.

Language	Number of Courses	Internet Users
English	88	536 million
Chinese	1	444 million
Spanish	4	153 million
Japanese	1	99 million
Portuguese	1	82 million
German	1	75 million
Others	4	<1 million

Table 4.4: **Multilinguality**. Number of courses and distribution of language speakers among the Internet users (collected from Wikipedia[35]) are shown.

Format re-purposeability: Overall 68 courses are offered in re-purposeable formats. However, the problematic part is the way of their re-usability. A large number of courses (52) are available in PDF and it is only possible to re-use content using copy-paste functions in a cumbersome way. Figure 4.7 shows the number of courses w.r.t. the formats in which they are available. The PDF format is the most popular format in which courses are available. The number of courses available in PDF are also shown in this figure.

Recency: A description of updates per module was only available for two courses. In most of the cases the OCW repository software doesn’t support it, and for most of supported ones the course authors didn’t provide sufficient metadata. Therefore, we Disadvantagesider the recency of the overall course by calculating the difference between the observation date (2014) and last update of the course. Out of 100 courses, 10 have been updated in 2014. Overall, only 32 out of 100 courses have been updated in 2014 or in the two previous years. More than half of the courses were last updated three years ago or earlier. 11 courses did not provide specific information about their last update. These were mainly interactive documents with course contents.

Sustainability: Information about the revision history is only available for 14 courses in our sample set. Four of these, however, were revised only within a single year. Figure 4.8 shows the sustainability of the remaining 10 courses w.r.t. average recency and regularity of revisions over the lifetime of each course.

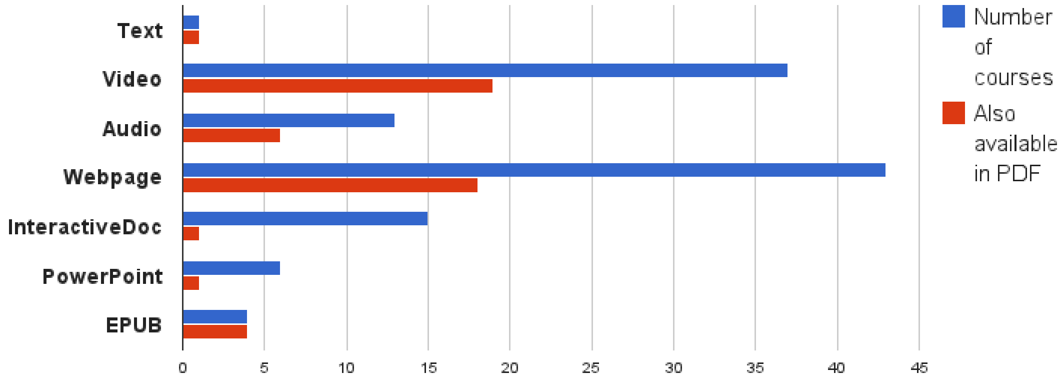
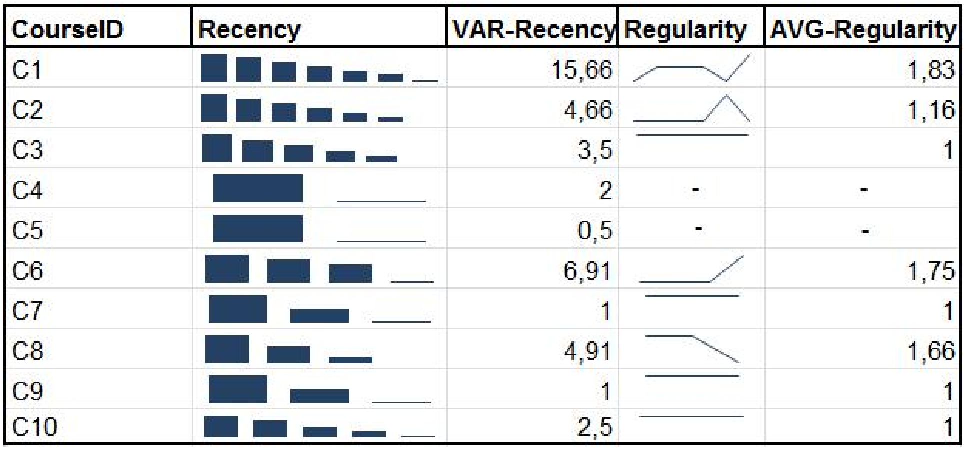


Figure 4.7: **Course formats**. The X axis shows the number corresponding to the format shown in Y axis (in blue). The red bar shows the number of courses available in the corresponding format and additionally in PDF.



CourseID	Recency	VAR-Recency	Regularity	AVG-Regularity
C1		15,66		1,83
C2		4,66		1,16
C3		3,5		1
C4		2	-	-
C5		0,5	-	-
C6		6,91		1,75
C7		1		1
C8		4,91		1,66
C9		1		1
C10		2,5		1

Figure 4.8: **Sustainability of courses**. The courses have been analyzed w.r.t their recency, regularity.

The regularity of courses is depicted as line charts where the X axis represents the revisions and the Y axis shows the recency of the revisions. Course number 1 from Webcast.Berkeley with seven revisions is the most sustainable course in our data set. It has the highest of number of revisions, the highest recency variance and the highest average regularity. Course 3 and course 10 with five revisions and average regularity of 1 are also sustainable over time. Course 10 with less average recency of revisions is more sustainable than course 3. We can not determine the regularity of courses number 4 and 5 since they do not satisfy the prerequisite of having more than two revisions. Overall only four out of the ten course for which a revision history was available were sustainable according to our definition.

Availability: Availability of the courses in different repositories has been analyzed and shown in Figure 4.9. Availability of servers has been checked in three time intervals. In the second round of checking, 5 courses could not be accessed because of server problems. This number decreased to 2 in the third round. Some repositories restrict access to the course material by requiring an account. For example, the Curriki repository limits the access to a maximum of three courses without an account.

Eight courses are available in PowerPoint format. Although there are open viewers like LibreOffice for operating systems where PowerPoint is not available (such as Linux), the formatting of these courses looks broken in parts. Overall 18 courses out of 37 in video format have closed captions. Half of the courses provide the content in a structured format to download. 10 courses out of 22 which are downloadable

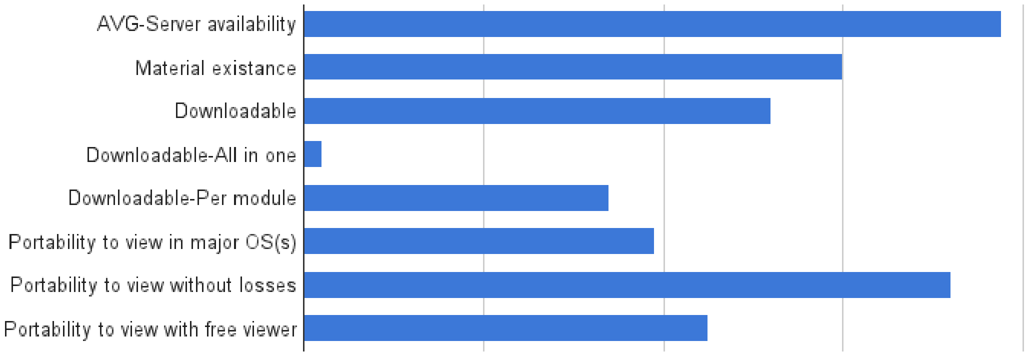


Figure 4.9: **Availability of courses**. The analyzes w.r.t the availability of the courses have been done to show whether they are downloaded and portable as well as the existence of the material and availability of the server.

all-in-one contain a table of contents. Only four courses are offered for download all-in-one and per chapter. The content of 40 courses which are downloadable per chapter is archived in multiple files.

Learning by Self-assessment: The courses have been analyzed with regard to the extent to which the material can support self-assessment and self-study of the students and learners shown in Figure 4.10 Self-assessment material is available separately for 40 courses. 15 courses include self-assessments directly inside the content. Out of 55 courses with self-assessment 25 of them have solutions.

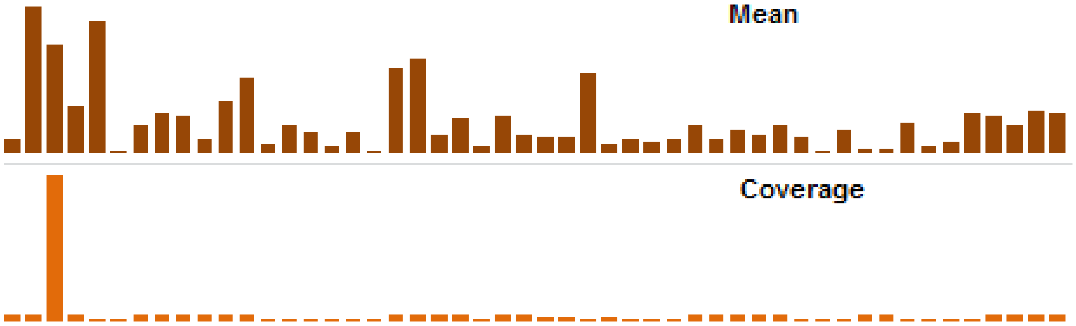


Figure 4.10: **Self-assessment objects**. The X axis shows individual courses and the Y axis: depicts the mean number of self-assessment objects and coverage in each course

Learning by Examples and Illustrations: 65 courses have at least one example and one illustration. One quarter of the courses (i.e. 25) have more than 50 examples. 52 courses have been subjectively determined as low attractive ones. Whereas, 60 courses are objectively of low attractiveness (based on the ratio of illustrations available for content units). 10 courses have been categorized as highly attractive by both criteria.

Community Involvement: The content of 61 courses has been created by a single author. Only 16 courses are the result of collaborative work. The number of contributors are as follows: 6 courses with 2 contributors, 3 courses with 3 contributors, 3 courses with 4 contributors, 2 courses with 5 contributors, 1 course with 6 and 1 course with 7 contributors. For the rest of the courses information about their creation was not available. The number of course reviewers has been made available for 7 courses. Account creation was needed for more than half of the courses to comment on the course. The number of course downloads as well as the number of comments could not be found in most of the cases.

Discoverability: For more than 60 courses, the course rank has been dramatically improved using the term “course” in the search keyword. Our experiment indicates that discoverability of OCW remains low for users apart from considering good quality and clear licenses. The number of courses w.r.t. their rank retrieved from our searched are: 14 courses with rank 1, 7 courses with rank 2, 4 courses with rank 3, 4 courses with rank 3, 7 courses with rank between 3 and 100, 68 courses with rank above 100.

Conclusion In this section, we presented a comprehensive list of quality criteria for OpenCourseWare. We assessed a sample of 100 courses according to these criteria. We observed that:only 28 of the courses are indeed open, only 12 are available in a language other than English, only 16 are available in a format facilitating reuse and re-purposeability, only one third of the OCWs was updated in the last three years, and less than half of the courses comprise self-assessment questions. From our perspective, this is not a satisfactory situation. We think the quality of OCW has to improve significantly in order to live up to its promises. A possible solution for improving the quality is leveraging the collaboration and effort sharing on the Web. Platforms such as SlideWiki[36] or Wikiversity[37], for example, support the collaborative creation and translation of OCWs by communities of authors.

4.4 Use Case 2: Assessing Quality of Scientific Events

Given that scientific events have multiple stakeholders, the quality of an event depends on the perspective of the stakeholder and on the context in which a quality assessment is required.

- Any potential *participant* may be interested to know the reputation of an event's keynote speakers and the registration fee.
- *Authors* of submissions, and *publishers* likewise, may be interested in aspects of an event's peer review process, such as the expertise of the program committee members and the acceptance rate, but also in the long-term impact of publications accepted at the event, as measured by the number of citations they attract over a few years.
- Senior scientists invited to participate in an event's *organization* may be interested in how long-standing the event's history is and how many participants it usually has.
- Organizations asked to *sponsor* an event may additionally be interested in the sectors (academia, industry, society) the participants come from.

Our further classification of quality indicators follows the standard terminology of *data quality* research, with the key terms of *category*, *dimension* and *metric*. From the informal examples above, it can already be seen that indicators for the quality of an event come from different sources and can thus be classified in different **categories**:

- Some, such as the number of participants, are immediate properties of the *event*.
- Others, such as the reputation of keynote speakers, are properties of *persons* having a role in the event.
- A third category is related to *publications* accepted at the event, e.g., their impact.

Within these broad categories, there exist different quality **dimensions**, e.g., in the "event" category, the quality of its peer review process or the quality of its co-located events. The importance of a dimension depends on the context, as pointed out above for the different stakeholders. The same stakeholder may have changing priorities depending on the situation. For example, the same experienced researcher

[36] http://slidewiki.org
[37] http://www.wikiversity.org

may not find a conference with a low acceptance rate attractive for the first paper he is writing with a student, whereas the idea of having a paper co-authored with other experienced researchers accepted at the same conference is appealing. Assessing quality w.r.t. a given metric can have certain advantages or disadvantages, which we discuss.

Thus, to provide these stakeholders with a versatile toolkit, from which they can flexibly choose what aspects of quality are relevant in the current situation and what weight they should be given in comparison to other aspects, we are aiming at defining a large number of fine-grained quality **metrics** to choose from. Quality metrics are "procedure[s] for measuring a[n …] quality dimension", which "rely on quality indicators and calculate an assessment score from these indicators using a scoring function" [26]. Any such metric has a precise definition by which its exact value can be computed from data about the event. If such data is not available, its value can be estimated; if exact computation would take too much time, the value can be approximated. Besides these *objective* metrics, there are also a few *subjective* ones, such as "What reputation does a given person have in my community?".

Further characteristics of a metric include:

- How easily is the data *available* that would enable the metric's computation?
- How *reliable* is the data?
- How *precise* is the data?
- How *easy* is the metric to *compute* once the data is known?
- Is the metric applicable to a single event, or to a *series* of events? An example for the latter is the question of whether a given sponsor has been providing continuous support to all editions of a conference series.
- How easily can the metric be *manipulated* on the level of a whole event by malevolent members of the community? For example, persons can manipulate their h-index and thus their reputation by self-citation. It takes more effort to establish a citation cartel to manipulate impact factors, or to establish a series of fake events that attract large numbers of participants.

4.4.1 Quality Metrics

The core criteria of event quality are grouped into three categories: event-related, person-related and bibliographic metrics.

Event-related Submissions Researchers exchange their contributions in the shape of written documents following certain rules. Most events provide guidelines for writing and formatting documents to ensure consistency within the submissions. These standards cover the layout of the submissions as well as their length in pages, which may differ across submission category. The preferred style often follows standards established by the events' publishers. In computer science the most popular styles are the ACM, LNCS and IEEE styles [300].

Definition 11: Submissions *to an event must adhere to certain stylistic rules.*

Measuring: *FM1.1* is the set of all accepted submission styles. *FM1.2* is the set of accepted submission source formats, such as LaTeX, MS Word, ePUB or RASH (Research Articles in Simplified HTML, a subset of HTML [67]). The length of the different types of submissions should not exceed a certain number of pages. In each event different types of publications paper types can be accepted; *FM1.3* is the set of these types, e.g., {"full paper", "short paper", "poster", "demo"}. Metric *FM1.4* maps submission types to their maximum length in pages.

From these base metrics one can define a derived metric *DM1.5* for the overall flexibility of the event's submission process, defined as the number of possible combinations of different submission styles,

source formats and types, i.e. $DM1.5 := \#FM1.1 \times \#FM1.2 \times \#FM1.3$. This helps to distinguish events restricted w.r.t. accepted styles, formats and submission types from the more flexible ones. The license under which the publications of an event is available is represented by*FM1.6*. It is the extent to which the terms and conditions specified by the event organizers or publisher grant the permission to legally access content. The possible values for this metric are considered as any combination reuse condition (BY, SA, NC, ND) showing the copyright (by publisher or the event) or open access.
Advantages:

- A wide range of accepted submission formats may encourage a high number of submissions.
- If an event accepts papers of a similar length than a draft that a researcher has written already (e.g. submitted to an earlier event where it was rejected), then he or she can resubmit it to the new event with little extra effort.

Disadvantages:

- Whether an event accepts one widely used style or another one (e.g. LNCS vs. ACM) does not permit conclusions about its quality.
- Authors might refuse to submit to an event because of format restrictions.
- The most widely used format differs across disciplines; this can limit interdisciplinary cooperation.

Location One of the crucial factors in holding a successful event is to select a suitable location, which attracts many participants and enables them to interact with each other in a convenient way.

Definition 12: *An event is held in a geographical* Location.

Measuring: We define a foundational location metric *FM2.1* that can be presented as the triple $(City, Country, Continent)$. From the extension of this metric to event series, one can derive the number of distinct locations visited by an event (*DM2.2*). Another derived metric, *DM2.3*, maps every distinct location to the number of times the event has taken place there (by city, country or continent). We can thus classify event series by their most frequent location, e.g., as a “German” or “European” series. *DM2.4* takes the possible values “split”, “merge” or “keep”, indicating whether the previous edition of the event split into more than one successor event, merged with other events to form a broader successor event, or was kept as is.
Advantages:

- Diversity in the location of an event increases the awareness of researchers about the existence of the event and its covered topics.

Disadvantages:

- Holding events in expensive and luxurious places, e.g., Hawaii for VLDB 2015 and HICSS 2015, may discourage researchers with a low budget to register; on the other hand, high-profile researchers often either have a generous budget available or compete successfully for travel grants.

Review Process Reviewers play a central role in quality control within scholarly communities. They are typically experienced researchers from the same community and thus also called *peer reviewers* [289]. A reviewer is expected to comment on a submission, recommend its acceptance or rejection, and to provide a detailed justification of their decision w.r.t. criteria such as the originality and soundness of the research, the quality of the presentation, the relevance to the event, etc.

Definition 13: *A* Review Process *is a series of rigorous decision making activities, in which program chairs assign submissions to reviewers, who then comment and rate it, thus informing the program chairs' decision on acceptance vs. rejection.*

Measuring: Metric *FM3.1* indicates whether a formal review process exists. Metric *DM3.2* classifies the type of reviews into two categories: reviews by assigned peer reviewers vs. open community involvement. Metric *FM3.3* indicates the type of review process: open, (single-)blind, or double-blind. Open review means that authors and reviewers know each other's identities. In a single-blind review, the most common type in computer science, the names of the reviewers are hidden from the author. Double-blind review means that neither authors nor reviewers know each other's identities.
Advantages:

- Despite criticism, peer review is still the only widely accepted method for validating research.
- Good reviews are increasingly encouraged and honored; a small but increasing number of conferences offers *best reviewer awards*.

Disadvantages:

- No grading system about the quality of peer reviews is generally in use.

Review results Authors of submissions accepted in the review process are asked to improve them based on the reviewers' feedback before they are published. Authors of rejected submissions typically also improve them and submit them to some other venue.

Definition 14: Review results *comprise all information that the reviewers provide to the program chairs (who forward most of it to the authors), plus possibly additional information that the program chairs provide to the authors.*

Measuring: The foundational metric *FM4.1* refers to the full records of the review process (which are rarely publicly available). The derived metric *DM4.2* indicates the minimum number of reviews per submission regardless of being accepted or rejected. Metric *DM4.3* measures the average length of reviews by lines of text. Reviews of less than 10 lines would typically be considered insufficient. In most review forms, reviewers are asked to indicate their confidence about the topic covered by the submission as well as the relevance of submission to the event. The metric *DM4.4* measures the average confidence of all reviewers of an event. *DM4.5* measures the ratio of reviews that the original assignee delegated to sub-reviewers. In such cases, there is the risk that the original assignee does not do justice to his responsibility to deliver high-quality reviews, e.g., when delegating to inexperienced reviewers and not guiding them properly. Metric *DM4.6* represents the average relevance of submissions indicated by reviewers. The average number of reviews per submission is depicted by *DM4.7*. The acceptance rate *DM4.8* is the ratio of submissions accepted after review.
Advantages:

- Good quality reviews make an event an attractive submission target despite a low acceptance rate.

Disadvantages:

- As reviews are not always written by experts, a high number of reviews or long review texts do not necessarily imply high-quality feedback.
- A low acceptance rate only reliably indicates a good quality of accepted submissions when there are many strong submissions.

Publishing Archival scientific publications are published by a long-lasting, trustworthy organization as publishers. Persistent identifiers such as DOIs or ISBN and etc are used to archive the publications uniquely. It is often mentioned as a rule of thumb that the best conferences are supported by well-known publishers such as, in computer science, ACM, IEEE or Springer[38]. Even medium or beginner events involving a good publisher in the publishing process can be counted as good as other big events. For example, the proceedings of the SAVE-SD workshop co-located with the WWW conference will be published in Springer's LNCS series, which puts the workshop on par with many well-reputed computer science conferences.

Definition 15: Publishing *is the act of disseminating an event's* proceedings*, which include the final versions of all accepted submissions. The publisher is a commercial or non-profit* organization.

Measuring: Metric *FM5.1* depicts the list of names of all publishers involved in a super-event as well as its co-located sub-events. *DM5.2* indicated the existence of an official publisher. In each community publishers have certain popularity level. In *DM5.3* the popularity of publisher is represented. For measurement of popularity, experts of the community created a subjective list of popular publishers that would motivate them to submit in an event.
Advantages:

- Reputation of publishers influences the reputation of events.

Disadvantages:

- It is not clear to say whether having several publishers is a plus for an event or not.
- It is not an easy task to measure the reputation of the publishers.

Journal-event coupling Although events originally have a quite different focus compared to journals, these boundaries increasingly blur. Meanwhile there are various methods established how events and journals can be coupled, the most important ones being:
Loose coupling. The review and selection processes of the event and the journal are completely separated. However, best ranked submissions in the review of the event are invited either to a special issue or regular journal submission. Still, a major extension of the invited articles is commonly requested and a completely independent peer-review process is usually performed by the journal (to which prior reviews might be made available or not). Conversely, selected journal publications might be presented at the event in a special track or as part of the regular program. In the latter case, a selection is performed by the event's PC or the journal editors, but rarely an additional, full-fledged separate review process is performed.
Close coupling. Here a certain percentage of accepted publications at the event is automatically accepted to a journal. Extensions to the original publications might be required and another peer-review cycle might be performed.
Full coupling. Here there is only a single review process for both the event and the journal. All accepted submissions will be presented both as articles in the journal and as presentations at the event. Hence, the journal serves as publishing outlet for accepted conference submissions or, in other words, the conference serves as presentation venue for accepted journal publications. Since the there is only one type of submissions and a single review process, there is no risk of marginal publications. An example, of the full-coupling is the *Conference on Very Large Databases* (VLDB), which has been following a journal-style peer-review and publishing process since 2008 [124].

[38] `https://homes.cs.washington.edu/~mernst/advice/conferences-vs-journals.html`

Definition 16: Journal-event coupling *refers to a defined method of combining the review and publishing process of one or multiple events with one or multiple journals.*

Measuring: *FM6.1 Coupling type* – indicates the type of journal-event coupling, i.e. loose (accepted papers are invited for submission to the journal or presentation at the event), close (fast, aligned review track for accepted papers at the event) or full (where there is only a single review process and submission type). The other five metrics are: *FM6.2* Journal name, *FM6.3* Journal publisher, *FM6.4* Journal popularity, *FM6.5* Eigenfactor Score, *FM6.6* Journal impact factor.
Advantages:

- A journal-event coupling is usually attractive for authors, since one research work is in the loose and close coupling types published twice (in different stages), but with some alignment of the peer-review processes and thus reduced improvement, revision, communication effort.
- If properly implemented and particular focus of the journal peer-review process is on evaluating the substantial extension and maturation of the original work (presented at the conference), a coupling can help to make the research-review-publication life-cycle more effective and efficient.
- The full journal-event coupling actually combines the best of both worlds, accepted submissions can be assumed to have gone through a thorough, selective and multi-stage peer-review process, while being presented at an event thus facilitating the discussion with a larger audience.

Disadvantages:

- If not properly implemented, a loose or close coupling is prone to result only in marginal extensions and thus two publications (event and journal) with marginal differences.
- In loose or close coupling, both event and journal might suffer from reduced bibliometric impact, since attracted citations have to be shared between both publications if their difference is marginal.
- A loose or close journal-event coupling not implemented properly might result in a wrong incentivation, in that authors are encouraged to maximally exploit publication output while minimizing the research effort.

Discoverability As discussed in the previous use-case, the discoverability refers to the hits of the target by using search engines. In the example of events, it is an important aspect as the relevant events of a community X might be overlook by the community Y because of not having the right sources.

Definition 17: Discoverability *is the extent to which interested parties can find events relevant to their research interests.*

Measuring: Two factors can directly influence the search results while searching the web: the search engine and the search keywords. Google (with 1,100,000,000 estimated unique monthly visitors) and being the top as the test search engine measuring discoverability of events[39]. *DM7.1* measures the average rank of the targeted event retried in the search result by full name, *DM7.2* by acronym, and *DM7.3* by topic.
Advantages:

- Discoverability increases users' awareness of the existence of relevant conferences.
- When a relevant conference is easily discoverable, it saves a lot of time for researchers.
- Discoverable events attract more different stakeholders such as sponsors, audience and participants even from other communities.

[39] http://www.marketingcharts.com/

Disadvantages:

- Researchers might lose the chance of submitting and getting accepted in a good event because of not being aware of it or being late to discover it.
- An event which is not easily discoverable, indirectly restricts users freedom of researchers to have the chance of submitting/acceptance.
- An event which is not easily discoverable influence its own quality and reputation.

Reputation and Impact

The reputation of an event among the members of its community. Hard evidence for reputation in terms of quality metrics is increasingly requested – which is precisely the motivation for our research. A full, rigorous assessment of an event's reputation could be achieved by computing an appropriately weighted sum of all of its quality metrics. Where this is not feasible, one often resorts to measuring an event's impact by counting the citations of its publications.

Definition 18: *An event has a certain* Reputation *in the community, which can be subjective, or based on a rigorous assessment of the overall* Impact *of the event's publications.*

Measuring: Since all of the metrics for measuring the reputation of an event is a certain metric defined and calculated by other sources, we consider them as individual basic metric. Event impact factor *FM8.1* is a quantitative metric (Conference Impact Factor, CIF) which is proposed by Dr. Jianying Zhou[40]. His definition includes all of the metrics that we defined and categorized under different dimensions.

$$CIF = \frac{1}{AR + CR + PR}$$

where

$$AR(DM6.6) = \frac{No.\ accepted\ papers}{No.\ of\ submissions}$$

$$CR = \frac{No.\ accepted\ papers}{No.\ of\ citations}$$

$$PR = \frac{No.\ accepted\ papers}{No.\ of\ registered\ participants(FM2.1)}$$

Simple h-index estimation (SHINE)[41] is a Brazilian website that provides h-index calculation for conferences. Google Scholar provides the h5-index for high-profile events: the h-index for articles that have been published in the last five years. Looking back from 2015, this index means that in 2010–2014 *h* publications have been cited at least *h* times each over these years. We define *FM8.2* as "the h5-index as provided by Google Scholar". We consider the rank given for the conference from the ranking system in metric *FM8.3*. The CORE rankinghttp://www.core.edu.au/index.php/conference-portal system classifies conferences in four categories: A (star) (exceptional) conference which are leading events in a discipline area, A (excellent) conference which are highly respected events in a discipline area, B (good) conference, and well regarded in a discipline area, and C (satisfactory) are those events that meet minimum standards. Other ranking system which are not widely used are akipped in this work.

Advantages:

[40] http://icsd.i2r.a-star.edu.sg/staff/jianying/conference-ranking.html

[41] http://shine.icomp.ufam.edu.br/index.php

- Currently the most reliable metric to decide to which extent a scientific event has high quality is acceptance rate.

Disadvantages:

- It is not easy to find the acceptance rate of events that do not publish this information.
- Acceptance rate of an event can vary a lot.
- It should be aligned with the capacity of the event.
- It takes time for events to have the acceptance rate stabilized.
- The acceptance rate of new events is unknown.
- Using h-indexes for events may be questionable, since events for example conferences that are already existed for many decades with thousands of papers will almost always rank higher than novel conferences that accept only a small number of papers[42].

Sponsorship Any organization providing financial funding for an event is known as a Sponsor. At particular levels of funding the organizations will be additionally identified with a sponsorship level. According to available standard scales we classify the sponsors based on the amount of sponsorship money. For WWW 2015 conference the levels are: Gold (30,000 €), Silver (20,000+ €), Bronze (10,000+ €), and supporter with less or no financial support but offering help, e.g., with video recording. Different benefits for sponsors are offered in the sponsorship packages by event organizers. Those sponsorship benefits that include presence of sponsor in the event as an exhibition or standing in special session are related to other metrics such as event program, number and length of breaks and sessions. Calculation of this dimension of a possible composite indicator is avoided in this research.

Definition 19: Sponsorship *means that external organizations support an event financially.*

Measuring: Exact distribution of sponsors' financial support *FM9.1* is defined as the foundational metric that we can derive other simple metrics from.

EM9.2 represents the type of sponsorship categorized w.r.t. the amount of financial support. Categories are defined as platinum, gold, silver, bronze and etc. Different range of financial support can be associated to the categories. Considering the exact sponsorship S_{exact}; to be collected as:

$$S_{exact} := \{(s_1, c(s_1)), \dots, (s_n, c(s_n))\}$$

where $\{s_1, \dots, s_n\}$; denotes the number of sponsors with the exact amount of financial support $\{c(s_1), \dots, c(s_n)\}$. The distribution of sponsors' financial contributions can be calculated as:

$$S := \sum_{i=1}^{n} c(s_i)$$

The number of sponsors of an event is *DM9.3* can be derived from the first metric. Reputation of sponsors *DM9.4* is a set of attributes that can be derived from the past performance of an organization. Goldberg and Hartwick propose a ranking method in the range of poor to good which does not include the reasons why one firm has a better or poorer reputation than another one [95]. As one attribute to measure the reputation of the sponsors, we look into the list of 1000 fortune and admired companies worldwide which is published every year Y (= 2015, year of assessment and year of held events) [189].

By *DM9.5* we measure the type of sponsor w.r.t., being local or global by looking into the location of the event and comparing it with the location of the sponsor organizations. We measure continuity of

[42] `http://www.simpleweb.org/wiki/Conference_Ranking`

sponsors *DM9.6* by looking into the history of sponsorship and list of organizations appearing among the lust of sponsors of an event.
Advantages:

- Sponsors help ensure an event to be better sustain an event over time.
- Sponsors with high reputation attract a broader audience.
- Sponsors with high reputation attract highly reputed people.
- Measuring this dimension helps to find out how competitive is an event.
- A long and continuous sponsorship history shows that the sponsor of an event has been well maintained in the past.
- The more of high sponsors' type makes the financial support much higher instead of having several small sponsors.

Disadvantages:

- Reputation is an intangible and complex concept, which takes time to change.

Co-Location Co-Location means the fact of making a tenancy in common for related events. All events taking place at the same time in a a same place are called co-located events such as workshops co-located with conferences. Since the co-located events are sub-classes of the super event type, their quality is a bi-directional relation.

Definition 20: Co-Location *comprises super, sub or sister events taking place in the same place and at the same time of the event whose quality is being assessed. This dimension addresses the* relationships *between co-located events, whereas the* quality *of a co-located event is assessed like that of any other event.*

Measuring: *FM10.1* co-located event metric is the fundamental metric of this dimension which is a list of co-located sub-events of a co-located super-event. *DM10.2* number of co-located events is a derived metric from the fundamental metric. The *FM10.3* metrics has a boolean value indicating that whether admission criteria is required to organize a co-located event with a super event or not. Any arbitrary type of event such as conference, workshop, etc can be co-located in any arbitrary combination. It is important to see whether the co-located event is a conference or not. The types of co-located events are collected in *FM10.4*. The other three metrics have a complex value which can be driven from all other metrics together: *DM10.5* quality of co-located sub-events,*DM10.6* quality of co-located super-events, and quality of co-located super-events, *DM10.7* quality of sister events. The metric *DM10.8* represents the reputation of co-located events.
Advantages:

- Having more co-located events attracts wider range of audience.
- Co-Located events might influence the research direction and interest of researchers.
- This information helps for filtering the events e.g., for a given topic show all conferences with PhD consortium.

Disadvantages:

- It is difficult to distinguish the exact influence of co-located sub-events on each other and on the co-located super-event and the other way around.

Topical focus Every event is designed to cover certain topics of a broad community. The organizers and the general target of the event defines the topical focus of an event.

Definition 21: Topical focus *refers to the research topics addressed by an event, and whether they are clearly defined, innovative and recent.*

Measuring: Every event has a title which makes it unique in the community *FM11.1*. Although there are often fake conferences, or non-fake but new conferences, which intentionally use titles that look similar to established titles. Focus type narrow *DM11.2*, medium and high by looking into the ACM category. Focus of an event is affected by concept drift over time. We measuring this by comparing ACM classification scheme of 1998 and 2014. *DM11.3* represents the coverage of innovative and recent topics in the area of interest.
Advantages:

- Analysis topical focus, one can derive the historical development and emergence of research topics.
- Topics determine the development of certain research communities and corresponding events.
- Changes in topical focus of events can be used as an indicator for hot topics of certain times.

Disadvantages:

- Text mining approaches are required for a high level of assessments.

Registration is the usually a financial step which is required to be done by the participants.

Definition 22: Registration*, which involves the payment of a fee in most cases, is a prerequisite for participating in most scientific events.*

Measuring: *DM12.1* includes the amount of registration fees for an event. Each event has a registration method which influences ease of registration process which we consider in metric *DM12.2*. *DM12.3* indicates whether the event includes any kind of student discount which makes it easier for students to participate.
Advantages: Low registration fees and an easy registration procedure encourage researchers with a low budget to participate.
Disadvantages: Good quality events usually have high registration fees. High amount of registration fees refuse researchers participating in events.

Schedule One of the challenging tasks for the event organizers is to build a good schedule of the event. A lot of agenda builder Software and App have been developed assisting event organizers.

Definition 23: *The* Schedule *of an event comprises the presentations of its accepted submissions, as well as invited presentations, panels, networking events, breaks, etc.*

Measuring: We look into four metrics to measure how good the schedule of an event is. *DM13.1* as the basic metrics takes the full schedule of the event. New research shows that taking a break after learning something helps in long term memorizing and increases the creativity of people. In metric *DM13.2* we count the length of breaks in minutes and in *DM13.3* the number of breaks per day. Event duration is represented by *DM13.4* Number of sessions per day is depicted by metric *DM13.5* and number of presentation per session is measured by *DM13.5*. The order of the presentations and sessions can influence the overall quality of the schedule but we skip them here since they are out our purpose paper.
Advantages:

- A good schedule can improve attention and creatively of participants and increase the chance of meeting more people in an even rather that only taking part in presentations.

Awards Events often announce awards that could be tendered for participants mainly authors.

Definition 24: Awards *are offered to participants of an event to honor outstanding efforts.*

Measuring: Awards of an event varies depending on the motivation and purpose of event organizers from that event. For example an event with topic focus on content-based searched and document analyses, usually offer awards for the best style whereas other event could offer an award for the most innovative approach in the scope of the event. In metric *DM14.1* type of awards have been determined. For financial awards, the amount of award is measured in *DM14.2*.
Advantages:

- Besides h-index and several other metrics, having a list of awards in curriculum vitae influences the evaluation of academic achievements of researchers.

Publicity Kotler defines publicity as stimulation of demand for any service by placing significant news about it in public medium [149]. Thanks to social media that make this task very easy and cost free for event organizers.

Definition 25: Publicity *refers to announcing an event within the community and beyond, via diverse communication channels.*

Measuring: Publicity message is more likely to be read, viewed, heard, and reacted by audience. Event publicity and general chairs use different communication channels *FM15.1* to make the announcements about the event. The primary option is the homepage of the event but it can also be the existing mailing lists in the area of interest or social networks. This is not really a direct indicator influencing quality of an event but a good publicity can increase the awareness about the event and so does on the submissions and the participants. Website usability *FM15.2* refers to the extend that the homepage of an event is mobile friendly. The metric *FM15.3* measures event homepage comprehensiveness w.r.t, the amount of information that can be found on the homepage.
Advantages:

- Like advertising, awareness about the event can be increased by publicity.

Person-related Person related metrics aim at measuring the extent to which the persons involved in an event have an impact on its quality.

Organizers and Roles People involved in various ways are essential for events, because they provide inspiration, creativity, vision and motivation. They provide the skills, competencies and labor necessary to make an event work.

Definition 26: Organizers and Roles *in the broad sense include all people involved in an event in different roles, regardless of physical presence.*

The responsibilities of people involved in different roles depends on the particular event structure and size. Large first-tier conferences, such as WWW have a four-layer program committee structure consisting of program chairs, track chairs, senior PC members and ordinary PC members. Smaller events lack the track and/or senior PC chair layer.
Measuring: People are involved in events with different roles. *FM16.1* is a set valued basic metric that maps the person names to roles. The roles in a scientific event can be: general chair, organizing/program committee chair/senior member/member, author, keynote/invited speaker and participant. Person reputation *DM16.2* is the extent to which a person involved in a scientific event is popular and active in the community. Reputation can be measured by a number of indicators, a detailed discussion of which is

out-of-scope here. For academics these can be, for example, no. of publications, no. of citations, h-index, i10-index etc. Frequency of involvement *DM16.3* is a derived metric from *DM16.1* which shows the number of times a person have had different roles in a particular event series. Frequency of same-role involvement *DM16.4* is a metric that shows the number of times a person have had the same roles in that event.
Advantages:

- The involvement of high-profile researchers in an event can improve the quality and raise enthusiasm among prospective authors and attendees for an event.

Disadvantages:

- It is difficult to measure the actual involvement and commitment and a pro-forma involvement will result in a lack of commitment of people involved in certain roles.
- Some measures in this dimension individually are only limited indicators for quality.

Person Backgrounds Background diversity of people participating in different events can be an indicator for broad coverage of teh relavant topics by an event,

Definition 27: Participant *indicates the quality or state of participants in terms of experience and workplace.*

Measuring: Diversity of experience level *DM17.1* can be measured as the average number of years of experience that shows whether people in different roles are experienced researchers or students. Countries of participants or organizers *DM17.2* and *DM17.3* shows who is involved in different roles in the event from different sectors.
Advantages: Event participants with diversity of experience level has an impact on other dimensions, e.g., topic coverage, ways of publicity, good quality reviews, etc. People involved from different countries increase the awareness about the event and its impact.

- Event participants with diversity of experience level has an impact on other dimensions, e.g., topic coverage, ways of publicity, good quality reviews, etc.
- People involved from different countries increase the awareness about the event and its impact.

Bibliographic data-related The outcome of a positive review is a set of accepted submissions, which will be published section 4.4.1 and presented in the event.
Publication Availability and Accessibility Availability of the artifacts published by an event directly impact the awareness of communities about the event and its quality.

Definition 28: Availability and Accessibility of Publications *of an event refers to its metadata as well as its full texts.*

Measuring: *FM18.1* measures the indexing of the individual publications of an event (actually of the *metadata* of its publications) in indexes such as the commercial Web of Science, Scopus, or the free DBLP. Accessibility of full texts is considered in *FM18.2*. We measure whether the publications of an event are published in institutional open access repositories.
Advantages:

- Easy availability and accessibility of publications gives a quick intuition about the covered topics and the type of accepted contribution by the event.

Dimension	**Metric**	**Type of Impl.**	**Ease of collect**	**Data avail.**	**Data Reliab.**	**Data precision**	**Ease of comput.**
M1. Submissions	FM1.1 Accepted styles	-	✓	✓	✓	✓	✓
	FM1.2 Accepted formats	-	✓	✓	✓	✓	✓
	FM1.3 Submission types	*(prop)*	✓	✓	✓	✓	✓
	FM1.4 Page count	-	✓	✓	✓	✓	✓
	DM1.5 Submission combinations	-	✓	✓	✓	✓	✓
	FM1.6 License	-	✓	✓	✓	✗	✗
M2. Location	FM2.1 Location	*(cate)*	✓	✓	✓	✓	✓
	DM2.2 Location-visited	*(temp)*	✓	✓	✓	✓	✗
	DM2.3 Distribution of locations	*(ASK)*	✗	✓	✓	✓	✗
	DM2.4 Event type (split/merge)	*(temp)*	✗	✓	✓	✗	✗
M3. Review process	FM3.1 Existence of review process	-	✓	✓	✓	✓	✓
	DM3.2 Review type	-	✓	✓	✗	✓	✓
	FM3.3 Reviews process type	-	✓	✓	✗	✓	✓
M4. Review results	FM4.1 Full review records	-	✗	✗	✗	✗	✗
	DM4.2 Minimum number of reviews	-	✗	✗	✗	✗	✗
	DM4.3 Avg. length of reviews	-	✗	✗	✗	✗	✗
	DM4.4 Avg. confidence of reviewers	-	✗	✗	✗	✗	✗
	DM4.5 Delegation ratio	-	✗	✗	✗	✗	✗
	DM4.6 Avg. relevance of submissions	-	✗	✗	✗	✗	✗
	DM4.7 Avg. no. of reviews	-	✗	✓	✗	✗	✗
	DM4.8 Acceptance rate	*(temp)*	✗	✗	✗	✗	✗
M5. Publishing	FM5.1 List of publishers	*(prop)*	✓	✓	✓	✓	✓
	DM5.2 Existence of official publisher	-	✓	✓	✓	✓	✓
	DM5.3 Publisher popularity	*(prop)*	✗	✗	✗	✗	✗
M6. Journal event coupling	FM6.1 Coupling type	-	✓	✓	✓	✓	✓
	FM6.2 Journal name	*(prop)*	✓	✓	✓	✓	✓
	FM6.3 Journal publisher	*(prop)*	✓	✓	✓	✓	✓
	FM6.4 Journal popularity	-	✗	✗	✗	✗	✗
	FM6.5 Eigenfactor score	-	✗	✗	✗	✗	✗
	FM6.5 Journal impact factor	*(prop)*	✗	✗	✗	✗	✗
M7. Discoverability	DM7.1 By full name	*(prop)*	✗	✗	✗	✗	✗
	DM7.2 By acronym	*(prop)*	✗	✗	✗	✗	✗
	DM7.3 By topic	*(cate)*	✗	✗	✗	✗	✗
M8. Reputation and Impact	FM8.1 Impact factor	-	✗	✓	✓	✓	✓
	FM8.2 h5-index	*(prop)*	✓	✓	✓	✓	✓
	FM8.3 Rank	*(prop)*	✓	✓	✓	✓	✓
M9. Sponsorship	FM9.1 Distribution of finance cont.	-	✗	✓	✓	✗	✗
	EM9.2 Distribution of sponsors/type	*(ASK)*	✗	✓	✓	✗	✗
	DM9.3 No. of sponsors	-	✓	✓	✓	✓	✗
	DM9.4 Reputation of sponsors	-	✗	✓	✓	✓	✗
	DM9.5 Type of sponsor local/global	*(prop)*	✗	✓	✓	✓	✗
	DM9.6 Continuity of sponsors	*(SPARQL)*	✗	✓	✓	✓	✗
M10. Co-Location	FM10.1 Co-Located events	*(temp)*	✓	✓	✓	✓	✓
	DM10.2 No. of co-located events	*(prop)*	✓	✓	✓	✓	✓
	FM10.3 Admission criteria:	-	✓	✓	✓	✓	✓
	FM10.4 Type of co-located events	*(prop)*	✓	✓	✓	✓	✓
	DM10.5 Qual. of co-loc.. sub-events	-	✗	✓	✓	✓	✓
	DM10.6 Qual. of co-loc.. super-events	-	✗	✓	✓	✓	✓
	DM10.7 Qual. of sister events	-	✗	✓	✓	✓	✓
	DM10.8 Repu. of co-located event	-	✗	✓	✓	✓	✓

Table 4.5: **Overview of quality metrics.** The defined metrics for person and bibliographic data related dimensions and of the characteristics of their metrics (✗means that metrics applies and ✓means does not apply, type = S means subjective, type = O means objective)

Dimension	Metric	Type of Impl.	Ease of collect	Data avail.	Data Reliab.	Data precision	Ease of comput.
M11. Topical focus	FM11.1 Event name	*(prop)*	✓	✓	✓	✓	✓
	FM11.2 Focus type	*(cate)*	✓	✓	✓	✓	✓
	FM11.3 Coverage of innovative topics	*(cate)*	✓	✓	✓	✗	✓
M12. Registration	DM12.1 Fees	*(prop)*	✓	✓	✓	✓	✓
	DM12.2 Ease of registration	-	✗	✗	✓	✓	✓
	DM12.3 Student discounts	*(prop)*	✓	✓	✓	✓	✓
M13. Schedule	FM13.1 Full schedule	*(temp)*	✓	✓	✓	✗	✓
	DM13.2 Avg. Length of breaks	*(SPARQL)*	✓	✓	✓	✗	✓
	DM13.3 No. of breaks	*(prop)*	✓	✓	✓	✓	✓
	DM13.4 Avg. Event duration	*(temp)*	✓	✓	✓	✓	✓
	DM13.5 No. of sessions per day	*(temp)*	✓	✓	✓	✓	✓
	DM13.6 No. Presentations per session	*(SPARQL)*	✓	✓	✓	✓	✓
M14. Awards	DM14.1 Type of awards	*(prop)*	✓	✓	✓	✓	✗
	DM14.2 Amount of awards	*(prop)*	✓	✗	✓	✗	✓
M15. Publicity	FM15.1 Communication channels	*(prop)*	✗	✗	✗	✗	✗
	FM15.2 Homepage usability	-	✓	✓	✓	✗	✓
	FM15.3 Homepage comprehensiveness	-	✓	✓	✓	✗	✓
M16. Organizers and Roles	FM16.1 Person Roles	*(prop)*	✓	✓	✓	✗	✗
	DM16.2 Person reputation	*(prop)*	✗	✗	✓	✗	✗
	FM16.3 Freq. of involvement	*(SPARQL)*	✓	✗	✓	✗	✗
	FM16.4 Freq. of same-role	*(SPARQL)*	✗	✗	✓	✗	✗

Table 4.6: **Overview of quality metrics.** The defined metrics person and bibliographic CONT.,

Bibliometrics Information related to bibliometrics of an event can be the papers accepted, cited or referenced in an event. This plays an important role as the publications are their quality of science directly impact the profile of events.

Definition 29: Bibliometrics *is the statistical analysis of the publications of an event.*

Measuring: Number publications *FM19.1* shows to which extent the event is productive. In order to assess the impact of event and its publications number of citation per publications *FM19.2* is studied. Assessing the impact of publication can have several other dimensions. The proposed measurement is only limited to the current most used metric, citations. A more systematic evaluation required to be done by using minding and artificial intelligence approaches. In such measurement, community involvement plays an important role. However, in order to stay with a least possible objective metric, we define the citation count as the only metric here.
Advantages:

- The quality and quantity of publications and number of citations provide a key measure of the event productivity.
- Bibliometrics of an event indicates the extent to which the event helped researchers and their contributions to be recognized in the community.

Disadvantages:

- Pure number of citations is not a good quality indicator, citations should be analyzed.

Social network impact Publicity of any announcement has been increased with the help of social media. Scientific events are not excluded from the impact of social media.

Dimension	**Metric**	**Type of Impl.**	**Ease of col-lect**	**Data avail.**	**Data Reliab.**	**Data preci-sion**	**Ease of com-put.**
M17. Person Backgrounds	DM17.1 Experience level	-	✗	✗	✗	✗	✗
	DM17.2 Countries	*(ASK)*	✗	✗	✗	✗	✗
	DM17.3 Sectors	-	✗	✗	✗	✗	✗
M18. Publication Avail. & Access.	FM18.1 Indexing	-	✓	✓	✓	✗	✗
	FM18.2 Accessibility	-	✓	✓	✓	✗	✗
M19. Biblio-metrics	FM19.1 No. of publications	*(prop)*	✓	✓	✓	✓	✗
	FM19.2 No. of citations per publication	*(SPARQL)*	✗	✓	✓	✗	✗
M20. Social network impact	FM20.1 No. Page view	-	✗	✓	✓	✗	✗
	FM20.2 No. Discussed	-	✗	✗	✗	✗	✗
	FM20.3 No. Twitter hashtags	*(SPARQL)*	✗	✓	✓	✗	✓
	FM20.4 No. Recommended	-	✗	✗	✗	✗	✗

Table 4.7: **Overview of quality metrics.** The defined metrics for person and bibliographic CONT.,)

Definition 30: Altmetrics *is considered as a composite indicator representing* impact *of the event with regard to publicity of its publications.*

In this paper, an adopted definition by ImpactStory and Public Library of Science is used for Altmetrics and will be measured with the three following metrics: *FM20.1* representing number of HTML page views of the publications and number of times the publications have been downloaded, *FM20.2* shows the number of times the published papers by the event is discussed in social media such as Twitter, Facebook, or science blogs and the number of times the papers are saved in social bookmarks depicted by *FM20.3* shows the number of twitter hashtags corresponding to the event, and *FM20.4* represents the number of times the publications of an event is recommended by systems such as F1000Prime[43].

Advantages:

- This metric concerns a different aspect than citation of the publications.
- It collects the publicity of the individual publications thus the event concerning the out-standing audience of the event.

Disadvantages:

- The quality of such measurement can not be precise.

Event History Almost all of the metrics defined in previous section 4.4.1 can also be applied to event *series* i.e. to each edition of an event (as pointed out in Table **??**). Even the length of the recorded history (which is closely related to the number of editions) is a valuable information about how much the respective scientific event has established in the respective community.

The following historical criteria have been defined by taking other metrics into consideration:

- *Regularity* is calculated by looking at all years of the event and computing the avg. distance of true values on this axis e.g., how regularly had the conference journal special issue.
- *Continuity* refers to how long an event has lasted since it has been established. Year is the granularity for this metric and counting the number of years since the first time the event was established.
- *Monotonisity* is defined as a growth measure of event submissions. To be more concrete, an event is called with increasing monotonisity if every year the event have had more submissions that the previous year.

[43] `http://f1000.com/prime`

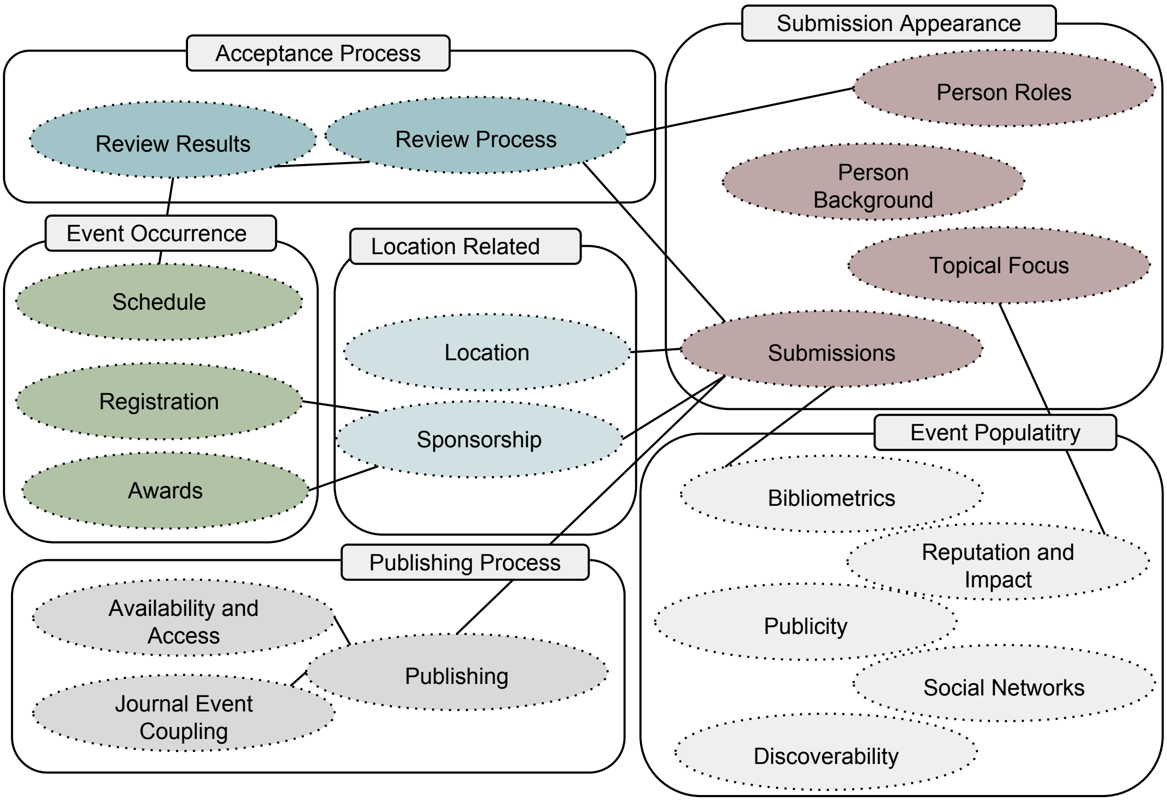


Figure 4.11: **Dependency between dimensions.** The dimensions and the following metrics are clustered ased on their dependency to each other.

- *Diversity* is defined as the degree of variation of one metric.

We used the previously proposed formula to calculate the percentage of continuity for a specific event:

$$continuity = \min\{100\%, (\#\ event\ editions * regularity)/age\}$$

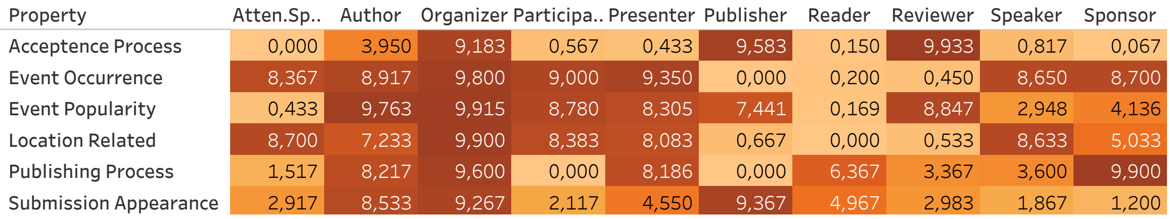
Overall, most of the defined metrics are related to individual events and also event series except *DM2.2 Location-visited*, *DM2.3 Distribution of locations*, *DM2.4 Event type (split/merge)*, *FM8.3 Rank*, *DM9.6 Continuity of sponsors*, *FM16.3 Freq. of involvement*, and *FM16.4 Freq. of same-role involvement*. All metrics are calculated objectively.

Dependency Graph of Dimensions The quality metrics for quality assessment of scientific events are not independent from each other. A higher level dependency graph is shown in figure 4.11. For example, more sponsors become active based on the location of the conference or dicscoverability of a an event can affect biliometrics of it.

4.4.2 Analysis and Assessment Results

Necessity Survey We asked 60 researchers from two scientific fields, Computer Science and Social Science to explain: the most important metrics for their community to find and select a scientific event and the current ways they use to explore scientific events based on these metrics. Findings of this study is explained below and the data is available online[44].

[44] https://goo.gl/DSSazs



Property	Atten.Sp..	Author	Organizer	Participa..	Presenter	Publisher	Reader	Reviewer	Speaker	Sponsor
Acceptence Process	0,000	3,950	9,183	0,567	0,433	9,583	0,150	9,933	0,817	0,067
Event Occurrence	8,367	8,917	9,800	9,000	9,350	0,000	0,200	0,450	8,650	8,700
Event Popularity	0,433	9,763	9,915	8,780	8,305	7,441	0,169	8,847	2,948	4,136
Location Related	8,700	7,233	9,900	8,383	8,083	0,667	0,000	0,533	8,633	5,033
Publishing Process	1,517	8,217	9,600	0,000	8,186	0,000	6,367	3,367	3,600	9,900
Submission Appearance	2,917	8,533	9,267	2,117	4,550	9,367	4,967	2,983	1,867	1,200

Figure 4.12: **Relevance of Quality Metrics to Stakeholders.** The relevance of the defined metrics to different stockholders of the scientific communication.

Finding Relevant and Good Events Participants indicated that they explore scientific events using search engines, mailing lists, social media, and personal contacts. Then, they assess the CfPs to find out whether an event satisfies their criteria. For selecting an event to participate in, all participants confirmed that they consider information that is not served directly by the current communication channels.
Agreement with Our Metrics More than half of the participants agree that, from the criteria defined in previous section, the main criteria that make an event the best event of its scientific community are: quality of reviews (M3, M4), reputation of organizers and keynote speakers, topical focus of the event, and high number of citations to the accepted papers of the previous years. Additionally, they confirmed the relevance of the following criteria: location (M2), 96% networking possibilities (M), review quality (M3, M4), the reputation of the organizers (M16), keynote speakers and sponsors (M16), acceptance rate (M), the quality of co-located events (M16), the accessibility of the location (M16), and citations counts for accepted papers of previous years (M16), i.e., the "impact factor"(M16).
Most Valuable Metrics Overall, 95% of the participants agreed that the quality metrics defined by us can lead to a quality-based ranking of scientific events. For some metrics, it is not straightforward to *interpret* whether the raw, measurable value is good or bad. Concretely, we asked the participants to characterize the metrics in the "review" dimension accordingly. 90% of the participants of the best conferences in their area of research received 3 or 4 reviews (including meta-reviews) per submission. Over 77% of the participants estimated the average length of the reviews between 10 and 30 lines of text.
Relevance of a Recommendation Service Over 36% of the participants agreed that having an event recommendation service is highly relevant for them. 46% of them answered it is somewhat relevant.
Relevance of a Defined Metrics to Person Roles We categorized the metric into six high level category including different dimensions and metrics as: submission appearance, acceptance appearance, location, event occurrence, event popularity and publishing process. Survey participants are asked to score (0 being lowest and 10 as highest rank) relevance of person roles to the defined property groups. Results are shown in figure 4.12[45].

Event Metadata Evaluation by Applying Quality Metrics We evaluate our quality metrics for scientific events by applying them to events in our own community, i.e., events of whose quality we have an intuitive understanding, and then discussing our observations. Table 4.8 lists the conference series we considered in providing the statistics. These are the either top ranked conferences or often targeted venues from the majority of the community participated in our surveys. For evaluating the application of metrics to event series as well, we studied the past 10 editions of the for WWW and VLDB conference series because of data availability. In the following we summarize the results of the assessment per dimension.
Submissions: All observed conferences accepted submissions in the LaTeX and MS Word formats. Only one conference, WI 2015, allowed submissions in a Web-based format providing better accessibility and

[45] DrawnbyTableau:https://www.tableau.com

Acronym	URL	Community Ranking	SuggestedBy**	Publisher
WWW	http://iw3c2.org/	A*	19	ACM
SIGMOD/PODS	http://sigmod2015.org/	A*	8	ACM
VLDB	http://vldb.org/conference	A*	6	VLDB
JCDL	http://jcdl.org/	A*	2	ACM
ISWC	http://swsa.semanticweb.org/	A	18	Springer
ESWC	http://eswc-conferences.org/	A	14	Springer
WI	http://wi-consortium.org/	B	2	IEEE
SEMANTiCS	http://semantics.cc/	-	6	ACM
KESW	http://kesw.ru/	-	2	Springer
WIMS	http://wims.vestforsk.no/	-	2	ACM
ICSC	http://ieee-icsc.org/	-	2	IEEE

Table 4.8: **List of the observed conferences** Conferences in 2015 (sorted by rank) – suggested by the participants of the first survey from computer science community.

potential for interactivity: ePUB. All of the conferences followed their publisher's document style. Six accepted submissions in ACM style, two of them in LNCS style, and the remaining two in IEEE style. VLDB has its own style.

Overall, flexibility w.r.t. style is limited by the publishers, and LaTeX and MS Word still dominate the accepted formats. This proves that scientists still write static documents, which do not use the possibilities of digitization, such as interactivity, easy accessibility, multimodality or semantic content annotation and representation [44, 67]. Besides the positive example of WI 2015 accepting ePUB, such innovations are typically pioneered by smaller events. The SAVE-SD workshop co-located with WWW 2015 allowed LaTeX and MS Word but preferred submissions in RASH.

As data about authors' affiliation is rarely available in an open, structured, reusable form but often hidden in the databases of submission management systems such as EasyChair[46], it is hard to determine the diversity of submissions w.r.t. countries. Determining diversity w.r.t. sectors is harder: while an author's country is usually recorded explicitly, one would have to infer the sector from the author's organization.

Location: Table 4.9 shows number of times event series considered in this study visited different continents. ESWC as a European conference obviously never has been held out of European. The same fact is true for KESW as a east European conference has been held in Russia seven times out of overall eight edition and once in Poland. ISWC, which had its 16th edition in 2017, has been held equally in Europe and the US for five times each, and four times in Asia and two time in Australia. VLDB has visited a wide range of visited locations, including many countries per continent, while WWW is selective about countries. Neither of them has visited the same continent for two successive editions.

Review process: The majority of computer science conference proceedings are peer-reviewed, in particular those in our evaluation. Some computer science journals do double-blind reviews, whereas conferences, in particular those in our evaluation, typically do a single-blind review. To the best of our knowledge none of the formal events in computer science has an open review process at the moment. The Semantic Web Journal publishes its formal reviews if the assigned reviewers agree (most of them do) but only welcomes informal comments from third parties.

Review results: The acceptance rate of WWW 2017 was 17%;in the previous editions it has always ranged between 11% and 19%. VLDB maintains a dedicated statistics page. The acceptance rate for the research track has always been below 20%. For the other conferences, the acceptance rate varied between 22% and 32%. Further information about the review result is not public for the assessed conferences. We therefore asked 10 researchers in the community who had submitted to any of these conferences to

[46] http://www.easychair.org/

Conference	Overall	America	Europe	Asia	Australia
ESWC	14	0	14	0	0
ICSC	11	10	1	0	0
ISWC	16	5	5	4	2
JCDL	15	13	1	0	1
KESW	6	0	6	0	0
PODS	34	30	2	1	1
SEMANTiCS	13	0	13	0	0
VLDB	41	11	19	9	2
WI	16	5	5	5	1
WIMS	7	0	7	0	0
WWW	24	9	9	5	1

Table 4.9: **Distribution of conferences** Statistics about the conferences distributed over continents (to reduce the number of zeros North America and South America are presented as one and Africa is skipped.)

provide us with information on the reviews they had received regardless of acceptance or rejection.

Based on this data, ESWC 2017 had 4 reviews plus one meta-review per submission. All the other conferences provided at least three reviews per submission, of which around two were of a sufficient quality. The average length of WWW 2017 and VLDB 2017 reviews is more than 100 lines of text, which indicates a high quality. Regardless of whether papers were accepted or rejected, the average review length in four more conferences was more than 50 lines per review, and authors considered them helpful, which emphasizes the expertise of the reviewers. For the remaining four conferences (ICSC, WI, WIMS, and SEMANTiCS) the average length of reviews was below 25 lines. Surprisingly, reviews from KESW conference have been ranked highly with regard to the length and quality of the reviews in the average of three reviews per submission. However its acceptance rate has been always over 35% and surprisingly dropped to 28% in 2017.

Publishing: The reputation of the publisher and the expected impact of a publications published in a community have a great influence on the decision of a researcher to submit to an event or to read its proceedings. All of the conferences publish with one of the major commercial computer science publishers: ACM, IEEE, Springer. Not all of the co-located events have been used the same publisher as the main event itself. In some cases other publishers such as Elsevier[47] or IOS Press[48] have been the option. VLDB conference series is using its own publishing process without having an external publishing house involved. ACM is the major used publisher for the conferences evaluated in this study ; however electronic versions are included in both the ACM and IEEE digital libraries.

Discoverability: The results we obtained when searching by topic proves that without being aware of the existence of a particular event, one will hardly discover it. We did find related *journals* while searching by topic (e.g. the Semantic Web Journal when searching for “semantic web”) but none of our evaluated conferences. The ranking of every conference improved significantly by adding the type of the event, i.e. “conference”, as a search keyword. In addition, the homepage of every event evaluated made it into the top 10 results by adding the year “2017” to the acronym of the event.

Sponsorship: The big players of the community, including Google, Facebook and Yahoo, typically sponsor big events. WWW 2015 sponsors have one local sponsors “TIM”, which aimed at increasing their popularity in academia by taking advantage of the proximity of the event’s location. Our data shows a high relation between sponsors and organizers;in most cases, at least one organizer has a role in the organization or the company or the project that becomes the sponsor of an event.

Reputation: According to Google Scholar, the h5-indexes of WWW, VLDB, ISWC are 77, 73, and 41,

[47] www.elsevier.com

[48] https://www.iospress.nl/

respectively. No such information is available for the other conferences. According to the CORE2017 ranking, WWW, VLDB, PODS and JCDL are A* conferences, ISWC and ESWC are A conferences, and WI is a B conference; the remaining ones are not ranked. However, community consider them as a serious target.

Participants: Every WWW conference from 2006 to 2017 has recruited at least half of their general chairs from the country where the conference was located. WWW generally has PC members with a high reputation, who demonstrate further commitment in that they often organize co-located sub-events. The h-index of people involved in the WWW series ranges from 15 to 90; their i10-index ranges from 20 to 500 with up to ~30,000 citations.

The frequency of involvement of PC members and keynote speakers is high, but the organizers vary. For example, Tim Berners-Lee has been the keynote speaker of six editions of the WWW conference. All academic keynote speakers of the WWW conferences evaluated had an h-index of over ~25 with more than ~1000 citations. Most of the above facts are similar for VLDB. Industrial keynote speakers of WWW and VLDB are founders of big players, heads of big companies, etc. Every edition of WWW and VLDB over the past 10 years had around 900 registered participants.

The other eight conferences evaluated pursue a different strategy. They have a core team of people frequently involved in the organization, whereas the frequency of involvement of PC members and keynote speakers in the same role varies.

Person Role: The most obvious finding from the analysis is that, major number of main organization committee members as well as keynote speakers are changing with event location. However core part of program committee remains the same with slight changes from local scientists be introduced. In WWW and VLDB since 2007 to 2017, people happen to take several roles wither in the same editions or different roles in different editions. In more than 20 cases out of 40, keynote speakers have been often in the role of program committee members also.

Looking at the list of keynote speakers of the above conferences during last 10 year, overall at WWW 30% have been from academia and 60% from industry and 10% from both academia and industry (based on the affiliation given in the conference homepage or Google Scholar profile). However, at VLDB conference 57% of the keynote speakers have been from academia and 38% from industry and the remaining 5% were from both categories. The data for this analysis is gathered from the affiliations provided on the homepage of each event edition. SEMANTiCS event series have more than 60% of its keynote speakers from industry and the opposite fact applied for JCDL.

4.5 An Alternative Approach: Bootstrapping a Value Chain

An alternative approach have been applied in order to assess quality of scholarly artifacts by producing semantically enriched dataset. Having such a dataset at hand lead to a *data value chain* producing value for the scientific community [184]. Bootstrapping and enabling such value chains is not an easy task. A solution that has proved to be successful in other communities is to run challenges, i.e. competitions in which participants are asked to complete tasks and have their results ranked, often in objective way, to determine the winner. Even a number of projects have been launched to accelerate this process, for instance LinkedUp[49] or Apps for Europe[50]. The success of the LAK[51] or Linked Up[52] challenges is worth mentioning here. However, these challenges focus on exploiting scholarly linked data for different

[49] http://linkedup-project.eu/

[50] http://www.appsforeurope.eu/

[51] http://meco.l3s.uni-hannover.de:9080/wp2/?page_id=18

[52] http://linkedup-challenge.org/

purposes (for instance, to monitor progress) but less on actually producing such datasets. To this end, a series of challenges have been designed with the following main objectives:

- bootstrap a value chain for scientific data,
- enable services on top of the collected and transformed dataset,
- perform quality assessment tests over data at the time of data production.

Semantic publishing is defined as the use of Semantic Web technologies to make scholarly publications and data easier to discover, browse and interact with [243]. Aligned with this concept, the challenge series were named Semantic Publishing Challenges (SemPub) [53], aiming at the production of datasets on scholarly publications [71]. It has been held along four editions of the Extended Semantic Web Conference series (ESWC) [54] starting in 2014 [158] till 2017 [272]. The key idea was to involve participants in extracting data from heterogeneous resources and producing datasets on scholarly publications, which can be exploited by the community itself. Differently from other challenges in the semantic publishing domain, whose focus is on exploiting semantically enriched data, SemPub focuses on producing Linked Open Datasets. This is done by extracting, annotating and sharing scientific data (by which, here, we mean standalone research datasets, data inside documents, as well as metadata about datasets and documents), up to building new research on them. To the best of our knowledge, this was the first and only challenge of its kind.

4.5.1 Definition of The Challenge: Tasks, Queries and Datasets

The challenge was defined as a call for participation to the Semantic Web community. The participants have been asked to extract data from scholarly papers and to produce an RDF dataset that could be used to answer some relevant queries: concretely, queries about the quality of scientific output. Challenge organizers were aware of other topics of interest for the community e.g., nanopublications, research objects, etc. but the challenge focused on papers only to bootstrap the initiative and to start collaboratively producing initial data. A list of different tasks has been designed, sharing the same organization, rules and evaluation procedure. All three editions used the same evaluation procedure, but the tasks were refined over time. In the first edition, having called for submissions, we received feedback from the community that mere information extraction, even if motivated by quality assessment, was not the most exciting task related to the future of scholarly publishing, as it assumed a traditional publishing model. Furthermore, in 2014 to address the primary target of the challenge, i.e. "publishing" rather than just "metadata extraction", we widened the scope by adding an *open task*, whose participants were asked to showcase data-driven applications that would eventually support publishing. Two tasks have been defined at the very beginning [158], and a third task on interlinking was added in 2015 [66]. The tasks of the challenge have been:

- Task 1: participants were asked to extract information from selected CEUR-WS[55] workshop proceedings volumes (HTML tables of content using different levels of semantic markup, plus PDF full text) to enable the computation of indicators for the workshops' quality assessment. They were asked to answer 20 different queries.
- Task 2: participants were asked to extract data about citations, to enable precise assessment of linking, sharing and evaluating research through citations. The dataset included a set of XML-encoded research papers, taken from PubMedCentral and Pensoft Open Access archives, and

[53] https://github.com/ceurws/lod/wiki
[54] https://en.wikipedia.org/wiki/Extended_Semantic_Web_Conference
[55] http://ceur-ws.org/

heterogeneous in terms of internal structure, styles and numbers. Both dataset and queries were completely disjoint from Task 1.

- Task 3: participants were asked to interlink scholarly entities in the CEUR-WS dataset with the same entities as they appear in other datasets.Persons acting e.g as authors of a publication or editors of a workshop, and their affiliations might already appear on other datasets of LOD, e.g. DBLP. Similarly, events, as conferences and workshops, might also appear on the aforementioned datasets. All those entities should be identified and interlinked.

In 2015 we were asked to include only tasks that could be evaluated in a fully objective manner, and thus discarded the open task. The distance between Tasks 1 and Task 2 was reduced by using the same dataset for both. We transformed Task 2 into a PDF mining task and thus moved all PDF-related queries there. The rationale was to differentiate tasks on the basis of the competencies and tools required to solve them, but to make tasks interplay on the same dataset.

Two types of datasets have been made available for challenge participants in different time period. A training dataset (TD) has been published on which the participants could test and train their extraction tools. A few days before the submission deadline, we published an evaluation dataset (ED): the input for the final evaluation. An overview of the datasets used for the above mentioned tasks are as following:

- Training and Evaluation dataset for Task 1: The CEUR-WS.org workshop proceedings volumes served as the source for selecting the training and evaluation datasets of Task 1 in all challenge editions. In this data source, which included data spanning over 20 years, workshop proceedings volumes were represented in different formats and at different levels of encoding quality and semantics. An HTML 4 main index page links to all workshop proceedings volumes, which have HTML tables of contents and contain PDF or PostScript full texts. A mixture of different HTML formats (no semantic markup at all, different versions of microformats, RDFa) were chosen for both the training and evaluation datasets. The training dataset comprised all volumes of several workshop series, including, e.g., the Linked Data on the Web workshop at the WWW conference, and all workshops of some conferences, e.g., of several editions of ESWC. In 2014 and 2015, the evaluation dataset was created by adding further workshops on top of the training dataset. To support the evolution of extraction tools, the training datasets of 2015 and 2016 were based on the unions of the training and evaluation datasets of the previous years. In 2015 and 2016, the Task 1 dataset of the previous year served as an input to Task 3.
- Training and Evaluation dataset for Task 2: In 2014, the datasets for Task 2 included XML files encoded in JATS[56] and TaxPub[57], an official extension of JATS customized for taxonomic treatments [49]. The training dataset consisted of 150 files from 15 journals, while the evaluation dataset included 400 papers and was a superset of the training dataset. In 2015, we switched to PDF information extraction: the training dataset included 100 papers taken from some of the workshops analyzed in Task 1, while the evaluation dataset included 200 papers from randomly selected workshops (uniform to the training dataset). In 2016, we reduced the number of papers increasing the cases for each query. Thus, we included 50 PDF papers in the training and 40 in the evaluation dataset. Again, the papers were distributed in the same way and used different styles for headers, acknowledgments and structural components.
- Training and Evaluation dataset for Task 3: The training dataset for Task 3 consists of the CEUR-

[56] JATS, http://jats.nlm.nih.gov/

[57] TaxPub, https://github.com/plazi/TaxPub

WS.org dataset produced by the 2014 winning tool of Task 1[58], COLINDA[59], DBLP[60], Lancet[61], SWDF[62], and Springer LD[63] in 2015 and the CEUR-WS.org datasets produced by the 2015 winning tools of Task 1[64] and Task 2[65], of COLINDA, DBLP, and Springer LD in 2016.

CEUR-WS.org data has become the central focus of the whole Challenge, for two reasons:on the one hand, the data provider (CEUR-WS.org) takes advantage of a broader community that builds on its data, which, before the SemPub Challenges, had not been available as linked data. On the other hand, data consumers gain the opportunity to assess the quality of scientific venues by taking a deeper look into their history, as well as the quality of the publications.

Tasks Evaluation Process The evaluation of the submitted solutions was conducted in a transparent and objective way by measuring precision and recall. To perform the evaluation, we relied on a gold standard and an evaluation tool which was developed to automate the procedure. A gold standard datasets have beengenerated *manually* and used for each task's evaluation. It consisted of a set of CSV files, each corresponding to the output of one of the queries used for the evaluation. Each file was built after checking the original sources – for instance HTML proceedings in case of Task 1 and PDF papers for Task 2 – and looking for the output of the corresponding query; then, it was double-checked by the organizers. Furthermore, we also made available the gold standard to the participants (after their submission) so as they have the chance to report inaccuracies or inconsistencies. The final manually-checked version of the CSV files was used as input for the evaluation tool. The evaluation tool[66] compares the queries output provided by the participants (in CSV) against the gold standard and measures precision and recall. It was not made available to the participants after the 2014 edition, it was only made available after the 2015 edition, while it was made available already by the end of the training for the 2016 edition. This not only increased transparency but also allowed participants to refine their tools and address output imperfections, increasing this way the quality of their results.

Queries as Quality Assessment Metrics For each task, a set of queries in natural language was published and participants were asked to translate them into SPARQL and to submit a dataset on top of which these queries would run. Common questions related to the quality of a scientific workshop or conference include whether a researcher should submit a paper to it or accept an invitation to its program committee, whether a publisher should publish its proceedings, or whether a company should sponsor it [41]. To test whether the produced dataset are suitable for quality assessment of scholarly artifacts, a set of queries have been designed to be performed immediately after generating the dataset [158]. Overall, we had 20 queries for task 1 each year. Here, we represent two examples in order to show how the participants were to produced data and use it for quality assessment:

Q. 1 (institutional diversity and internationality of chairs) Identify the affiliations and countries of all editors of the proceedings of workshop W.

[58] 2014 CEUR-WS dataset, https://github.com/ceurws/lod/blob/master/data/ceur-ws.ttl
[59] COLINDA, http://www.colinda.org/
[60] DBLP, http://dblp.l3s.de/dblp++.php
[61] Lancet, http://www.semanticlancet.eu/
[62] SWDF, http://data.semanticweb.org/
[63] Springer LD, http://lod.springer.com/
[64] 2015 CEUR-WS Task 1 dataset, http://rml.io/data/SPC2016/CEUR-WS/CEUR-WStask1.rdf.gz
[65] 2015 CEUR-WS Task 2 dataset, http://rml.io/data/SPC2016/CEUR-WS/CEUR-WStask2.rdf.gz
[66] SemPubEvaluator, https://github.com/angelobo/SemPubEvaluator

The query was asked for a given list of workshops. The markup for the affiliations in the proceedings tables of contents is not quite uniform which makes it challenging to extract the right information with a solid solution. A list of certain workshops has been given as an input and the participants were supposed to design the queries considering the institutional diversity and internationality of chairs.

```
PREFIXes skipped...

SELECT ?workshop ?editor_name ?affil ?country WHERE {
  VALUES ?workshop {
    <http://ceur-ws.org/Vol-1085/>
    <http://ceur-ws.org/Vol-800/>
    <http://ceur-ws.org/Vol-540/>
    <http://ceur-ws.org/Vol-1/>}
  [...]
  ?proc bibo:presentedAt ?workshop .
  ?proc swrc:editor ?editor .
  ?editor foaf:name ?editor_name .
  ?editor swrc:affiliation ?affil .
  OPTIONAL {
    ?editor bibo:place ?country . }
} ORDER BY ?workshop
```

Listing 4.3: **Query. 1**. (institutional diversity and internationality of chairs).

The participants were asked to include the full affiliation including department names, but without details of the addresses. For country names, it was recommended to follow the DBpedia naming conventions. Therefore, the expected output would be:

```
Expected output format (CSV):
workshop-iri ,editor-name,affiliation ,country
<IRI>,rdfs:Literal ,rdfs:Literal ,<IRI>
[...]
 <.../Vol-1/>,"Manfred A. Jeusfeld","RWTH Aachen" ,<.../Germany>
 <.../Vol-1085/>,"Marc Aiguier","Laboratoire MAS" ,<.../France>
 <.../Vol-540/>,"Florian Urmetzer","SAP" ,<.../Switzerland>
 <.../Vol-800/>,"Manfred A. Jeusfeld","Tilburg University" ,<.../Netherlands>
 <.../Vol-1515/>,"Janna Hastings"," European Bioinformatics Institute" ,<.../
    United_Kingdom>
 <.../Vol-1499/>,"Henry Muccini","University degli Studi dell'Aquila" ,<.../Italy>
 <.../Vol-540/>,"Ismael Rivera","DERI/NUIG" ,<.../Republic_of_Ireland>
 <.../Vol-1499/>,"Martin Gogolla","University of Bremen" ,<.../Germany>

[...]
```

Listing 4.4: **Output. 1**. institutional diversity and internationality of chairs.

The solutions should have covered the cases that authors may move to different countries. For example, *Manfred Jeusfeld* was based in *Germany* in `http://ceur-ws.org/Vol-1/`, but in the *Netherlands* in `http://ceur-ws.org/Vol-800/`.

Q. 2 (continuity of authors) Identify the full names of those authors of papers in the workshop series titled T that have so far (co-)authored a paper in every edition of the workshop published with CEUR-WS.org.

The query was asked to be performed over the whole datasets with the following two input values for T: Mathematical User Interfaces and Ontology Matching.

```
PREFIXes skipped [...]
SELECT ?search ?author_name WHERE {
  { VALUES ?search {"Mathematical User Interfaces" [...]}
 {     SELECT ?search ?workshop ?author WHERE {
       ?workshop a swc:WorkshopEvent. [...]
       ?paper  dc:title ?paper_name .
 { SELECT ?search (COUNT(?workshop) AS ?count) WHERE {
       ?workshop a swc:WorkshopEvent.  [...]
       FILTER( strStarts(?title , ?search ) )   }
    GROUP BY ?search } ?author foaf:name ?author_name .}
GROUP BY ?search ?author_name
HAVING (COUNT(?search) = MAX(?count) ) ORDER BY ?search
```

Listing 4.5: **Query. 2(a)**. continuity of authors with regard to event participation.

A different submission for the same query was:

```
PREFIXes skipped ...
SELECT ?T ?author_name WHERE {{
SELECT distinct ?author_name ?T ?author (count(distinct ?W) as ?author_count)
WHERE {?S a bibo:Series; rdfs:label ?T.
          [...]
     ?author rdfs:label ?author_name.
     FILTER (?T = "Ontology Matching") }}{
SELECT (count(distinct ?W) as ?total)
WHERE {?S a bibo:Series; rdfs:label ?T.
        ?W a bibo:Workshop; dcterms:isPartOf ?S.
        FILTER (?T = "Ontology Matching") }}
FILTER( ?author_count = ?total)}
```

Listing 4.6: **Query. 2(b)**. continuity of authors with regrad to location.

The participants were asked to implement the query by looking for workshop title strings that start with the value given for T. In order to have a complete output, it should be considered that a workshop can be held in a multi-workshop volume. The following list is the expected output:

```
workshop-title ,author-name
rdfs:Literal ,rdfs:Literal
[...]
"Mathematical User Interface","Andrea Kohlhase"
"Ontology Matching","Jerome Euzenat"
```

Listing 4.7: **Output. 2**. continuity of authors.

4.5.2 Solutions and Produced Datasets

There were four distinct solutions in total for Task 1 during the three editions of the challenge, eight distinct solutions in total for Task 2 and none for Task 3 during the last two editions. All solutions for each task are briefly summarized here.

Task 1. There were four distinct solutions proposed to address Task 1 in 2014 and 2015 editions of the challenge. Three participated in both editions, whereas the fourth solution participated only in 2015.

Solution 1.1 [146] and [147] presented a case-specific crawling based approach for addressing Task 1. It relies on an extensible template-dependent crawler that uses sets of special predefined templates based on XPath and regular expressions to extract the content from HTML and convert it in RDF. The RDF is then processed to merge resources using fuzzy-matching. The use of the crawler turns the system tolerant to invalid HTML pages. This solution improved its precision in 2015 as well the richness of the data model.

Solution 1.2 [112] and [72] exploited a generic tool for generating RDF data from heterogeneous data. It uses the RDF Mapping Language (RML)[67] to define how data extracted from CEUR-WS.org Web pages should be semantically annotated. RML extends R2RML[68] to express mapping rules from heterogeneous data to RDF. CSS3 selectors[69] are considered to extract the data from the HTML pages. The RML mapping rules are parsed and executed by the RML Processor[70]. In 2015 the solution reconsidered its data model and was extended to validate both the mapping documents and the final RDF, resulting in an overall improved quality dataset.

Solution 1.3 [228, 229] designed a case-specific solution that relies on chunk-based and sentence-based Support Vector Machine (SVM) classifiers which are exploited to semantically characterize parts of CEUR-WS.org proceedings textual contents. Thanks to a pipeline of text analysis components based on the GATE Text Engineering Framework[71], each HTML page is characterized by structural and linguistic features: these features are then exploited to train the classifiers on the ground-truth provided by the subset of CEUR-WS.org proceedings with microformat annotations. A heuristic-based annotation sanitizer is applied to fix classifiers imperfections and interlink annotations. The produced dataset is also extended with information retrieved from external resources.

Solution 1.4 [183] presented an application of the FITLayout framework[72]. This solution participated in the Semantic Publishing Challenge only in 2015. It combines different page analysis methods, i.e. layout analysis and visual and textual feature classification to analyze the rendered pages, rather than their code. The solution is quite generic but requires domain/case-specific actions in certain phases (model building step).

Task 2 There were eight distinct solutions proposed to address Task 2 in the 2015 and 2016 editions of the challenge. Three participated in both editions, three only in 2015 and two only in 2016. As the definition of Task 2 changed fundamentally from 2014 to 2015, the only solution submitted for Task 2 in 2014 [24] is not comparable to the 2015 and 2016 solutions and therefore not discussed here.

Solution 2.1 [264] relied on CERMINE[73], an open source system for extracting structured metadata and references from scientific publications published as PDF files. It has a loosely captured architecture and a modular workflow based on supervised and unsupervised machine-learning techniques, which simplifies the system's adaptation to new document layouts and styles. It employs an enhanced Docstrum algorithm for page segmentation to obtain the document's hierarchical structure, Support Vector Machines (SVM) to classify its zones, heuristics and regular expressions for individual and Conditional Random Fields (CRF) for affiliation parsing and thus to identify organization, address and country in affiliation. Last, K-Means clustering was used for reference extraction to divide references zones into individual reference strings.

Solution 2.2 [141, 142] implemented a processing pipeline that analyzes a PDF document structure incorporating a diverse set of machine learning techniques. To be more precise, they employ unsupervised machine learning techniques (Merge-&-Split algorithm) to extract text blocks and supervised (Max Entropy and Beam search) to extend the document's structure analysis and identify sections and captions. They combine the above with clustering techniques to obtain the article's hierarchical table of content and

[67] RML, `http://rml.io`

[68] R2RML, `https://www.w3.org/TR/r2rml/`

[69] CSS3, `https://www.w3.org/TR/selectors/`

[70] RMLProcessor, `https://github.com/RMLio/RML-Mapper`

[71] GATE, `https://gate.ac.uk/`

[72] FITLayout framework, `http://www.fit.vutbr.cz/~burgetr/FITLayout/`

[73] CERMINE, `http://cermine.ceon.pl/`

	Sol 1.1	Sol 1.2	Sol 1.3	Sol 1.4	Sol 2.1	Sol 2.2	Sol 2.3	Sol 2.4	Sol 2.5	Sol 2.6	Sol 2.7	Sol 2.8
bibo	✓	✓		✓				✓	✓			
co			✓							✓		
DBO	✓		✓	✓		✓			✓			
DC	✓	✓	✓	✓	✓				✓			✓
DCterms	✓			✓	✓			✓		✓		
event		✓							✓			
FOAF	✓	✓	✓	✓				✓	✓	✓		✓
schema						✓		✓				
SKOS	✓											
SWC	✓			✓				✓				
SWRC	✓	✓	✓	✓					✓	✓		
timeline	✓			✓								
vcard			✓	✓	✓							
custom								✓	✓		✓	✓

Table 4.10: **Task 1 and 2 solutions**. The vocabularies used to annotate the data.

classify blocks into different meta-data categories. Heuristics are applied to detect the reference section and sequence classification to categorize the tokens of individual references to strings. Last, Named Entity Recognition (NER) is used to extract references to grants, funding agencies, projects, figure and table captions.

Solution 2.3 [197, 198] relied on the Metadata And Citations Jailbreaker (MACJa – IPA) in 2015, which was extended to the Article Content Miner (ACM) in 2016. The tool integrates hybrid techniques based on Natural Language Processing (NLP, Combinatory Categorial Grammar, Discourse Representation Theory, Linguistic Frames), Discourse Reference Extraction and Linking, and Topic Extraction. It also employs heuristics to exploit existing lexical resources and gazetteers to generate representation structures. Moreover, it incorporates FRED[74], a novel machine reader, and includes modules to query external services to enhance and validate data.

Solution 2.4 [235, 236], relying on LODeXporter[75], proposed an iterative rule-based pattern matching approach. The system is composed of two modules: (i) a text mining pipeline based on the GATE framework to extract structural and semantic entities. It leverages existing NER-based text mining tools to extract both structural and semantic elements, employing post-processing heuristics to detect or correct the authors affiliations in a fuzzy manner, and (ii) a LOD exporter, to translate the document annotations into RDF according to custom rules.

Solution 2.5 [150] relies on a rule-based and pattern matching approach, implemented in Python. Some external services are employed for improving the quality of the results (for instance, DBLP for validating author's data), as well as regular expressions, NLP methods and heuristics for HTML document style and standard bibliographic description. It also relies on an external tool to extract the plain text from PDFs.

Solution 2.6 [229] extended their framework used for Task 1 (and indicated as Solution 1.3 above) to extract data from PDF as well. Their linear pipeline includes text processing and entity recognition modules. It employs external services for mining PDF articles and heuristics to validate, refine, sanitize

[74] FRED, `http://wit.istc.cnr.it/stlab-tools/fred`

[75] LODeXporter, `http://www.semanticsoftware.info/lodexporter`

and normalize the data. Moreover, linguistic and structural analysis based on chunk-based & sentence-based SVM classifiers are employed, as well as enrichment by linking with external resources such as Bibsonomy, DBpedia Spotlight, DBLP, CrossRef, FundRef & FreeCite.

Solution 2.7 [1] proposed a heuristic-based approach that uses a combination of tag-/rule-based and plain text information extraction techniques combined with generic heuristics and patterns (regular expressions). Their approach identifies patterns and rules from integrated formats.

Solution 2.8 [223] proposed a solution based on a sequential three-level Conditional Random Fields (CRF) supervised learning approach. Their approach follows the same feature list as [141]. However, they extract PDF to an XML that conforms to the NLM JATS DTD, and generate RDF using an XSLT transformation tool dedicated for JATS.

	Sol 2.1	**Sol 2.2**	**Sol 2.3**	**Sol 2.4**	**Sol 2.5**	**Sol 2.6**	**Sol 2.7**	**Sol 2.8**
Extraction								
PDF-to-XML	✓	✓		✓(2016)			✓	✓
PDF-to-HTML					✓			
PDF-to-text			✓	✓(2015)		✓	✓	
Machine Learning								
supervised	✓	✓	✓	✓		✓		✓
unsupervised	✓	✓						
CRF	✓	✓						✓
Text recognition								
NLP/NER		✓	✓	✓	✓	✓		
heuristics	✓	✓	✓	✓	✓	✓	✓	✓
	✓	✓	✓	✓	✓	✓		✓
Evaluation								
best performing	✓(2015)						✓ (2016)	
most innovative		✓ (2016)		✓ (2015)				

Table 4.11: **Task 2 solutions**. The primary analysis methods, their methodologies (i) in general as well as with respect to (ii) extraction, (iii) text recognition and (iv) use of machine learning techniques, and evaluation results.

4.5.3 Lessons learned from the Challenge Organization

In this section we discuss lessons learned from the experience in organizing the challenge. The lessons are grouped in four categories for clarity, even though there is some overlap between them.

1. Lessons learned on defining tasks The definition of the tasks is the most critical part of organizing a challenge.In our case, it was difficult to define appealing tasks that bridge the gap between building up initial datasets and exploring possibilities for innovative semantic publishing.As discussed before, we refined the tasks over the years according to the participants' and organizers' feedback.Overall, we think that tasks could have been improved in some parts – and undeniably other interesting ones could have been defined – but they were successful.There are other less evident issues which are worth discussing.

L1.1. Continuity: allow users to re-submit the improved version of their tool over different editions.One of the goals of the first edition of the challenge was also to explore the interest of the participants. Exploiting such feedback and creating a direct link between different editions is a success key factor. In 2015, in fact, the Challenge was re-organized aiming to commit participants to re-submit overall improved versions of their first year submissions.Results were very good, as the majority of first year's participants competed for the second year too.Continuity is also a key aspect of SemPub2016, whose tasks are the same as last year's edition, allowing participants to reuse their tools to adapt to the new call after some tuning.

L1.2. Split tasks with a clear distinction of the competencies required to complete them.One of the main problems we faced was that some tasks were too difficult. In particular the Task 2 – extraction from XML and PDF – showed unexpectedly low performance.The main reason, in our opinion, is that the task was actually composed of two sub-tasks that required very different tools and technologies: some queries required participants to basically map data from XML/PDF to RDF, while the others required additional processing on the content. Some people were discouraged to participate as they only felt competitive for the one and not for the other. Our initial goal was to explore a larger amount of information and to give participants more options but, in retrospect, such heterogeneity was a limitation. A sharper distinction between tasks would have been more appropriate. In particular, it is important to separate tasks on plain data extraction from those on natural language processing and semantic analysis.

L1.3. Involve participants in advance in the task definition. Though we collected some feedback when designing the tasks, we noticed that such preliminary phase was not given enough relevance. The participants' early feedback can help to identify practical needs of researchers and to shape tasks. Talking with participants, in fact, we envisioned alternative tasks, such as finding high-profile venues for publishing a work, summarizing publications, or helping early career researchers to find relevant papers. Proposing tasks emerged from the community can be a winning incentive to participate.

2. Lessons learned on building input datasets The continuity between tasks (L1.1) can be applied to the datasets as well:

L2.1. Use the same data source for multiple editions. We noticed benefits of using the same data sources across multiple editions of the Challenge.From the task 1 of the 2014 edition, in fact, we obtained an RDF dataset that served as the foundation to build the same task in 2015 and 2016.Participants were able to reuse their existing tools and to extend the previously-created knowledge-bases with limited effort. For the other tasks, which were not equally stable, we had to rebuild the competition every year without being able to exploit the past experience.

L2.2. Designing all three tasks around the same dataset is valuable. First of all, for the participants: they could extend their existing tools to compete for different tasks, with a quite limited effort. This also opens new perspectives for future collaboration: participants' work could be extended and integrated in a shared effort for producing useful data. It is also worth highlighting the importance of such uniformity for the organizers. It reduces the time needed to prepare and validate data, as well as the risk of errors and imperfections. Last but not least, it enables designing interconnected tasks and producing richer output.

L2.3. Provide an exhaustive description of the expected output on the training dataset. An aspect that we underestimated in the first editions of the Challenge was the description of the training dataset. While we completely listed all papers we did not provide enough information on the expected output: we went into details for the most relevant and critical examples but we did not provide the exact expected output for all papers in the training dataset. Such information should instead be provided as it impacts directly the quality of the submissions and help participants to refine their tools.

3. Lessons learned on evaluating results All three editions of the Challenge shared the same evaluation procedure.The workflow presented some weaknesses, especially in the first two years, which we subsequently addressed.Three main guidelines can be derived from these issues.

L3.1. Consider all papers in the final evaluation. Even though we asked participants to run their tools on the whole evaluation dataset, we considered only some exemplary papers for the final evaluation.These papers have been randomly selected from clusters representing different cases, which participants were required to address.Since these papers were representative of these cases we received a fair indication of the capabilities of each tool.On the other hand, some participants were penalized as their tool could have worked well on other values, which were not taken into account for the evaluation.In the third edition, we will radically increase the coverage of the evaluation queries and their number in order to assure that greatest part of the dataset (or the whole dataset) is covered.

L3.2. Make evaluation tool available during the training phase. The evaluation was totally transparent and all participants received detailed feedback about their scores, together with links to the open source tool used for the final evaluation.However we were able to release the tool only after the Challenge.It is instead more helpful to make it available during the training phase, as participants can refine their tool and improve the overall quality of the output.Such an approach reduces the (negative) impact of output imperfections.Though the content under evaluation was normalized and minor differences were not considered as errors, some imperfections were not expected and were not handled in advance.Some participants, for instance, produced CSV files with columns in a different order or with minor differences in the IRI structure.These all could have been avoided if participants received feedback during the training phase, with the evaluation tool available as a downloadable stand-alone application or as a service.

L3.3. Use disjoint training and evaluation datasets. A 2015 participant raised the issue that we underestimated when designing the evaluation process: the evaluation dataset was a superset of the training one.This resulted in some over-training of the tools, and caused imbalance in the evaluation. It is more appropriate to use completely disjoint datasets, a solution we are implementing for the last edition.

	Solution 1.1		**Solution 1.2**		**Solution 1.3**		**Solution 1.4**
year	**2014**	**2015**	**2014**	**2015**	**2014**	**2015**	**2015**
dataset size	1.5M	25M	1.7M	7.2M	2.7M	9.1M	9.7M
# triples	32,088	177,752	14,178	58,858	60,130	62,231	79,444
# entities	4,770	11,428	1,258	11,803	9,691	11,656	19,090
# properties	60	46	43	23	45	48	23
# classes	8	30	5	10	10	19	6

Table 4.12: **Statistics**. about the produced dataset (Task 1 – 2014 and 2015 editions)

4. Lessons learned on expected output and organizational aspects Further suggestions can also be derived from the Challenge's organizational aspects, in particular regarding the expected outcome:

L4.1. Define clearly the license of produced output.Some attention should be given to the licensing of the output produced by the participants.We did not explicitly say which license they should use: we just required them to use an open license on data (at least permissive as the source of data) and we encouraged open-source licenses on the tools (but not mandatory). Most of the participants did not declare which exact license applies to their data.This is an obstacle for the reusability: especially when data come from heterogeneous sources and are heterogeneous in content and format, as in the case of CEUR-WS papers, it is very important to provide an explicit representation of the licensing information.

L4.2. Define clearly how the output of the challenge will be used.The previous observation can be generalized into a wider guideline about reusability.It is in fact critical to state how the results of the challenge will be eventually used, in order to encourage and motivate participants.The basic idea of the Challenge was to identify the best performing tool on a limited number of papers and to use the winning tool (or a refined version) to extract the same data on the whole CEUR-WS corpus. The production of the CEUR-WS Linked Open Dataset was actually slower than expected and we are finalizing it in these days.This is a critical issue: participants' work should not target the challenge only, but it should produce an output that is directly reusable by the community.

L4.3. Study conflicts and synergies with other events. The last guideline is not surprising and was confirmed by our experience as well.In 2015, in fact, we introduced a task on interlinking.The community has been studying interlinking for many years and a lot of research groups could have participated in the task (and produced very good results).However we did not receive enough submissions. One of the issues not the only one, communication might be another – is the conflict with events like OAEI (Ontology Alignment Evaluation Initiative). Even though Task 3 of SemPub2015 did not intend to cover the specialized scope of OAEI, but rather put the interlinking task in a certain use case scope that merely serves in aligning the tasks output among each other and with the rest LOD cloud.The study of overlapping and similar events should always be kept in mind.Not only to identify potential conflicts but also to generate interest: the fact that the SePublica workshop was at ESWC 2014, for instance, was positive since we had fruitful discussions with the participants.

4.5.4 Lessons Learned from Submitted Solutions

In this section we discuss lessons learned from the participants' solution. We start with an overview of the solutions; next, we group the lessons into four categories: lessons on submitted tools, used ontologies, submitted data and evaluation process; even though there is some overlap between these aspects.

	Sol 2.1	**Sol 2.2**		**Sol 2.3**	**Sol 2.4**	**Sol 2.5**	**Sol 2.6**	**Sol 2.7**	**Sol 2.8**
year	2015	2015	2016	2016	2015	2015	2015	2016	2016
dataset size	2.6M	1.5M	285	184K	3.6M	2.4M	17M	152	235
# triples	21,681	10,730	2,143	1,628	15,242	12,375	98,961	1,126	1,816
# entities	4,581	1,300	334	257	3,249	2,978	19,487	659	829
# properties	12	23	23	15	19	21	36	571	23

Table 4.13: **Statistics**. about the produced dataset (Task 2 – 2015 and 2016 editions)

Lessons learned from the tools **L5.1. There are both generic and ad hoc solutions**. All solutions were methodologically different among each other.For Task 1, for instance, two solutions (1.1 and 1.3) primarily consisted of a tool developed specific to this task, whereas the other two solutions wrote task-specific templates in the otherwise generic implementations (adaptive to other domains).In the later case, Solution 1.2 abstracted the case-specific aspects from the implementation, whereas Solution 1.4 kept them inline with the implementation.It becomes, therefore, clear that there are alternative approaches which can be used to produce RDF datasets.

L5.2. There are HTML code and content-based approaches to information extraction. Even though solutions were methodologically different, two main approaches for dealing with the HTML pages prevailed: HTML-code-based and content-based.

Lessons learned from models and ontologies **L6.1. All solutions used almost the same data model (Task 1)**. All solutions of Task 1 tend to converge regarding the model of the data.The same occurs but on a higher level in the case of Task 2.In particular for Task 1, Solution 1.4 domain modeling was inspired by the model used in Solution 1.1, with some simplifications. Note also that Solution 1.2 was the winner solution in 2014.Based on the aforementioned, we observe a trend of converging regarding the model the CEUR-WS data set should have, as most of the solutions converge on the main identified concepts in the data (Conference, Workshop, Proceedings, Paper and Person).

L6.2. All solutions used almost the same vocabularies for the same data (Task 1). There is a wide range of vocabularies and ontologies that can be used to annotate scholarly data. However, most of the solutions preferred to (re)use almost the same existing ontologies and vocabularies (see table6 for Task 1). This is a good evidence that the spirit of vocabulary reuse gains traction. However, it is interesting that different solutions used the same ontologies to annotate the same data differently.

CHAPTER 5

Publishing Linked Open Scholarly Metadata

Scholarly metadata on the Web have been published by different sources and data providers. In addition to the huge volume, such datasets are represented in various formats in terms of data type and schema. Therefore, the Web contains heterogeneous and disconnected scholarly datasets as well as the other domains. It is required to have homogeneous data in order to integrate with other sources and use semantic-based technologies. Therefore, immediately after (or simultaneously) the data acquisition from different resources a (semi-)automated procedure for data transformation is needed. As the aim of this research is to build services based on semantic technologies, the uniform data type considered here is the Resource Description Framework (RDF) (explained in subsection 3.4.1) which is a W3C standard language that organizes data into a set of triples. Having data in this format enables the interlinking with other datasets. This chapter addresses these three steps of the metadata life cyclesection 3.4: *Exraction*, *Transformation* and *Interlinking*. The following sections of this chapter are based on the research contributions related to these two steps that have been previously published as research articles [1].

Several data gathering methods have been implemented to mature OpenResaerch.org (mainly introduced in chapter 6). One of the main resources that has been used to gather event-related metadata has been the emails of calls for papers distributed through certain mailing lists. The corresponding work is introduced in the following publication that will be explained in 5.1 the data gathering section of this chapter. As explained in 1, this has been a teamwork led by the author. Rebaz Omar have done the modeling of metadata distributed in mailing lists and implemented the proposed approach by the author namely SAANSET. The integration of SAANSET with OpenResearch.org ontology and the collected data was also the contribution of the author.

> Rebaz Omar, **Sahar Vahdati**, Christoph Lange, Maria-Esther Vidal and, Andreas Behrend, *SAANSET: Semi-Automated Acquisition of Scholarly Metadata using OpenResearch.org Platform*, ICSC 2018.

Section 5.2 describes the work done for transformation of different metadata formats to machine-readable RDF.

[1] **Own Manuscript Contributions:** The author contributed to the conception and design of the research work, transformation of CSV to RDF and comparison of the results and finally making the RDFization based on the winner approach, CSV to RDF in the context of the OpenAIRE project. The work co-authored by Alexiou and et al. is a join work of the OpenAIRE LOD team. The University of Bonn is coordinating the effort of publishing the OpenAIRE data as Linked Open Data (LOD) and the effort is further supported by the Athena Research and Innovation Center and CNR-ISTI (Alexiou and Papastefanatos). Vahdati has mainly contributed in generating, constructing and improving the interlinking patterns with assessed links and Alexiou aligned it to the OA infrastructure. The work by Ameri et al. has been a master thesis mainly supervised by Vahdati. In both articles, the author had main role in drafting and final approval of the published versions.

- **Sahar Vahdati**, Farah Karim, Jyun-Yao Huang, Christoph Lange. *Mapping Large Scale Research Metadata to Linked Data: A Performance Comparison of HBase, CSV and XML* In Metadata and Semantics Research Conference 2015..

Representation of data in such formats makes it interoperable and easily reusable. More details will be discussed in section 5.3. In the following publications the we discuss the design and implementation of scholarly metadata interlinking.

- Giorgos Alexiou, **Sahar Vahdati**, Christoph Lange, George Papastefanatos, Steffen Lohmann. *LOD services: Scholarly Communication Data as Linked Data*, SAVE-SD Workshop of WWW2016, LNCS post-proceedings 2016;
- Shirin Ameri, **Sahar Vahdati**, Christoph Lange. *Interlinking OpenAIRE LOD and related Datasets*, Theory and Practice of Digital Libraries 2017.

5.1 Extraction

In our era open access to scientific literature has become widespread. The overall process of scientific communication, e.g., preparation of manuscripts, organization of conferences, and a peer review process have become considerably efficient. This results in an enormous amount of research output and information about research activities. Researchers spend a lot of time in finding information about other researchers, scientific events, journals, scientific papers and research topics related to their interest. Although there exist a lot of services, such as data and content repositories, digital libraries or metadata catalogues to assist researchers, it is often a time-consuming task to find information such as:

- Which scientific events covering topic X and including a PhD Consortium will be held near location Y during the next Z months? (a community calendar)
- Where does the next event of an event series X take place?
- Which countries have the research groups that have been most active in organizing events (considering roles in events, e.g., PC membership) over a period of X years?
- What upcoming events on topic X have a high networking potential in terms of interesting participants (e.g., keynote speakers) and its schedule (e.g., social events)?

Mailing lists are used as a popular way [250] of exchanging announcements or spreading discussions easily among researchers. They form one of the most reliable sources of information about upcoming events because of the large coverage of events by Calls for Papers (CfPs) disseminated in those mailing lists. The principal reasons for using email as a scientific communication channel are the known target group, speed and immediacy it offers. However, the sheer amount of emails sent through those mailing lists makes it difficult for one individual to keep track of them.

Although data from mailing lists is a reliable source of information about upcoming events, it is hard for one individual to extract specific information from them. To obtain the information they are interested in, subscribers are required to first filter a huge amount of emails by relevance, and then, in the worst case, read the full text of the relevant ones. In this section, we present a semi-automatic approach for relevance filtering and metadata extraction from CfPs and expose the extracted data in a useful way in the OR information portal.

Motivating Example We motivate the problem of filtering and extracting metadata about scientific events from CfP emails of mailing lists with the following scenario. Our focus is on mailing lists, i.e., a communication medium often used by research communities as a specific channel for distributing, e.g., announcements of releases of software packages or datasets, CfPs of upcoming scientific events, and research related opinions and questions. Active Researchers receive a vast amount of emails about conferences and scientific progress every day. Subscribing to such mailing lists increases the enormous number of announcements every day. Suppose a researcher who has subscribed to such a mailing list needs to identify upcoming related scientific events. Figure **??** depicts a pipeline that can be followed to achieve this goal using mailing lists. The upper part of the figure shows researchers in the role of an event organizer, who are concerned with preparing CfPs and are seeking ways and channels to distribute them to the relevant communities. A researcher in our scenario has to trace the emails on a list and to decide which ones to have a closer look into. Although this process looks straightforward and is one of the favorite communication channels for researchers, a lot of relevant information might either be overlooked or overwhelm recipients. We therefore present SAANSET (Semi-Automated AcquisitioN of Scholarly mETadata), a method to support researchers with these tasks; the proposed method is not only able to filter emails but is also able to capture knowledge encoded in CfP emails and to represent this metadata as structured data in OR for further reuse.

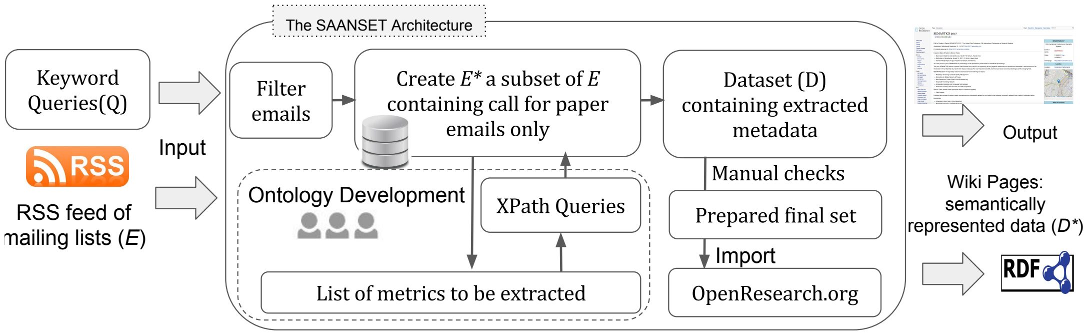

The SAANSET Architecture
Keyword Queries(Q)
RSS
RSS feed of mailing lists (E)
Input
Filter emails
Create E* a subset of E containing call for paper emails only
Dataset (D) containing extracted metadata
Manual checks
Prepared final set
Import
OpenResearch.org
Ontology Development
XPath Queries
List of metrics to be extracted
Output
Wiki Pages: semantically represented data (D*)
RDF

Figure 5.1: **The Architecture of SAANSET.** SAANSET receives as input a set E of emails and a keyword query Q and outputs an RDF (Resource Description Framework) dataset D^*. A keyword query Q is used to select a set E^* of relevant emails containing CfPs. The RDF dataset D^* is composed of the RDF triples that describe the scientific events described in E^*.

5.1.1 Semi-Automated Acquisition of Scholarly Metadata

This section focuses on collecting data from mailing lists that expose their archives in an accessible way via RSS feeds. As mentioned initially, announcing CfPs through different mailing lists is a traditional but still the most popular way of disseminating information about scientific events. It is one of the main and still most reliable sources for different research communities to share information about upcoming relevant events. To make better use of the critical mass of information being transferred through mailing lists, we aim at adding it to the research knowledge graph underlying our OpenResearch.org (OR) platform (see chapter 6).

Metadata of scientific events, including their research community, event name, topics covered, the date and location of the event, deadlines and list of organizers are knowledge encoded in CfP emails. Given the set E of all emails in the archive of a mailing list and keyword query Q, the problem of *capturing scholarly* metadata according to Q corresponds to:

- Finding a subset of emails E^* that contains only CfP emails and *satisfy* the keywords in Q, and
- Extracting relevant knowledge from these CfPs.

The emails in this subset E^* contain unstructured scholarly event metadata. The problem of *extracting* relevant knowledge from the email contents and *transforming* them to structured data requires a second set Q^* of queries. Whereas query Q is used for filtering purposes, query Q^* extracts predefined knowledge from the filtered emails. From the unstructured content of CfP emails, relevant data D is extracted and represented as structured knowledge. Importing it into OR and thus giving it a semantic structure finally results in an RDF dataset D^*.

Mailing Lists As a proof of concept, out of 25 active mailing lists of the Semantic Web and Database communities, we have selected three that receive more than 40 messages per month. Table **??** shows the features of the selected mailing lists of the three research communities of Semantic Web (SW), Databases (DB) and Information Retrieval (IR).

Each record comprises event title, deadline, event homepage, and the full-text description. The *Semantic Web*[2] (SemWeb) mailing list addresses semantic web related topics; the *Linked Open Data*[3] (LOD) list covers a specific sub-field of the Semantic Web. The *Database World*[4] (DBworld) mailing list covers research in the broader field of Databases. *Information Retrieval Specialist Group*[5] (IRSG) contains information about events in the field of information retrieval.

The SAANSET Framework To extract data from CfPs, both full texts of and email-level metadata from the RSS feeds are extracted.[6] RSS feeds [276] provide queryable structured data, thus minimizing the time needed for the data collection. Filtering and subsequent metadata extraction are performed as follows: 1) filtering for emails that contain CfPs, and then 2) extracting information about events from these emails.

We propose SAANSET, a framework for capturing scholarly metadata and importing it into OR. Figure 5.1 shows the components of our proposed approach, while algorithm **??** sketches a semi-automated method implemented by SAANSET. Manual steps are marked with an m while automatic steps are labeled with an a.

Algorithm 1 Extraction. Extracting information algorithm

```
Import the RSS feed of the mailing list as an XML file (a)
Create a new spreadsheet (a)
Extract the given information of all emails from the RSS feed (a)
for each email do
    Check if the email is a CfP email (a)
    if email is a CfP email then
        Add the extracted information to the spreadsheet (a)
Return the spreadsheet (a)
Complete the extracted information (m)
Import the file to OR (m)
```

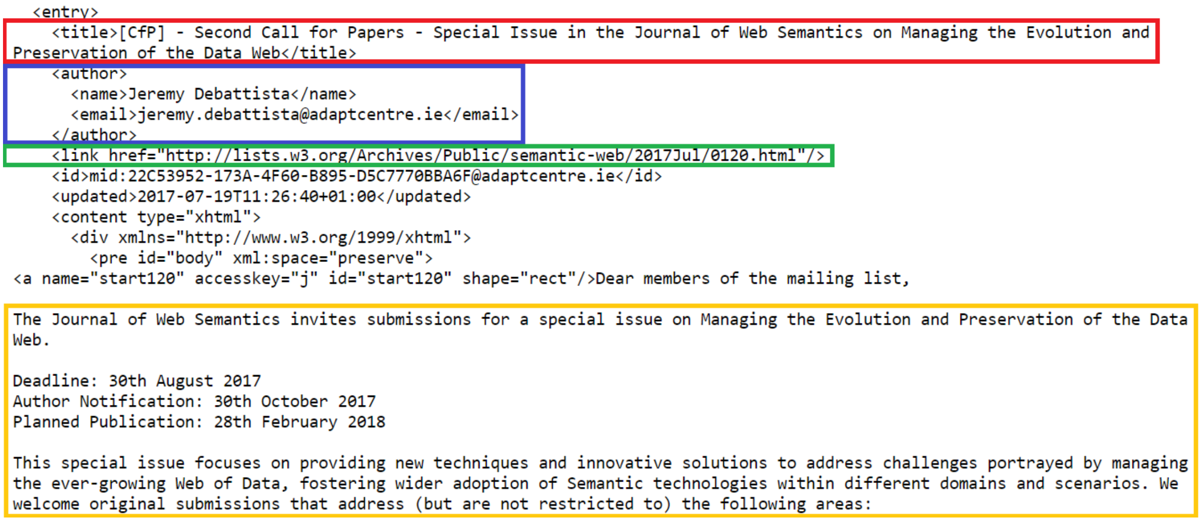
[2] https://lists.w3.org/Archives/Public/semantic-web/

[3] https://lists.w3.org/Archives/Public/public-lod/

[4] https://research.cs.wisc.edu/dbworld/register.html

[5] https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=IR

[6] See, e.g., https://lists.w3.org/Archives/Public/semantic-web/feed.rss for the SemWeb and https://lists.w3.org/Archives/Public/public-lod/feed.rss for the LOD mailing list.



```
  <entry>
    <title>[CfP] - Second Call for Papers - Special Issue in the Journal of Web Semantics on Managing the Evolution and
Preservation of the Data Web</title>
    <author>
      <name>Jeremy Debattista</name>
      <email>jeremy.debattista@adaptcentre.ie</email>
    </author>
    <link href="http://lists.w3.org/Archives/Public/semantic-web/2017Jul/0120.html"/>
    <id>mid:22C53952-173A-4F60-B895-D5C7770BBA6F@adaptcentre.ie</id>
    <updated>2017-07-19T11:26:40+01:00</updated>
    <content type="xhtml">
      <div xmlns="http://www.w3.org/1999/xhtml">
        <pre id="body" xml:space="preserve">
<a name="start120" accesskey="j" id="start120" shape="rect"/>Dear members of the mailing list,

The Journal of Web Semantics invites submissions for a special issue on Managing the Evolution and Preservation of the Data
Web.

Deadline: 30th August 2017
Author Notification: 30th October 2017
Planned Publication: 28th February 2018

This special issue focuses on providing new techniques and innovative solutions to address challenges portrayed by managing
the ever-growing Web of Data, fostering wider adoption of Semantic technologies within different domains and scenarios. We
welcome original submissions that address (but are not restricted to) the following areas:
```

Figure 5.2: **Example of a CfP**. The XML structure of an RSS feed.

Processing the RSS Feed RSS feeds are XML documents. A typical RSS feed of a mailing list only contains data from mails from the beginning of the current month until the time of processing. For each month an RSS is stored in the list archive. The advantage of using XML as the input for our program is its uniform structure. Thus, the specific objective is to analyze the structure and identify those elements in the XML input that can be used for extracting relevant information. We can then use XPath [18] expressions to access these elements. Figure 5.2 shows an excerpt from the RSS feed of the SemWeb list.

The subject of the email, information about the author, the link to the email and the content of the email can be found in the XML file. Any of these pieces of information can be addressed in a unique way. Thus, straightforward XPath expressions can be used to address and extract this information. The content of the email is not accessible via a single, unique element. Therefore another approach has to be followed to extract the content of the email. Since the actually relevant content of the emails (in our example: the text following the salutation) given in the XML document cannot generally be accessed via a single element, and any XPath expression employed to address it might also refer to other, irrelevant parts of the email, it is a difficult task to use XPath to identify and extract the content. Further, there is no way to specify which content belongs to which email because there is no link. To solve this problem, we will use the following approach:

1. Find the unique link to the full text of each email in the XML document using XPath.
2. Access the link and get the full text of each email.
3. Apply another XPath query to the full text of the email to extract its content.

By using this approach we can iterate over all emails and get the content and specify to which email it belongs because we are using the unique link of the email to access it. Using XPath we can identify the content and access it via the following query

```
/html/body/div[@class='mail']/pre[@id='body']
```

Listing 5.1: **XPath.** XPath to access the email body

As shown in table 5.2, several emails are not about CfPs, and thus irrelevant for our use case, as these emails can be assumed not to provide information about events. Accessing the link of every email and loading the content would be unnecessary and lead to a *poor performance* of SAANSET. Therefore our goal is to first identify all the CfP emails and iterate through their links using the query presented above to get the email body.

Month (2017)	Total	CFPs	Announcements	Discussions
March	96	38	16	10
April	78	16	2	37
May	138	34	10	71

Table 5.2: **Email Distribution.** Total number of emails and distribution of types of emails on the semantic web mailing list.

Filtering out Irrelevant Emails In this section, we present a way of identifying irrelevant emails to be able to filter them out before carrying out further analysis steps. SAANSET uses characteristic features of CfP emails to identify these as relevant, rather than explicitly identify irrelevant emails, as there are many diverse types of irrelevant ones (e.g., discussions or release announcements). It is clear that all CfP emails have characteristic subjects; however, once more, there are several variants, such as “CFP”, “Call for Papers” and “Call for Paper”. We also have to consider different capitalizations.

Table 5.3 presents the distribution of these labels over five months. It shows: a) what labels the authors of CfP emails use in their subjects and b) how we can search for them. Besides those emails that are not CfPs, we also have to filter out responses to those that are (e.g., someone pointing out a mistake or asking a question regarding that event). These response emails contain the same information as the original CfP email, i.e., these emails are redundant.

Month (2017)	CfP emails	CFP	CfP	Cfp	Call for Papers	Call for papers
March	38	13	16	0	8	1
April	16	6	8	0	0	2
May	34	8	17	0	4	5
June	35	12	17	0	6	0
July	46	11	32	0	3	0

Table 5.3: **CfP Description.** Distribution of different labels in the subject of the emails in the semantic web mailing list.

Although we want to filter as much as possible to obtain only relevant emails, we do not want to eliminate duplicates of CfP emails with our implementation. There are two reasons behind this decision: Second or third CfP emails for the same event are frequently sent to mailing lists. Most of the time, subsequent emails contains changes of the important dates e.g., submission deadline of the event. If we filtered out such emails, we would miss such changes. When importing the generated spreadsheet containing information about events to OR via its CSV import interface[7], Semantic

[7] http://openresearch.org/Special:ImportCSV

MediaWiki (the software on which OR is built) provides the option to deal with duplicates. We can choose to ignore them: if an event already exists in OR and we want to import it again nothing will happen. We can also choose the option to add our information to existing content. Using this we can add information such as an updated date to existing events. Given the existing data about an event e, represented as a set of triples $(e, p_1, o_1), (e, p_2, o_2), (e, p_3, o_3) \ldots$, the information from the 2nd CfP looks like $(e, p_1, o_1), (e, p_2, o_2'), (e, p_4, o_4), \ldots$, i.e., some information (here p_2, for example the submission deadline) might change, other information might disappear (here p_3, for example the name of a PC member); finally, some information might be entirely new (here p_4, for example, keynote speakers). With an import as explained above, the triple $(e, p_1, o_1), (e, p_2, o_2'), (e, p_3, o_3), (e, p_4, o_4), \ldots$ would be kept after importing the 2nd CfP.

Saving the Extracted Information Data extracted by SAANSET is temporarily stored in a database; eventually, SAANSET generates RDF triples describing all the data from this temporal database. This database is composed of tables only containing information about CfP emails. For each such email the following information is inserted into the table with the attributes: email title, author name, author email, link to the full text of the email and the full text of the email.

Because of differences in email templates, in some cases manual checks are required to complete the metadata extracted from emails, as mentioned in section 5.1.1. After running the automated part SAANSET, the user currently has to manually complete the metadata, arriving at the overall pieces of information about an event: title, date (start and end date), location, field, event type, and homepage. This metadata is also repeated in the unstructured text of the content. As of this step, the dataset (D) is ready to be imported into OpenResearch.org (OR) to create the corresponding event wiki pages. OR uses the semantic extension of MediaWiki, which supports automatic transformation of data into RDF. Currently OR produces downloadable RDF data[8] weekly, and an active SPARQL endpoint enables users to run complex queries[9].

5.1.2 Implementation

In this section, the implementation of SAANSET is described.

Extracting Information from the RSS Feed First of all, we have to load the RSS feed into the script and come up with an approach to access all entries and retrieve the information from these. This will be done by the following code:

```
xml = loadXML("https://lists.w3.org/.../semantic-web/feed.rss");
XML[] children = xml.getChildren("entry");
```

Listing 5.2: **Information Loading.** Loading the RSS feed into the script and setting up access to the entries

After loading the RSS feed we need to go over all entries and extract the information discussed in section 5.1.1.

```
for (int i = 0; i < children.length; i++) {
    XML[] titles = xml.getChildren("entry/title"); title = titles[i].getContent();
    XML[] links = xml.getChildren("entry/link"); link = links[i].getString("href")
    ;}
```

Listing 5.3: **Information Acquisition** Getting the information from the RSS feed

[8] http://openresearch.org/Special:ExportRDF

[9] http://openresearch.org/Sparql_endpoint/Examples

Extracted data is stored in arrays; the first entry of the array *authormails* is the email address of the person in the first position in the array *authors*.

Filtering out Irrelevant emails In the next step, we want to filter out emails that are not CfP emails. We achieve that by checking the subjects of the emails containing keywords from our dictionary, i.e., the *titles* of the RSS feed entries, for the CfP labels discussed in the section 5.1.1.

```
    String[] m = match(title, "CfP");
    if (m != null ) {
      hasCfP = true;
    } else
      hasCfP = false;
```

Listing 5.4: **Information Filtering** Checking if the subject of the email contains CfP labels

In this example, we are checking if the subject contains the string "CfP". We use the same method and test if the subject contains other CfP labels. If it contains any such CfP label, then the email is a CfP. Otherwise we consider it irrelevant.

Writing the Extracted Information To obtain the content of the CfP emails, we access the link of each email. To have access to each link of the corresponding email, we store the information we currently have (title, author name, author email, link) in a table. By doing so we can use a simple loop going through the link column and save the content in another column. To get the content we have to find a XPath query we can run. To build this query we analyzed the XML structure and found the query *html/body/div[@class='mail']/pre[@id='body']* to work. It is used to access the email body and extract it. Thus, we end up with a table holding all the extracted information.

```
x = table.getRowCount();
for (int i = 0; i<x; i++) {
  // Access each row of the table one at a time, in a loop.
  TableRow row = table.getRow(i);
  n = row.getString("link");
  xml2 = loadXML(n);
  String query = "html/body/div[@class='mail']/pre[@id='body']";
  String content s= xml2.getContent(query);
  table.setString(i, "content", contents);
}

saveTable(table, "data/new.csv");
```

Listing 5.5: **Information Extractions** Extracting the content of each CfP email

5.1.3 Evaluation

To assess the usability of our semi-automatic approach, we compare it to a fully manual metadata extraction and import workflow. We present an approximate metric to measure the manual effort required for importing event metadata using our approach vs. the fully manual one. The following algorithm shows the steps required when performing this task fully manually:

We assume that the user has the level of experience required to carry out these steps. To approximately measure the user's effort, we count each atomic action the user has to perform manually as one unit To determine the difference between manually gathering information and using SAANSET over a fixed

Algorithm 2 Extraction. Steps for Manual Extraction and Import of Information

```
Prepare a CSV file with the following columns: event title, event date, city, country, field, event type,
  event homepage (each information counts as one unit in our measurement)
for each email do
    Check (manually) if the email is a CfP email
    Open the relevant email
    Read the content and write down the following information in the CSV file: event title, event
      date, city, country, field,event type, event homepage (each information counts as one unit in our
      measurement)
Import the CSV file into OpenResearch.org
```

period of time, we calculate the amount of required units. Thus, we will finally be able to tell the improvement that SAANSET provides over the fully manual workflow.

Showcase: Semantic Web Emails over Three Months We applied the evaluation methodology introduced previously to the task of importing event metadata from the posts of the semantic web mailing list from March to May 2017. Table 5.2 shows 312 emails over that period, 88 of them being CfPs.

The effort to perform all required actions – let their number be n – is computed as follows:

$$overallEffort(a) = \sum_{i=0}^{n} effortPerAction(i)$$

The following enumeration shows the effort required for each manual action of performing the overall task of extracting event metadata from the given 312 emails. The effort is given in units (u):

1. $1u$: create a CSV file
2. $312u$: checking each email if it is a CfP
3. $88u$: open each CfP email
4. $616u$: from each CfP email, extract 7 pieces of information
5. $1u$: import the CSV file into OR
6. Thus, the total effort of gathering and importing information amounts to $1018u$.

Using SAANSET, the first two steps from this list, i.e., creating a CSV file and checking each email whether it is a CfP or not, are automated, i.e., the remaining manual effort required amounts to $88u + 616u + 1u = 705u$. This means in our particular example the user needs to perform $\sim$31% less manual actions for the complete process when using SAANSET.

Formalization and Generalization From the previous calculation, the benefit gained by using SAANSET in the scenario explained above can be quantified as 31%, which serves as a reference point. For a general benefit analysis, we define a formalization as follows. Let N be the number of total emails received through one or more mailing lists and N_C the amount of CfP emails among them. Following the same steps as for the specific scenario above, we arrive at the following expression for the effort of manually gathering information:

$$2u + Nu + 8N_C u \tag{5.1}$$

$2u$ corresponds to the initial, one-time actions of creating a CSV file and importing the metadata to OR. For each of the N emails, the user has to check whether it is a CfP or not (effort Nu). For each of the

N_C CfP emails, the user has to open it and has to extract 7 pieces of information from it to the CSV file (event title, event date, city, country, field, event type, event homepage), resulting in an effort of $8N_Cu$.

Assuming a large number of emails, we can neglect the constant summand $2u$, resulting in an approximate effort of $Nu + 8N_Cu$. Using SAANSET, which automates the checking of whether or not an email is a CfP, the user's remaining manual effort reduces to $8N_Cu$. The ratio of the effort with SAANSET vs. the all-manual workflow is therefore approximated by the following expression:

$$\frac{8N_C}{N + 8N_C} \tag{5.2}$$

Equivalently, the following *benefit function* answers the user's question of what percentage of his or her effort will be saved thanks to SAANSET:

$$savedEffort = (1 - (\frac{8 \cdot N_C}{N + 8 \cdot N_C})) \cdot 100 \tag{5.3}$$

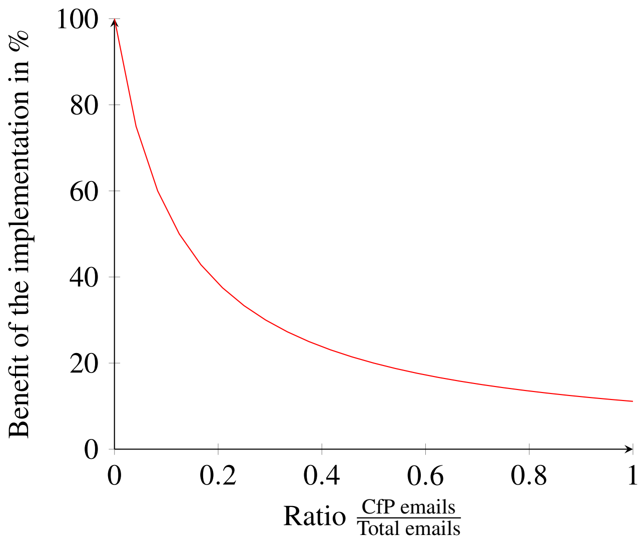


Figure 5.3: **Effectiveness of SAANSET.** The behavior of SAANSET is reported in terms of the *benefit* function; it is computed in terms of the ratio between the number of CfP emails to total number of emails in a mailing list.

Figure 5.3 visualizes the benefit function (5.3). We can clearly see that SAANSET provides greater benefit the lower the ratio of CfP emails is, or, in other words, the higher the ratio of irrelevant emails is. If no CfP emails are sent via a mailing list, SAANSET eliminates all mails and there would be no need for manual user actions. If all emails were CfPs, SAANSET would automate $\frac{1}{9}$ of the work, leaving the manual extraction of information to the user.

Adaptation to the specific structure of a given mailing list is as easy as changing one XPath expression, as shown by the placeholder `XPATH_QUERY` in the following listing:

```
// get some information from the XML
for (int i = 0; i < children.length; i++) {
    XML[] informations = xml.getChildren(XPATH_QUERY);
    information = informations[i].getContent();}
```

Listing 5.6: **Template.** Template for extracting any information using XPath

Adaptability is another feature of SAANSET; it is able to adjust its behavior to the structure of the RSS feed by allowing expert users to define *generic templates* for knowledge extraction. Flexibility is another advantage of SAANSET. A user can choose the time interval in which she runs the script. Daily execution typically yields the latest CfPs, while weekly execution results in importing the metadata of several events at once.

5.2 Transformation

The diverse data formats and means to access or query the already existing scholarly metadata, the use of duplicate identifiers, and the heterogeneity of metadata schemas pose practical limitations on reuse. As discussed before, Linked Data, based on the RDF graph data model, is by now increasingly accepted as a lingua franca to overcome such barriers. There is a recognized gap of a comprehensive metadata management for all kinds of research outputs, artifacts, events across disciplines and countries. The goal is to explore possibilities with the already existing tools to build a big scholarly knowledge graph including metadata about OCW, Events and Papers. This can increase technical interoperability of the offered services, engagement with additional user communities, explore synergies with and evaluate the added values to related initiatives. Mapping large-scale research metadata to linked data requires different data models to come to an agreement. We gather information from different sources. Such data are available in structured, semi-structured and unstructured formats.

The data formats in each of these categories are diverse for example data are available in Tab-Separated Values (TSV) , Comma-Separated Values (CSV), tuple (relational tables), Extensible Markup Language (XML) etc. Depending on the basic requirements of an underlying framework and limits of the data format, even a RDF dataset which is available in a certain format is required to go through tranformation step. A dataset can be available in any of RDF syntaxes [10] e.g., RDF/XML, Turtle, Notation 3, nTriples, RDFa and yet required to be transformed to the other. The transformation step can be general of very specific deepening on the data format of the original artifact. For example, the tool ocw2rdf [11] harvests metadata from the platforms for OpenCourseWare and transforms it into an RDF representation. This tools is not easily adoptable to other formats. Whereas Triplify [12], Sparqlify [13] and Tarql [14] that are originally made for relational to RDF, can easily be adopted to transformation of JSON format.

In the scope of this research, in several occasions there was a need for data transformation. In preparation of the input datasets for the Semantic Publishing Challenge (see section 4.5), we needed to go through the data transformation from relational format to RDF (Turtle syntax). In order to do so, in 2015 both *Triplify* and *Sparqlify* tools have been used. However, with a better performance, *Sparqlify* was the favorite tool to provide the datasets for all the editions of the challenge. In the context of OCW, there was hardly any available dataset in any format. Most of the platforms offering OCW was manually harvested in order to collect metadata (see section 4.3). For OpenResearch.org platform (see chapter 6), the main data acquisition is done with crowdsurcing. Apart from all the above attempts or possibilities for data transformation, this step is mainly done in the context of OpenAIRE project. To provide LOD services and interlink the OpenAIRE data with related data on the Web, the prerequisite was transferring the OA data to Linked Open Data (LOD). Concrete steps towards this vision are (1) mapping the OpenAIRE data model to suitable standard LOD vocabularies, (2) exporting the objects in the OpenAIRE information

[10] https://www.w3.org/wiki/RdfSyntax

[11] http://simile.mit.edu/repository/RDFizers/ocw2rdf/

[12] https://web.archive.org/web/20150208024727/http://triplify.org:80/Overview

[13] https://github.com/SmartDataAnalytics/Sparqlify

[14] https://github.com/tarql/tarql/wiki/TARQL-Mapping-Language

space as a LOD graph and (3) facilitating integration with related LOD graphs. Expected benefits include:

- enabling semantic search over the outputs of European research projects,
- simplifying the way the OA data can be enriched by third-party services, and consumed by interested data or service providers,
- facilitated outreach to related open content and open data initiatives, and
- enriching the OA information space itself by exploiting how third parties will use its LOD graph.

The following sections will focus on transformation of OpenAIRE data into RDF.

5.2.1 Input Data Formats

The specifically tailored nature of the OpenAIRE infrastructure, its large amount of data (covering more than 11 million publications) and the frequent updates of the more than 5000 repositories from which the data is harvested pose high requirements on the technology chosen for mapping the OpenAIRE data to LOD. As explained before (see subsection 2.2.1), OpenAIRE, the Open Access Infrastructure for Research in Europe, comprises a database of all EC FP7 and H2020 funded research projects, including metadata of their results (publications and datasets). These data are stored in an HBase NoSQL database, post-processed, and exposed as HTML for human consumption, and as XML through a web service interface. As an intermediate format to facilitate statistical computations, CSV is generated internally.

We thus faced the challenge of identifying the best performing conversion approach with high maintenance. We evaluated the performances of creating LOD by a MapReduce job on top of HBase, by mapping the intermediate CSV files, and by mapping the XML output. We therefore compared in depth three alternative mapping methods, one for each source format in which the data are available: HBase, CSV and XML. For each possible approach, i.e. mapping HBase, CSV or XML to RDF, we briefly review the state of the art to give an overview of technology we could potentially reuse or build on, whereas We assess reusability w.r.t. the OpenAIRE-specific requirements stated above.

HBase, currently, is the master source of all OpenAIRE data. It is a column store based on HDFS (Hadoop Distributed File System) [245]. HBase was introduced in 2012 when data integration efforts pushed the original PostgreSQL database to its limits: joins became inefficient and parallel processing, as required for deduplication, was not supported. HBase is a sparse, distributed and multidimensional format sorted map, and provides dynamic control over the data format and layout.

```
message Person {
  optional Metadata metadata = 2;
  message Metadata {
    optional StringField firstname = 1;
    repeated StringField secondnames = 2;
    optional Qualifier nationality = 9; ... }
  repeated Person coauthors = 4; }
```

Listing 5.7: **HBase**. An example of OpenAIRE Data stored in HBase. morekeywords

Each row of the HBase table has a unique row key and stores a main entity and a number of related linked entities. The attribute values of the main entities are stored in the *<family>:body* column, where the *<family>* is named after the type of the main entity, e.g., *result*, *person*, *project*, *organization* or *datasource*. The attribute values of linked entities, indicating the relationship between main entities, are stored in dedicated column families *<family>:<column>*, where *<family>* is the class of the linked entity and *<column>* is the row key of the target entity. Both directions of a link are represented. Cell values are serialized as byte arrays according to the Protocol Buffers [217] specification. example:

RowKey	result:	person:	…hasAuthor:		…isAuthorOf:
	body	body	30|…001::9897…	30|…001::ef29…	50|…001::39b9…
50|…001::39b9…	resulttype=“publication”; title=“The Data Model of …”; dateofacceptance=“2012-01-01”; language=“en”; publicationDate=“2012”; publisher=“Springer”;		ranking=1;	ranking=2;	
30|…001::9897…		firstname=“Paolo”; lastname=“Manghi”;			ranking=1;
30|…001::ef29…		firstname=“Nikos”; lastname=“Houssos”;			ranking=2;

Table 5.4: **HBase.** An example of OpenAIRE Data stored in HBase

The Table 5.4 shows a publication and its authors. For readability, we abbreviated row keys and spelled out key-value pairs rather than showing their binary serialization.

Several works have therefore explored the suitability of HBase as a triple store for semi-structured and sparse RDF data. Sun et al. adopted the idea of the Hexastore indexing technique for storing RDF in HBase [257]. Khadilkar et al. focused on a distributed RDF storage framework based on HBase and Jena to gain scalability[138]. Others have provided MapReduce implementations to process SPARQL queries over RDF stored in HBase [107, 204].

We are only aware of one work on exposing data from column-oriented stores as RDF. Kiran et al. provide a method for generating a SPARQL endpoint, i.e. a standardized RDF query interface, on top of HBase [128]. They map tables to classes, rows to resources, and columns to properties. Their approach do not scale well with increasing numbers of HBase entries, as the results show that the time taken to map HBase data to RDF is in hours for a few million rows [128].

CSV (Comma Separated Values), is widely used for publishing tabular data [163]. The CSV format on the Web W3C Working Group[15] provides technologies for data dependent applications on the Web working with CSV. CSV files aid the computation of statistics on the OpenAIRE information space. HBase is a sparse key value-store designed for data with little or no internal relations. Therefore, it is impossible to run complex queries directly on top of HBase, for example a query to find all results of a given

[15] http://www.w3.org/2013/05/lcsv-charter.html

project. It is thus necessary to transform the data to a relational representation, which is comprehensible for statistics tools and enables effective querying. Via an intermediate CSV representation, the data is imported into a relational database, which is queried for computing the statistics. In this generation process, each main entity type (result, project, person, organization, datasource) is mapped to a CSV file of the same name, which is later imported into a relational database table. Each single-valued attribute of any entity (id, title, publication year, etc.) becomes a field in the corresponding table for each entity. Multi-valued attributes, such as the publication languages of a result, are mapped to relation tables (e.g. `result_languages`) that represent a one-to-many relation between entity and attributes. Linked entities, e.g. the authors of a *result*, are represented similarly. As the data itself includes many special characters, for example commas in publication titles, the OpenAIRE CSV files use ! as a delimiter and wrap cell values into leading and trailing hashes:

```
#dedup_wf_001::39b91277f9a2c25b1655436ab996a76b#!#The Data Model of the OpenAIRE
Scientific Communication e-Infrastructure#!#null#!#null#!#Springer#!#null#!#null
#!#null#!#null#!#2012#!#2012-01-01#!#Open Access#!#Open Access#!#Access#!#null#!#
0#!#null#!#nulloai:http://helios-eie.ekt.gr:!#publication#10442/13187oai:pumaoai.
isti.cnr.it:cnr.isti/cnr.isti/2012-A2-040#!#1#!
```

Listing 5.8: **CSV**. An example of OpenAIRE Data stored in CSV.

Finally, using CSV has the advantage that existing tools such as Sqoop can be used, thus reducing the need to develop and maintain customly implemented components on the OpenAIRE production system.

Customizable mappings are more suitable for our purpose. In Tarql (Transformation SPARQL)[16], one can define such mappings in SPARQL; Tabels (Tabular Cells)[17] and Sparqlify[18] use domain-specific languages similar to SPARQL. Tabels provides auxiliary machinery to filter and compare data values during the transformation process. Sparqlify is mainly designed to map relational databases to RDF but also features the sparqlify-csv module.

XML is used for various data and document exchange purposes. OpenAIRE features a set of HTTP APIs[19] for exporting metadata as XML for easy reuse by web services. These APIs use an XML Schema implementation of the OpenAIRE data model called OAF (OpenAIRE Format)[20], where each record represents one entity. There is one API for searching, and one for bulk access. For example, the listing below shows an excerpt of the metadata of a publication that has been searched for.

The API for bulk access uses OAI-PMH (The **O**pen **A**rchives **I**nitiative **P**rotocol for **M**etadata **H**arvesting)[21] to publish metadata and its corresponding endpoint is at `http://api.openaire.eu/oai_pmh`. The bulk access API lets developers fetch the whole XML files step by step. For our experiments, we obtained the XML data directly from the OpenAIRE server, as an uncompressed Hadoop SequenceFile[22] comprising 500 splits of ∼300 MB each.

Like for CSV→RDF, there are generic and domain-specific XML→RDF approaches. Breitling implemented a direct, schema-independent transformation, which retains the XML structure [39].

Turning this generic RDF representation into a domain-specific one requires post-processing on the RDF side, e.g., transformations using SPARQL CONSTRUCT queries. On the other hand, the current version of Breitling's approach is implemented in XSLT 1.0, which does not support streaming and is therefore not suitable for the very large inputs of the OpenAIRE setting. Klein uses RDF Schema to map

[16] `https://tarql.github.io`

[17] `http://idi.fundacionctic.org/tabels`

[18] `https://github.com/AKSW/Sparqlify` [79]

[19] `http://api.openaire.eu/`

[20] `https://www.openaire.eu/schema/0.2/doc/oaf-0.2.html`

[21] `http://www.openarchives.org/OAI/openarchivesprotocol.html`

[22] `http://wiki.apache.org/hadoop/SequenceFile`

XML elements and attributes to RDF classes and properties [143]. It does not automatically interpret the parent-child relation between two XML elements as a property between two resources, but a lot of such relationships exist in the OpenAIRE XML. XSPARQL can transform XML to RDF and back by combining the XQuery and SPARQL query languages to [25]; authoring mappings requires good knowledge of both.

```
<oaf:result>
  <title schemename="dnet:dataCite_title" classname="main title"
   schemeid="dnet:dataCite_title" classid="main title">The Data Model of the
    OpenAIRE Scientific Communication e-Infrastructure</title>
  <dateofacceptance>2012-01-01</dateofacceptance>
  <publisher>Springer</publisher>
  <resulttype schemename="dnet:result_typologies" classname="publication"
   schemeid="dnet:result_typologies" classid="publication"/>
  <language schemename="dnet:languages" classname="English"
   schemeid="dnet:languages" classid="eng"/>
  <format>application/pdf</format>
  ...
</oaf:result>
```

Listing 5.9: **XML**. An example of OpenAIRE Data stored in XML.

By supporting XQuery's expressive mapping constructs, XSPARQL requires access to the whole XML input via its DOM (Document Object Model), which results in heavy memory consumption. A subset of XQuery[23] is suitable for streaming but neither supported by the XSPARQL implementation nor by the free version of the Saxon XQuery processor required to run XSPARQL.

Comparisons of different approaches of mapping data to RDF have mainly been carried out for relational databases as a source [181, 260]. Similarly to our evaluation criteria, the reference comparison framework of the W3C RDB2RDF Incubator Group covers mapping creation, representation and accessibility, and support for data integration [232]. Hert et al. compared different RDB2RDF mapping languages w.r.t. syntactic features and semantic expressiveness [110]. For other linked datasets about research, we refer to the "publication" and "government" sectors of the LOD Cloud, which comprises, e.g., publication databases such as DBLP, as well as snapshots of funding databases such as CORDIS. From this it can be seen that OpenAIRE is a more comprehensive data source than those published as LOD before.

5.2.2 Mapping Large Scale Research Metadata to Linked Data

As the schema of the OpenAIRE LOD we specified an RDF vocabulary by mapping the entities of the ER data model to RDF classes and its attributes and relationships to RDF properties. We reused suitable existing RDF vocabularies identified by consulting the Linked Open Vocabularies search service[24] and studying their specifications. Reused vocabularies include Dublin Core for general metadata, SKOS[25] for classification schemes and CERIF[26] for research organizations and activities. We linked new, OpenAIRE-specific terms to reused ones, e.g., by declaring *Result* a superclass of `http://purl.org/ontology/bibo/Publication` and `http://www.w3.org/ns/dcat#Dataset`.

[23] cf. "Streaming in XQuery", `http://www.saxonica.com/html/documentation/sourcedocs/streaming/streamed-query.html`

[24] `http://lov.okfn.org`

[25] `http://www.w3.org/2004/02/skos/`

[26] Common European Research Information Format; see `http://www.eurocris.org/cerif/main-features-cerif`

We keep the URIs of the LOD resources (i.e. entities) in the `http://lod.openaire.eu/data/` namespace. We modelled them after the HBase row keys. In OpenAIRE, these are fixed length identifiers of the form {*typePrefix*}|{*namespacePrefix*} ::*md5hash*. *typePrefix* is a two digit code, 10, 20, 30, 40 or 50, corresponding to the main entity types datasource, organization, person, project and result. The *namespacePrefix* is a unique 12-character identifier of the data source of the entity. For each row, *md5hash* is computed from the entity attributes.

The following listing shows our running example in RDF/Turtle syntax. It represent metadata about a publication (a result entity) entitled as *The Data Model of the OpenAIRE Scientific Communication e-Infrastructure*. Other metadata about the publication year, authors, publisher are shown using the already existing vocabularies.

```
@prefix oad: <http://lod.openaire.eu/data/> .
@prefix oav: <http://lod.openaire.eu/vocab#> .
# further prefixes omitted; see http://prefix.cc for their standard bindings.

oad:result/...001::39b9... rdf:type oav:Result, bibo:Publication;
    dcterms:title "The Data Model of the OpenAIRE e-Infrastructure"@en ;
    dcterms:dateAccepted "2012-01-01"^^xsd:date ;
    dcterms:language "en";
    oav:publicationYear 2012 ;
    dcterms:publisher "Springer";
    dcterms:creator oad:person/...001::9897..., oad:person/...001::ef29... .
oad:person/...001::9897... rdf:type foaf:Person;
    foaf:firstName "Paolo"; foaf:lastName "Manghi";
    oav:isAuthorOf oad:result/...001::39b9... .
oad:person/...001::ef29... rdf:type foaf:Person;
    foaf:firstname "Nikos"; foaf:lastName "Houssos";
    oav:isAuthorOf oad:result/...001::39b9... .
```

Listing 5.10: **NT**. An example of OpenAIRE Data stored in NT.

Requirements In cooperation with the other technical partners in the OpenAIRE2020 consortium, most of whom had been working on the infrastructure in previous projects for years, we established the following requirements for the LOD export:

R1 The LOD output must follow the vocabulary specified for OA [27].

R2 The LOD must be generated from one of the three existing data sources, to avoid extra pre-processing costs.

R3 The mapping to LOD should be maintainable w.r.t. planned extensions of the OpenAIRE data model (such as linking publications and data to software) and the evolution of linked data vocabularies.

R4 The mapping to LOD should be orchestrable together with the other existing OpenAIRE data provision workflows, always exposing a consistent view on the information space, regardless of the format.

R5 To enable automatic and manual checks of the consistency and correctness of the LOD before its actual publication, it should be made available in reasonable time in a private space.

To prepare an informed decision on the preferred input format to use for the LOD export, we realised one implementation for each of HBase, CSV and XML.

[27] `http://lod.openaire.eu/vocab`

Implementation As the only existing **HBase→RDF** implementation does not scale well (cf. section 5.2.1), we decided to follow the MapReduce paradigm for processing massive amounts of data in parallel over multiple nodes. We implemented a single MapReduce job. Its mapper reads the attributes and values of the OpenAIRE entities from their protocol buffer serialization and thus obtains all information required for the mapping to RDF. Hence no reducer is required. The map-only approach performs well thanks to avoiding the computationally intensive shuffling. RDF subjects are generated from row keys, predicates and objects from attribute names and cell values or, for linked entities, from column families/qualifiers.

Mapping the OpenAIRE **CSV→RDF** is straightforward: files correspond to classes, columns to properties, and each row is mapped to a resource. We initially implemented mappings in Tarql, Sparqlify and Tabels (cf. section 5.2.1) and ended up preferring Tarql because of its good performance[28] and the most flexible mapping language – standard SPARQL[29] with a few extensions. As we *map* CSV→RDF, as opposed to *querying* CSV like RDF, we implemented *CONSTRUCT* queries, which specify an RDF template in which, for each row of the CSV, variables are instantiated with the cell values of given columns.

To enable easy maintenance of **XML→RDF** mappings by domain experts, and efficient mapping of large XML inputs, we implemented our own approach[30]. It employs a SAX parser and thus supports streaming. Our mapping language is based on RDF triple templates and on the XPath[31] language for addressing content in XML. XPath expressions in the subjects or objects of RDF triple templates indicate where in the XML they obtain their values from. To keep XPath expressions simple and intuitive, we allow them to be ambiguous, e.g., by saying that *oaf:result/publisher/text()* (referring to the text content of the *publisher* element of a result) maps to the *dcterms:publisher* property of an *oav:Result*, and that *oaf:result/dateofacceptance/text()* maps to *dcterms:dateAccepted*. In theory, any combination of *publisher* and *dateofacceptance* elements would match such a pattern; however in reality only those nodes that have the shortest distance in the XML document tree represent attributes of the *same* OpenAIRE entity. XML Filters [68] efficiently restrict the XPath expressions to such combinations.

5.2.3 Performance Comparison of HBase, CSV and XML

In this section, we represent the results of the comparisons. The aim is to find a reasonable and easily maintainable way of transformation for OA large scale scholarly metadata transformation.

Comparison Metrics The **time** it takes to transform the complete OpenAIRE input data to RDF is the most important performance metric. The **main memory usage** of the transformation process is important because OpenAIRE2020 envisages the development of further services sharing the same infrastructure, including deduplication, data mining to measure research impact, classification of publications by machine learning, etc. One objective metric for **maintainability** is the size of the mapping's source code – after stripping comments and compression, which makes the comparison "independent of arbitrary factors like lengths of identifiers and amount of whitespace" [290].[32] The "cognitive dimensions of

[28] Tabels failed to handle large CSV files because it loads all the data from the CSV into main memory; Sparqlify works similar to Tarql but with almost doubled execution time (7,659 s) and more than doubled memory usage.

[29] `http://www.w3.org/TR/sparql11-query/`

[30] See source code and documentation at `https://github.com/allen501pc/XML2RDF`.

[31] `http://www.w3.org/TR/xpath20/`

[32] We used `tar cf - <input files> | xz -9`. For HBase, we considered the part of the Java source code that is concerned with declaring the mapping, whereas our CSV and XML mappings are natively defined in high-level mapping languages.

Objective Comparison Metrics	HBase	CSV	XML
Mapping Time(s)	1,043	4,895	45,362
Memory (MB)	68,000	103	130
Compressed Mapping Source Code (KB)	4.9	2.86	1.67
Number of Input rows/records	20,985,097	203,615,518	25,182,730
Number of Generated RDF Triples	655,328,355	654,193,273	788,953,122

Table 5.5: **Measurements.** Result of performance comparison of transformations.

notation" (CD) evaluation framework provides further criteria for systematically assessing the "usability of information artefacts" [31].

The following dimensions are straightforward to observe here: *closeness* of the notation to the problem (here: mapping HBase/CSV/XML to RDF), *terseness* (here measured by code size; see above), *error-proneness*, *progressive evaluation* (i.e. whether one can start with an incomplete mapping rule and evolve it to further completeness), and *secondary notation and escape from formalism* (e.g. whether reading cues can be given by non-syntactic means such as indentation or comments).

Measurements and Observations: The **HBase→RDF** evaluation ran on a Hadoop cluster of 12 worker nodes operated by CNR.[33] As our **CSV→RDF** and **XML→RDF** implementations required dependencies not yet installed there, we evaluated them locally: on a virtual machine on a server with an Intel Xeon E5-2690 CPU, having 3.7 GB memory and 250 GB disk space assigned and running Linux 3.11 and JDK 1.7. As we did not have a cluster available, and as the tools employed did not natively support parallelization, we ran the mappings from CSV and XML sequentially. The Table 5.5 table lists our measurements; further observations follow below.

For **HBase→RDF**, the peak memory usage of the cluster was 68 GB, i.e. ∼5.5 GB per worker node. No other MapReduce job was running on the cluster at the same time; however, the usage figure includes the memory used by the Hadoop framework, which schedules and monitors job execution.

The 20 **CSV** input files correspond to different entities but also to relationships. This, plus the way multi-valued attributes are represented, causes the high number of input rows. The size of all files is 33.8 GB. The **XML→RDF** memory consumption is low because of stream processing. The time complexity of our mapping approach depends on the number of rules (here: 118) and the size of the input (here: 144 GB). With the complexity of the XML representation, this results in an execution time of more than 12 hours. The size of the single RDF output file is ∼91 GB. Regarding *cognitive dimensions*, the different notations expose the following characteristics; for lack of space we focus on selected highlights. *Terseness*: the high-level CSV→RDF and XML→RDF languages fare better than the Java code required for HBase→RDF. Also, w.r.t. *closeness*, they enable more intuitive descriptions of mappings. As the CSV→RDF mappings are based on SPARQL, which uses the same syntax for RDF triples than the Turtle RDF serialization, they look closest to RDF. *Error-proneness*: Syntactically correct HBase→RDF Java code may still define a semantically wrong mapping. In Tarql's CSV→RDF mappings, many types of syntax and semantics errors can be detected easily. *Progressive evaluation*: one can start with an incomplete Tarql mapping rule CSV→RDF mapping rule and evolve it towards completeness. *Secondary notation*: Tarql and Java support flexible line breaks, indentation and comments, whereas our current XML→RDF mapping implementation requires one (possibly long) line per mapping rule. Overall, this strongly suggests that CSV→RDF is the most maintainable approach.

In conclusion, we have mapped a recent snapshot of the OpenAIRE data to RDF. A preliminary

[33] https://issue.openaire.research-infrastructures.eu/projects/openaire/wiki/Hadoop_Clusters#section-3

dump as well as the definitions of the mappings are available online at `http://tinyurl.com/OALOD`. Mapping from HBase is fastest, whereas mapping from CSV promises to be most maintainable. Its slower execution time is partly due to the less powerful hardware on which we ran it; comparing multiple CSV→RDF processes running in parallel to the HBase→RDF implementation on the CNR Hadoop cluster seems promising. Based on these findings the OpenAIRE2020 LOD team will decide on the preferred approach for providing the OpenAIRE data as LOD; we will then make the data available for browsing from their OpenAIRE entity URIs, and for querying via a SPARQL endpoint.

Having implemented almost the whole OpenAIRE data model, future steps include interlinking the output with other existing datasets. E.g., we so far output countries and languages as strings, whereas DBpedia and Lexvo.org are suitable linked open datasets for such terms. Link discovery tools will further enable large-scale linking against existing "publication" and "government" datasets.

5.3 Interlinking

Linked Open Data (LOD) is a popular approach for maximizing both legal and technical reusability of data, and enabling its connection with further datasets [16]. However, without further work, LOD datasets do not yet provide added value to end users, as they are only accessible for service and application developers familiar with Semantic Web technology and the datasets' vocabularies.

There are several related work on interlinking scholarly metadata. Rajabi has studied the exploitation of educational metadata using interlinking methods [219]. His work objectives closely related to ours; however its application domain is eLearning services and therefore he discusses the benefits of interlinking educational (meta)data in practice. Rajabi et al. provide a comparison of interlinking tools as well as interlinking rules [221] and a method for identification of duplicate links [220]. Hallo et al. follow the same objective as we do, i.e., publishing Open Access metadata as LOD [104]. Their work focuses on providing better search services on top of open journal datasets, but their data could be used as a candidate dataset for our interlinking. Recent work by Purohit et al. addresses the problem of scholarly resource discovery [218]. They also reviewed tools providing such services and present a framework for Resource Discovery for Extreme Scale Collaboration (RDESC)[34] which has common objectives with OA. However, they have not yet initiated interlinking of research metadata and the provision of a comprehensive knowledge graph.

This section focuses on enriching the OpenAIRE LOD by interlinking, and utilizing this interlinked data to provide added value to users in situations where they need scholarly communication metadata, e.g., when they are looking for a publication to cite, or for all publications of a given author. OpenAIRE (OA), covers more than 23M publications, 12M authors and scientific datasets. OA metadata has been exposed as LOD [274], aiming at maximizing its reusability and technical interoperability by:

- providing an infrastructure for data access, retrieval and citation (e.g., a SPARQL endpoint or a LOD API),
- interlinking with popular LOD datasets and services (DBLP, ACM, CiteSeer, DBpedia, etc.),
- enriching the OpenAIRE Information Space with further information from other LOD datasets.

OpenAIRE aims at increasing interoperability and reusability of this data collection by exposing it as Linked Open Data (LOD). Therefore, the main motivation for exposing OA as LOD is to provide wider data access, and easier and broader metadata retrieval by enabling interlinking with relevant and popular LOD datasets [274]. By following the LOD principles, it is now possible to further increase interoperability and reusability by connecting the OpenAIRE LOD to other datasets about projects,

[34] `https://tw.rpi.edu/web/project/RDESC`

publications, people and organizations. Doing so required us to identify link discovery tools that perform well, as well as candidate datasets that provide comprehensive scholarly communication metadata, and then to specify linking rules. Metadata about different types of entities – research results (publications and datasets), persons, projects and organizations – that the OA infrastructure aggregates is being exposed as LOD. OA LOD uses terms from existing vocabularies and, where necessary, defines new terms. Existing ontologies reused include SKOS, CERIF, DCMI Terms, FOAF [273, 274]. Two prefixes/namespaces are OA specific: `oav: http://lod.openaire.eu/vocab/` for the OA vocabulary, and `oad: http://lod.openaire.eu/data/` for OA instance data.

The data has been exposed in three ways: (1) small fragments of RDF, accessible by dereferencing the URI that identifies a particular entity, (2) a downloadable all-in-one dump[35], and (3) a SPARQL endpoint, i.e. a standardized query interface accessible over the Web[36].

It is envisaged to extend the OA LOD by enriching and interlinking it with the following types of data:

- data that has not (yet) been collected by OA's existing mechanisms, e.g., certain types of persistent identifiers of publications or people (e.g., ORCID),
- data that is expensive to collect and/or not included in the OA data model, e.g., data about scientific events, and
- data that is related to open research but out of the scope of the OA infrastructure itself and therefore not targeted to be ever collected, e.g., biographies of persons, or geodata about the locations of organizations.

The primary objectives are (1) providing added value to users, by enabling those who develop user-oriented applications and services to access a richer collection of relevant data than just OA's own, and (2) facilitating internal data management, e.g., by aiding the resolution of duplicates resulting from metadata being harvested from different repositories by linking to external reference points. In the following sections, we demonstrate the added value that interlinking provides for end users by implementing visual frontends for looking up publications to cite, and publication statistics, and evaluating their usability on top of interlinked vs. non-interlinked data.

5.3.1 Identifying Properties and Target Datasets Suitable for Interlinking

Investigating Existing interlinking tools: Not all properties of an OA entity are suitable for the purpose of interlinking to other entities, as Rajabi et al. have investigated in the related domain of metadata about educational resources [221]. Following their method, we analyzed all OpenAIRE entities and their properties to discover linkable elements. We filtered out properties that potentially cannot be linked due to their specific values, for example Booleans (Yes/No), format values (PDF, JPEG), or language codes (en, de), and properties whose meaning is local to some source repository according to its policy, for example local identifiers or version numbers. This left us with properties such as "publication title" and "author name", "published year", "description", "subject", etc., which have string or integer values. Where initial interlinking tests yielded subjectively satisfactory results, we chose the respective properties for interlinking – i.e. the following: **Title** and **Digital Object Identifier** of **Publication**, **Full name**, **First name** or **Last name** of **Person**s, and **Label** or **Homepage** of **Organizations**. Table 5.6 lists the ten most relevant datasets according to these criteria.

Identifying Interlinking Target Datasets: To identify appropriate target datasets to be interlinked with OA, we examined several datasets from the LOD Cloud, in the following steps:

[35] `http://tinyurl.com/OALOD`
[36] `http://lod.openaire.eu/sparql`

Datasets	Size	Endpoint	Dump	Covered OA entity types
DBpedia	1B	Available	NT	Person, Organization
DBLP	55M	–	NT	Publication, Person
ACM	12M	Available	RDF/XML	Publication, Person
CiteSeer	8M	Available	RDF/XML	Publication, Person
BibBase	200K	–	RDF/XML	Person, Publication, Organization
IEEE	200K	Available	RDF/XML	Publication, Person
OpenCitations	3M	Available	JSON-LD	Person, Publication, Organization
SWDF	242K	–	RDF/XML	Person, Publication, Organization
BNB	109M	–	NT, RDF/XML	Person, Publication
COLINDA	149K	Available	RDF/XML	Publication
GeoNames	93M	–	RDF/XML	Organization

Table 5.6: **Target Datasets**. List of candidate Datasets containing scholarly metadata about different artifacts for interlinking with OA.

1. **Identifying publication-related datasets in DataHub**: our aim is to find datasets tagged with the same domain as that of OA or a related one. We therefore searched the DataHub portal[37] for datasets tagged with "publication" or related domains. This search yielded more than 900 datasets.
2. **Checking data endpoint availability**: we filtered the datasets identified previously by checking their SPARQL endpoints' or RDF dumps' availability.
3. **Retrieving datasets specification**: of the remaining datasets (still more than 60), we next retrieved each dataset's specification (size, metadata schema, etc.). From an interlinking point of view, we considered data volume, frequent updates, and matches with the entity types and properties identified previously as the most important characteristics of a dataset. Moreover, we considered available links to other related datasets desirable.

5.3.2 Identifying Tools and Algorithms Suitable for Interlinking

Identifying Interlinking Tools: There exist a number of tools for creating semi-automatic links between datasets by running some matching techniques. These linking tools identify similarities between entities and generate links (e.g.owl:sameAs) that connect source and target entities. Rajabi et al. conducted a study that suggests that data publishers can trust interlinking tools to interlink their data to other datasets; accordingly, LIMES and Silk are the most promising frameworks [221]. Simperl et al. have compared various linking tools by addressing aspects such as required input, resulting output, considered domain and matching techniques used [246]. This allowed for a comparison from several perspectives: degree of automation (to what extent the tool needs human input) and human contribution (the way in which users are required to do the interlinking.

In summary, these comparisons point out the two well-known open source interlinking frameworks that we also used: LIMES[38] (Link Discovery Framework for Metric Spaces) and Silk[39] (Link Discovery Framework for the Web of Data). In an evaluation of the two frameworks, the LIMES developers showed that LIMES considerably outperforms Silk in terms of running time, with a comparable quality of the

[37] https://datahub.io/

[38] http://aksw.org/Projects/LIMES.html

[39] http://silkframework.org/

Metric	Description
Trigrams	uses the number of matching triples in both strings as $s = 2 \times \frac{m}{(a \times b)}$ where m is the number of matching trigrams, a is the number of trigrams in string 1, and b is the number of trigrams in string 2 [227].
Levenshtein	is based on the minimum number of insertion, deletion or replacement operations to transform string 1 into string 2.
Jaro	is a measure of characters in common, being no more than half the length of the longer string in distance, with consideration for transpositions; it is best suited for short strings such as person names [255].
Jaro-Winkler	is an optimized version of Jaro designed and best suited for short strings such as person names
Cosine	is the cosine of the angle between string vectors; for equal strings the angle between them will be 0 and the cosine will be 1 [227].

Table 5.7: **String matching algorithms.** The string matching algorithms are shows with a description of their characteristics.

output. Moreover, LIMES can be downloaded as a standalone tool for carrying out link discovery locally and consists of modules that can be extended easily to accommodate new or improved functionality.

Our comparative evaluation of Silk and LIMES, which finally made us choose LIMES based on the quality of the output.

Identifying String Matching Algorithms: One of the most important factors in discovering links effectively is choosing the right string matching algorithm. The results of our heuristic experiments shows that both tools supports string matching according to trigrams, Levenshtein [40], Jaro, Jaro-Winkler and cosine (all of them normalized); cf. Table 5.7. It shows detailed definition of the algorithms. In our initial experiments, Jaro and Levenshtein proved most reliable for identifying equivalent names and titles. Thus, we chose Levenshtein for long string values, i.e., publication titles, and Jaro for short string values, i.e., person names.

We constructed the configuration files with the metrics defined above and perform a test interlinking based on author name and publication title matching between publication resources which has been published in year 2008 along with their author resources of OA and SWDF dataset. The test interlinking performed on all the mentioned metrics. Based on the result we got, Jaro and Levenshtein were most reliable by means of identifying equivalent names and titles. Thus, we chose Levenshtein metric for long string values, i.e., publications title and Jaro metric for short string values, i.e., person name since this metric is best suited for short string. An example of a metric definition in LIMES is shown below.

```
<METRIC>AND(Jaro(x.foaf:name, y.foaf:name)|0.8, levenshtein(
  x.dcterms:creator/cerif:name, ^y.dblp:hasAhutor/dblp:title)|0.8)
</METRIC>
```

Listing 5.11: **Metric definition in LIMES** LIMES takes certain metrics as an input and combines them for the matching instances in the sources and target datasets.

Set a threshold can be used to find the exact matching and our aim is to correctly identify links. To this end, the threshold in the work-flow was set to 0.95, which means that two concepts are considered as matched if their syntax similarity is more than 95%. An example of an acceptance definition in LIMES is shown below.

```
<ACCEPTANCE><THRESHOLD>0.95</THRESHOLD>
<FILE>openaire_dblp_accept.nt</FILE><RELATION>owl:sameAs</RELATION></ACCEPTANCE>
```

Listing 5.12: **Acceptance definition in LIMES.** The sump of the datasets given for interlinking is accepted through a certain rule.

[40] https://wikipedia.org/Levenshtein_distance

We chose 0.75 as a threshold value for review condition. These review links should be verified by manual evaluation. We have chosen a low value as threshold for review in order to observe how well linking specification works. We observed that similar but different entities appeared as links in review files. There can be a situation where a correct link appears in review file, this can be due to spelling differences. An example of a review definition in LIMES is shown below.

```
<REVIEW><THRESHOLD>0.75</THRESHOLD>
  <FILE>openaire_dblp_review.nt</FILE><RELATION>owl:sameAs</RELATION></REVIEW>
```

Listing 5.13: **Review definition in LIMES** For finding teh exact matches a review of the metrics is done by taking into account a certain threshold.

5.3.3 Results from Interlinking Scholarly Metadata

Here, we outline a use case that demonstrate how the result of interlinking OpenAIRE with related datasets can support scholarly communication. Listing 5.14 We imagine such services to be integrated into environments for reading or writing scholarly papers. A difficult and time consuming task for peer reviewers is to get a quick overview of the state of the art of the field covered by the paper or dataset under review. The OpenAIRE LOD itself has information about the subject of a paper or a dataset, which can be linked to subject classification schemes such as the ACM CCS. Furthermore, CiteSeer provides citation graphs of papers. We can thus offer to peer reviewers a service that finds papers or datasets similar to the one under review. One of the critical facts in the process of writing and publishing is the comprehensiveness of citations inside scholarly data.

```
<SOURCE>
  <ID>source1</ID>
  <ENDPOINT>http://beta.lod.openaire.eu/sparql</ENDPOINT>
  <VAR>?x</VAR>
  <PAGESIZE>10000</PAGESIZE>
  <RESTRICTION>?x a oav:Person</RESTRICTION>
  <PROPERTY>foaf:name AS lowercase RENAME name</PROPERTY>
  <PROPERTY>dcterms:creator/cerif:name AS lowercase->
  regexreplace("[^A-Za-z0-9]"," ") RENAME title</PROPERTY>
</SOURCE>
<TARGET>
      <ID>source2</ID>
      <ENDPOINT>C:\dblp2.nt</ENDPOINT>
      <VAR>?y</VAR>
      <PAGESIZE>-1</PAGESIZE>
        <RESTRICTION>?y rdf:type dblp:Person</RESTRICTION>
      <PROPERTY>dblp:primaryFullPersonName AS
      lowercase RENAME dname</PROPERTY>
      <PROPERTY>^dblp:authoredBy/dblp:title AS lowercase->
      regexreplace("[^A-Za-z0-9]"," ") RENAME dtitle</PROPERTY>
      <TYPE>NT</TYPE>
</TARGET>
<METRIC>AND(Jaro(x.name,y.dname)|0.85,
  Levenshtein(x.title,y.dtitle)|0.7)</METRIC>
```

Listing 5.14: **LIMES configuration** A configurations file for interlinking of the Person entity is represented with certain metrics and metadata.

A service similar to the one for peer reviewers explained above could be offered to authors. Research dynamics could be understood better by analyzing how people who publish on certain topics move in the community, e.g., to other organizations. Having access to the networks of a papers and authors and their organizations, and furthermore taking into account the events in which people participate enables new indicators for measuring the quality and relevance of research that are not just based on counting citations.

Having access to the networks of a papers and authors and their organizations, and furthermore taking into account the events in which people participate enables new indicators for measuring the quality and relevance of research that are not just based on counting citations. To enrich content of openAIRE dataset, we carried out interlinking between different concepts from OpenAITRE and the corresponding concept in four candidate datasets; namely: DBLP, DBpedia, ACM and SWDF. The Person entity in OpenAITRE is defined as oav:Person vocabulary in OpenAITRE data schema and as dblp:Person vocabulary in DBLP.

It should be highlighted that the comparison of person entities and associated properties in OpenAIRE. While running the tool, this configuration file will construct and execute two different SPARQL queries from source and target datasets to get the selected properties values and apply string similarity matching on. The result of this interlinking is number of links in RDF, shown in Listing 5.15 , which connect OpenAITRE and DBLP Person entities using “owl:sameAs” relationship. We can follow a similar approach in Silk linkage rule for Person Interlinking.

```
<\href{http://lod.openaire.eu/data/result/doajarticles::65803
    a423ca8b7cc411d97c008b1b4ec}{http://lod.openaire...b1b4ec}> owl:sameAs <\href{
    http://dblp.org/rec/journals/entropy/ZengeyaBC15}{http://dblp.org.../
    ZengeyaBC15}>\newline
```

Listing 5.15: **Sample Interlinking Result.** The *sameAs* relations are constructed based on LIMES configuration.

Evaluation of Interlinking Tools To find the common and individual links created by selected interlinking tools, we wrote a script [5, Appendix C], which compares the contents of results obtained by two tools and returns the number of common links and also the number of links found by one tool but not by the other. In an experiment with considering publications of OA data and publications of DBLP data LIMES was able to match 432 entities, i.e. more than Silk. The number of common records discovered by both Silk and LIMES is 358. 74 links were found by LIMES but not by Silk, and 3 links were found by Silk but not by LIMES.

In addition to the number of discovered links, reliability of the obtained links is also important. Thus, to evaluate the quality and reliability of the links obtained via each tool, we created a reference linkset (gold standard) consisting of 100 publication resource selected from OA and by manual research found 38 links to SWDF. We then ran Silk and LIMES to find only links from these 100 selected OA resources to SWDF and then compared their output to the gold standard. We computed precision, recall and F-measure to check completeness and correctness of the links found; Table 5.8 shows the results. Precision is the ratio of the number of relevant items to the number of retrieved items, i.e.:

$$Precision = \frac{\text{true positive}}{\text{true positive + false positive}}$$

In our case, this means

$$Precision = \frac{\text{(Number of created links – Number of incorrect links)}}{\text{Number of created links}}$$

and indicates the correctness of links discovered.

Recall is the ratio of the number of retrieved relevant items to the number of relevant items, i.e.:

Tool	Number of created links	Number of missing links	Number of incorrect discovered links	Precision	Recall	F-measure
LIMES	37	1	0	1	0.97	0.98
Silk	29	9	1	0.96	0.76	0.85

Table 5.8: **Evaluation.** The evaluation of interlinking tools result against a gold standard.

$$Recall = \frac{\text{true positive}}{\text{true positive + false negative}}$$

In our case, this means

$$Recall = \frac{\text{(Number of created links – Number of incorrect links)}}{\text{(Number of correct links + number of missing links)}}$$

and indicates the completeness of links discovered. F-measure is a combined measure of accuracy defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

The evaluation revealed 9 missing links and one incorrectly discovered link in Silk and 1 missing in LIMES. This corresponded to a Precision of 1, a Recall of 0.97 and an F-measure of 0.98 for LIMES and a Precision of 0.96, a Recall of 0.76 and an F-measure of 0.84 for Silk. The main advantage for LIMES within this small evaluation is the execution time. However we consider the best practices so far which showed that LIMES outperforms Silk dealing with big data. Therefore, due to the fact that we got more relevant, reliable and accurate results from LIMES compared to Silk, we chose LIMES for further interlinking OpenAIRE with other datasets.

Links between	Target dataset	Target instances	Generated links	Sample of generated links	Verified links	Precision
Publication	DBLP	164890	2276	150	147	0.98
Publication	SWDF	5009	432	150	150	1.0
Publication	ACM	10378	1082	150	136	0.9
Person	SWDF	11184	2000	200	180	0.9
Person	DBLP	932000	6852	200	111	0.55
Person	DBpedia	23373	1088	200	80	0.40
Organization	SWDF	3212	866	30	30	1.0
Organization	DBpedia	3472	38	30	30	1.0

Table 5.9: **Evaluation**. Number of links and precision values obtained between OA and DBLP, SWDF, ACM and DBpedia for publications, persons and organizations.

5.3.4 Evaluation of Interlinking Results

We configured LIMES to generate *owl:sameAs* links between resources with a similarity of above 95%. However, the question is to what extent resources linked in this way are actually the same. Given the size of the linkset, manually assessing and analyzing each link would have been too time-consuming. We therefore picked a number of sample links from each linkset based on its size, aiming at feasibility of a manual inspection (150 samples of publication links, 200 samples of person links and 25 samples of organization links). We then manually verified the correctness of each link and computed precision as “number of correct links” / “number of sample links”. In the absence of a gold standard, we did not compute recall. The number of links obtained between OA and DBLP, SWDF, ACM and DBpedia for publications, persons and organizations is displayed in Table 5.9 along with the precision for each linkset. We obtained high precision in Publication and Organization interlinking, but not in Person interlinking. This is because initially we carried out Person interlinking by just comparing the names, which was not sufficient, as different persons may have the same name. In future work, we should improve the linking rule for persons taking into account not only their names but also the titles of their publications.

CHAPTER 6

Utilization of a Crowdsourced Scholarly Knowledge Graph

The aim of this chapter is to explain the utilization of the created and curated scholarly knowledge graph. The base of the following chapters are the publications in which the author majorly contributed [1].

Scholars often need to search for matching, high-profile scientific events to publish their research results. Information about topical focus and quality of events is not made sufficiently explicit in the existing communication channels where events are announced. Therefore, scholars have to spend a lot of time on reading and assessing calls for papers but might still not find the right event. Additionally, events might be overlooked because of the large number of events announced every day. We introduce OpenResearch, a crowd sourcing platform that supports researchers in collecting, organizing, sharing and disseminating information about scientific events in a structured way. It enables quality-related queries over a multidisciplinary collection of events according to a broad range of criteria such as acceptance rate, sustainability of event series, and reputation of people and organizations. Events are represented in different views using map extensions, calendar and time-line visualizations. We have systematically evaluated the timeliness, usability and performance of OpenResearch. The curation section of this chapter is based on the work presented in the following publication:

> **Sahar Vahdati**, Natanael Arndt, Sören Auer, Christoph Lange. *OpenResearch: Collaborative Management of Scholarly Communication Metadata* In Knowledge Engineering and Knowledge Management 2016.

The following work uses the event knowledge graph and graph mining techniques in order to provide co-authorship recommendation.

> **Sahar Vahdati**, Rahul Jyoti Nath, Guillermo Palma, Maria-Esther Vidal, Christoph Lange, Sören Auer, *Unveiling Scholarly Communities of Researchers using Knowledge Graph Partitioning*, TPDL 2018.

6.1 Curation

There is currently an era of departure to investigating how scholarly work and communication can be taken to the digital world. Much attention is devoted to new forms of publishing (e.g. semantic papers,

[1] **Own Manuscript Contributions.** The author of this thesis has been the main author of the publications. The article co-authored withNath et al. is the result of a master thesis mainly supervised by Vahdati where she guided the students through a successful research and mainly contributed to the writing of the papers.

micro-publications), open access, and free availability of publication metadata. Still, a large number of scholarly communication processes and artifacts (other than publications) are not currently well supported. This includes in particular information about events (conferences, workshops), projects, tools, funding calls etc. In particular for young researchers and interdisciplinary work it is of paramount importance to be able to easily identify venues, actors and organizations in a certain field and to assess their quality.

Research results are published as scientific papers in journals and events such as conferences, workshops etc. Each component of this communication needs to be open and easily accessible. Besides conducting their actual research, scholars often need to search for scientific events to submit their research results to, for projects relevant to their research, for potential project partners and related research schools, for funding possibilities that support their particular research agenda, or for available tools supporting their research methodology. For lack of better support, scholars rely a lot on individual experience, recommendations from colleagues and informal community wisdom, they do simple Web searches or subscribe to mailing lists and are stuck with simplistic rankings such as calls for papers (CfPs) sorted by deadline. Domain specific mailing lists are a medium often used by conference and workshop organizers for posting initial, second, final calls for papers, as well as deadline extensions. But this situation leads to discussions on whether to allow calls for papers on the lists or treat them as spam[2]. It is especially hard for subscribers to filter those calls according to their individual interests, or maybe explicitly subscribe to important information, such as deadline extensions or subsequent calls, on a specific event or an event series.

On the other hand, the quality of scientific events is directly connected to the research impact and the rankings of the scientific *papers* published by them. For example, the *Research Excellence Framework* (REF) for assessing the quality of research in UK higher education institutions, classifies publications by the venues they are published in. This facilitates assessing every researcher's impact based on the number of publications in conferences and journals. Providing such information to researchers supports them with a broader range of options and a comprehensive list of criteria while they are searching for events to submit their research contributions. To provide comprehensive information about scientific venues, projects, results etc., we present `OpenResearch.org`. OpenResearch is a platform for automating and crowd-sourcing the collection and integration of semantically structured metadata about scholarly communication. In particular, with regard to events, OpenResearch.org . . .

1. reduces the effort for researchers to find 'suitable' events (according to different metrics) to present their research results,
2. supports event organizers in visibly promoting their event,
3. establishes a comprehensive ranking of events by quality,
4. provides a cross-domain service, recommending suitable submission targets to authors, and
5. supports easy and flexible data exploration using Linked Data technology: a structured dataset of conferences facilitates selection regarding fields of interest or quality of events.

The core of the OpenResearch.org approach is to balance manual/crowd-sourced contributions and automated methods. OpenResearch empowers researchers of any field to collect, organize, share and disseminate information about scientific events, projects, organizations, funding sources and available tools. It enables the community to define views as queries over the collected data; assuming sufficient data, such queries can enable rankings by relevance or quality. Driven by Semantic MediaWiki (SMW), OpenResearch provides a user interface for creating and editing semantically structured event profiles, tool and project descriptions, etc. in a collaborative wiki way. OpenResearch is part of a greater research

[2] Note a recent survey on calls for papers on the W3C mailing lists: `https://lists.w3.org/Archives/Public/semantic-web/2016Mar/0108.html`

and development agenda for enabling true open access to all types of scholarly communication metadata (beyond bibliographic ones) not just from a legal but also from a technical perspective. The work on OpenResearch is aligned with *OpenAIRE*, the Open Access Infrastructure for Research in Europe.

6.1.1 Collaborative Management of Scholarly Communication Metadata

Problem Statement With a focus on the scholarly metadata on the example of scientific events, a list of management challenges have been identified: *Challenge 1: Communication.* Research communities use different communication channels to distribute event announcements and CfPs. Announcing CfPs through different mailing lists is the traditional but still most popular way of disseminating information about an event. Exploring the calls for papers posted on mailing lists of the Semantic Web community shows that 500 to 700 event announcements have been posted every year between 2006 and 2016 (approx. 15-30% of the overall traffic). This shows that a large and widely spread amount of unstructured data about scientific events is increasingly being published via communication channels not specifically designed for this purpose. Due to the interdisciplinary nature of research, event organizers easily overlook relevant channels to announce their event. In addition, browsing through the CfPs in several channels to identify events that might be of interest is a time and effort consuming task.

Challenge 2: Structure. There are structural differences across events, for example, events with many co-located events or sub-events, or new events emerged from multiple smaller ones. One example for the latter is the Conference on Intelligent Computer Mathematics (CICM), which results from the convergence of four conferences that used to be separate but now are tracks of a single conference.[3] Scholars who want to find out whether an event matches their research interests therefore have to understand its structure; if they cannot find the desired information for the super-event, they will have to study the sub-events.

Challenge 3: Series. Most scientific events occur in series, whose individual editions take place in different locations with narrow topical changes. Researchers often need to explore several resources to obtain an overview of the previous editions of an event series to be able to estimate the quality of the next upcoming event in this series.

Challenge 4: Addressing Different Stakeholders. Event organizers aim to attract as many submitters as possible to their events. Publishers want to know whether they should accept a particular event's proceedings in their renowned proceedings series. Potential PC members want to decide whether it is worth spending time in the reviewing process of an event. Similarly, sponsors and invited speakers need to decide whether a certain event is worth sponsoring or attending. Researchers receiving CfP emails have to distinguish whether the event is appropriate for presenting their work. Researchers searching for events through various communication channels assess events based on criteria such as thematic relevance, feasibility of the deadline, close location, low registration fee etc. The organizers of smaller events who plan to organize their event as a sub-event of a bigger event have to decide whether this is the *right* venue to co-locate with. These examples prove the importance of filtering events by topic and quality from the point of view of different stakeholders. Currently, the space of information around scientific events is organized in a cumbersome way, thus preventing events' stakeholders from making informed decisions, and preventing a competition of events around quality, economy and efficiency.

Strategies. Event organizers employ a number of strategies to cope with the challenges of advertising their event and engaging with the potential audience. They use multiple channels (mailing lists, social networks, homepages) to distribute CfPs. Some organizers plan deadline extensions in advance, as a strategy to attract more submissions. Some communities employ databases on top of mailing lists for

[3] `http://www.cicm-conference.org/`

announcing scientific events e.g., researchers in information systems and databases use the DBWorld database (cf. chapter 2). The strategies mentioned so far target authors of submissions, whereas event organizers also have to find sponsors, high-profile program committee members and keynote speakers. This is currently done by contacting researchers or companies that the organizers know already. An approach for a centralized and holistic infrastructure for managing the information about scientific events was missing so far.

Requirements A collaborative and partially decentralized environment is required to enable community-based scientific data curation and extension, and to tap into the 'wisdom of the crowd' for elicitation and representation of metadata associated to scholarly communication. In particular, such a system is aimed to address the following requirements as services, which we have derived from the challenges C1–C4 pointed out in the problem statement and from the review of related work (R):

R1 It should be easily possible to create various views on the resulting data (addressing various communities), also in a collaborative way. (C1)

R2 Fine-grained and user extensible semantic representation of the (meta)data should be supported. (C1)

R3 The resulting ontological model should capture the relationships between various types of entities (e.g. event series, sub/super events, roles in event organization, etc.). (C2, C3)

R4 Different stakeholders of scholarly communication (event organizers, PC members, developers, etc.) have to be supported adequately. (C4)

R5 The data representation and view generation mechanisms should support fine-grained analyses (e.g. about the quality of events according to various indicators). (C4)

R6 The collaborative authoring and curation interfaces should be user friendly and enable novices to participate in the data gathering and curation processes.(C4)

R7 The system architecture should support automatic as well as manual/crowd-sourced data gathering from a variety of information sources. (R)

R8 All changes should be versioned to support tracking particular users' contributions and their review by the community. (R)

R9 The collected data should be easily reusable by application and service developers. (R)

6.1.2 The Architecture of the OpenResearch.org Platform

OpenResearch uses semantic descriptions of scientific events based on a comprehensive ontology; this enables distributed data collection by embedding markup in conference websites aligned with schema.org, and links to other portals and services. Semantic MediaWiki (SMW) serves as data curation interface employing semantic forms, templates various extensions and semantic annotations in the wiki markup. In the remainder, we describe the architecture of OpenResearch. Figure 6.1 depicts the three layers of OpenResearch's architecture:

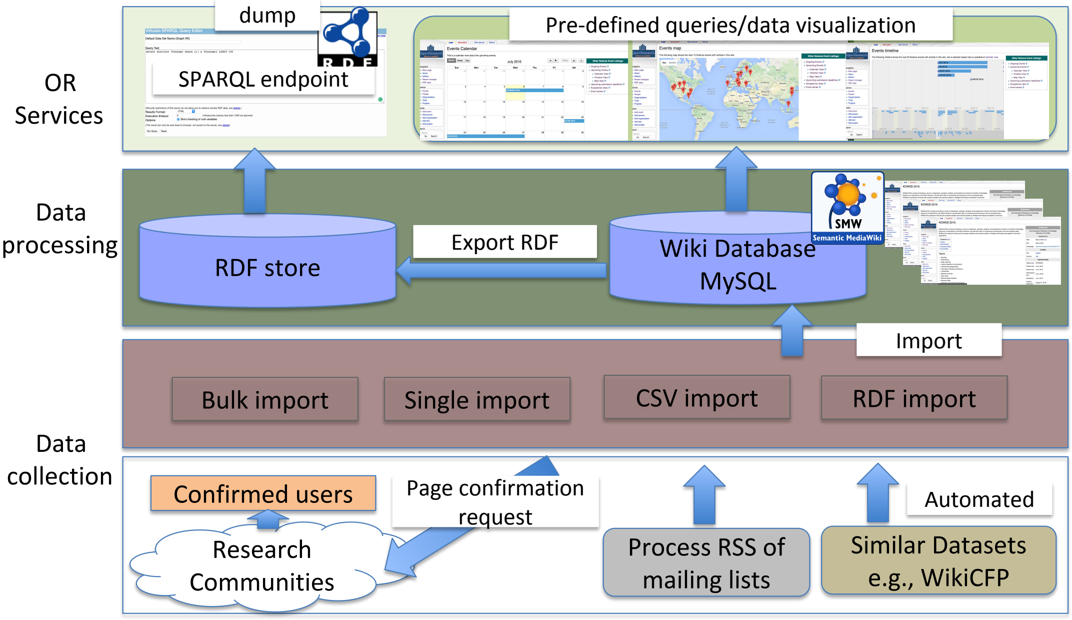


Figure 6.1: **OpenResearch Architecture.** Three main parts of the the OR architecture including the crowdsourcing components is shown.

- **Data Gathering and Scrapers** This layer supports ingestion, semantic lifting and integration of relevant information from various sources. To populate the OpenResearch knowledge base in addition to crowd-sourcing, we gather information from different sources. Sources can be available as Linked Data already, or structured, semi-structured and unstructured. SMW itself provides two options for importing data: creation of individual pages/resources and bulk import[4] using the MediaWiki export format. Structured and semi-structured information can be imported as CSV and RDF: CSV files, prepared manually or obtained from WikiCFP via a crawler that we have implemented[5], can be transformed to the MediaWiki export format using the MediaWiki CSV Import[6] and then imported using the bulk importer; RDF datasets can be imported using the RDFIO MediaWiki extension[7].
- **Data processing** This layer enables the storing and management of unstructured (text markup), semi-structured (annotations and infoboxes), structured data (RDF data adhering to an ontology) and schema data (the underlying ontology) Two database management systems are used in the OpenResearch architecture: one to store the schema-level information, the other to store the generated semantic triples. SMW supports multiple triple stores for storing the RDF graph, e.g., Blazegraph or Virtuoso. We use Blazegraph as it has been selected Wikimedia Foundation based on

[4] https://www.mediawiki.org/wiki/Help:Export
[5] https://github.com/EIS-Bonn/OpenResearch/tree/master/wikiCFP
[6] http://mwcsvimport.pronique.com/; usage described at http://openresearch.org/OpenResearch:HowTo
[7] https://www.mediawiki.org/wiki/Extension:RDFIO

a performance and quality.[8] A MySQL relational database is used to store the templates, properties and, form names.

- **Data exploring** This layer comprises various means for human and machine-readable consumption of the data. Several types of data representation are made possible by data exploration. CfPs are represented as individual wiki pages for each event instance, including a semantic representation of their metadata. SMW provides a full-text search facility and supports semantic queries. Furthermore, the RDF triple store can be accessed using a SPARQL endpoint or downloadable RDF dump.

Accessing OpenResearch.org Knowledge Graph All data created within OpenResearch is published as Linked Open Data (LOD). In the sequel, we describe ways for accessing OpenResearch LOD. Afterwards, we outlines how the LOD approach enables building further services on top by sketching two possible ways of consuming the OpenResearch LOD: interlinking with relevant datasets, and using OpenResearch LOD as external plug-in for the Fidus Writer scientific authoring platform[9].

An updated version of the OpenResearch dataset is produced daily and available for download and query.[10]. The data is also queryable via a SPARQL endpoint[11]. In addition, the semantic representation of the metadata for each event is represented as an RDF feed in each page. The RDF feed for the EKAW 2016 resource is available at `http://openresearch.org/Special:ExportRDF/EKAW_2016`. To expose dereferenceable resources conforming with Linked Data best practices, the URI resolver provides URIs with content negotiation; e.g., for the EKAW 2016 resource the URI is `http://openresearch.org/Special:URIResolver/EKAW_2016`.

6.1.3 Performance Measurements and Usability Analysis

The main objective of this work is to introduce a comprehensive approach for collaborative management of scholarly communication metadata with a special focus on events. We are for now mainly interested in collecting data, as this allows to provide more interesting analysis services.Nevertheless, we evaluated three aspects of OpenResearch including two surveys, performance measurements of the system as well as a usability analysis.

Timeliness Questionnaire: In a survey, we asked 40 researchers from different fields including Computer Science, Social Science to explain how they explore scientific events[12]. Over 75% of the participants agree that having an event recommendation service is very relevant for them. For selecting an event to participate, all participants confirmed that they consider information that is not served directly by the current communication channels. Some of these criteria are networking possibilities, review quality, high-profile organizers, keynote speakers and sponsors, low acceptance rate, having high quality co-located events, close location, citations counts for accepted papers of previous years. Participants indicated that they explore scientific events using: search engines, mailing lists, social media and personal contacts. Then, they assess the CfPs to find out whether that event satisfies their criteria. Over 85% of the participants supported the idea of using a knowledge base for this purpose.

[8] `https://goo.gl/NNm407`

[9] `https://www.fiduswriter.org/`

[10] `https://zenodo.org/record/57899`

[11] `http://openresearch.org/sparql`

[12] `https://goo.gl/L02UU5`

Objective Comparison Metrics	Data Import	Complex Queries
Time(s)	32.6	0.31
Memory (MB)	24.44	2.89
Number of pages	100	n/a
Number of queries	n/a	10

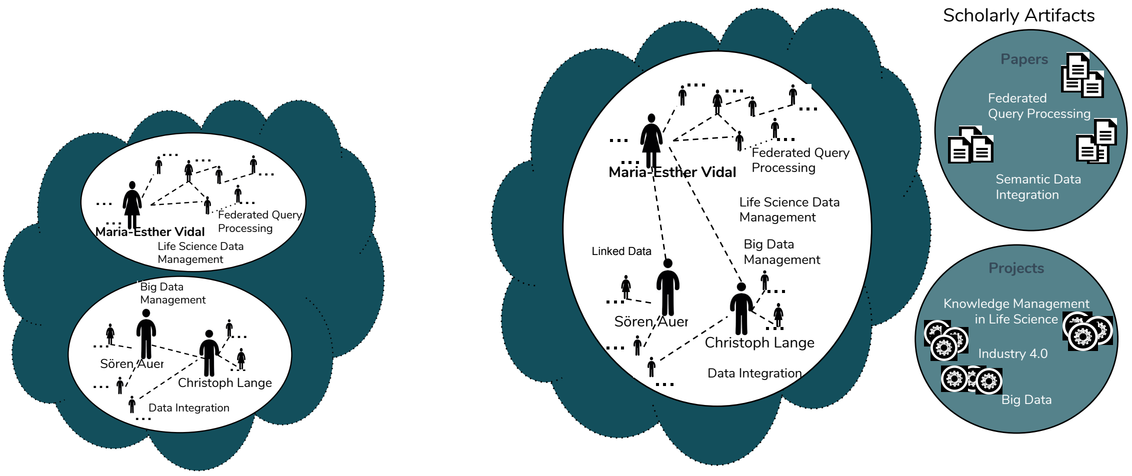
Usability survey: We asked users to tell us about their experience wrt. the ease and usability of the system[13]. Overall 12 users participated in the survey; they have had several roles in scientific events (participant, PC member, event organizer and keynote speaker). 75% of the users replied they had basic knowledge about wikis in general, however, half of them did not know about SMW. 66% got familiarized easily with OpenResearch which shows its suitability for researchers of different fields. Again 66% answered that they needed less than 5 minutes to add a single event which is relatively low time wrt. the time organizers need to announce their event in several channels. The average number of single events created by individual users is 10. More than half of the participants needed less than 5 minutes for a bulk upload.The participants largely agreed that these times are reasonable.

Performance measurement: Currently, OpenResearch is running on a Debian server at the University of Bonn with 8 GB of RAM allocated. By private invitation (OpenResearch has not yet been publicly announced at a large scale), 70 users have been added during the last two months. Above 300 events have been added by the users during last two months and several bulk uploads of data are performed every week by the admins; each time 100 pages were created. The measured time for bulk import varies with the content of CfPs and reduces when events exist already in the system. The table below shows a performance measurements of OR w.r.t. the average time and memory usage for several bulk imports and complex queries running over the event query form.

6.2 Mining

Knowledge semantically represented in knowledge graphs can be exploited to solve a broad range of problems in the respective domain. For example, in scientific domains, such as bio-medicine or, on the meta level, scholarly communication, or even in industry, knowledge graphs enable not only the description of the meaning of data, but the integration of data from heterogeneous sources and the discovery of previously unknown patterns. With the rapid growth in the number of publications, scientific groups, and research topics, the availability of scholarly datasets has considerably increased. This generates a great challenge for researchers, particularly, to keep track of new published scientific results and potential future co-authors. To alleviate the impact of the explosion of scholarly data, knowledge graphs provide a formal framework where scholarly datasets can be integrated and diverse knowledge-driven tasks can be addressed. Nevertheless, to exploit the semantics encoded in such knowledge graphs, a deep analysis of the graph structure as well as the semantics of the represented relations, is required. There have been several attempts considering both of these aspects. However, the majority of previous approaches rely on the topology of the graphs and usually omit the encoded meaning of the data. Most of such approaches are also mainly applied on special graph topologies, e.g., ego networks rather than general knowledge graphs. To provide an effective solution to the problem of representing scholarly data in knowledge graphs, and exploiting them to effectively support knowledge-driven tasks such as pattern

[13] `https://goo.gl/HIIeEh`



((a)) Researchers working on similar topics were in two co-authorship communities.

((b)) Researchers working on similar topics constitute a co-authorship community and produce a large number of scholarly artifacts.

Figure 6.2: **Motivating Example**. Co-authorship communities from the Semantic Web area working on data-centric problems. Researchers were in different co-authorship communities (2016) (a) started a successful scientific collaboration in 2016 (b), and as a result, produced a large number of scholarly artifacts.

discovery, we propose Korona, a knowledge-driven framework for scholarly knowledge graphs. Korona enables both the creation of scholarly knowledge graphs and knowledge discovery. Specifically, Korona resorts to community detection methods and semantic similarity measures to discover hidden relations in scholarly knowledge graphs. We have empirically evaluated the performance of Korona in a knowledge graph of publications and researchers from the Semantic Web area. As a proof of concept, we studied the accuracy of identifying co-author networks. Further, the predictive capacity of Korona has been analyzed by members of the Semantic Web area. Experimental outcomes suggest the next conclusions:

- Korona identifies co-author networks that include researchers that both work on similar topics, and attend and publish in the same scientific venues.
- Korona allows for uncovering scientific relations among researchers of the Semantic Web area.

The contributions of this paper are as follows:

- A scholarly knowledge graph integrating data from DBLP datasets;
- The Korona knowledge-driven framework, which has been implemented on top of two graph partitioning tools, semEP [203] and METIS [132], and relies on semantic similarity to identify patterns in a scholarly knowledge graph;
- Collaboration suggestions based on co-author networks; and
- An empirical evaluation of the quality of Korona using semEP and METIS.

6.2.1 Unveiling Scholarly Communities over Knowledge Graphs

Motivating Example We motivate the problem of knowledge discovery. We present an example of co-authorship relation discovery between researchers working on data-centric problems in the Semantic Web area. We checked the Google Scholar profiles of three researchers between 2015 and 2017, and compared

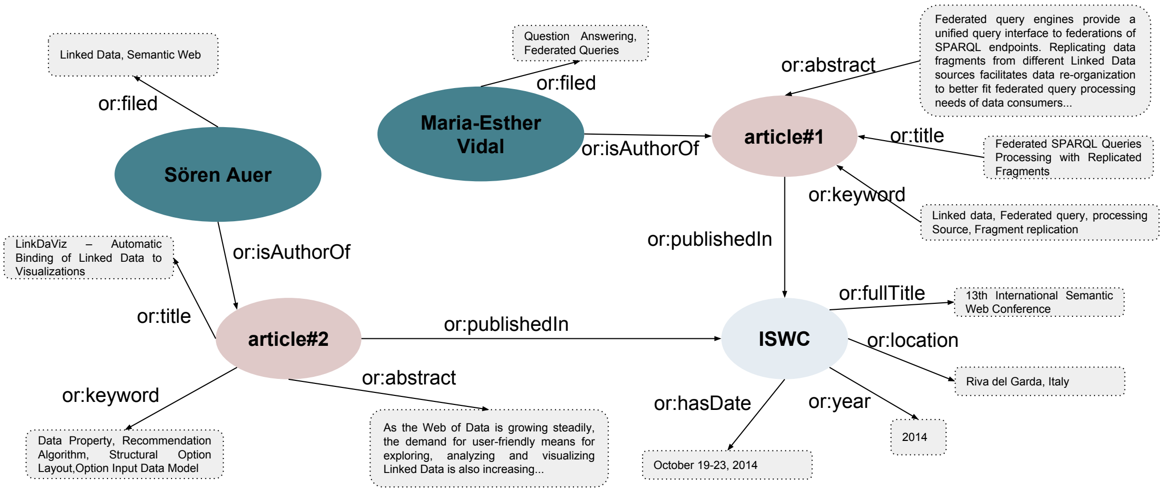

Linked Data, Semantic Web
or:filed
Sören Auer
or:isAuthorOf
LinkDaViz – Automatic Binding of Linked Data to Visualizations
or:title
article#2
or:keyword
Data Property, Recommendation Algorithm, Structural Option Layout,Option Input Data Model
or:abstract
As the Web of Data is growing steadily, the demand for user-friendly means for exploring, analyzing and visualizing Linked Data is also increasing...
or:publishedIn
Question Answering, Federated Queries
or:filed
Maria-Esther Vidal
or:isAuthorOf
or:abstract
Federated query engines provide a unified query interface to federations of SPARQL endpoints. Replicating data fragments from different Linked Data sources facilitates data re-organization to better fit federated query processing needs of data consumers...
article#1
or:title
Federated SPARQL Queries Processing with Replicated Fragments
or:keyword
Linked data, Federated query, processing Source, Fragment replication
or:publishedIn
ISWC
or:fullTitle
13th International Semantic Web Conference
or:location
Riva del Garda, Italy
or:hasDate
October 19-23, 2014
or:year
2014

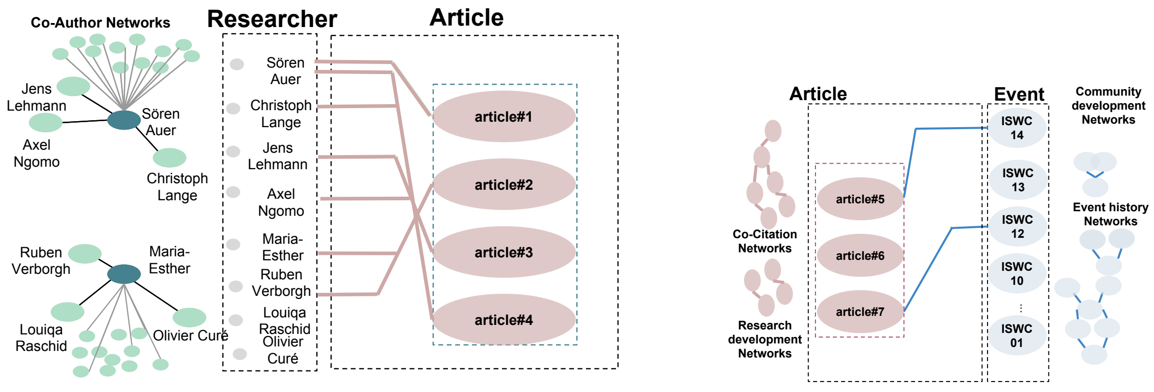
Figure 6.3: **Korona Knowledge Graph.** Scholarly entities and relations.

their networks of co-authorship. By 2016, Sören Auer and Christoph Lange were part of the same research group and wrote a large number of joint publications. Similarly, Maria-Esther Vidal, also working on data management topics, was part of a co-authorship community. Figure Figure 6.2(b) illustrates the two co-authorship communities, which were confirmed by the three researchers. After 2016, these three researchers started to work in the same research lab, and a large number of scientific results, e.g., papers and projects, was produced. An approach able to discover such potential collaborations automatically would allow for the identification of the best collaborators and, thus, for maximizing the success chances of scholars and researchers working on similar scientific problems. In this paper, we rely on the natural intuition that successful researchers working on similar problems and producing similar solutions can collaborate successfully, and propose Korona, a framework able to discover unknown relations between scholarly entities in a knowledge graph. Korona implements graph partitioning methods able to exploit semantics encoded in a scholarly knowledge graph and to identify communities of scholarly entities that should be connected or related.

Preliminaries The definitions required to understand our approach are presented in this sectiom. First, we define a scholarly knowledge graph as a knowledge graph where nodes represent scholarly entities of different types, e.g., publications, researchers, publication venues, or scientific institutions, and edges correspond to an association between these entities, e.g., co-authors or citations. Scholarly Knowledge Graph. Let U be a set of RDF URI references and L a set of RDF literals. Given sets V_e and V_t of scholarly entities and types, respectively, and given a set P of properties representing scholarly relations, a scholarly knowledge graph is defined as $\mathcal{SKG}$=$(V_e \cup V_t, E, P)$, where:

- Scholarly entities and types are represented as RDF URIs, i.e., $V_e \cup V_t \subseteq U$;
- Relations between scholarly entities and types are represented as RDF properties, i.e., $P \subseteq U$ and $E \subseteq (V_e \cup V_t \times P \times V_e \cup V_t \cup L)$

Figure 6.3 shows a portion of a scholarly knowledge graph describing scholarly entities, e.g., papers, publication venues, researchers, and different relations among them, e.g., co-authorship, citation, and collaboration.



((a)) Network of Researchers and Articles.

((b)) Networks of Events and Articles.

Figure 6.4: **Scholarly networks**. (a) Co-authors networks from researchers and articles.(b) Co-citation networks from discovered from events and articles.

Co-author Network. A co-author network $\mathcal{CAN}$=(V_a, E_a, P_a) corresponds to a subgraph of $\mathcal{SKG}$=$(V_e \cup V_t, E, P)$, where

- Nodes are scholarly entities of type *researcher*,

$$V_a = \{a \mid (a \textit{ rdf:type :Researcher}) \in E\}$$

- Researchers are related according to co-authorship of scientific publications, $E_a = \{(a_i \textit{ :co-author } a_j) \mid \exists p \,.\, a_i, a_j \in V_a \land (a_i \textit{ :author } p) \in E \land (a_j \textit{ :author } p) \in E \land (p \textit{ rdf:type :Publication}) \in E\}$

Figure 6.4 shows scholarly networks that can be generated by Korona. Some of these networks are among the recommended applications for scholarly data analytics in [298]. However, the focus on this work is on co-author networks.

Problem Statement Let $\mathcal{SKG}'$=$(V_e \cup V_t, E', P)$ and $\mathcal{SKG}$=$(V_e \cup V_t, E, P)$ be two scholarly knowledge graphs, such that $\mathcal{SKG}'$ is an *ideal* scholarly knowledge graph that contains all the *existing and successful relations* between scholarly entities in V_e, i.e., an oracle that knows whether two scholarly entities should be related or not. $\mathcal{SKG}$=$(V_e \cup V_t, E, P)$ is the *actual* scholarly knowledge graph, which only contains a portion of the relations represented in $\mathcal{SKG}'$, i.e., $E \subseteq E'$; it represents those relations that are known and is not necessarily complete. Let $\Delta(E', E) = E' - E$ be the set of relations existing in the ideal scholarly knowledge graph $\mathcal{SKG}'$ that are not represented in the actual scholarly knowledge graph $\mathcal{SKG}$. Let $\mathcal{SKG}_{\text{comp}}$=$(V_e \cup V_t, E_{\text{comp}}, P)$ be a *complete* knowledge graph, which includes a relation for each possible combination of scholarly entities in V_e and properties in P, i.e., $E \subseteq E' \subseteq E_{\text{comp}}$. Given a relation $e \in \Delta(E_{\text{comp}}, E)$, the problem of discovering scholarly relations consists of determining whether $e \in E'$, i.e., whether a relation r=$(e_i \; p \; e_j)$ corresponds to an existing relation in the ideal scholarly knowledge graph $\mathcal{SKG}'$.

In this research work, we specifically focus on the problem of discovering *successful co-authorship relations* between researchers in scholarly knowledge graph $\mathcal{SKG}$=$(V_e \cup V_t, E, P)$. Thus, we are interested in finding the co-author network $\mathcal{CAN}$=(V_a, E_a, P_a) composed of the maximal set of relationships or

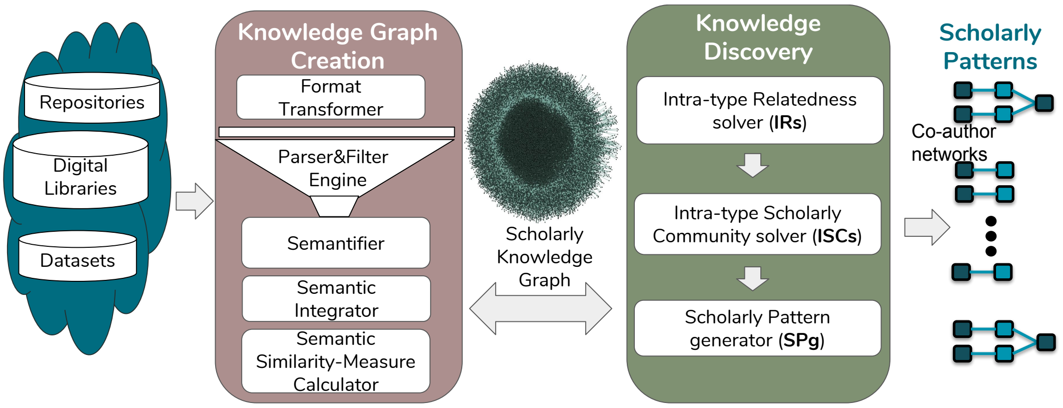


Figure 6.5: **The Korona Architecture**. Korona receives scholarly datasets and outputs scholarly patterns, e.g., co-author networks. First, a scholarly knowledge graph is created. Then, community detection methods and similarity measures are used to compute communities of scholarly entities and scholarly patterns.

edges that belong to the ideal scholarly knowledge graph, i.e., the set E_a in $\mathcal{CAN}$ that corresponds to a solution of the following optimization problem:

$$_{E_a \subseteq E_{comp}} |E_a \cap E'| \tag{6.1}$$

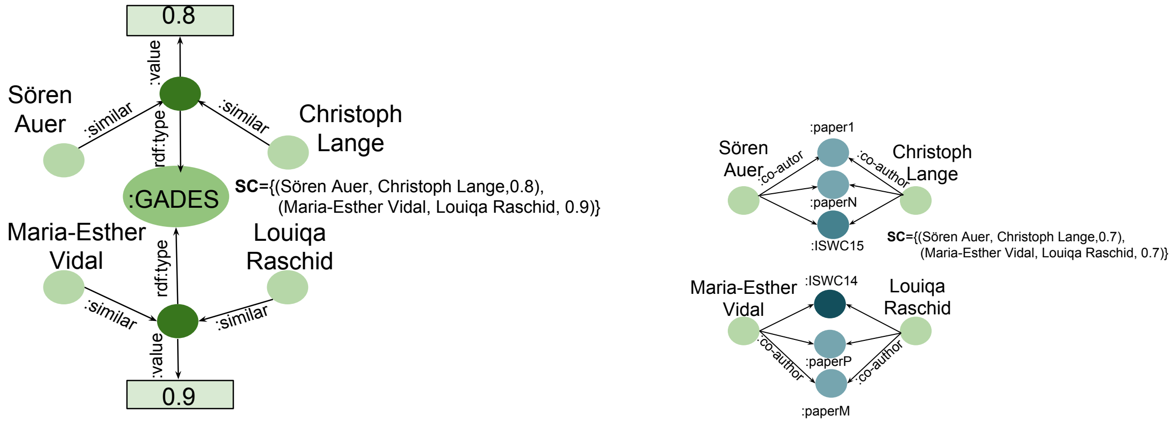
6.2.2 Discovering Hidden Relations in the Knowledge Graph

We propose Korona to solve the problem of discovering meaningful co-authorship relations between researchers in scholarly knowledge graphs. Korona relies on information about relatedness between researchers to identify communities composed of researchers that work on similar problems and publish in similar scientific events. Korona is implemented as an unsupervised machine learning method able to partition a scholarly knowledge graph into subgraphs or communities of co-author networks. Moreover, Korona applies the *homophily* prediction principle over the communities of co-author networks to identify successful co-author relations between researchers in the knowledge graph. The *homophily* prediction principle states that similar entities tend to be related to similar entities [167]. Intuitively, the application of the *homophily* prediction principle enables Korona to relate two researchers r_i and r_j whenever they work on similar research topics or publish in similar scientific venues. The relatedness or similarity between two scholarly entities, e.g., researchers, research topics, or scientific venues, is represented as RDF properties in the scholarly knowledge graph. Semantic similarly measures, e.g., GADES [226] or Doc2Vec [162], are utilized to quantify the degree of relatedness between two scholarly entities. The identified degree shows the relevance of entities and returns the most related ones.

Figure 6.5 depicts the Korona architecture; it implements a knowledge-driven approach able to transform scholarly data ingested from publicly available data sources into patterns that represent discovered relationships between researchers. Thus, Korona receives scholarly data sources and outputs co-author networks; it works in two stages:

- Knowledge graph creation and
- Knowledge graph discovery.

During the knowledge graph creation stage, a semantic integration pipeline is followed in order to create a scholarly knowledge graph from data ingested from heterogeneous scholarly data sources. It utilizes mapping rules between the Korona ontology and the input data sources to create the scholarly knowledge graph. Additionally, semantic similarity measures are used to compute the relatedness between scholarly



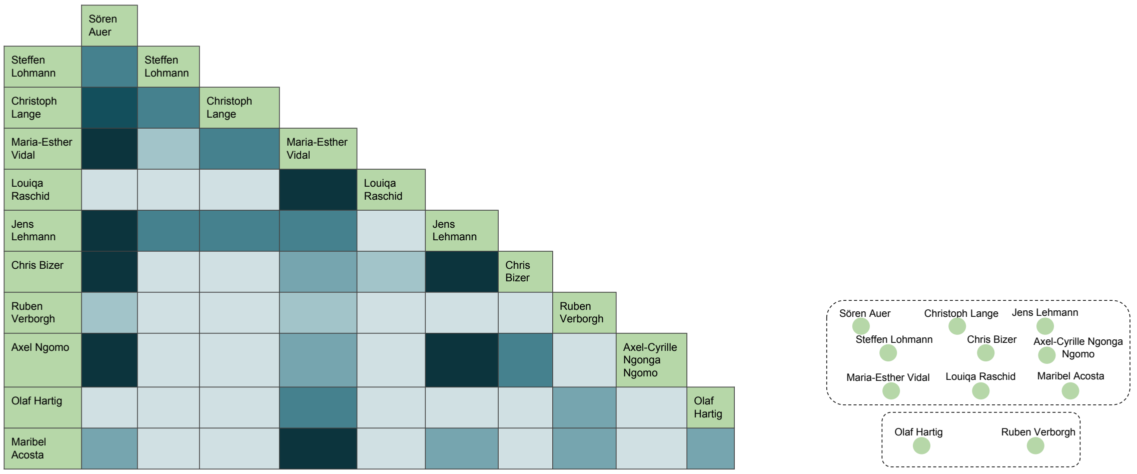
((a)) Similarity-based Relatedness

((b)) Path-based Relatedness

Figure 6.6: **Intra-type Relatedness solver (IRs)**. Relatedness across scholarly entities. (a) Relatedness is computed according to the values of a semantic similarity metrics, e.g., GADES. (b) Relatedness is determined based on the number of paths between two scholarly entities, i.e., Sören Auer and Christoph Lange, and Maria-Esther Vidal and Louiqa Raschid have same values of relatedness.

entities; the results are explicitly represented in the knowledge graph as scores in the range of 0.0 and 1.0. The knowledge graph creation stage is executed offline and enables the integration of new entities in the knowledge graph whenever the input data sources change. On the other hand, the knowledge graph discovery step is executed *on the fly* over an existing scholarly knowledge graph. During this stage, Korona executes three main tasks:

- Intra-type Relatedness solver (**IRs**);
- Intra-type Scholarly Community solver (**IRSCs**); and
- Scholarly Pattern generator (**SPg**).



((a)) Relatedness Across Researchers

((b)) Communities of Researchers

Figure 6.7: **Intra-type Relatedness solver (IRs)**.Communities of similar researchers are computed. (a) The tabular representation of $\mathcal{SC}$; lower and higher values of similarity are represented by lighter and darker colors, respectively. (b) Two communities of researchers; each one includes highly similar researchers.

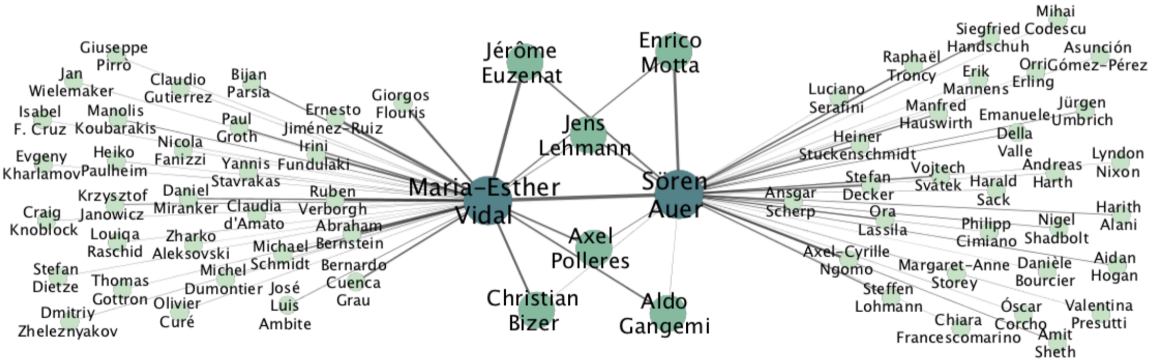


Figure 6.8: **Co-author network.** A network generated from scholarly communities.

Intra-type Scholarly Community solver (IRSCs). Once the relatedness between the scholarly entities has been computed, communities of highly related scholarly entities are determined. **IRSCs** resorts to unsupervised methods such as METIS or semEP, and to relatedness values stored in $\mathcal{SC}$, to compute the scholarly communities. Figure 6.7 depicts scholarly communities computed by **IRSCs** based on similarity values; as observed, each community includes researchers that are highly related; for readability, $\mathcal{SC}$ is shown as a heatmap where lower and higher values of similarity are represented by lighter and darker colors, respectively. For example, in FigureFigure 6.7(a), Sören Auer, Christoph Lange, and Maria-Esther Vidal are quite similar, and they are in the same community.

Scholarly Pattern generator (SPg). **SPg** receives communities of scholarly entities and produces a network, e.g., a co-author network. **SPg** applies the *homophily* prediction principle on the input communities, and connects the scholarly entities in one community in a network. Figure 6.8 shows a co-author network computed based on a scholarly knowledge graph created from DBLP; as observed, Sören Auer, Christoph Lange, and Maria-Esther Vidal are included in the same co-author network. In addition to computing the scholarly networks, **SPg** scores the relations in a network and computes the *weight of connectivity* of a relation between two entities. For example, in Figure 6.8, thicker lines represent strongly connected researchers in the network. **SPg** can also filter from a network the relations labeled with higher values of weight of connectivity. All the relations in a network correspond to solutions to the problem of discovering *successful co-authorship relations* defined in Equation 6.1. To compute the weights of connectivity, **SPg** considers the values of similarity of the scholarly entities in a community C; weights are computed as aggregated values using an aggregation function $f(.)$, e.g., average or triangular norm. For each pair (e_i, e_j) of scholarly entities in C, the weight of connectivity between e_i and e_j, $\phi(e_i, e_j \mid C)$, is defined as:

$$\phi(e_i, e_j \mid C) = \{f(score) \mid e_z, e_q \in C \wedge (e_z, e_q, score) \in \mathcal{SC}\}$$

Empirical Evaluation

Knowledge Graph Creation A scholarly knowledge graph has been crafted using the DBLP collection (7.83 GB in April 2017[14]); it includes researchers, papers, and publication year from the International Semantic Web Conference (ISWC) 2001–2016. The knowledge graph also includes similarity values between researchers who have published at ISWC (2001–2017). Let PC_{e_i} and PC_{e_j} be the number of papers published by researchers e_i and e_j together (as co-authors), respectively at ISWC (2001–2017). Let

[14] http://dblp2.uni-trier.de/e55477e3eda3bfd402faefd37c7a8d62/

TP_{e_i} and TP_{e_j} be the total number of papers they each have in all conferences of the scholarly knowledge graph. The similarity measure is defined as:

$$SimR(e_i, e_j) = \frac{PC_{e_i} \cap PC_{e_j}}{TP_{e_i} \cup TP_{e_j}}$$

The similarities between ISWC (2002–2016) are represented as well. Let RC_i and RC_j the number of the authors with papers published in conferences c_i and c_j respectively. The similarity measure corresponds to:

$$SimC(c_i, c_j) = \frac{RC_i \cap RC_j}{RC_i \cup RC_j}$$

Thus, the scholarly knowledge graph includes both scholarly entities enriched with their values of similarity.

6.2.3 Experimental Study

The effectiveness of Korona has been evaluated in terms of the quality of both the generated communities of researchers and the predicted co-author networks.

The assessment is done in order to answer two questions:

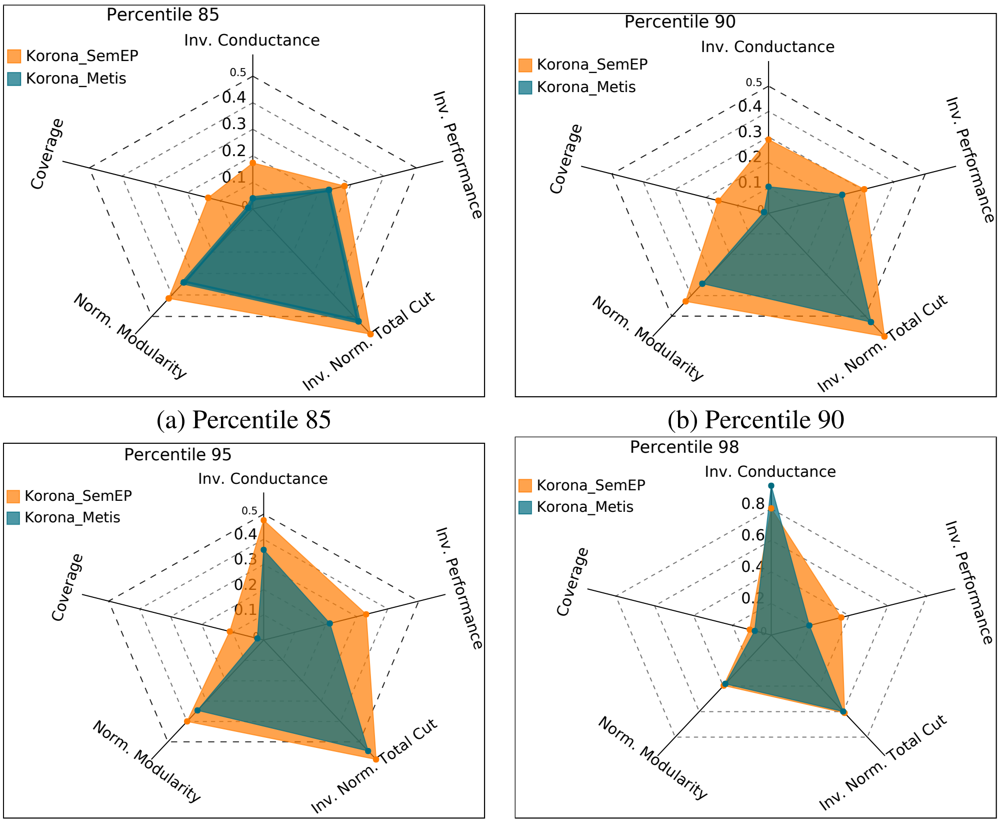
- Does the semantics encoded in scholarly knowledge graphs impact the quality of scholarly patterns?
- Does the semantics encoded in scholarly knowledge graph allow for improving the quality of the predicting co-author relations?

Evaluation metrics: Let $Q = \{C_1, \dots C_n\}$ be the set of communities obtained by Korona: *Conductance*: measures relatedness of entities in a community, and how different they are to entities outside the community [90]. The inverse of the conductance $1 - Conductance(S)$ is reported. *Coverage*: compares the fraction of intra-community similarities among entities to the sum of all similarities among entities [90]. *Modularity*: is the value of the intra-community similarities among the entities divided by the sum of all the similarities among the entities, minus the sum of the similarities among the entities in different communities, in the case they were randomly distributed in the communities [195]. The value of the modularity lies in the range $[0.5, 1]$, which can be scaled to $[0, 1]$ by computing:

$$\frac{Modularity(Q) + 0.5}{1.5}$$

Performance: sums the number of intra-community relationships, plus the number of non-existent relationships between communities [90]. *Total Cut*: sums all similarities among entities in different communities [42]. Values of total cut are normalized by dividing by the sum of the similarities among the entities; inverse values are reported, i.e., $1 - NormTotalCut(Q)$.

Experiment 1: Evaluation of the Quality of Collaboration Patterns. Prediction metrics are used to evaluate the quality of the communities generated by Korona using METIS and semEP; relatedness of the researchers is measured in terms of *SimR* and *SimC*. Communities are built according to different similarity criteria; percentiles of 85, 90, 95, and 98 of the values of similarity are analyzed. For example, in percentile 85 only 85% of all similarity values among entities have scores lower than the similarity value in the percentile 85. Figure 6.9 presents the results of the studied metrics. In general, in all percentiles, the communities include closely related researchers. However, both implementations of Korona exhibit quite



(a) Percentile 85

(b) Percentile 90

(c) Percentile 95

(d) Percentile 98

Figure 6.9: **Quality of Korona.** Communities evaluated in terms of prediction metrics (higher values are better); percentiles 85, 90, 95, and 98 are reported. Korona exhibits the best performance at percentile 95 and groups similar researchers according to research topics and events where they publish.

good performance at percentile 95, and allow for grouping together researchers that are highly related in terms of the research topics on which they work, and the events where their papers are published.

Experiment 2: Survey of the Quality of the Prediction of Collaborations among Researchers. Results of an online survey[15] among 10 researchers are reported; half of the researchers are from the same research area, while the other half was chosen randomly. Knowledge subgraphs of each of the participants are part of the Korona research knowledge graph; predictions are computed from these subgraphs. The predictions for each were laid out in an online spreadsheet along with 5 questions and a comment emph.

Table 6.1 lists the five questions that the survey participants were asked to validate the answers, while Table 6.2 reports on the results of the study. The analysis of results suggests that Korona predictions represent potentially *successful co-authorship relations*; thus, they provide a solution to the problem tackled in this paper.

There are several related works in this regard. Xia, Wang, Bekele and Liu [298] provides a comprehensive survey of tools and technologies for scholarly data management, as well as a review of data analysis techniques, e.g., social networks and statistical analysis. However, all the proposals have been made over raw data and knowledge-driven methods were not considered. Wang, Xu, Wu and Zhou [282] present a comprehensive survey of link prediction in social networks, while Paulheim [207] presents a

[15] https://bit.ly/2ENEg2G

Q1. Do you know this person? Have you co-authored before? To avoid confusion, the meaning of "knowing" was kept simple and general. The participants were asked to only consider if they were aware of the existence of the recommended person in their research community.

Q2. Have you co-authored "before" with this person at any event of the ISWC series? With the same intent of keeping the survey simple, all types of collaboration on papers in any edition of this event series were considered as "having co-authored before".

Q3. Have you co-authored with this person after May 2016? Our study considered scholarly metadata of publications until May 2016. The objective of this question was to find out whether a prediction had actually come true, and the researchers had collaborated.

Q4. Have you ever planned to write a paper with the recommended person and you never made it and why? The aim is to know whether two researchers who had been predicted to work together actually wanted to but then did not and the reason, e.g., geographical distance.

Q5. On a scale from 1–5, (5 being most likely), how do you score the relevance of your research with this person? The aim is to discover how close and relevant are the collaboration recommendations to the survey participant.

Table 6.1: **Survey**. Questions to validate the recommended collaborations.

Korona	%	Precision	Q.1(a)	Q.1(b)	Q.2	Q.3	Q.4	Q.5
Korona-METIS	85	0.19	4.37/2.62	1.57/6.42	0.14/6.8	0.5/7.0	0.42/6.85	3.11
Korona-semEP	85	0.08	6.0/3.0	0.71/8.14	1.57/5.16	0.0/6.2	0.85/6.85	3.35
Korona-METIS	90	0.04	2.7/2.5	0.33/6.66	0.0/6.75	0.0/7.33	0.2/7.2	2.97
Korona-semEP	90	0.04	5.0/1.0	0.33/7.5	0.0/7.5	0.0/7.5	1.8/6.2	3.24
Korona-METIS	95	0.09	4.25/1.62	0.8/7.6	0.0/7.0	0.0/7.0	0.5/7.25	3.08
Korona-semEP	95	0.11	3.0/0.65	0.5/4.0	0.2/3.5	0.2/3.5	0.6/4.4	3.84

Table 6.2: **Survey results.** Precision and aggregated results with a standard division of answers for each question validating the recommended collaborations.

survey of methodologies used for knowledge graph refinement; both works show the importance of the problem of knowledge discovery. Traverso-Ribón, Palma, Flores and Vidal [265] introduces a relation discovery approach, $\mathcal{KOI}$, able to identify hidden links in TED talks; it relies on heterogeneous bipartite graphs and on the link discovery approach proposed in [203]. In this work, Palma, Vidal and Raschid present semEP, a semantic-based graph partitioning approach, which was used in the implementation of Korona-semEP. Graph partitioning of semEP is similar to $\mathcal{KOI}$ with the difference of only considering isolated entities, whereas $\mathcal{KOI}$ is desired for ego networks. However, it is only applied to ego networks, whereas Korona is mainly designed for knowledge graphs. Sachan and Ichise [231] propose a syntactic approach considering dense subgraphs of a co-author network created from the DBLP dataset. They discover relations between authors and propose pairs of researchers belonging to the same community. A link discovery tool is developed for the biomedical domain by Kastrin, Rindflesch and Hristovski [133]. Albeit effective, these approaches focus on the graph structure and ignore the meaning of the data.

6.3 Query Analysis

6.3.1 Analysis on OpenResearch.org

On top of the basic architectural layers, OpenResearch offers services for different stakeholders of scientific communication. As a semantic wiki, it offers initial LOD services and semantic representation of metadata about events. We address the issues discussed in section 6.1.1 by establishing a set of quality metrics for scientific events and implementing them as properties. We adopt the definition of quality as *fitness for use*, which, here, means the extent to which the specification of an event satisfies its stakeholders [127, 144]. In the remainder of this section, the current services are explained in three categories: wiki pages, LOD services and queries.

Semantic Wiki Pages SMW powers OpenResearch to provide semantic representation of CfPs as one wiki page per event. In OpenResearch, specific semantic forms have been designed for each type of entities to make content creation and revision as easy as possible for users.

```
{{Event series
|Acronym=ESWC
|Title=Extended Semantic Web Conference
|has Twitter=@eswc_conf
|has CORE2014 Rank=A
|has CORE2017 Rank=A
|has CORE2018 Rank=A
|Field=Semantic Web
|Homepage=eswc-conferences.org
|has Bibliography=dblp.uni-trier.de/db/conf/esws/}}
[[Category:Conference series]]
```

Listing 6.1: **Metadata Representation on OR.** Metadata of the ESWC conference series.

Properties of each semantic object are populated via fields in these semantic forms. The following example shows the generated SMW wiki markup containing general information about an event. Further information about committee members, extensions and other important dates can also be provided in other parts of the form. The complete textual representation of the CfPs can also be added as content of the wiki page with embedded semantic annotations.

All data created within OpenResearch is published as Linked Open Data (LOD). In the sequel, we describe ways for accessing OpenResearch LOD. Afterwards, we outlines how the LOD approach enables building further services on top by sketching two possible ways of consuming the OpenResearch LOD: interlinking with relevant datasets, and using OpenResearch LOD as external plug-in for the Fidus Writer scientific authoring platform[16].

Implementation of the defined metrics and dimensions has been done with an on-demand decision making process. Some of the metrics suited to be defined as a raw property. The derived metrics have been computed by queries over the data (using MediaWiki expressions) for example acceptance rate := accepted/submitted; average acceptance rate over series):

$$AcceptanceRate = \frac{No.\ accepted\ papers}{No.\ submissions}$$

The implementation of this composite that can be calculated from the raw properties has been done in the template of the corresponding entity(here event):

```
{{#ifeq:{
{{Submitted papers|}}|||{{Tablerow|Label=Papers:
|Value=Submitted([[Submitted papers::{{{Submitted papers}}}]])/
   ([[Accepted[[Accepted papers:={{{Accepted papers}}}]]])
([[Acceptance rate::{{#expr:{{{Accepted papers}}}/
                {{{Submitted papers}}}
                * 100 round 1}}]]])\%}}}}
```

[16] https://www.fiduswriter.org/

Accessing OpenResearch LOD An updated version of the OpenResearch dataset is produced daily and available for download and query.[17]. The data is also queryable via a SPARQL endpoint[18]. In addition, the semantic representation of the metadata for each event is represented as an RDF feed in each page. The RDF feed for the EKAW 2016 resource is available at `http://openresearch.org/Special:ExportRDF/EKAW_2016`. To expose dereferenceable resources conforming with Linked Data best practices, the URI resolver provides URIs with content negotiation; e.g., for the EKAW 2016 resource the URI is `http://openresearch.org/Special:URIResolver/EKAW_2016`.

To support the creation of various views, recommendations and ranked lists (by quality indicators), queries can be defined and executed using all defined properties and classes and the results can be embedded in wiki pages. For example, events can be ranked by acceptance rate using the corresponding properties in queries:

```
{{#ask:[[Category:Event]]
 | ?title = Name
 | ?Event in series = Series
 | ?Category | ?Acceptance rate
 | format = table
 | limit=10
 | sort=Acceptance rate
 | order=desc
}}
```

Listing 6.2: **ASK Query on OR.** Top 10 event series sorted with their acceptance rate.

It is also possible to capture the relationships between various types of entities (e.g. event series, sub/super events, roles of a person in event organization, etc.). Many popular views have been implemented in OpenResearch as pre-defined queries. Various display formats provided by SMW extensions are used to visualize the query results. Figure 6.10 shows a map view of the upcoming events using location-based filtering. Similarly, calendar and timeline views show upcoming submission and notification deadlines as well as the events themselves.

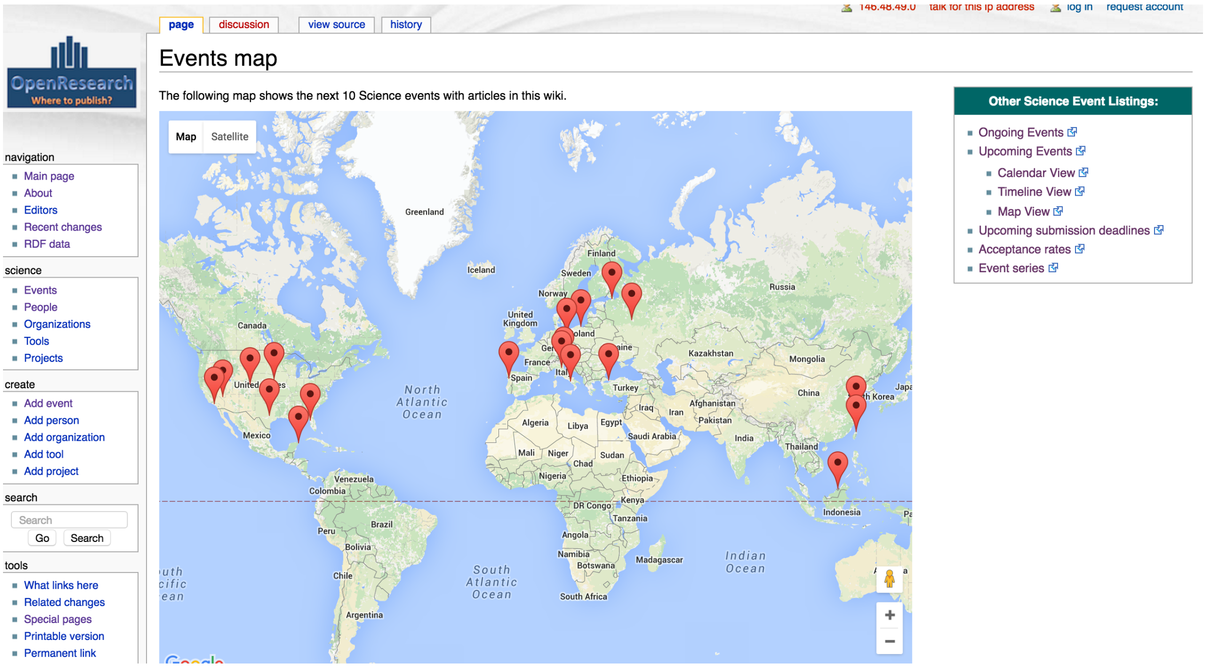
In addition, taking, for example, participation figures into account enables new indicators for measuring the quality and relevance of research that are not just based on citation counts [121]. Based on semantically enriched indicators, predefined SPARQL queries as well as form-based search facilities will be implemented for recommendation services.

Integration with an Authoring Platforms In this section we introduce our approach to improve the workflow of authoring processing [177]. The OpenResearch LOD will be plugged into the Fidus Writer authoring platform to improve the workflow in the following use cases:

1. *Venue recommendation:* One of the critical aspects in the process of writing and publishing is to find a suitable event to submit the scientific results. The OpenResearch dataset contains data about events annotated with corresponding scientific field as *:category* and keywords. We also annotate keywords from the content of the under-production scholarly document in the OSCOSS project that could be imported to the OpenResearch search services.For example, *Find all events in the computer science field that focus on data analysis, big data, knowledge engineering, linked data.* The result of queries can be shown to the authors with a user-friendly interface and filtering metrics such as deadline and location distance.

[17] `https://zenodo.org/record/57899`

[18] `http://openresearch.org/sparql`



Events map

The following map shows the next 10 Science events with articles in this wiki.

Figure 6.10: **Geographic Data with Dynamic.** Location-related distribution of upcoming events on a map view.

2. *Direct link to submission pages:* The OpenResearch data contains a property named *submission link* that provides a direct link to paper submission pages of events. The submission page of the targeted event can be made accessible easily from the authoring platform.

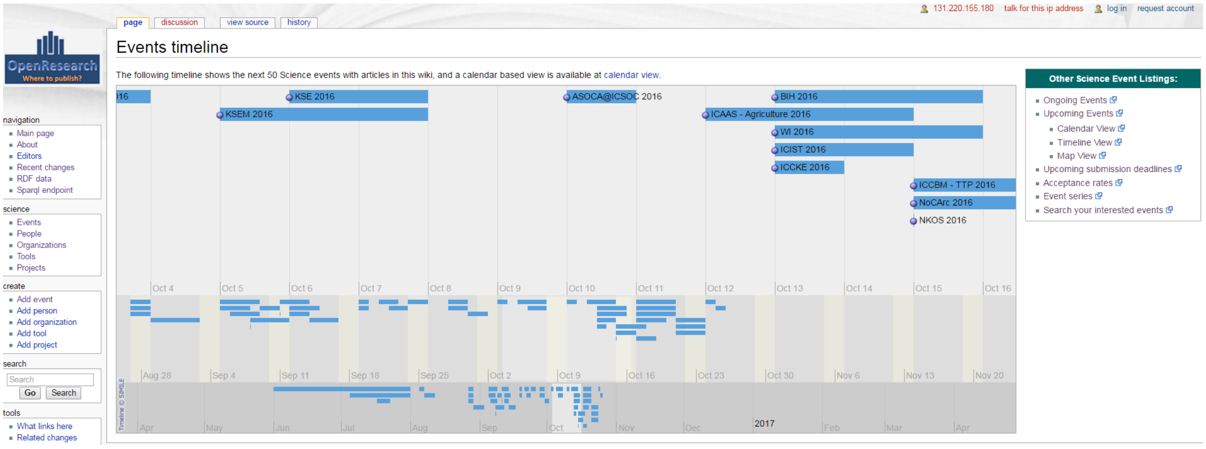
3. *Notification services:* there are different deadlines attached to the events that should be considered by authors such as abstract deadline, submission deadline or registration deadline as well as deadline extensions. Enabling notification services in the authoring platform will support both organizers and researchers.

The OR knowledge graph is built upon a combination of data captured by SAANSET and crowd sourcing, which is utilized to define and execute complex queries. Many convenient views, e.g., calendar view, map view, time line, have been implemented in OR as pre-defined queries. Various display formats provided by SMW extensions are also used to visualize the results of query. The following query utilizes the information captured by SAANSET (title, date, city, country, field, homepage) to answer a question that is otherwise cumbersome or impossible to answer for researchers.

Sample SPARQL queries are defined in the example page of OR[19]. The following visualization of metadata related to scholarly events have been implemented using MediWiki extensions. Many popular views have been implemented in OpenResearch as pre-defined queries. Various display formats provided by SMW extensions are used to visualize the query results. Figure 6.10 shows a map view of the upcoming events using location-based filtering. Users are enables to pin the locations and add them to their personal pages. Filtering can easily work in different granularity e.g., city, country, and continent.

Similarly, calendar and time line Figure 6.11 views show upcoming submission and notification deadlines as well as the events themselves. Every individual person or group can crate a list of events as agenda for a year. Listing 6.3 shows the corresponding query for these visualizations.

[19] http://openresearch.org/Sparql_endpoint/Examples



Events timeline

The following timeline shows the next 50 Science events with articles in this wiki, and a calendar based view is available at calendar view.

Figure 6.11: **Timeline View.** Time line of upcoming events.

```
SELECT ?event ?endDate ?startDate ?city ?country ?homepage
WHERE {
  ?e property:Has_location_country category:Germany.
  ?e rdfs:label ?event.
  ?e property:Has_location_city ?city.
  ?e property:Has_location_country ?country.
  ?country rdfs:subClassOf ?partContinent.
  ?partContinent rdfs:subClassOf ?continent.
  ?continent rdfs:isDefinedBy site:Category:Europe.
  ?e a category:Semantic_Web.
  ?e icaltzd:dtend ?endDate.
  ?e icaltzd:dtstart ?startDate.
  ?e foaf:homepage ?homepage.
  FILTER (?startDate >= "2018-01-01"^^xsd:date &&
          ?endDate < "2019-01-01"^^xsd:date).
} BINDINGS ?EventTypes {(smwont:ConferenceEvent)
                        (smwont:WorkshopEvent)}
```

Listing 6.3: **SPARQL Query on OR.** Upcoming events wrt. specific criteria

6.3.2 LOD Services for OpenAIRE.eu

One can use SPARQL queries to access results from the OpenAIRE LOD endpoint. Based on semantically enriched indicators, predefined SPARQL queries as well as form-based search facilities are implemented for OA LOD services. Listing 6.4 is an example of query to find the top 100 funders regarding the number of funded projects. Several other samples of the pre-defined queries can be found on the project homepage for LOD [20].

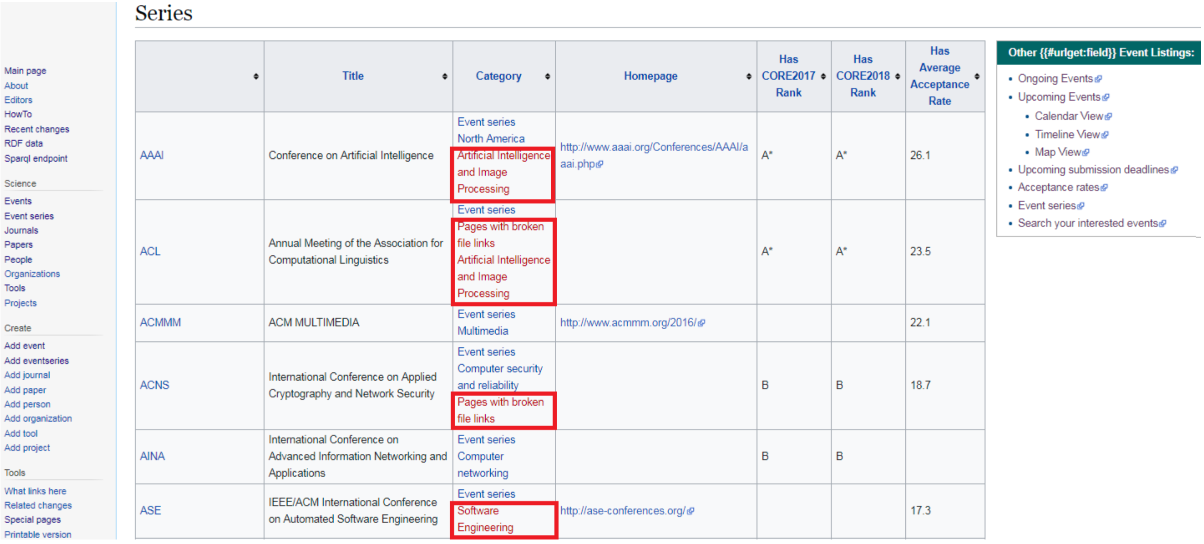
```
SELECT ?funder, count(?project) as ?number_of_projects
WHERE {
  ?project a oav:ProjectEntity.
  ?project oav:funder ?funder.}
GROUP BY ?funder
ORDER BY DESC (count (?project))
LIMIT 100
```

Listing 6.4: **SPARQL Query on OA LOD.** Top 100 funders regarding the number of funded projects.

[20] http://lod.openaire.eu/

6.3.3 Unknown Metadata Identification – Completing Cycle

Visualization of the already existing metadata shows which parts of the collection have not been harvested or integrated properly. For example, the MediaWiki differentiates the missing or not existing metadata and values in red color. Figure 6.12 shows the missing metadata and values for event series. The users of the wiki and the contributors can easily select the missing information and add them to the knowledge graph. Now only this way, there are other visualization techniques and tools to improve data comprehensiveness. In this way, the cycle is complete by going from *visualization* to the first step of *selection*.



Main page
About
Editors
HowTo
Recent changes
RDF data
Sparql endpoint

Science
Events
Event series
Journals
Papers
People
Organizations
Tools
Projects

Create
Add event
Add eventseries
Add journal
Add paper
Add person
Add organization
Add tool
Add project

Tools
What links here
Related changes
Special pages
Printable version

Series

	Title	Category	Homepage	Has CORE2017 Rank	Has CORE2018 Rank	Has Average Acceptance Rate
AAAI	Conference on Artificial Intelligence	Event series North America Artificial Intelligence and Image Processing	http://www.aaai.org/Conferences/AAAI/aaai.php	A*	A*	26.1
ACL	Annual Meeting of the Association for Computational Linguistics	Event series Pages with broken file links Artificial Intelligence and Image Processing		A*	A*	23.5
ACMMM	ACM MULTIMEDIA	Event series Multimedia	http://www.acmmm.org/2016/			22.1
ACNS	International Conference on Applied Cryptography and Network Security	Event series Computer security and reliability Pages with broken file links		B	B	18.7
AINA	International Conference on Advanced Information Networking and Applications	Event series Computer networking		B	B	
ASE	IEEE/ACM International Conference on Automated Software Engineering	Event series Software Engineering	http://ase-conferences.org/			17.3

Other {{#urlget:field}} Event Listings:

- Ongoing Events
- Upcoming Events
  - Calendar View
  - Timeline View
  - Map View
- Upcoming submission deadlines
- Acceptance rates
- Event series
- Search your interested events

Figure 6.12: **Timeline View.** Time line of upcoming events.

Use of a chart or graph to summarize complex data ensures faster comprehension of relationships than cluttered reports or spreadsheets. In order to draw insights from Big data, visualization tools can provide real-time information. The interactive visualization approaches can be employed in order to assist contributors of the platform to identify the missing metadata and their values. One benefit of big data visualization is that it allows users to track relations and entities. In addition, it supports finding correlations and dependencies between the metrics related to quality of artifacts. Data visualization is also used for monitoring key indicators. With the increasing use of machine learning approaches and tightening with visualization tools, more automatic interpretations can be generated from data.

CHAPTER 7

Conclusion and Outlook

In this thesis, we tackle the problem of integrating and managing heterogeneous scholarly metadata on the Web as explained in chapter 1. The aim was to facilitate services for scholarly communication by following the FAIR principles and making (meta)data Findable, Accessible, Interoperable, and Reusable. In order to systematically identify the needs of the stakeholders and the gaps of the already existing support, the development of the scholarly communication and the state-of-the-art of the current services were discussed in chapter 2. A structured methodology has been defined in chapter 3 for organizing management activities of the challenges. A set of distinct steps of actions proposed for the management of scholarly metadata, namely: Selection, Extraction, Transformation, Interlinking, Enrichment, Curation, Mining, Quality Assessment, Analysis, and Visualization. Overall, the research presented in this thesis has contributed to six of the nine steps. The *Selection* step as the starting point of the cycle together with the *Quality Assessment* (*fitness for use*) of the artifacts as research output, venues and stakeholders have been discussed in chapter 4. The other two steps of metadata acquisition and integration are *Transformation* and *Interlinking*, which are presented in chapter 5. As an important step for refinement of the metadata, *Curation*, is presented in chapter 6. The same chapter contains the results of utilization over the scholarly knowledge graph of the collected and curated metadata followed in the *Analysis* step. In this chapter, a conclusion of the contributions and results of each step are divided into the following subsection: summary and impact in section 7.1 and the future work in section 7.2 .

7.1 Summary and Impact

To the end of providing a comprehensive facilitation of scholarly metadata management using Semantic Web technologies, four different challenges have been identified:

Challenge 1: collecting and curating metadata from multiple distributed sources as well as communities

Challenge 2: integrating heterogeneous metadata resources

Challenge 3: assessing the quality of scholarly artifacts

Challenge 4: addressing the information needs of many different kinds of stakeholders

The research questions addressed in this thesis were derived from the challenges and the required metadata management activities. The contributions of this research are considered in the context of the three main projects in which the doctoral candidate was involved: OpenResearch.org platform, OpenAIRE.eu project and the SemPub challenge series. The main approach and technical product that stand in most of the life cycle steps is the OpenResearch.org platform. Therefore, each of the research questions and the contributions will be discussed with the main focus on this platform.

Research Question 1: How can we leverage semantic technologies to facilitate the acquisition and the collaborative curation of scholarly metadata?

The OpenResearch.org platform (OR) facilitates two options for metadata import: (1) creating individual pages using semantic forms and (2) bulk import of metadata collected from multiple sources. Semantic forms enable collaboratively collected and curated scholarly metadata by community members and domain experts as OR content contributors. The collected and curated metadata are directly stored in machine-readable and interoperable format. Thus, the contribution to the state-of-the-art in the area of scholarly knowledge creation and curation was reactivating[1] a crowdsourcing system with a representation of scientific events in semantic wiki pages. As a result, a knowledge graph of scholarly event metadata has been generated with the involvement of the research communities in the metadata acquisition and curation. These features have been discussed in section 6.1.

Currently, OpenResearch.org covers metadata of artifacts in the field of computer science. However, its user-friendly interface enables the involvement of other research communities (with less technical background) and provides an easy way of creating the semantic representation of metadata. By creating new semantic forms to collect metadata about new concepts, the platform is easily extendable and adaptable. This can reduce the effort that every community makes in order to collect and curate such scholarly metadata. OpenResearch.org has the potential to become a global gateway of scholarly metadata. In order to attract other communities to become users for OR, there is a need for more systematic engagement of stakeholders.

In the context of the SemPub challenge series, the tasks were designed to investigate creating an enriched metadata collection together with the community. Queries aimed at collection and curation of CEUR-WS.org dataset. For example, in task 2 the participants collected extra information by metadata extraction from PDF files and, in task 3, to curate metadata about the same entity in different datasets. However, the community involvement in the challenge was limited to a number of experts from the semantic web research domain.

Research Question 2: To what extent can we increase the coherence of scholarly communication artifacts by semi-automatic linking?

The OpenAIRE.eu (OA) information space contains metadata (at the time of writing this thesis) about 21 million publications, 606 thousand research dataset from 13 thousand content providers and 16 funders. Those repositories are harvested into an integrated metadata portal. As a preparatory technical step, the OpenAIRE data transformed to the RDF format. The metadata are represented in three formats in the OA information space (HBase, CSV, XML) and finding the best maintainable way of doing transformation was a challenging step. Several tools have been used for transforming each of these formats to RDF and finally, CSV to RDF was identified as the best performing method; see section 5.2. The RDFized version of the OA dataset was interlinked with a list of candidate datasets (e.g., DBLP, DBpedia, OpenCitations, WikiData). Based on the performance analysis, the LIMES link discovery tool was selected to interlink OpenAIRE LOD to these datasets. We achieved a high precision for interlinking on publications and organizations, whereas the interlinking of persons requires further improvement. Interlinking increased the coherence of OA dataset by adding data that have not (yet) been collected, data that are expensive

[1] OpenResearch.org was created in 2008 by Sören Auer (the main supervisor of this thesis), however, it was passive till 2015. With the contributions of this research, OR is reactivated and it is moving toward further steps.

to collect, and data that are out of the scope of the OA infrastructure; see section 5.3. For example, the OpenAIRE information space does not cover certain scholarly metadata, e.g., events, OCW which could be easily gained through interlinking processes rather than difficult metadata harvesting mechanisms.

Research Question 3: How can the quality of scholarly artifacts be assessed systematically?

As a preliminary work, a quality assessment framework was defined for OpneCourseWare with 10 dimensions and 36 quality metrics; see section 4.3. A sample of 100 courses was assessed with regard to quality (*fitness for use*) according to these criteria. The results showed the quality of OCW has to improve significantly in order to live up to its promises of being freely and easily accessible and usable for global communities.

The metadata on OR, which been mainly collected to provide user analytics, could also be used for quality assessment of scholarly artifacts, mainly events. Following a systematic methodology, a framework was proposed that allows for a differentiated quantification of the quality of scientific events and publications. More than 40 quality metrics in 20 dimensions were defined for scientific events; see section 4.4. The metrics have been defined under three categories: Event related metrics evaluate the extent to which the entities involved in the submission, acceptance, and organization process have an impact on the quality of the event. Person related metrics aim at measuring the extent to which the persons involved in an event have an impact on its quality. Bibliographic related metrics refer to the outcome of the review process of the research results published and presented in the event and their influence on their quality.

After defining the metrics, OR was used as an an executable platform in order to provide services for end-user. OR users can define derivative metrics by flexible combination of the quality metrics. By using the general pre-defined queries and customizing them to the specific values, the users are enabled to compute the quality of events by different weights. The metrics on OR are no longer limited to the already in-use metrics such as citation counts and take into account the perspective of other stakeholders. For example, in evaluating the quality of events, we take into account the reputation of the people involved in a conference, sponsorship, registration fee etc., not just the number of papers and citations they got.

In addition, through the SemPub challenge organization (see section 4.5), a list of queries have been defined for assessing the quality of scholarly artifacts, people, events, etc. The challenge was designed with a combination of information extraction and quality assessment tasks. For example, a query was asked to be designed by the participants to list the systems in a special events edition with the best precision or the worst recall. Such information can only be collected from the publications of an event. However, having them at hand opens up a new horizon of quality assessments for scholarly artifacts.

Research Question 4: What analytic services can fulfill the information needs of the stakeholders in scholarly communication and how can they be realized?

OpenResearch.org enables a comprehensive support for a diversity of the stakeholders in the whole system of scholarly communication having a variety of skills and information needs. For the passive users of OR, the metadata are represented with different visualizations such as a default metadata view in table format. This enables users to sort the information based on their column of interest. Some special metadata are visualized using Semantic MediaWiki extensions, e.g., the map extension for locations, calendar extensions for temporal information. More precisely, OR enables computing a numerical quality-related metric using underlying metadata (e.g., the internationality of an event) and visualize them in a way that satisfies user needs. For the users consuming information passively, OR makes it easier to understand insights from metadata and explore and assess analytics from different points of view.

Every member of the community, once becaming a user or contributor of the OR content, can freely reuse the underlying metadata. They can create customized wiki pages and add them to their own page. For example, a researcher can create a customized calendar of events with certain criteria and share it with other community members. Flexible and easy combination of the metrics provides innovative ways for assessments of artifacts with different information needs. OR follows the FAIR principles. Since the OpenResearch.org dataset is offered in a downloadable version together with a SPARQL endpoint, reusability of the datasets are high. This enables other communities with the objective of offering other analytical services to easily reuse the dataset.

Furthermore, we have shown a clustering example for co-authorship recommendation on a dataset sourced from DBLP and OR as a showcase of possible recommendation services and visual analytics in research of research section 6.2. The system was able to identify possible co-authorships for researchers of different scholarly domains. In an evaluation with participation of 10 high-profil researchers the results were validated with positive comments.

Impact on the Scientific Communities This thesis contributed a comprehensive facilitation of scholarly metadata management that enabled many benefits for stakeholders involved in scholarly communication. Table 7.1 summarizes the list of impact with regard to the platform or activity on which the contributions are done.

Impact	**OpenAIRE**	**OpenResearch**	**SemPub**	**OCW**
facilitating educational resource management	✗	✗	✗	✓
enabling a multidisciplinary aggregation of scholarly metadata	✓	✓	✗	✗
publishing FAIR metadata	✓	✓	✓	✓
fostering more efficient, effective metadata by community involvement	✗	✓	✓	✗
increasing efficiency of curating scholarly metadata	✗	✓	✓	✗
capturing hidden information from knowledge graph	✗	✓	✗	✗
providing quality assessment methods and analysis on top of enriched metadata	✗	✓	✓	✓
reducing redundant search effort for information in multiple sources	✓	✓	✓	✗
enabling comprehensive and flexible analysis of knowledge graphs	✓	✓	✗	✗

Table 7.1: **Impact.** Following the steps of the life cycle, a list of impact on scholarly communities are shows. The reported impact was achieved using the presented knowledge graphs as test bed platforms in this thesis.

7.2 Future Research Directions

The methodology of this research was a life cycle of scholarly metadata management. Some parts of the life cycle for metadata management are yet to be implemented or improved. Considerably more work will need to be done for future directions of this research.

This section provides the following insights for future research per each life cycle step.

Selection: Currently, this process is done manually and has the potential to be implemented as a (semi-)automated feature. Web crawlers are ideal help for such activities. With a specific definition of criteria, such machines can be implemented to explore the Web and harvest required information. With the help of AI, content discovery and natural language processing methods can be employed to further empower this feature. In this way, huge quantities of target sources and high-quality metadata can be gained. As a complementary step to ensure the relevance and quality of sources and metadata, admin users only need to confirm.

Extraction: Metadata extraction was not the main focus in the scope of this research. Preliminary work has been done for extracting metadata about research datasets linked or mentioned in the footnote part of the scholarly articles. However, a variety of scholarly artifact types (e.g., videos, audios, software codes etc.) could be considered in future work to extract metadata from. Such metadata can strongly support the quality assessment of individual artifacts, events, and scholars. In addition, more analytical studies and recommendation services can be provided having such metadata at hand.

Transformation: Metadata represented in several formats such as CSV, HBase, XML, and data represented in Tuples have been converted to RDF. In several occasions, different serialization of RDF had to be converted to other serialization. Although the process of RDFization is automated in the context of OpenAIRE, the other actions have been done through multiple manual steps. Together with automating the metadata source selection, transformation can also be deployed in the infrastructure of the OpenResearch.org platform. In this way, transforming from unstructured representations of metadata towards a knowledge-based infrastructure can change the dominant paradigm of current scholarly communication.

Interlinking: Deployment of OpenAIRE interlinking with already examined datasets in the infrastructure of OA is a future work. Interlinking OA dataset with other relevant datasets is an ongoing task for the OA LOD team and further improvements are required in order to serve users with a cross-dataset utilization of the metadata. Based on the current observations, we also plan to enhance the interlinking results between OA and other candidate datasets related to other fields such as biology and astronomy. The current plan is to adopt the implemented interlinking services into the infrastructure of the OA and have them publicly available with a user-friendly interface.

As discussed through the thesis, each research community has its own formal and informal rules of communication, which are often barely usable by other research domains. The knowledge graph of the OpenResearch.org platform could easily be connected to the OA LOD and other relevant datasets such as Springer LOD. Linked data can support analysis in a broader interdisciplinary range.

Enrichment: Metadata enrichment features need to be implemented from scratch. OpenResearch.org benefits from MediaWiki's feature of showing the missing and not existing properties, pages and values in red. This feature is currently used for enrichment of event wiki pages, but manually and mainly by the admin users. By identifying such cases, a systematic enrichment could be implemented using Web crawlers and AI content discovery approaches from certain sources only seeking for such information e.g., exploring CfP mailing lists or conference home pages.

Curation: In addition to the traditional publishing through event homepages, CfP emails etc. a new channel has emerged for scholarly event metadata management. However, the OR platform has a limited number of users and the curation step is mainly done by admins. In order to take metadata curation to the next level in the OpenResearch.org platform, more community involvement is required. The other steps of the life cycle also heavily depend on curation and community involvement. Once it has a massive amount of metadata and coverage goes beyond the computer science (current focus of OR), it has the potential to be an effective platform serving easy exploration over FAIR metadata. The validation and preciseness of analytics also depend on the amount of metadata being curated in OR knowledge graph. As one of the ways that OR uses for metadata curation is the semantic forms, it is easily extendable to other types of artifacts such as blogs or micropublications. OR is planned to be extended with a video recording and broadcasting environment for scientific events. Metadata Curation and annotation of such artifact types can increase the usage of the platform for multiple objectives.

Mining: The scholarly knowledge graphs such as the one created in OR or OpenAIRE or the SemPub Challenge series with variety and complexities in nodes and relations are a perfect application domain for graph mining and clustering algorithms together with artificial intelligence models. In this thesis, graph mining has been done for co-authorship recommendation using semantic similarities to have a proof of concept. As an ongoing task, this research is further being extended with employing AI algorithms. It is planned to apply it for citation, event, sponsor etc. recommendation.

Quality Assessment: Quality assessment of data is often considered with a list of predefined quality metrics such as Accuracy and Reliability, Serviceability, Accessibility, Methodological soundness, and Assurances of integrity. Usually this step is a distinct phase within the data quality life cycle that is used to verify the source, quantity and impact of any data items. The quality of data can quickly decay over time, even with stringent data capture methods cleaning the data as it enters your database. For this step, we require to do a systematic follow up in future work of this thesis.

Analysis: The analysis over the collected metadata is mainly focused on quality assessment of scholarly artifacts. The future of OR is envisioned as a multidisciplinary platform for different types of information needs for a variety of stakeholders. Educational resources are one of the important means for educators or students. OCW, as well as many other types of scholarly artifacts, are planned to be added to the OR knowledge graph. The metrics defined for quality assessment of OCW is planned to be added in the OR platform. The properties of events in OR are mainly defined based on the established quality assessment framework that can serve a broad range of metadata consumers with different information needs. Importing more metadata that is relevant to these metrics would help us to establish a greater degree of accuracy in quality assessment services. The metrics defined for quality assessment are similar to altmetrics and vary from traditional bibliometrics measurements. OR has a focus on scientific events and publications. However, extending its coverage to other scholarly artifacts such as research datasets, software etc. makes OR more comprehensives system with regards to the quality assessment. In addition, artifact creators and organizers are supported in visibly promoting their research activities and results.

In the context of the SemPub challenge, all the defined queries with their corresponding solutions remain to be integrated into the OpenResearch.org knowledge graph.

Visualization: So far the main focus has been in creating and interlinking of the knowledge graph, curation metadata, providing quality assessment and analysis. Each of these steps, combined with a suitable visualization platform can ease the utilization of the information by users. Beside the default table view, currently OR is using MediaWiki extensions for visualization of data in a map view, a calendar view and a timeline view. Already existing plug-ins such as D3.js can be adapted to provide visual analytics on top on the queries and quality assessment results. In the context of OpenAIRE, the whole workflow follows certain design decisions. The scholarly metadata knowledge graph is a perfect application domain for Machine Learning approaches and Artificial Intelligence approaches.

With regard to scholarly communication, we are currently at a crossroad: On the one hand, there are commercial publishers and new incumbents such as social networks for researchers (e.g. ResearchGate, Academia.edu), which provide commercial services to the research community. Researchers either pay directly for these services by means of publication and access fees or indirectly (such as in the case of social networks) with their data. Either way, these commercial services strive to create a lock-in effect, which forces researchers to continue using these services without being able to migrate and choose competing services. On the other hand, there is an increasing push towards more open-access and open platforms for scholarly communication. Examples are open-access repositories such as arXiv, Zenodo, bibliographic metadata services such as DBLP and OpenAIRE, journal and conference management software and services such as Open Journal Systems and EasyChair or OpenCourseWare platforms such as SlideWiki.org. We see the work on OpenResearch as a first step towards tighter interlinking and integrating of open services for scholarly communication. The future vision is to establish a service to provide comprehensive scholarly metadata management about different types of artifact, events, and scholars, which enables researchers to find, promote and archive information. The starting point would be maturing of the existing OpenResearch.org platform with all the proposed future work.

Bibliography

[1] R. A. Ahmad, M. T. Afzal and M. A. Qadir, “Information Extraction for PDF Sources based on Rule-based System using Integrated Formats”, *ESWC 2016 Challenges*, Springer, 2016 (cit. on p. 121).

[2] G. Alexiou, S. Vahdati, C. Lange, G. Papastefanatos and S. Lohmann, “OpenAIRE LOD services: scholarly communication data as linked data”, *2nd Workshop, SAVE-SD. LNCS. Springer*, 2016 (cit. on p. 10).

[3] M. Allen, *Relational databases are not designed for scale*, relational-databases-scale (2015) (cit. on p. 32).

[4] A. Alshamrani and A. Bahattab, *A comparison between three SDLC models waterfall model, spiral model, and Incremental/Iterative model*, International Journal of Computer Science Issues (IJCSI) **12**.1 (2015) 106 (cit. on p. 45).

[5] S. Ameri, *Exploiting Interlinked Research Metadata to Provide Recommendations for Authors of Scientific Papers*, MA thesis: University of Bonn, 2017, URL: `http://eis-bonn.github.io/Theses/2017/Shirin_Ameri/thesis.pdf` (cit. on p. 150).

[6] C. P. Antonopoulos and N. S. Voros, “Data Management Processes”, *Cyberphysical Systems for Epilepsy and Related Brain Disorders*, Springer, 2015 111 (cit. on p. 42).

[7] M. Assante, L. Candela, D. Castelli, P. Manghi, P. Pagano and C. Nazionale, *Science 2.0 repositories: time for a change in scholarly communication*, D-Lib Magazine **21**.1/2 (2015) 1 (cit. on p. 3).

[8] D. Atkins, J. S. Brown and A. L. Hammond, “A review of the open educational resources (OER) movement: Achievements, challenges, and new opportunities”, *Creative common*, 2007 (cit. on p. 80).

[9] D. Atkinson, *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675-1975*, Routledge, 1998 (cit. on p. 16).

[10] S. Auer, V. Bryl and S. Tramp, *Linked Open Data–Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project*, vol. 8661, Springer, 2014 (cit. on pp. 43, 44).

[11] S. Auer, S. Dietzold and T. Riechert, “OntoWiki–a tool for social, semantic collaboration”, *International Semantic Web Conference*, Springer, 2006 736 (cit. on p. 56).

[12] R. Baeza-Yates, C. Castillo and F. Saint-Jean, “Web Dynamics, Structure, and Page Quality, Adapting to Change in Content, Size, Topology and Use”, *Web Dynamics*, Springer, 2004 93 (cit. on p. 83).

[13] A. Baglatzi, T. Kauppinen and C. Keßler, *Linked science core vocabulary specification*, tech. rep., available at, `http://linkedscience.org/lsc/ns/`, 2011 (cit. on p. 21).

[14] A. Ball, *Review of data management lifecycle models*, (2012) (cit. on p. 42).

[15] A. Bardi and P. Manghi, *Enhanced publications: Data models and information systems*, Liber Quarterly **23**.4 (2014) (cit. on p. 37).

[16] F. Bauer and M. Kaltenböck, *Linked open data: The essentials*, 2011 (cit. on p. 145).

[17] A. Bequai, W. L. Foundation and U. S. of America, *Computers+ Business: Liabilities: a Preventive Guide for Management*, Constitutional Institute of America, 1984 (cit. on p. 33).

[18] A. Berglund, S. Boag, D. Chamberlin, M. F. Fernández, M. Kay, J. Robie and J. Siméon, *XML path language (XPath)*, World Wide Web Consortium (W3C) (2003) (cit. on p. 131).

[19] M. Bergman, *AI3 Adaptive Information Innovation Adaptive Infrastructure*, 2011, URL: `http://www.mkbergman.com/232/sources-and-classification-of-semantic-heterogeneities/` (cit. on p. 38).

[20] M. Bergman, *AI3 Adaptive Information Innovation Adaptive Infrastructure*, 2014, URL: `http://www.mkbergman.com/1778/what-is-big-structure/` (cit. on p. 35).

[21] T. Berners-Lee, *Is your linked open data 5 star*, BERNERS-LEE, T. Linked Data. Cambridge: W3C (2010) (cit. on p. 53).

[22] T. Berners-Lee and R. Cailliau, *WorldWideWeb: Proposal for a HyperText Project*, (1990), URL: `http://www.w3.org/Proposal` (cit. on pp. 1, 14).

[23] T. Berners-Lee, J. Hendler and O. Lassila, *The semantic web*, Scientific american **284**.5 (2001) 34 (cit. on pp. 2, 34).

[24] M. Bertin and I. Atanassova, "Extraction and Characterization of Citations in Scientific Papers", *Semantic Web Evaluation Challenges*, Springer, 2014 120, URL: `http://dx.doi.org/10.1007/978-3-319-12024-9_16` (cit. on p. 119).

[25] S. Bischof, S. Decker, T. Krennwallner, N. Lopes and A. Polleres, *Mapping between RDF and XML with XSPARQL*, English, Journal on Data Semantics **1**.3 (2012), ISSN: 1861-2032 (cit. on p. 141).

[26] C. Bizer and R. Cyganiak, *Quality-driven information filtering using the WIQA policy framework*, J. Web Sem. **7**.1 (2009) (cit. on pp. 60, 75, 95).

[27] C. Bizer, *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*, PhD thesis: FU Berlin, 2007 (cit. on p. 85).

[28] C. Bizer, T. Heath and T. Berners-Lee, *Linked data-the story so far*, International journal on semantic web and information systems **5**.3 (2009) 1 (cit. on pp. 32, 41, 53).

[29] C. Bizer, R. Cyganiak, T. Heath et al., *How to publish linked data on the web*, (2007) (cit. on p. 67).

[30] B.-C. Björk, *A lifecycle model of the scientific communication process*, Learned Publishing **18**.3 (2005) 165 (cit. on p. 29).

[31] A. Blackwell and T. Green, *Cognitive Dimensions of Notations Resource Site*, 2010, URL: `http://www.cl.cam.ac.uk/~afb21/CognitiveDimensions/` (cit. on p. 144).

[32] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab and M. G. Strintzis, "Semantic annotation of images and videos for multimedia analysis", *European Semantic Web Conference*, Springer, 2005 592 (cit. on p. 49).

[33] R. Blumberg and S. Atre, *The problem with unstructured data*, Dm Review **13**.42-49 (2003) 62 (cit. on p. 32).

[34] J. Blumenfeld, *The Global Change Master Directory: Data, Services, and Tools Serving the International Science Community*, EOSDIS Science Writer; accessed 23 May 2018, URL: `https://earthdata.nasa.gov/gcmd-retrospective-and-future` (cit. on p. 34).

[35] B. Boehm, J. A. Lane, S. Koolmanojwong and R. Turner, *The incremental commitment spiral model: Principles and practices for successful systems and software*, Addison-Wesley Professional, 2014 (cit. on p. 45).

[36] R. Borchardt, C. Moran, S. Cantrill, S. A. Oh, M. R. Hartings et al., *Perception of the importance of chemistry research papers and comparison to citation rates*, PloS one **13**.3 (2018) e0194903 (cit. on p. 18).

[37] L. Bornmann and R. Mutz, *Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references*, Journal of the Association for Information Science Technology **66**.11 (2015) 2215, URL: `https://EconPapers.repec.org/RePEc:bla:jinfst:v:66:y:2015:i:11:p:2215-2222` (cit. on p. 37).

[38] D. Boud, *Enhancing Learning Through Self-assessment*, RoutledgeFalmer, 1995, ISBN: 0749413689 (cit. on p. 87).

[39] F. Breitling, *A standard transformation from XML to RDF via XSLT*, Astronomische Nachrichten **330**.7 (2009) (cit. on p. 140).

[40] V. Bryl, A. Birukou, K. Eckert and M. Kessler, "What's in the proceedings? Combining publisher's and researcher's perspectives", *Proceedings of the 4th Workshop on Semantic Publishing (SePublica)*, 2014 (cit. on pp. 6, 22, 23).

[41] V. Bryl, A. Birukou, K. Eckert and M. Kessler, "What's in the proceedings? Combining publisher's and researcher's perspectives.", *SePublica*, 2014 (cit. on p. 116).

[42] A. Buluç, H. Meyerhenke, I. Safro, P. Sanders and C. Schulz, "Recent Advances in Graph Partitioning", *Algorithm Engineering: Selected Results and Surveys*, Cham: Springer, 2016 (cit. on p. 166).

[43] P. Buneman, "Semistructured data", *Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, ACM, 1997 117 (cit. on p. 32).

[44] S. Capadisli, R. Riedl and S. Auer, "Enabling Accessible Knowledge", *CeDEM*, 2015, URL: `http://csarven.ca/enabling-accessible-knowledge` (cit. on p. 111).

[45] D. A. H. Carney and J. E. Readence, *Transfer of Learning from Illustration–Dependent Text*, Educational Research **76**.4 (1983) (cit. on p. 88).

[46] R. N. Carney and J. R. Levin, *Pictorial Illustrations Still Improve Students Learning From Text*, Educational Psychology Review **14**.1 (2002) 5 (cit. on p. 88).

[47] D. Castelli, P. Manghi and C. Thanos, *A vision towards scientific communication infrastructures*, International Journal on Digital Libraries **13**.3-4 (2013) 155 (cit. on p. 5).

[48] T. Caswell, S. Henson, M. Jensen and D. Wiley, *Open content and open educational resources: Enabling universal education*, International Review of Research in Open and Distance Learning **9**.1 (2008) (cit. on p. 77).

[49] T. Catapano, *TaxPub: An Extension of the NLM/NCBI Journal Publishing DTD for Taxonomic Descriptions*, Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010 (2010) (cit. on p. 115).

[50] P. Checkland, *Systems thinking*, Rethinking management information systems (1999) 45 (cit. on p. 31).

[51] I.-M. A. Chen, V. M. Markowitz et al., *An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools.*, Information Systems **20**.5 (1995) 393 (cit. on p. 42).

[52] P. Ciccarese and T. Groza, *Ontology of Rhetorical Blocks (ORB). Editor's Draft, 5 June 2011*, World Wide Web Consortium. (last visited March 12, 2012) (2011) (cit. on p. 20).

[53] E. F. Codd, *A relational model of data for large shared data banks*, Communications of the ACM **13**.6 (1970) 377 (cit. on pp. 32, 34).

[54] U. N. S. Commission and E. C. F. Europe, *Terminology on Statistical Metadata*, (2000) (cit. on p. 33).

[55] A. Constantin, S. Peroni, S. Pettifer, D. Shotton and F. Vitali, *The document components ontology (DoCO)*, Semantic Web **7**.2 (2016) 167 (cit. on p. 21).

[56] E. Cortez, A. S. da Silva, M. A. Gonçalves, F. Mesquita and E. S. de Moura, "FLUX-CIM: flexible unsupervised extraction of citation metadata", *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, ACM, 2007 215 (cit. on p. 22).

[57] M. H. Cragin, P. B. Heidorn, C. L. Palmer and L. C. Smith, *An Educational Program on Data Curation*, 2013 (cit. on p. 56).

[58] K. Crowston and J. Qin, *A capability maturity model for scientific data management: Evidence from the literature*, Proceedings of the Association for Information Science and Technology **48**.1 (2011) 1 (cit. on p. 43).

[59] M. Dabrowski, M. Synak and S. R. Kruk, *Bibliographic ontology*, 2009 (cit. on p. 21).

[60] P. T. Daniels and W. Bright, *The World's Writing Systems*, Oxford University Press (1996) (cit. on p. 13).

[61] *Data Collection Methods for Program Evaluation: Observation*, Evaluation Briefs 16, Center for Diseases Control and Prevention, 2008, URL: `http://www.cdc.gov/healthyyouth/evaluation/pdf/brief16.pdf` (cit. on p. 78).

[62] *Data Life Cycle Models and Concepts*, Review, Working Group on Information Systems and Services, 2012, URL: `http://www.cdc.gov/healthyyouth/evaluation/pdf/brief16.pdf` (cit. on p. 42).

[63] T. H. Davenport, P. Barth and R. Bean, *How'big data'is different*, MIT Sloan Management Review, 2012 (cit. on p. 2).

[64] U. M. Dholakia, W. J. King and R. Baraniuk, "What makes an open education program sustainable? The case of Connexions", *Open Education Conference (OEC)*, 2006 1 (cit. on p. 84).

[65] A. Di Iorio and C. Lange, eds., *Semantic Publishing Challenge (Extended Semantic Web Conference, Semantic Web Evaluation Track)*, (Anissaras, Greece, 25th May 2014), 2014, URL: `http://2014.eswc-conferences.org/program/semwebeval` (cit. on p. 76).

[66] A. Di Iorio, C. Lange, A. Dimou and S. Vahdati, "Semantic publishing challenge–assessing the quality of scientific output by information extraction and interlinking", *Semantic Web Evaluation Challenge*, Springer, 2015 65 (cit. on p. 114).

[67] A. Di Iorio, A. G. Nuzzolese, F. Osborne, S. Peroni, F. Poggi, M. Smith, F. Vitali and J. Zhao, "The RASH Framework: enabling HTML+RDF submissions in scholarly venues", *Proceedings of the Posters & Demonstrations Track of the 14th International Semantic Web Conference (ISWC)*, 2015 (cit. on pp. 95, 111).

[68] Y. Diao, P. Fischer, M. J. Franklin and R. To, "Yfilter: Efficient and scalable filtering of XML documents", *Data Engineering*, IEEE, 2002 (cit. on p. 143).

[69] C. Dichev, B. Bhattarai, C. Clonch and D. Dicheva, "Towards Better Discoverability and Use of Open Content", *3rd Int. Conf. on Software, Services and Semantic Technologies S3T*, Springer, 2011 (cit. on p. 89).

[70] M. W. Dictionaries, *Data*, Dictionary; accessed 24 May 2018, URL: `https://www.merriam-webster.com/dictionary/data` (cit. on p. 31).

[71] A. Dimou, S. Vahdati, A. Di Iorio, C. Lange, R. Verborgh and E. Mannens, *Challenges as enablers for high quality Linked Data: insights from the Semantic Publishing Challenge*, PeerJ Computer Science **3** (2017) e105 (cit. on p. 114).

[72] A. Dimou, M. Vander Sande, P. Colpaert, L. De Vocht, R. Verborgh, E. Mannens and R. Van de Walle, "Extraction and Semantic Annotation of Workshop Proceedings in HTML Using RML", *Semantic Web Evaluation Challenges*, Springer, 2014 114, URL: `http://dx.doi.org/10.1007/978-3-319-12024-9_15` (cit. on p. 119).

[73] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao and F. Diaz, "Towards recency ranking in web search", *Web Search and Data Mining*, ACM, 2010 11 (cit. on p. 83).

[74] S. Downes, "Models for sustainable open educational resources", *NRC*, 2007 1 (cit. on p. 84).

[75] D. Dunning, C. Heath and J.-M. Suls, *Flawed Self-Assessment Implications for Health, Education, and the Workplace*, Psychological Science **5**.3 (2004) 69 (cit. on p. 87).

[76] O. for Economic Co-operation and D. (OECD), *OECD Guidelines for the Security of Information Systems*, 1992 (cit. on p. 33).

[77] M. Elias, A. James, E. Ruckhaus, M. C. Suárez-Figueroa, K. A. de Graaf, A. Khalili, B. Wulff, S. Lohmann and S. Auer, *SlideWiki–Towards a Collaborative and Accessible Platform for Slide Presentations*, () (cit. on p. 72).

[78] A. Elliott, *Melvil Dewey: A Singular and Contentious Life.*, Wilson Library Bulletin **55**.9 (1981) 666 (cit. on p. 13).

[79] I. Ermilov, S. Auer and C. Stadler, "CSV2RDF: User-Driven CSV to RDF Mass Conversion Framework", *I-Semantics*, 2013 (cit. on p. 140).

[80] *ESWC 2016 Challenges*, Springer, 2016.

[81] S. Fathalla and C. Lange, "EVENTSKG: A Knowledge Graph Representation for Top-Prestigious Computer Science Events Metadata", Springer International Publishing, 2018 53 (cit. on p. 76).

[82] S. Fathalla, S. Vahdati, S. Auer and C. Lange, "Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles", *International Conference on Theory and Practice of Digital Libraries*, Springer, 2017 315 (cit. on p. 21).

[83] B. Fidler and A. Acker, *Metadata and infrastructure in internet history: Sockets in the arpanet host-host protocol*, Proceedings of the Association for Information Science and Technology **51**.1 (2014) 1 (cit. on p. 34).

[84] T. Finne, *Information systems risk management: key concepts and business processes*, Computers & Security **19**.3 (2000) 234 (cit. on p. 32).

[85] J. Fitzpatrick, *Mendeley Manages Your Documents on Your Desktop and in the Cloud*, See http://lifehacker. com/5334254/Ravindra (2009) (cit. on p. 25).

[86] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi et al., *Science of science*, Science **359**.6379 (2018) eaao0185 (cit. on p. 6).

[87] J. Frey, K. Müller, S. Hellmann, E. Rahm and M.-E. Vidal, *Evaluation of Metadata Representations in RDF stores*, () (cit. on p. 38).

[88] N. Friesen, "Open Educational Resources: Innovation, Research and Practice", Commonwealth of Learning, Athabasca University, 2013, chap. Realising the open in Open Educational Resources: Practical concerns and solutions (cit. on p. 80).

[89] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna and T. Charnois, "Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers", *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018 679 (cit. on p. 21).

[90] M. Gaertler, "Clustering", *Network Analysis: Method. Found.* (Cit. on p. 166).

[91] A. Gangemi and V. Presutti, "Ontology design patterns", *Handbook on ontologies*, Springer, 2009 (cit. on pp. 68, 72).

[92] B. Gastel and R. A. Day, *How to write and publish a scientific paper*, ABC-CLIO, 2016 (cit. on p. 49).

[93] P. Ginsparg, *First steps toward electronic research communication*, MIT Press, Cambridge, MA, 1997 (cit. on p. 14).

[94] *GND Ontology*, Deutsche Nationalbibliothek, 2016, URL: `http://d-nb.info/standards/elementset/gnd#ConferenceOrEvent` (cit. on p. 72).

[95] M. Goldberg and J. Hartwick, *The effect of advertiser reputation and extremity of advertising claim on advertising effectiveness*, Consumer Research **17**.2 (1990) (cit. on p. 101).

[96] B. Golshan, A. Halevy, G. Mihaila and W.-C. Tan, "Data integration: After the teenage years", *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, ACM, 2017 101 (cit. on pp. 1, 5, 38).

[97] J. Gray, *Data Management: Past, Present, and Future*, arXiv preprint cs/0701156 (2007) (cit. on p. 42).

[98] J. Greenberg, "Metadata and the World Wide Web", *Encyclopedia of Library and Information Science*, Marcel Dekker, 2002 44 (cit. on pp. 32–34).

[99] J. Greenberg, "Metadata and the World Wide Web", *Encyclopedia of Library and Information Science*, Marcel Dekker, 2003 1876 (cit. on p. 34).

[100] U. Guardian, *A Guardian guide to your metadata*, 2011 (cit. on p. 31).

[101] S. C. Guptill, *Metadata and data catalogues*, Geographical information systems **2** (1999) 677 (cit. on pp. 33, 35, 36).

[102] T. Habermann, *Metadata Life Cycles, Use Cases and Hierarchies*, Geosciences **8**.5 (2018) 179 (cit. on p. 45).

[103] R. E. Hall, *Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence*, Journal of political economy **86**.6 (1978) 971 (cit. on p. 42).

[104] M. Hallo, S. Luján-Mora and C. Chávez, “An Approach to Publish Scientific Data of Open-Access Journals using Linked Open Data Technologies”, *EDULEARN*, 2014 (cit. on p. 145).

[105] M. Hamilton and S. Manuel, *Metadata is a Love Note to the Future: UK Higher Education Research Data Management Survey*, (2013) (cit. on pp. 35, 44).

[106] H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, C. L. Giles and X. Zhang, “Rule-based word clustering for document metadata extraction”, *Proceedings of the 2005 ACM symposium on Applied computing*, ACM, 2005 1049 (cit. on p. 48).

[107] A. Haque and L. Perkins, *Distributed RDF Triple Store Using HBase and Hive*, University of Texas at Austin (2012) (cit. on p. 139).

[108] S. Harris, A. Seaborne and E. Prud’hommeaux, *SPARQL 1.1 query language*, W3C recommendation **21**.10 (2013) (cit. on p. 39).

[109] M. S. Hart, *Gutenberg:Project Gutenberg Mission Statement by Michael Hart*, 2004 (cit. on p. 14).

[110] M. Hert, G. Reif and H. C. Gall, “A comparison of RDB-to-RDF mapping languages”, *I-Semantics*, ACM, 2011 (cit. on p. 141).

[111] L. K. Hessels, H. Van Lente and R. Smits, *In search of relevance: the changing contract between science and society*, Science and Public Policy **36**.5 (2009) 387 (cit. on p. 18).

[112] P. Heyvaert, A. Dimou, R. Verborgh, E. Mannens and R. Van de Walle, “Semantically Annotating CEUR-WS Workshop Proceedings with RML”, *Semantic Web Evaluation Challenges*, Springer, 2015 165, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_14` (cit. on p. 119).

[113] D. Hicks, P. Wouters, L. Waltman, S. d. Rijcke and I. Rafols, *Bibliometrics: the Leiden Manifesto for research metrics*, (2015) (cit. on p. 18).

[114] S. Higgins, *The DCC curation lifecycle model*, International Journal of Digital Curation **3**.1 (2008) (cit. on pp. 43, 44).

[115] D. Hoffmann, C. Standish, M. Garcıa-Diez, P. Pettitt, J. Milton, J. Zilhão, J. Alcolea-González, P. Cantalejo-Duarte, H. Collado, R. de Balbın et al., *U-Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art*, Science **359**.6378 (2018) 912 (cit. on p. 12).

[116] D. Howe, *Metadata*, Free online dictionary; accessed 14 May 2018, URL: `http://foldoc.org/metadata` (cit. on pp. 33, 34).

[117] J. Huang, Z. Zhuang, J. Li and C. L. Giles, “Collaboration over time: characterizing and modeling network evolution”, *Proceedings of the 2008 international conference on web search and data mining*, ACM, 2008 107 (cit. on p. 22).

[118] G. Hurtado Martin, S. Schockaert, C. Cornelis and H. Naessens, *An Exploratory Study on Content-based Filtering of Call for Papers*, Multidisciplinary Information Retrieval (2013) (cit. on p. 26).

[119] D. C. M. Initiative et al., *DCMI Metadata Basics*, Dublin Core Metadata Initiative (2014) (cit. on p. 34).

[120] J. P. Ioannidis, *How to make more published research true*, Revista Cubana de Información en Ciencias de la Salud (ACIMED) **26**.2 (2015) 187 (cit. on p. 16).

[121] A. D. Iorio, C. Lange, A. Dimou and S. Vahdati, "Semantic Publishing Challenge, Assessing the Quality of Scientific Output by Information Extraction and Interlinking", *Semantic Web Evaluation Challenges*, Springer, 2015 (cit. on pp. 22, 170).

[122] ISO, *Formulation of data definitions*, International Organization for Standardization; accessed 22 May 2018, URL: `https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-4:ed-2:v1:en` (cit. on pp. 33, 35).

[123] L. Issertial and H. Tsuji, *Information Extraction for Call for Paper*, Int. Journal of Knowledge and Systems Science (IJKSS) **6**.4 (2015) (cit. on p. 26).

[124] H. V. Jagadish, *The conference reviewing crisis and a proposed solution.*, SIGMOD Record **37**.3 (2008) (cit. on p. 98).

[125] S. Jeong and H.-G. Kim, *SEDE: An ontology for scholarly event description*, Journal of Information Science **36**.2 (2010) 209 (cit. on p. 21).

[126] K. Jones, "Research360: Managing data across the institutional research lifecycle", *Poster Presented the 7th International Digital Curation Conference, Bristol, UK, 5–8 Dec*, 2011 (cit. on p. 43).

[127] J. M. Juran, *Juran's Quality Control Handbook*, 4th, McGraw-Hill (Tx), 1974, ISBN: 0070331766 (cit. on pp. 6, 65, 78, 169).

[128] K. V. K. and D. G. S. Sadasivam, *A Novel Method For Dynamic SPARQL Endpoint Generation In NoSQL Databases*, Australian Journal of Basic and Applied Sciences **9**.6 (2015) (cit. on p. 139).

[129] B. K. Kahn, D. M. Strong and R. Y. Wang, *Information Quality Benchmarks: Product and Service Performance*, Commun. ACM **45**.4 (2002) 184, ISSN: 0001-0782, URL: `http://doi.acm.org/10.1145/505248.506007` (cit. on pp. 65, 78).

[130] S. Kaisler, F. Armour, J. A. Espinosa and W. Money, "Big data: Issues and challenges moving forward", *System sciences (HICSS), 2013 46th Hawaii international conference on*, IEEE, 2013 995 (cit. on p. 5).

[131] A. Kalyanpur, J. Golbeck, M. Grove and J. Hendler, "An RDF editor and portal for the semantic web", *Semantic Authoring, Annotation & Knowledge Markup Workshop*, vol. 32, 2002 (cit. on p. 56).

[132] G. Karypis and V. Kumar, *A fast and high quality multilevel scheme for partitioning irregular graphs*, Scientific Computing (1998) (cit. on p. 160).

[133] A. Kastrin, T. C. Rindflesch and D. Hristovski, "Link Prediction on the Semantic MEDLINE Network - An Approach to Literature-Based Discovery", *The Discovery Science Conference*, 2014 (cit. on p. 168).

[134] V. Katukoori, *Standardizatiing Availability Definitio*, (1995) (cit. on p. 85).

[135] A. Kazim, *A Study of Software Development Life Cycle Process Models*, International Journal of Advanced Research in Computer Science **8**.1 (2017) (cit. on p. 45).

[136] R. Kern, K. Jack, M. Hristakeva and M. Granitzer, *TeamBeam-meta-data extraction from scientific literature*, D-Lib Magazine **18**.7 (2012) 1 (cit. on p. 22).

[137] M. Khabsa and C. L. Giles, *The number of scholarly documents on the public web*, PloS one **9**.5 (2014) e93949 (cit. on p. 37).

[138] V. Khadilkar, M. Kantarcioglu, B. Thuraisingham and P. Castagna, "Jena-HBase: A distributed, scalable and efficient RDF triple store", *ISWC Posters & Demonstrations*, 2012 (cit. on p. 139).

[139] S. Khatchadourian and M. P. Consens, “ExpLOD: summary-based exploration of interlinking and RDF usage in the linked open data cloud”, *Extended Semantic Web Conference*, Springer, 2010 272 (cit. on p. 53).

[140] C. King, *Multiauthor papers: onward and upward*, Sciencewatch newsletter (2012) 1 (cit. on p. 15).

[141] S. Klampfl and R. Kern, “Machine Learning Techniques for Automatically Extracting Contextual Information from Scientific Publications”, *Semantic Web Evaluation Challenges*, Springer, 2015 105, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_9` (cit. on pp. 119, 121).

[142] S. Klampfl and R. Kern, “Reconstructing the Logical Structure of a Scientific Publication using Machine Learning”, *ESWC 2016 Challenges*, Springer, 2016 (cit. on p. 119).

[143] M. Klein, “Interpreting XML documents via an RDF schema ontology”, *DEXA*, 2002 (cit. on p. 141).

[144] S.-A. Knight and J. M. Burn, *Developing a framework for assessing information quality on the World Wide Web*, Informing Science: International Journal of an Emerging Transdiscipline **8**.5 (2005) 159 (cit. on pp. 6, 65, 78, 169).

[145] M. Kobayashi and K. Takeda, *Information retrieval on the web*, ACM Computing Surveys (CSUR) **32**.2 (2000) 144 (cit. on p. 1).

[146] M. Kolchin, E. Cherny, F. Kozlov, A. Shipilo and L. Kovriguina, “CEUR-WS-LOD: Conversion of CEUR-WS Workshops to Linked Data”, *Semantic Web Evaluation Challenges*, Springer, 2015 142, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_12` (cit. on p. 118).

[147] M. Kolchin and F. Kozlov, “A Template-Based Information Extraction from Web Sites with Unstable Markup”, *Semantic Web Evaluation Challenges*, Springer, 2014 89, URL: `http://dx.doi.org/10.1007/978-3-319-12024-9_11` (cit. on p. 118).

[148] R. N. Kostoff and M. F. Shlesinger, *CAB: Citation-Assisted Background*, Scientometrics **62**.2 (2005) 199 (cit. on p. 5).

[149] P. Kotler, *Marketing Management*, Pearson Education, 2000 (cit. on p. 104).

[150] L. Kovriguina, A. Shipilo, F. Kozlov, M. Kolchin and E. Cherny, “Metadata Extraction from Conference Proceedings Using Template-Based Approach”, *Semantic Web Evaluation Challenges*, Springer, 2015 153, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_13` (cit. on p. 120).

[151] M. Krötzsch and D. Vrandecic, “Semantic wikipedia”, *Social Semantic Web*, Springer, 2009 393 (cit. on p. 56).

[152] T. Kuhn, M. Perc and D. Helbing, *Inheritance patterns in citation networks reveal scientific memes*, Physical Review X **4**.4 (2014) 041036 (cit. on p. 22).

[153] A. Kunjithapatham, S. J. Gibbs, P. Rathod, P. Nguyen and M. Sheshagiri, *Semantic metadata creation for videos*, US Patent 8,145,648, 2012 (cit. on p. 49).

[154] M. Kurth, D. Ruddy and N. Rupp, *Repurposing MARC metadata: using digital project experience to develop a metadata management design*, Library Hi Tech **22**.2 (2004) 153 (cit. on p. 42).

[155] R. Lämmel, *Google’s MapReduce programming model—Revisited*, Science of computer programming **70**.1 (2008) 1 (cit. on p. 33).

[156] J. Lanagan and A. F. Smeaton, *Video digital libraries: contributive and decentralised*, International Journal on Digital Libraries **12**.4 (2012) 159 (cit. on p. 14).

[157] B. C. Landry and J. E. Rush, *Toward a theory of indexing—ii*, Journal of the Association for Information Science and Technology **21**.5 (1970) 358 (cit. on pp. 32, 33).

[158] C. Lange and A. Di Iorio, "Semantic publishing challenge–assessing the quality of scientific output", *Semantic Web Evaluation Challenge*, Springer, 2014 61 (cit. on pp. 114, 116).

[159] V. Lariviere, V. Kiermer, C. J. MacCallum, M. McNutt, M. Patterson, B. Pulverer, S. Swaminathan, S. Taylor and S. Curry, *A simple proposal for the publication of journal citation distributions*, BioRxiv (2016) 062109 (cit. on p. 18).

[160] P. O. Larsen and M. vonIns, *The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index*, Scientometrics **84**.3 (2010) 575 (cit. on p. 18).

[161] P. A. Lawrence, *The mismeasurement of science*, Current Biology **17**.15 (2007) R583 (cit. on p. 18).

[162] Q. V. Le and T. Mikolov, *Distributed Representations of Sentences and Documents*, CoRR **abs/1405.4053** (2014) (cit. on p. 163).

[163] T. Lebo and G. T. Williams, "Converting governmental datasets into linked data", *I-Semantics*, ACM, 2010 (cit. on p. 139).

[164] L. Leydesdorff and C. S. Wagner, *International collaboration in science and the formation of a core group*, Journal of Informetrics **2**.4 (2008) 317 (cit. on p. 15).

[165] G. Li, B. C. Ooi, J. Feng, J. Wang and L. Zhou, "EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data", *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ACM, 2008 903 (cit. on p. 32).

[166] X. Li, G. Harbottl, J. Zhang and C. Wang, *The earliest writing? Sign use in the seventh millennium BC at Jiahu, Henan Province, China*, Antiquity **77**.295 (2003) (cit. on p. 12).

[167] D. Liben-Nowell and J. Kleinberg, *The link-prediction problem for social networks*, JASIST **58**.7 (2007) (cit. on p. 163).

[168] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong and F. Xia, *A Survey of Scholarly Data Visualization*, IEEE Access **6** (2018) 19205 (cit. on p. 6).

[169] G. M. Liumbruno, C. Velati, P. Pasqualetti and M. Franchini, *How to write a scientific manuscript for publication*, Blood Transfusion **11**.2 (2013) 217 (cit. on p. 13).

[170] C. Lnage and A. D. Iorio, "Semantic Publishing Challenge, Assessing the Quality of Scientific Output", *Semantic Web Evaluation Challenges*, Springer, 2014 (cit. on p. 22).

[171] P. Lord and A. Macdonald, *e-Science Curation Report*, tech. rep., 2003 (cit. on p. 44).

[172] R. Lubas, A. Jackson and I. Schneider, *The Metadata Manual: A Practical Workbook*, Oxford, UK: Chandos Publishing, 2013 (cit. on p. 36).

[173] J. Madhavan, D. Ko, Ł. Kot, V. Ganapathy, A. Rasmussen and A. Halevy, *Google's deep web crawl*, Proceedings of the VLDB Endowment **1**.2 (2008) 1241 (cit. on p. 1).

[174] P. Manghi, N. Houssos, M. Mikulicic and B. Jörg, "The data model of the OpenAIRE scientific communication e-infrastructure", *Metadata and Semantics Research*, Springer, 2012 (cit. on p. 68).

[175] P. Manghi, N. Manola, W. Horstmann and D. Peters, *An infrastructure for managing EC funded research output-The OpenAIRE Project*, The Grey Journal (TGJ): An International Journal on Grey Literature **6**.1 (2010) (cit. on p. 23).

[176] P. Manghi, M. Mikulicic and C. Atzori, *De-duplication of aggregation authority files*, International Journal of Metadata, Semantics and Ontologies **7**.2 (2012) 114 (cit. on p. 38).

[177] P. Mayr, F. Momeni and C. Lange, "Opening Scholarly Communication in Social Sciences: Supporting Open Peer Review with Fidus Writer", EA Conference, 2016 (cit. on p. 170).

[178] J. L. McCarthy, *Metadata management for large statistical databases*, (1982) (cit. on pp. 33, 35).

[179] J. P. McDonough, *METS: standardized encoding for digital library objects*, International journal on digital libraries **6**.2 (2006) 148 (cit. on p. 20).

[180] P. N. Mendes, H. Mühleisen and C. Bizer, "Sieve: Linked Data Quality Assessment and Fusion", *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, Berlin, Germany: ACM, 2012 116, ISBN: 978-1-4503-1143-4, URL: `http://doi.acm.org/10.1145/2320765.2320803` (cit. on p. 6).

[181] F. Michel, J. Montagnat and C. Faron-Zucker, *A survey of RDB to RDF translation approaches and tools*, Research report, I3S, 2014 (cit. on p. 141).

[182] W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse and G. Janée, *DataONE: Data Observation Network for Earth—Preserving data and enabling innovation in the biological and environmental sciences*, D-Lib Magazine **17**.1/2 (2011) 12 (cit. on p. 44).

[183] M. Milicka and R. Burget, "Information Extraction from Web Sources Based on Multi-aspect Content Analysis", *Semantic Web Evaluation Challenges*, Springer, 2015 81, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_7` (cit. on p. 119).

[184] H. G. Miller and P. Mork, *From data to decisions: a value chain for big data*, It Professional **15**.1 (2013) 57 (cit. on p. 113).

[185] MIT OpenCourseWare, *2005 Program Evaluation Findings Report*, tech. rep., 2006, URL: `http://ocw.mit.edu/ans7870/global/05_Prog_Eval_Report_Final.pdf` (cit. on p. 77).

[186] MITNews, *MIT to make nearly all course materials available free on the World Wide Web*, 2001, URL: `http://newsoffice.mit.edu/2001/ocw` (cit. on p. 77).

[187] G. Moise, M. Vlădoiu and Z. Constantinescu, "MASECO: A Multi-agent System for Evaluation and Classification of OERs and OCW Based on Quality Criteria", *E-Learning Paradigms and Applications*, Studies in Computational Intelligence 528, Springer, 2014 (cit. on p. 27).

[188] K. Möller, S. Bechhofer and T. Heath, *Semantic Web Dog Food Ontology*, 2009, URL: `http://data.semanticweb.org/ns/swc/swc_2009-05-09.html` (visited on 20/01/2016) (cit. on p. 72).

[189] C. D. Morais, *Fortune 1000 Companies List for 2015*, 2015, URL: `http://www.geolounge.com/fortune-1000-companies-list-for-2015/` (visited on 18/01/2016) (cit. on p. 101).

[190] N. Moxham, *Authors, Editors and Newsmongers: Form and Genre in the Philosophical Transactions under Henry Oldenburg*, (2016) (cit. on p. 16).

[191] M. Moyle and P. Polydoratou, *Investigating overlay journals: introducing the RIOJA Project*, D-Lib Magazine **13**.Septem (2007) (cit. on p. 17).

[192] M. A. Musen, *The protégé project: a look back and a look forward*, AI Matters **1**.4 (2015) 4, URL: `http://doi.acm.org/10.1145/2757001.2757003` (cit. on p. 72).

[193] P. Naur, *Concise survey of computer methods*, Studentlitteratur Lund, Sweden, 1974 (cit. on p. 42).

[194] M. Nentwig, M. Hartung, A.-C. Ngonga Ngomo and E. Rahm, *A survey of current link discovery frameworks*, Semantic Web **8**.3 (2017) 419 (cit. on p. 3).

[195] M. E. Newman, *Modularity and community structure in networks*, Proceedings of the national academy of sciences **103**.23 (2006) (cit. on p. 166).

[196] A.-C. N. Ngomo and S. Auer, "Limes-a time-efficient approach for large-scale link discovery on the web of data.", *IJCAI*, 2011 2312 (cit. on p. 53).

[197] A. G. Nuzzolese, S. Peroni and D. R. Recupero, "ACM: Article Content Miner for Assessing the Quality of Scientific Output", *ESWC 2016 Challenges*, Springer, 2016 (cit. on p. 120).

[198] A. G. Nuzzolese, S. Peroni and D. Reforgiato Recupero, "MACJa: Metadata and Citations Jailbreaker", *Semantic Web Evaluation Challenges*, Springer, 2015 117, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_10` (cit. on p. 120).

[199] C. O'Carroll, B. Rentier, C. Cabello Valdès, F. Esposito, E. Kaunismaa, K. Maas, J. Metcalfe, K. Vandevelde, I. Halleux, C. L. Kamerlin et al., *Evaluation of Research Careers fully acknowledging Open Science Practices-Rewards, incentives and/or recognition for researchers practicing Open Science*, tech. rep., Publication Office of the Europen Union, 2017 (cit. on p. 17).

[200] S. Oreg and O. Nov, *Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values*, Computers in Human Behavior **24**.5 (2008) 2055 (cit. on p. 88).

[201] T. O'reilly, *What is web 2.0*, 2005 (cit. on p. 1).

[202] A. G. Ororbia II, J. Wu, M. Khabsa, K. Williams and C. L. Giles, "Big scholarly data in CiteSeerX: Information extraction from the web", *Proceedings of the 24th International Conference on World Wide Web*, ACM, 2015 597 (cit. on p. 22).

[203] G. Palma, M. Vidal and L. Raschid, "Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning", *ISWC*, 2014 (cit. on pp. 160, 168).

[204] N. Papailiou, I. Konstantinou, D. Tsoumakos and N. Koziris, "H2RDF: adaptive query processing on RDF data in the cloud.", *International conference on World Wide Web*, ACM, 2012 (cit. on p. 139).

[205] M. Patel, *I2S2 idealised scientific research activity lifecycle model*, (2011) (cit. on p. 43).

[206] M. Q. Patton, *Qualitative research*, Wiley Online Library, 2005 (cit. on p. 78).

[207] H. Paulheim, *Knowledge graph refinement: A survey of approaches and evaluation methods*, Semantic web **8**.3 (2017) 489 (cit. on pp. 41, 167).

[208] C. Pedrinaci, J. Cardoso and T. Leidig, "Linked USDL: a vocabulary for web-scale service trading", *ESWC*, Springer, 2014 (cit. on pp. 68, 72).

[209] S. Peroni and D. Shotton, *FaBiO and CiTO: ontologies for describing bibliographic resources and citations*, Web Semantics: Science, Services and Agents on the World Wide Web **17** (2012) 33 (cit. on p. 21).

[210] S. Peroni, D. Shotton and F. Vitali, *Freedom for bibliographic references: OpenCitations arise*, (2016) (cit. on p. 24).

[211] D. Porcello and S. Hsi, *Crowdsourcing and Curating Online Education Resources*, Science **341**.6143 (2013) 240 (cit. on p. 78).

[212] L. Pouchard, *Revisiting the data lifecycle with big data curation*, International Journal of Digital Curation **10**.2 (2016) 176 (cit. on p. 44).

[213] A. Powell, M. Nilsson, A. Naeve and P. Johnston, *DCMI abstract model*, (2005) (cit. on p. 20).

[214] A. Prabhune, *GENERIC AND ADAPTIVE METADATA MANAGEMENT FRAMEWORK FOR SCIENTIFIC DATA REPOSITORIES*, PhD thesis, 2018 (cit. on p. 44).

[215] N. Press, *Understanding Metadata*, tech. rep., National Information Standards Organization, 2010 (cit. on pp. 31, 34, 35, 37).

[216] J. Priem, *Scholarship: Beyond the paper*, Nature **495**.7442 (2013) 437 (cit. on p. 11).

[217] *Protocol Buffers*, 2015, URL: `https://developers.google.com/protocol-buffers/` (cit. on p. 138).

[218] S. Purohit, W. Smith, A. Chappell, P. West, B. Lee, E. Stephan and P. Fox, "Effective Tooling for Linked Data Publishing in Scientific Research", *International Conference on Semantic computing*, 2016 (cit. on p. 145).

[219] E. Rajabi et al., "Interlinking educational data to web of data", *Big Data Optimization: Recent Developments and Challenges*, 2015 (cit. on p. 145).

[220] E. Rajabi, M.-A. Sicilia and S. Sanchez-Alonso, *Discovering duplicate and related resources using an interlinking approach: The case of educational datasets*, Journal of Information Science **41** (3 2015) (cit. on p. 145).

[221] E. Rajabi, M.-A. Sicilia and S. Sanchez-Alonso, *Interlinking educational resources to web of data through IEEE LOM*, Computer Science and Information Systems **12**.1 (2015) (cit. on pp. 145–147).

[222] S. Ram and J. Liu, *A semiotics framework for analyzing data provenance research*, Journal of computing Science and Engineering **2**.3 (2008) 221 (cit. on p. 35).

[223] S. H. Ramesh, A. Dhar, R. R. Kumar, V. Anjaly, K. Sarath, J. Pearce and K. Sundaresan, "Automatically Identify and Label Sections in Scientific Journals using Conditional Random Fields", *ESWC 2016 Challenges*, Springer, 2016 (cit. on p. 121).

[224] *Records in DBLP*, `http://dblp.org/statistics/recordsindblp`, 2016 (cit. on p. 37).

[225] *Research Excellence Framework: Second consultation on the assessment and funding of research*, tech. rep., 2009 (cit. on p. 17).

[226] I. T. Ribón, M. Vidal, B. Kämpgen and Y. Sure-Vetter, "GADES: A Graph-based Semantic Similarity Measure", *SEMANTICS*, 2016 (cit. on p. 163).

[227] A. Rodichevski, *Approximate String Matching Algorithms*, `http://www.morfoedro.it/doc.php?n=223&lang=en` (cit. on p. 148).

[228] F. Ronzano, G. C. del Bosque and H. Saggion, "Semantify CEUR-WS Proceedings: Towards the Automatic Generation of Highly Descriptive Scholarly Publishing Linked Datasets", *Semantic Web Evaluation Challenges*, Springer, 2014 83, URL: `http://dx.doi.org/10.1007/978-3-319-12024-9_10` (cit. on p. 119).

[229] F. Ronzano, B. Fisas, G. C. del Bosque and H. Saggion, “On the Automated Generation of Scholarly Publishing Linked Datasets: The Case of CEUR-WS Proceedings”, *Semantic Web Evaluation Challenges*, Springer, 2015 177, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_15` (cit. on pp. 119, 120).

[230] J. Rowley, *The wisdom hierarchy: representations of the DIKW hierarchy*, Journal of information science **33**.2 (2007) 163 (cit. on p. 32).

[231] M. Sachan and R. Ichise, *Using semantic information to improve link prediction results in network datasets*, IJET **2**.4 (2010) (cit. on p. 168).

[232] S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. T. Jr, S. Auer, J. Sequeda and A. Ezzat, *A survey of current approaches for mapping of relational databases to RDF*, W3C RDB2RDF Incubator Group Report, 2009 (cit. on p. 141).

[233] S. A. Salloum, M. Al-Emran, A. A. Monem and K. Shaalan, *A survey of text mining in social media: facebook and twitter perspectives*, Adv. Sci. Technol. Eng. Syst. J **2**.1 (2017) 127 (cit. on p. 33).

[234] I. Sartori and A. G. Hestnes, *Energy use in the life cycle of conventional and low-energy buildings: A review article*, Energy and buildings **39**.3 (2007) 249 (cit. on p. 42).

[235] B. Sateli and R. Witte, “An Automatic Workflow for the Formalization of Scholarly Articles’ Structural and Semantic Elements”, *ESWC 2016 Challenges*, Springer, 2016 (cit. on p. 120).

[236] B. Sateli and R. Witte, “Automatic Construction of a Semantic Knowledge Base from CEUR Workshop Proceedings”, *Semantic Web Evaluation Challenges*, Springer, 2015 129, URL: `http://dx.doi.org/10.1007/978-3-319-25518-7_11` (cit. on p. 120).

[237] S. Schaffert, “IkeWiki: A semantic wiki for collaborative knowledge management”, *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2006. WETICE’06. 15th IEEE International Workshops on*, IEEE, 2006 388 (cit. on p. 56).

[238] J. Schaible, T. Gottron and A. Scherp, “Survey on common strategies of vocabulary reuse in linked open data modeling”, *European Semantic Web Conference*, Springer, 2014 457 (cit. on p. 67).

[239] R. Schatz and H. Chen, *Digital libraries: technological advances and social impacts*, Computer **32**.2 (1999) 45 (cit. on p. 14).

[240] F. Schmedding, C. Hanke and T. Hornung, “Rdf authoring in wikis”, *Third Workshop on Semantic Wikis–The Wiki Way of Semantics 5 th European Semantic Web Conference Tenerife, Spain, June 2008*, Citeseer, 2008 87 (cit. on p. 56).

[241] *Semantic Web Evaluation Challenges*, Springer, 2014.

[242] *Semantic Web Evaluation Challenges*, Springer, 2015.

[243] D. Shotton, *Publishing: open citations*, Nature News **502**.7471 (2013) 295 (cit. on p. 114).

[244] D. Shotton, *Semantic publishing: the coming revolution in scientific journal publishing*, Learned Publishing **22**.2 (2009) 85 (cit. on p. 22).

[245] K. Shvachko, H. Kuang, S. Radia and R. Chansler, “The hadoop distributed file system”, *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*, Ieee, 2010 1 (cit. on p. 138).

[246] E. Simperl, S. Wölger, S. Thaler, B. Norton and T. Bv̈rger, *Combining human and computation intelligence: the case of data interlinking tools*, International Journal of Metadata, Semantics and Ontologies **7** (2 2012) (cit. on p. 147).

[247] J. Singh, *FigShare.*, Journal of pharmacology & pharmacotherapeutics **2**.2 (2011) 138 (cit. on p. 23).

[248] A. Singhal, *Introducing the knowledge graph: things, not strings*, Official google blog **5** (2012) (cit. on pp. 2, 41).

[249] U. Sivarajah, M. M. Kamal, Z. Irani and V. Weerakkody, *Critical analysis of Big Data challenges and analytical methods*, Journal of Business Research **70** (2017) 263 (cit. on p. 37).

[250] D. of Sociology, *Social Research Update*, 1998, URL: `http://sru.soc.surrey.ac.uk/SRU21.html` (visited on 06/08/2017) (cit. on pp. 49, 128).

[251] L. N. Soldatova and R. D. King, *An ontology of scientific experiments*, Journal of the Royal Society Interface **3**.11 (2006) 795 (cit. on p. 21).

[252] N. Solntseff and A. Yezerski, "A survey of extensible programming languages", *International Tracts in Computer Science and Technology and Their Application*, vol. 7, Elsevier, 1974 267 (cit. on p. 34).

[253] G. Spence, J. Heer and C. Manning, "The efficacy of human post-editing for language translation", *SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, 2013 439, ISBN: 978-1-4503-1899-0 (cit. on p. 82).

[254] R. Spier, *The history of the peer-review process*, TRENDS in Biotechnology **20**.8 (2002) 357 (cit. on p. 16).

[255] Stackoverflow, *Difference between Jaro-Winkler and Levenshtein distance*, `http://stackoverflow.com/questions/25540581/difference-between-jaro-winkler-and-levenshtein-distance` (cit. on p. 148).

[256] I. W. Stats, *Internet Users by Language*, `https://www.internetworldstats.com/stats7.htm` (cit. on p. 55).

[257] J. Sun and Q. Jin, "Scalable RDF store based on HBase and MapReduce", *Advanced Computer Theory and Engineering (ICACTE)*, vol. 1, IEEE, 2010 (cit. on p. 139).

[258] B. Sundgren, *An infological approach to data bases*, National Central Bureau of Statistics, Sweden; University of Stockholm, 1973 (cit. on p. 34).

[259] Y. Sure, S. Bloehdorn, P. Haase, J. Hartmann and D. Oberle, "The SWRC ontology-semantic web for research communities", *EPIA*, Springer, 2005 218 (cit. on p. 20).

[260] M. Svihla and I. Jelinek, "Benchmarking RDF production tools", *DEXA*, Springer, 2007 (cit. on p. 141).

[261] *The Big Data Lifecycle*, 2018, URL: `https://www.pinkelephantasia.com/big-data-lifecycle/` (cit. on p. 43).

[262] *The Ontology Alignment Evaluation Initiative*, 2015, URL: `http://oaei.ontologymatching.org` (visited on 20/01/2016) (cit. on p. 72).

[263] J. B. Thompson, *Books in the digital age: The transformation of academic and higher education publishing in Britain and the United States*, Polity, 2005 (cit. on p. 13).

[264] D. Tkaczyk and Ł. Bolikowski, “Extracting Contextual Information from Scientific Literature Using CERMINE System”, *Semantic Web Evaluation Challenges*, Springer, 2015 93, URL: http://dx.doi.org/10.1007/978-3-319-25518-7_8 (cit. on p. 119).

[265] I. Traverso-Ribón, G. Palma, A. Flores and M.-E. Vidal, “Considering semantics on the discovery of relations in knowledge graphs”, *EKAW*, 2016 (cit. on p. 168).

[266] UNESCO, *UNESCO Promotes New Initiative for Free Educational Resources on the Internet*, 2002, URL: http://www.unesco.org/education/news_en/080702_free_edu_ress.shtml (visited on 17/06/2014) (cit. on p. 77).

[267] United Nations, *Universal Declaration of Human Rights-Right To Education, Article 26*, 1948, URL: http://www.un.org/cyberschoolbus/humanrights/declaration/26.asp (cit. on p. 77).

[268] P. S. Unwin, G. Unwin and D. H. Tucker, *History of publishing*, 2017, URL: https://www.britannica.com/topic/publishing (cit. on p. 13).

[269] S. Vahdati, F. Karim, J. Huang and C. Lange, “Mapping Large Scale Research Metadata to Linked Data: A Performance Comparison of HBase, CSV and XML”, *MTSR*, Springer, 2015 (cit. on p. 10).

[270] S. Vahdati, C. Lange and S. Auer, “OpenCourseWare Observatory – Does the Quality of OpenCourseWare Live up to its Promise?”, *LAK*, ACM, 2015 (cit. on pp. 10, 68).

[271] S. Vahdati, N. Arndt, S. Auer and C. Lange, “OpenResearch: Collaborative Management of Scholarly Communication Metadata”, *EKAW*, Springer, 2016 (cit. on p. 10).

[272] S. Vahdati, A. Dimou, C. Lange and A. Di Iorio, “Semantic Publishing Challenge: Bootstrapping a Value Chain for Scientific Data”, *SAVE-SD Workshop at WWW conference*, SAVE-SD, 2016 73 (cit. on pp. 10, 114).

[273] S. Vahdati, F. Karim, J.-Y. Huang and C. Lange, “Mapping Large Scale Research Metadata to Linked Data: A Performance Comparison of HBase, CSV and XML”, *Metadata and Semantics Research*, Springer, 2015, ISBN: 978-3-319-24128-9, arXiv: 1506.04006 [cs.DB] (cit. on p. 146).

[274] S. Vahdati, C. Lange, G. Alexiou and G. Papastefanatos, *LOD Services*, Deliverable D8.2, OpenAIRE2020, 2015 (cit. on pp. 69, 145, 146).

[275] S. Vahdati, C. Lange, S. Auer, A. Behrend and I. Grangel-Gonzalez, *Towards a Comprehensive Quality Model for Scientific Events and Publications*, tech. rep., 2016 (cit. on p. 10).

[276] J. Vayssiere, *System and method for accessing RSS feeds*, US Patent App. 11/024,382, 2006, URL: https://www.google.com/patents/US20060155698 (cit. on p. 130).

[277] M. Vladoiu, *State-of-the-Art in Open Courseware Initiatives Worldwide*, Informatics in Education **10**.2 (2011) (cit. on p. 77).

[278] M. Vlădoiu, *Toward Increased Discoverability of Open Educational Resources and Open Courseware*, International Journal of Computer Trends and Technology **4**.8 (2013) (cit. on p. 89).

[279] M. Vlădoiu, “Towards a Quality Model for Open Courseware and Open Educational Resources”, *New Horizons in Web Based Learning*, LNCS 7697, Springer, 2014 213 (cit. on p. 27).

[280] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov, *Silk-a link discovery framework for the web of data.*, LDOW **538** (2009) (cit. on p. 53).

[281] H.-D. Wang and J. Wu, “Collaborative Filtering of Call for Papers”, (7th Dec. 2015), IEEE, 2015 963 (cit. on p. 27).

[282] P. Wang, B. Xu, Y. Wu and X. Zhou, *Link prediction in social networks: the state-of-the-art*, Link Prediction in Social Networks(SCIS) **58**.1 (2015) (cit. on p. 167).

[283] Z. Wang, J. Zhang, J. Feng and Z. Chen, “Knowledge Graph Embedding by Translating on Hyperplanes.”, *AAAI*, 2014 1112 (cit. on p. 41).

[284] M. Ware, *Peer review: benefits, perceptions and alternatives*, Publishing Research Consortium **4** (2008) 1 (cit. on p. 16).

[285] L. Wayne, *Tetadata In Action: Expanding the Utility of Geospatial Metadata*, Federal Geographic Data Committee, 2005 (cit. on p. 42).

[286] J. Weckert and R. Lucas, *Professionalism in the information and communication technology industry*, ANU Press, 2013 (cit. on p. 33).

[287] W. B. Weeks, A. E. Wallace and B. S. Kimberly, *Changes in authorship patterns in prestigious US medical journals*, Social Science & Medicine **59**.9 (2004) 1949 (cit. on p. 15).

[288] S. Weibel, J. Godby, E. Miller and R. Daniel, *OCLC/NCSA metadata workshop report*, (1995) (cit. on p. 34).

[289] *What is peer review?*, Elsevier, 2016, URL: `https://www.elsevier.com/reviewers/what-is-peer-review` (visited on 24/01/2016) (cit. on p. 96).

[290] F. Wiedijk, *The “de Bruijn factor”*, 2012, URL: `http://cs.ru.nl/~freek/factor/` (cit. on p. 143).

[291] A. Wierse and T. Riedel, *Smart Data Analytics: Mit Hilfe von Big Data Zusammenhänge erkennen und Potentiale nutzen*, Walter de Gruyter GmbH & Co KG, 2017 (cit. on p. 33).

[292] Wikipedia, *Artificial neural network*, [Online; accessed 23-August-2018], 2018, URL: `%5Curl%7Bhttps://en.wikipedia.org/wiki/Artificial_neural_network%7D` (cit. on p. 22).

[293] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne et al., *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific data **3** (2016) (cit. on pp. 15, 35).

[294] A. Wright, *Cataloguing the World: Paul Otlet and the Birth of the Information Age (in Introduction)*, Oxford University Press, 2014 (cit. on p. 14).

[295] Z. Wu, J. Wu, M. Khabsa, K. Williams, H.-H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra et al., “Towards building a scholarly big data platform: Challenges, lessons and opportunities”, *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, IEEE Press, 2014 117 (cit. on p. 21).

[296] S. Wuchty, B. F. Jones and B. Uzzi, *The increasing dominance of teams in production of knowledge*, Science **316**.5827 (2007) 1036 (cit. on p. 15).

[297] F. Xia, W. Wang, T. M. Bekele and H. Liu, *Big scholarly data: A survey*, IEEE Transactions on Big Data **3**.1 (2017) 18 (cit. on pp. 5, 6, 37, 38).

[298] F. Xia, W. Wang, T. M. Bekele and H. Liu, *Big Scholarly Data:A Survey*, IEEE Big Data (2017) (cit. on pp. 162, 167).

[299] J. Xia, K. Wen, R. Li and X. Gu, “Optimizing Academic Conference Classification Using Social Tags”, *CSE*, (Hong Kong, China), IEEE, 2010 289 (cit. on p. 26).

[300] Z. Yang and B. D. Davison, "Venue Recommendation: Submitting Your Paper with Style", *ICMLA*, vol. 1, IEEE, 2012 (cit. on p. 95).

[301] O. Zamazal, M. Cheatham and A. Solimando, *Conference Track*, 2015, URL: `http://oaei.ontologymatching.org/2015/conference/index.html` (visited on 23/01/2016) (cit. on p. 72).

[302] S. Zamore, K. Ohene Djan, I. Alon and B. Hobdari, *Credit Risk Research: Review and Agenda*, Emerging Markets Finance and Trade **54**.4 (2018) 811 (cit. on p. 16).

[303] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer, *Quality Assessment for Linked Data*, Semantic Web Journal **7**.1 (2016) 63, URL: `http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey` (cit. on p. 80).

[304] D. Zhao and A. Strotmann, *The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis*, Journal of the Association for Information Science and Technology **65**.5 (2014) 995 (cit. on p. 22).

[305] P. Ziegler and K. R. Dittrich, "Data integration—problems, approaches, and perspectives", *Conceptual modelling in information systems engineering*, Springer, 2007 39 (cit. on p. 1).

Acknowledgements

I thank my supervisor Prof. Sören Auer for providing me the opportunity to develop my academic life by joining (the former) Enterprise Information Systems (EIS) group at the Institute for Computer Science III and being involved in the Open Research vision of him. I am forever grateful for Prof. Rainer Manthey who have always supported and encouraged me in the hard and challenging times and also in achievements.

I also want to thank Dr. Christoph Lange with whom I had many interesting discussions which matured this research work. Particular thanks are due to Prof. Maria-Esther Vidal for her guidance, encouragement and substantial discussions during the development of this thesis.

I also appreciate the great opportunity to be part of the Smart Data Analytics (SDA) research group at the Institute for Computer Science III led by Prof. Jens Lehmann who has always inspired me with his outstanding ideas and research results. I have been fortunate to receive extremely valuable feedback and guidance from Dr. Maria Maleshkova and Dr. Ioanna Lytra, and Dr. Andreas Behrend whom I am forever grateful for. I want to thank all my colleagues in the SDA and EIS groups for all the fun and inspiring discussions we had as it helped me to acquire and improve further academic skills.

During the last years, I met a number of excellent researchers who supported me by opening up new research opportunities. In particular, I thank to Dr. Silvio Peroni and Dr. Andrea Scharnhorst involving me deeply in the activities and events mainly in the Semantic Publishing and e-research communities.

Most of this research was financially supported by the OpenAIRE H2020 European Union project. I am grateful to have the opportunity to collaborate with our technical colleagues in this project from the National Research Council of Italy, Dr. Poalo Manghi and Dr. Claudio Atzori, and Linked Open Data team of Athena Research Center from Greece, Giorgos Alexiou.

Finally, I would like to thank my family and friends to be in my life and making it a magnificent journey for me by always keeping me motivated, dedicated, and inspired.