Linked Research on the Decentralised Web

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

vorgelegt von
Sarven Capadisli
aus
Istanbul, Turkey

Bonn, 2019-07-29

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

1. Gutachter: Prof. Dr. Sören Auer
2. Gutachter: Dr. Herbert Van de Sompel

Tag der Promotion 2020-03-03
Erscheinungsjahr 2020

Abstract

This thesis is about research communication in the context of the Web. I analyse literature which reveals how researchers are making use of Web technologies for knowledge dissemination, as well as how individuals are disempowered by the centralisation of certain systems, such as academic publishing platforms and social media. I share my findings on the feasibility of a decentralised and interoperable information space where researchers can control their identifiers whilst fulfilling the core functions of scientific communication: registration, awareness, certification, and archiving.

The contemporary research communication paradigm operates under a diverse set of sociotechnical constraints, which influence how units of research information and personal data are created and exchanged. Economic forces and non-interoperable system designs mean that researcher identifiers and research contributions are largely shaped and controlled by third-party entities; participation requires the use of proprietary systems.

From a technical standpoint, this thesis takes a deep look at semantic structure of research artifacts, and how they can be stored, linked and shared in a way that is controlled by individual researchers, or delegated to trusted parties. Further, I find that the ecosystem was lacking a technical Web standard able to fulfill the awareness function of research communication. Thus, I contribute a new communication protocol, *Linked Data Notifications* (published as a W3C Recommendation) which enables decentralised notifications on the Web, and provide implementations pertinent to the academic publishing use case. So far we have seen decentralised notifications applied in research dissemination or collaboration scenarios, as well as for archival activities and scientific experiments.

Another core contribution of this work is a Web standards-based implementation of a clientside tool, *dokieli*, for decentralised article publishing, annotations and social interactions. dokieli can be used to fulfill the scholarly functions of registration, awareness, certification, and archiving, all in a decentralised manner, returning control of research contributions and discourse to individual researchers.

The overarching conclusion of the thesis is that Web technologies can be used to create a fully functioning ecosystem for research communication. Using the framework of Web architecture, and loosely coupling the four functions, an accessible and inclusive ecosystem can be realised whereby users are able to use and switch between interoperable applications without interfering with existing data.

Technical solutions alone do not suffice of course, so this thesis also takes into account the need for a change in the traditional mode of thinking amongst scholars, and presents the *Linked Research* initiative as an ongoing effort toward researcher autonomy in a social system, and universal access to human- and machine-readable information. Outcomes of this outreach work so far include an increase in the number of individuals self-hosting their research artifacts, workshops publishing accessible proceedings on the Web, in-the-wild experiments with open and public peer-review, and semantic graphs of contributions to conference proceedings and journals (the Linked Open Research Cloud).

Some of the future challenges include: addressing the social implications of decentralised Web publishing, as well as the design of ethically grounded interoperable mechanisms; cultivating privacy aware information spaces; personal or community-controlled on-demand archiving services; and further design of decentralised applications that are aware of the core functions of scientific communication.

Author: Sarven Capadisli <https://csarven.ca/#i>

Identifier: https://csarven.ca/linked-research-decentralised-web

Created: 2016-04-13

Modified: 2019-07-29

Published: 2019-07-29

Latest Version: https://csarven.ca/archives/linked-research-decentralised-web/ce36de40-64a7-4d57-a189-f47c364daa74

TimeMap: https://csarven.ca/linked-research-decentralised-web.timemap

Derived From:

- *Linked SDMX Data*
- *Linked Research*
- *Linked Statistical Data Analysis*
- *Semantic Similarity and Correlation of Linked Statistical Data Analysis*
- *Call for Linked Research*
- *Enabling Accessible Knowledge*
- *This ‘Paper’ is a Demo*
- *Cooling Down Web Science*
- *dokieli*
- *Linked Research: An Approach for Scholarly Communication*
- *Linked Data Notifications*
- *Sparqlines: SPARQL to Sparkline*
- *Where is Web Science? From 404 to 200*
- *Embedding Web Annotations in HTML*
- *Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli*
- *Full Article, Immediate, Permanent, Discoverable, and Accessible to Anyone Free of charge*
- *LDN Test Reports and Summary*
- *Linking Specifications, Test Suites, and Implementation Reports*

Language: English

License: Creative Commons Attribution 4.0 International

Notifications Inbox: https://csarven.ca/archives/linked-research-decentralised-web/inbox/

In Reply To: Call for Linked Research

Document Status: Published

Declaration

I declare that the work covered by this thesis is composed by myself, and that it has not been submitted for any other degree or professional qualification except as specified.

Acknowledgements

This thesis is the result of interconnected ideas and people coming together. I would like the following to be carved in a hyperdimensional stone:

Brigitte Schuster, Emilian Capadisli, Junes Capadisli. My parents and brother.

Sören Auer for his supervision, endurance and support; the force that kept my work and initiatives real – a massive understatement.

Captain Amy Guy for her advice and friendship. Immensely grateful; my work would not be where it is today without her collaboration as well as justifiable distractions of epic proportions.

Herbert Van de Sompel for helping me to connect the fundamental dots to better situate my work in context of research communication. A mentor.

Tim Berners-Lee, *the* Web developer that I look up to for inspiration and model perseverance to continue fighting for the Web. I am grateful to collaborate with Tim on the challenges ahead of us.

Kingsley Idehen, my practise what you preach partner in crime. I have learned a lot from Kingsley as we aimed to realise cool Web stuff.

Amy van der Hiel for ethical Web checks and having me remind myself to assess the value of what I do with respect to society and adjusting my aims accordingly. Her influence will continue to shape the fabric of the projects that I am involved in.

Bernadette Hyland for supporting me to hang in there for the long game.

Stian Soiland-Reyes for brainstorming and experimenting on various Linked Data stuff.

Ruben Verborgh for being an exemplary researcher and developer to learn from, and support to better integrate our work out there in the wild.

Andrea Scharnhorst for helping me to develop a sense of academic collaboration.

Axel Polleres for his guidance and supporting initiatives to evolve research communication.

Dame Wendy Hall for keeping the spirit of the Web in scholarly communication alive.

Ilaria Liccardi for her mentorship and care to help me move my research forward.

Henry Story for providing context to undoubtedly interconnected ideas in technology and philosophy.

Melvin Carvalho for bouncing ideas on the intersection of McLuhan’s media theories and the Web.

Christophe Lange for helping to frame my work as research where I thought it was just development.

Jodi Schneider for helping to improve my argumentation in research communication.

Miel Vander Sande for pushing me in the right direction to frame the design science of my work.

Albert Meroño-Peñuela for collaborating on LSD and LDN, and getting interesting results.

Paul Groth is partly to blame for all this work as he (indirectly) challenged me to make it so.

Alexander Garcia Castro for initially nudging me to formulate my vision for scholarly communication which later set the core of my research and development.

The (anti?) Social Web: from your encouragement to your dismissal of radical and lunatic scholarly endeavours; being the widest sounding board possible.

Contents

Figures

Tables

Abbreviations

?
Cannot tell

⌙
Inapplicable

○
Untested

✔

Passed

✗
Failed

a11y
Accessibility

AAA
Anyone can say Anything about Anything

ABS
Australian Bureau of Statistics

ACL
Access Control List

AP
ActivityPub

APC
Article Processing Charge

ARK
Archival Resource Keys

AS2
Activity Streams 2.0

ATAG
Authoring Tool Accessibility Guidelines

AWWW
Architecture of the World Wide Web

BFS
Bundesamt für Statistik

BIS
Bank for International Settlements

BOAI
Budapest Open Access Initiative

CC
Creative Commons

CC0 1.0
Creative Commons CC0 1.0 Universal

CiTO
Citation Typing Ontology

CORS
Cross-Origin Resource Sharing

CRUD
Create, Read, Update, and Delete

CSS
Cascading Style Sheets

CWA
Closed-world Assumption

DID
Decentralized Identifiers

Disco
DDI-RDF Discovery Vocabulary

DNS
Domain Name System

DOAP
Description of a Project

DOI
Digital Object Identifier

DOM
Document Object Model

DPUB-ARIA
Digital Publishing WAI-ARIA Module

DRY
Don’t Repeat Yourself

DSSP
DataSpace Support Platform

DWBP
Data on the Web Best Practices

EARL
Evaluation and Report Language

EAV
entity–attribute–value

ECB
European Central Bank

EEA
European Economic Area

EPUB
Electronic Publication

ETL
Extract Transform Load

EU
European Union

FAIR
Findable, Accessible, Interoperable, Resuable

FAO
Food and Agriculture Organization of the United Nations

FOAF
Friend of a Friend

FOI
Freedom of Information

FRB
Federal Reserve Board

GDPR
General Data Protection Regulation

GRDDL
Gleaning Resource Descriptions from Dialects of Languages

HTML
Hypertext Markup Language

HTTP
Hypertext Transfer Protocol

HTTPS
Hypertext Transfer Protocol Secure

i18n
Internationalization

ICANN
Internet Corporation for Assigned Names and Numbers

IIIF
International Image Interoperability Framework

IMF
International Monetary Fund

IMRAD
Introduction, Methods, Results, and Discussion

IP
Internet Protocol

IRI
Internationalized Resource Identifier

ISBN
International Standard Book Number

ISSN
International Standard Serial Number

JATS
Journal Article Tag Suite

JIF
Journal Impact Factor

l10n
Localization

LANL
Los Alamos National Laboratory

LDF
Linked Data Fragments

LDN
Linked Data Notifications

LDP
Linked Data Platform

LDP-UCR
Linked Data Platform Use Cases and Requirements

LDT
Linked Data Templates

LOD
Linked Open Data

LORC
Linked Open Research Cloud

LR
Linked Research

LSD
Linked Statistical Data

MathML
Mathematical Markup Language

OA
Open Access

OAI
Open Archives Initiative

ODRL
Open Digital Rights Language

OECD
Organisation for Economic Co-operation and Development

OIDC
OpenID Connect

OPR
Open Peer Review

ORCID
Open Researcher and Contributor ID

OWA
Open-world Assumption

OWP
Open Web Platform

PAYGO
pay as you go

PDF
Portable Document Format

PDS
Personal Data Service

PGP
Pretty Good Privacy

PID
Persistent IDentifier

PII
Personally Identifiable Information

PIM
Personal Information Management

POD
Personal Online Datastore

PROV-O
The PROV Ontology

PURL
Persistent Uniform Resource locator

PuSH
PubSubHubbub

RACA
Registration, Awareness, Certification, Archiving

RASH
Research Articles in Simplified HTML

RDF
Resource Description Framework

REST
Representational State Transfer

RFC 1123
Requirements for Internet Hosts -- Application and Support

RFC 2818
HTTP over TLS

RFC 3986
Uniform Resource Identifier (URI): Generic Syntax

RFC 5246
The Transport Layer Security Protocol Version 1.2

RFC 5261
XML Patch

RFC 5280
Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile

RFC 5829
Link Relation Types for Simple Version Navigation between Web Resources

RFC 6455
The WebSocket Protocol

RFC 6749
The OAuth 2.0 Authorization Framework

RFC 6920
Naming Things with Hashes

RFC 7089
HTTP Framework for Time-Based Access to Resource States -- Memento

RFC 7159
The JavaScript Object Notation (JSON) Data Interchange Format

RFC 7230
Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing

RFC 7231
Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content

RFC 7351
A Media Type for XML Patch Operations

RFC 7617
The 'Basic' HTTP Authentication Scheme

RFC 8030
Generic Event Delivery Using HTTP Push

ROL
Republic of Letters

RS
The Royal Society

SAN
Subject Alternative Name

SCA
Semantic Content Authoring

SCTA
Scholastic Commentaries and Texts Archive

SDMX
Statistical Data and Metadata eXchange

SKOS
Simple Knowledge Organization System

Solid
Social Linked Data

SPAR
Semantic Publishing and Referencing Ontologies

SPARQL
SPARQL Protocol and RDF Query Language

SSN
Semantic Sensor Network

STM
Science, Technology, and Medicine

SVG
Scalable Vector Graphics

SWWG
Social Web Working Group

TAG
Technical Architecture Group

TEI
Text Encoding Initiative

TLS
Transport Layer Security

TPF
Triple Pattern Fragments

UDI
Universal Document Identifier

UIS
UNESCO Institute for Statistics

URI
Uniform Resource Identifier

URL
Uniform Resource Locator

UUAG
User Agent Accessibility Guidelines

UX
User Experience

VoID
Vocabulary of Interlinked Datasets

WA
Web Annotation

WAC
Web Access Control

WAI-ARIA
Accessible Rich Internet Applications

WAP
Web Annotation Protocol

WCAG
Web Content Accessibility Guidelines

WebID
Web Identity and Discovery

WoT
Web of Trust

XMP
Extensible Metadata Platform

XPointer
XML Pointer Language

XSLT
XSL Transformations

> The past went that-a-way. When faced with a totally new situation we tend always to attach ourselves to the objects, to the flavor of the most recent past. We look at the present through a rear-view mirror. We march backward into the future. Suburbia lives imaginatively in Bonanza-land.
>
> *The Medium is the Massage: An Inventory of Effects* [1], p. 74-75, Marshall McLuhan, 1967

1. Introduction

> 'Why' is the only real source of power, without it you are powerless.
>
> *The Matrix Reloaded* [2], Merovingian, 2003

1.1 Motivation

> Would you do me a favour? I'd like to stop talking for a minute and when I do, take a look at the room you're in and above all at the man-made objects in that room that surround you – the television set, the lights, the phone and so on – and ask yourself what those objects do to your life just because they're there. Go ahead.
>
> *Connections* [3], James Burke, 1979

Well, that is what this thesis is going to be all about. It is about the modern research communication, and just because it is there, shapes the way we think and behave, and why it exists in the form it does, and who or what was responsible for it to exist at all. What is an alternative way of change?

The idiosyncratic relationships between technological breakthroughs and societal transformations throughout history are intrinsically intertwined. It was not until I came across *Connections* [4], the TV series by James Burke that I had a glimpse of the phenomenon about the increasingly interlinked human endeavours over time. After encountering Marshall McLuhan’s theories on communication and how media have the power to shape and transform human nature, what to do next was mostly clear (in my head). Given these realisations or perspectives as a foundation, being a Web technologist only enabled me to build the necessary connections between what was previously missing or underdeveloped. The rest was mostly a matter of struggling with the ghost in the machine and working alongside our shared social challenges.

Ironically, even today, academics essentially operate within a paper-centric framework to create and disseminate publicly-funded knowledge by usually delegating it to third-party publishers which still operate in the 15th century. Meanwhile, the Web – if we can anthropomorphise for a moment – is disappointed by the distracted academics’ practices. I took the liberty to test the boundaries and demonstrate some of the aspects of what the native affordances of the Web provided – an extension of our central nervous system.

I live in an *information society* [5], which makes use of variety of communication media to meet societal needs. The Web has shaped social policies and practices around the world, as well as influence the creation and use of other technologies. It has also reconfigured individuals to be active participants in global information exchange, as well as passive consumers partly due to the abundance of instantaneously available multimedia. To date – 30 years old as of this writing – the Web is considered to be a net positive for society all meanwhile open and complex challenges remain. Like any human technology, what we devise and how we use them has societal implications.

We can start this story from anywhere. I argue that in order to best contextualise this thesis, it is essential to operate under the understanding of communication mediums and scientific revolutions, as well as their effects on society. This is especially because the research goals of this thesis at its core is *sociotechnical*, and that a clear division between them will not do its justice. I believe that in order to push the boundaries of our knowledge further, it necessitates a multidisciplinary undertaking – taken “with a grain of salt”. One overarching goal is to foster the social machinery from different perspectives that is required to build a knowledge Web for any type of user; human, machine, or other. While I set the problem context to scholarly communication, applications of the knowledge and generated artifacts are potentially applicable to other initiatives in society.

I assume that researchers in scholarly communication are motivated to some extent personally, whether that is with the goal to take part in advancing collective knowledge, society and life, career advancement, prestige, or financial gain. Researchers are intrinsically interested in making their impact by way of sharing their findings in a manner that is most accessible to potential users. As for the underlying technical machinery for information exchange, I assume that it is useful to aim in designing interoperable information systems to help us address problems in context of societal goals. All of these aspects are part of the “social machine” of scholarly communication, and are explored through the sections.

As many have argued, for scholarly communication to be more effective; accessible, usable, inclusive, as well as equitable, we need to adopt methods that are better at identifying and making use of the native affordances of the Web. This is particularly important today since academic and social activities use the Web as the primary medium to communicate (in contrast to systems that mimic paper media). Hence, I examine the core characteristics of the Web, discuss the state of affairs in scholarly communication, and share my findings on how a way to fulfill the forces and functions in scientific communication can be met by a socially-aware decentralised client application along with an ocean of open Web standards at its disposal.

In order to frame the “design science” of this thesis, the sections on *Research Goals* and *Research Questions* are based on *Research Goals and Research Questions* in *Design Science Methodology for Information Systems and Software Engineering* [6], Wieringa, 2014.

1.2 Research Goals

This section contains the *outlines* for social and technical research goals of this thesis. Each goal is further explored and related material is reviewed in relevant sections.

1.2.1 Stakeholders

In research projects, there are various stakeholders (in)directly acting as internal and external forces in scholarly communication. In order to address societal concerns, research sponsors fund projects to attain reusable technical designs and acquire new knowledge in the process. Knowledge workers like researchers and scholars are driven by curiosity to advance our understanding of the universe and share their findings. The public is interested in accessing applications of research to educate and for the well-being of life. The industry needs efficient functioning of scientific communication in order to translate the return of research capital.

I am a member of stakeholders acting in particular as a knowledge worker, a part of public, an operator (“end user”, maintainer, operational supporter).

1.2.2 Problem Context

There are several sociotechnical limitations in the contemporary paradigm that the scholarly communication operates in. I will summarise a subset of problem areas pertaining to interoperability of systems and autonomy of actors. Each problem space will be reviewed in relevant sections of this thesis.

There is an abundance of scientific information on the Web. However, humans and machines have restricted access, are required to pay, or hindered to efficiently discover, make sense of, and create new connections between information.

Data on the Web tends to be locked into non-interoperable Web systems or be only usable by applications which created it. Actors – creators and readers – in these systems often do not have the means to switch between applications and share data between them.

There is a lack of diverse value chains for different units of scholarly communication due to tightly coupled proprietary systems. The actors in these systems create and exchange information by (being forced into) using packaged systems and workflows, give up their rights, and not have control over how their creations can be disseminated and reused.

The actors in the system commonly use print-centric tools and data representations to exchange content on the Web and neglect to take advantage of the available affordances of the Web (media). Actors' possibility to shape their contributions in Web media is typically constrained to third-party service provider's centralised and non-interoperable solutions.

Actors' use of personal and professional identifiers, as well as data storage from third-parties are subject to privacy issues. Actors typically manage multiple disconnected profiles, social connections, and data spaces.

These class of problems in the scholarly ecosystem have implications on access, interoperability, findability, cost and privacy. Consequently, resulting in global deficiencies in the ecosystem. The research questions; *Technical Research Problems* and *Knowledge Questions*, are aimed at understanding the internal mechanisms of these type of problems and address them for stakeholders.

1.2.3 Technical Research Goals

Given the *Problem Context*, the technical goals of this thesis are to identify gaps in the Web standards landscape and to create new technical standards where necessary. A further goal is to demonstrate how existing and new standards and technologies can be used in combination as part of a user-facing application for publishing and interacting with scholarly artifacts.

The *Research Questions* are geared to meeting these goals.

1.2.4 Knowledge Goals

The knowledge goals of this thesis describe the theoretical underpinnings necessary to accomplish the *Technical Research Goals*.

We need to understand the various technological mediums in which scholarly communication has historically operated, and how and why this has changed, in order to have a frame of reference for new work. Our investigation must examine scholarship with respect to the media employed for publication, dissemination and feedback, and identify some of the constraints and limitations. This research can serve to describe the effects and artifacts of scholarly communication, as well as conceptual and technical designs that are used to exchange units of information by both humans and machines.

The scholarly communication ecosystem is a complex sociotechnical system, which cannot be understood from either a social or a technical perspective alone. Understanding how these aspects feed into and influence on another is crucial for the furtherance of this work.

The corpus of knowledge collected in this thesis is driven by understanding the current state of knowledge in the field based on scientific, technical, and professional literature. *Literature Review and Citations* describes the constraints on how research knowledge will be collected.

1.3 Research Questions

I first outline the *Technical Research Problems* and then a set of corresponding *Knowledge Questions*. The questions refine the *Research Goals*.

1.3.1 Technical Research Problems

The technical research problems here consist of artifacts to (re)design, requirements to satisfy, stakeholders' goals to achieve, and the problem context. The overarching technical research problem is defined as follows:

Central Design Problem
How can we design decentralised Web applications for information exchange in an open and decoupled scholarly communication ecosystem?

This research problem can be refined into two sub-problems:

Mechanisms
What technical mechanisms, standards or protocols are necessary for decentralised information exchange on the Web? Which already exist and what is missing?

Artifacts
How can Web technologies be employed to fulfill the core functions of scholarly communication in an open and interoperable way?

1.3.2 Knowledge Questions

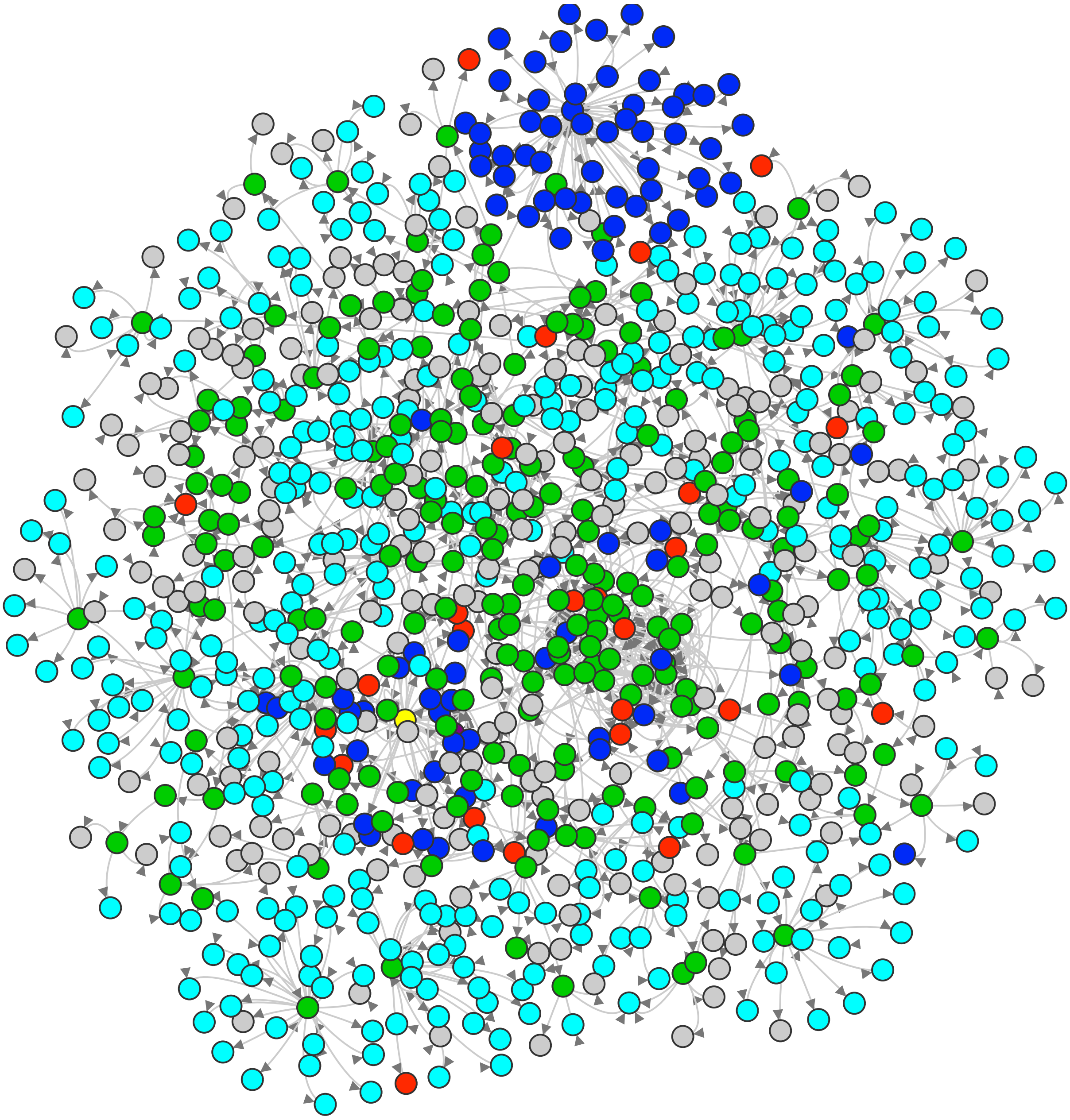
The *Knowledge Goals* can be refined into following knowledge questions:

- How can we situate scholarly communication using the Web with respect to technological communication mediums and scientific paradigms?
- What are the core components of scientific communication, how are they configured, and what are the effects of the interactions between them?
- What are contemporary practices around knowledge creation and dissemination, and how are communities working to improve these?

1.4 Thesis Overview

> Static stand-alone paper is not an appropriate habitat for a thesis on linking. Therefore, the paper is included with the CD-ROM, not vice versa.
>
> *Dynamic and context-sensitive linking of scholarly information* [7], Herbert Van de Sompel, 2000

This Thesis is a “Knowledge Graph”. *Linked Data* aware applications can use the resource’s underlying structured data to offer # visualisation and interaction possibilities.



All information of significance in this thesis is made available as *Linked Data* on the Web and archived.

1.4.1 Structure

This thesis is sectioned into following main topic areas:

Scholarly Communication on the Web

Answers knowledge questions to describe and explain characteristics of the Web and the state of scholarly communication on the Web.

Structure of Scholarly Information

Answers knowledge questions to identify common structural and semantic patterns for human- and machine-readable information. Contributions: *Linked Statistics* on publishing and exchanging statistical Linked Data.

Decentralising Scholarly Communication

Answers knowledge questions pertaining to design of socially-aware decentralised systems, as well as classification of Web specifications for forces and functions in scientific communication. Contributions: *Degree of Control*, *Forces and Functions in Specifications*, *The Effects and Artifacts of Autonomous Engagement*.

Linked Data Notifications

Answers technical research problems on how to design a decentralised notification protocol. Contributions: Design of *A Decentralised Notifications Protocol* to exchange reusable decentralised notifications.

Decentralised Linked Research Application

Answers technical research problems on how to design a decentralised application. Contributions: *Linking the Decentralised Information Space*, *Implementing a Read-Write Application*, *Forces and Functions in dokieli*.

Linked Research

A description and explanation of the effects of applying Web-centric standards and practices in order to enable creators and consumers of open knowledge, as well as a generalisation of core principles that can serve as an exemplar to set forth a shift in scholarly communication. Contributions: *Design Principles*, *Call for Linked Research*, *Forces and Functions of Linked Research, Linked Research as a Paradigm*.

In *Conclusions*, I will discuss the thesis findings in *Research Questions Review*, offer my *Interpretations*, and share *Perspectives* for the future.

1.4.2 Literature Review and Citations

The five review characteristics; focus, goal, perspective, coverage, and audience, are borrowed from the summary at *Practical Assessment, Research & Evaluation* [8], Randolph, 2009, which is based on *Organizing knowledge syntheses: A taxonomy of literature reviews*, Cooper, 1988.

Focus: The core focus of the reviews and cited material was to establish a connection between the available theories and their underlying mechanisms, as well as the application of certain interventions to help identify and apply a practical need. Hence, majority of the analysis and synthesis focuses on practices or applications in the field.

Goal: My intention is to present this thesis in a way that is technically and meaningfully connected with the knowledge out there, as well as to connect my contributions. Hence, I would like any reader (human, machine, or other) in the future to be able to observe and verify the connections that I have put together as much as possible. The main interest for the reviews and citations were information integration for the big picture, identification and discussion of central issues, and explicating arguments.

Perspective: As this thesis is about connecting and endorsing open scholarly knowledge in context of decentralised and social Web, there is a selection bias in favour of (research) information based on the following criteria:

- Publicly accessible URL in the case of digital objects.
- Referencable URI in the case of printed material.
- Article and data including full text and media.
- Accessible by anyone (humans and machines).
- Available to anyone without financial, legal, or technical barriers (besides an Internet connection, free of charge).
- Preference for source materials (as opposed to third-party re-publications).
- Preference for research artifacts that can be publicly archived and recalled with free on-demand services.
- Preference for research artifacts made available with declarative languages (as opposed to dynamically generated as part of an application).

At the time of this writing, HTTP URIs of research objects resolving to “paywall” or “access toll” landing pages were excluded. The rationale is that if information is *only* available to a subset of humans, then it is not deemed to be part of the observable and reproducible universal knowledge – a fundamental requirement of applying the scientific method. Fortunately, I was able to gather plenty of open and free *resources* to support the narrative of this thesis, and addressing its knowledge goals.

Coverage: The reviewed and cited works are composed of a purposive sample, examining only pivotal works and the source of concepts in the field to tie them together. While this selection is naturally not exhaustive in and itself, I have considered, located, and discussed material that was available at the time of my investigation. While there may be natural imperfections in data collection, evaluation, analysis, and interpretation in any literature review and bibliographic citations, I have aimed to be systematic in what’s covered. To that end, the works – text, data, media, or other – mentioned in this thesis were included or excluded based on the following criteria:

- Information in English.
- Web accessible digital objects.
- Printed material available through a public library loan.
- Referencing information that is legally made available to the public (to the best of my knowledge!)

Organisation: Historical and conceptual formats were used as organisation schemes. My observations and categorisations are in *Scholarly Communication on the Web*, *Structure of Scholarly Information* and *Decentralising Scholarly Communication* are first organised conceptually, and then chronologically ordered.

Audience: for anyone interested in traversing the connections and building on the underlying findings. See also *Audience*.

1.4.3 Document Convention

Unless otherwise noted, all of the quotations are *sic*.

With the exception of referring to existing literature, references made to *research*, *scientific* or *scholarly* communication in this thesis is discipline agnostic. Hence, I mean any academic discipline, from humanities, social, natural, formal, to applied sciences.

The prefixes and namespaces that are used in this article are as follows:

`acl`
http://www.w3.org/ns/auth/acl#

`as`
https://www.w3.org/ns/activitystreams#

`cert`
http://www.w3.org/ns/auth/cert#

`cito`
http://purl.org/spar/cito/

`contact`
http://www.w3.org/2000/10/swap/pim/contact#

`dcterms`
http://purl.org/dc/terms/

`doap`
http://usefulinc.com/ns/doap#

`earl`
http://www.w3.org/ns/earl#

`foaf`
http://xmlns.com/foaf/0.1/

`ldn`
https://www.w3.org/TR/ldn/#

`ldp`
https://www.w3.org/ns/ldp#

`mem`
http://mementoweb.org/ns#

`oa`
http://www.w3.org/ns/oa#

`owl`
http://www.w3.org/2002/07/owl#

`pim`
http://www.w3.org/ns/pim/space#

`prov`
http://www.w3.org/ns/prov#

`qb`
http://purl.org/linked-data/cube#

`rdf`
http://www.w3.org/1999/02/22-rdf-syntax-ns#

`rdfs`
http://www.w3.org/2000/01/rdf-schema#

`solid`
http://www.w3.org/ns/solid/terms#

`schema`
http://schema.org/

`skos`
http://www.w3.org/2004/02/skos/core#

`void`
http://rdfs.org/ns/void#

`xhv`
http://www.w3.org/1999/xhtml/vocab#

2. Scholarly Communication on the Web

In this section I describe and explain characteristics of the Web and the state of scholarly communication on the Web.

2.1 Mediums and Paradigms

The effects of technological mediums on society have some similarities to the evolution of scientific paradigms. This will serve as the historical framing for this thesis and helps to contextualise the application of core methods and values of the Web on scholarly communication.

2.1.1 On Mediums

> Gutenberg had, in effect, made every man a reader. Today, Xerox and other forms of reprography tend to make every man a publisher. This massive reversal has, for one of its consequences, elitism. The nature of the mass production of uniform volumes certainly did not foster elites but rather habits of universal reading. Paradoxically, when there are many readers, the author can wield great private power, whereas small reading elites may exert large corporate power.
>
> *Understanding Me* [9], The Future of the Book, p. 179, Marshall McLuhan, 1972

In the 1962 book, *The Gutenberg Galaxy* [10], Marshall McLuhan explains how the adoption of new technological mediums shifted human culture in Europe through four eras, with emphasis placed on extending our senses:

i. acoustic age (c. until 800 BCE); audile-tactile, preliterate tribal culture,
ii. age of writing (c. 800 BCE to 1500 CE); audile/visual, manuscript culture,
iii. age of print (c. 1500 to 1850 CE); visual, mechanical mass communication,
iv. age of electronic media (c. 1850 CE to present); central nervous system, “global village”

Acoustic age: According to McLuhan, the acoustic world depended on speech and hearing for communication. Information was an instantaneous and simultaneous experience for speakers and listeners with a collective identity. Existence was about narrated events and survived through story-telling, songs, and poems.

Writing age: When the phonetic alphabet came along, information could be defined, classified, and referenced, as well as better preserved. The written word made human knowledge tangible, stretching itself across time and space. While the written word increased the need of our visual system, it brought along the acoustic world with it – early reading was still an oral, and by and large an external (social) activity, as opposed to internal in one’s mind.

Print age: Through the age of writing to the age of print, the phonetic alphabet transformed from being an audile-tactile experience to purely visual. Gutenberg’s printing press (c. 1439 CE), through the process of mechanical mass production, had an extensive impact on the circulation of information, and triggered reconfiguration of societies’ structures. Books have boundaries, uniformity, continuity, linearity, and repeatability; this had a profound influence on the organisation of social systems. Print further “detribalised” social organisations, enabling individualism, fragmentation, government centralism, and nationalism. As a work was mechanically copied, it maintained its consistency and accuracy in comparison to hand copied manuscripts. It was easier to transmit information without personal contact. Then came the responsibility for authors to be consistent with their definitive statements. In *Connections* [11], 1978, Burke, posits that once a piece of knowledge was publicly shared, its creator was identifiable and received recognition. Print media democratised knowledge as soon as it became a commodity, mass produced and distributed to populations. The general public had a reason to not only access or acquire media like books, but also had a reason to learn to read for themselves, where previously it was necessary to go through a third-party; the wealthy, scholar, or theologian, in order to be informed. The increase in literacy consequently helped the public to be better knowledgeable about world matters without being subject to official or unofficial censorship. As there were more readers, the ability to write fostered the formulation and preservation of individual thought and culture, as well as enabling the public to voice themselves and question authority. In *The Printing Press as an Agent of Change* [12], Eisenstein, 1979, contends that the print media also played a key role in facilitating scientific publishing and the scientific revolution.

Electronic age: The electric *mass*-age made it possible to have instantaneous communication in different forms across the globe. Electronic media did not have the same kinds of boundaries that the print had on society and so society was able to re-organise itself differently. Unlike print media, electronic media like telephone, radio, and television, did not require literacy, and consequently the support of a stable education system. As a whole, it had the same effects on society as the acoustic space created in the tribal world. The speed of information increased, and with it recovered sociality and involvement which was previously abandoned as the primary mode of information exchange. Perhaps more interestingly, as the flow of information through various electronic media accelerated, it facilitated our natural tendency to recognise patterns and (ir)regularities in data – in comparison to print media which happened to isolate, classify, and immobilise items in fixed space.

Later I will revisit McLuhan’s studies on the psychic and social consequences of technological innovation and apply them to scholarly communication on the Web.

Next, to continue to paint the historical backdrop against which this thesis is placed, I look at the European-centric scientific revolution which overlapped with the transition from the manuscript to the print-centric mode of information organisation and dissemination.

2.1.2 On Paradigms

> I simply assert the existence of significant limits to what the proponents of different theories can communicate to one another.
>
> *Objectivity, Value Judgment, and Theory Choice* [13], Thomas Kuhn, 1973

In *The Structure of Scientific Revolutions* [14], Thomas Kuhn, 1962, proposes an explanation for the process of discovery and progress in the history and philosophy of science. Kuhn contended that science did not progress linearly through history, rather it starts with a period that’s disorganised, and then a paradigm – a collection of *exemplar* agreed by scientists on how problems are to be understood – develops to organise things for day-to-day “normal science” to take place under the governed paradigm. After some time, “anomalies” in the paradigm accumulate and lead to a “crisis” phase where multiple competing paradigms attempt to resolve the issues. This is a particular point in which the community struggles to come to an agreement on the fittest way to practise future science. Finally a “paradigm shift” – revolution – refers to the last phase where transforming ideas and assertions mature enough to bring order to earlier anomalies, and cause a change in world view.

In *Objectivity, Value Judgment, and Theory Choice* [15], Kuhn, 1973, sets out five characteristics as a basis to explain which paradigms and methods have been adopted in the past, or may be considered for adoption in the future:

Accuracy
Application of quantitative and qualitative agreements in a domain.

Consistency
Internal and external consistency in relation to related theories.

Scope
Coverage and consequences of theories in what they can explain.

Simplicity
Computational labour required for explanations.

Fruitfulness
Explaining previously unnoted relationships or potential to disclose new phenomenon.

Kuhn reasoned that progress in science was not a mere line leading to truth, but the notion of moving away from less adequate concepts of and interactions with the world in favour of more *fruitful*. As to which paradigm can be objectively determined to be better, Kuhn argued that scientists' decision was ultimately inter-subjective and required value judgement because paradigms – the ideas and assertions within – were not necessarily directly comparable. For example, early versions of the heliocentric model proposed by Copernicus in *Commentariolus* [16] (1514), as well as the matured version in *De revolutionibus orbium coelestium* [17] (1543) was neither quantitatively or qualitatively more accurate than Ptolemy's *geocentric model* [18] (c. 2nd century CE) based on the knowledge and instruments available at the time. While they were equally internally consistent, geocentrism was regarded to be more consistent with the other accepted theories and doctrines in comparison. Copernicus' case however offered a simpler explanation and broader scope of application which ultimately was preferred. These were attractive characteristics for Kepler and Galileo when they came to investigate. With Newton's laws for motion and gravitation, the heliocentric model matured enough to set a new course for science and research.

Another example demonstrating the complexity of choosing a theory was from the perspective of language and communication. For instance, the term "mass" has radically different meanings in Newton's theory of gravity and Einstein's relativity theory because the definitions are isolated to their respective paradigms. However, General Relativity gave a new lens to understanding the interaction of mass in space-time. While the definition for mass changed, Einstein's theory was still able to precisely predict Newton's theory, and further provided an explanation for how gravity worked.

Contemporary scholarly communication is situated within a particular paradigm. In *Linked Research as a Paradigm* I will use Kuhn's theory of choosing paradigms as a heuristic device to *Contextualise Linked Research*.

2.1.3 Rear-View Mirror

The Gutenberg Galaxy [19] and *The Structure of Scientific Revolutions* [20] were coincidently published in 1962. While the accomplishments of McLuhan and Kuhn had differences in methodology and structure, they reveal the overwhelming effects of patterns of change when new technologies and paradigms are adopted by society. These major communication shifts in society as a whole, and specifically in the mode of scientific inquiry, are rare and significant events in history. Content follows form; mediums and paradigms shape the space.

As the characteristics of the spoken word carried over to writing, and writing to print, print media also influenced electronic media like the Web and hypermedia. Print's typography and knowledge organisation affects the standards, tools, and interfaces that are still used today to exchange knowledge on the Web, as well as certain social expectations or assumptions. For example, authoring tools help us to design and create electronic documents that resemble print representations even if they are not intended to be printed. McLuhan described this phenomenon as though looking at the "rear-view mirror"; "the old medium is always the content of the new medium”, *This is Marshall McLuhan: The Medium is the Massage* [21], McLuhan, 1967. While scholarly communication is transforming due to the effects of the Web, its content, shape, and way of communicating is still based on the characteristics of print media.

We need to examine our assumptions about scholarly communication on the Web in order to understand its social implications. Building on the previous brief histories of media evolution and scientific communication, the next review is of the history of the Web (a “social machine”) and then specific role the Web plays in contemporary scholarly communication. The focus of the rest of this section is on structural changes in scholarly communication, while the evolution of technical advancements runs in parallel.

2.2 Web: A Social Machine

> ```
> From: timbl@info .cern.ch (Tim Berners-Lee)
> Newsgroups: alt.hypertext
> Subject: WorldWideWeb: Summary
> Date: 6 Aug 91 16:00:12 GMT
>
> The WWW project merges the techniques of information retrieval and hypertext to
> make an easy but powerful global information system.
>
> The project started with the philosophy that much academic information should
> be freely available to anyone. It aims to allow information sharing within
> internationally dispersed teams, and the dissemination of information by
> support groups.
> ```
>
> *WorldWideWeb: Summary* [22], alt.hypertext, Tim Berners-Lee, 1991

The World Wide Web is inherently social and comprises various abstract *social machines* [23] – “processes in which the people do the creative work and the machine does the administration”, Berners-Lee, 1999, p. 172. In *Weaving the Web* [24], Berners-Lee, 1999, p. 123, states that he “designed it for a social effect – to help people work together”. The Web’s remarkable growth is fundamentally due to its social characteristics above anything else. The technical challenge was to define the basic common rules of protocol for machines to exchange information in a global system. In the remainder of this section, I outline the core components of the *Architecture of the Web*, discuss the notion and practice of *Linked Data*, and postulate how the *Web Science* framework can help towards a holistic understanding of the interplay between social and technical challenges.

2.2.1 Architecture of the Web

In *Evolution of the Web* [25], Berners-Lee, 1998, mentions the mandate to maintain the Web as an interoperable information space while evolving with the needs of society and technology. Interoperability meant that different systems without prior out-of-band knowledge should be able to communicate if they agree to operate based on open standards. Evolvability meant that languages and interfaces need to handle extensions, mixing, accommodate diverse scenarios for information exchange, as well as co-evolve.

The Web architecture initially included three independent components:

- the *Universal Document Identifier* [26] (UDI) to refer to information on the Web – now known as *Uniform Resource Identifier* [27] (URI) to identify things, and *Uniform Resource Locators* [28] (URL) to locate them on the Web.
- the *Hypertext Transfer Protocol* [29] (HTTP) as the request-response mechanism for systems – “client–

server model” – to communicate and exchange hypermedia documents on the Web.
- the *Hypertext Markup Language* [30] (HTML) as the lingua franca to describe and navigate hypertext and multimedia documents on the Web.

The W3C *Technical Architecture Group* [31] (TAG) codified the design principles, constraints, and good practice notes of the Web architecture in W3C TAG Finding, *Architecture of the World Wide Web, Volume One* [32] (AWWW), 2004. It documents the three architectural bases of the Web: “Identification”, “Interaction”, “Data Formats”, and states that the specifications emerging from each may evolve independently – according to the principle: *Orthogonality* [33] – in order to increase the flexibility and robustness of the Web.

2.2.1.1 Identification

In the context of the Web, anything – a resource – can be globally identified or named with a URI. As per *Uniform Resource Identifier (URI): Generic Syntax* [34] (RFC 3986) the “generic URI syntax” is defined as follows:

```
URI         = scheme ":" hier-part [ "?" query ] [ "#" fragment ]
```

The hierarchical part (`hier-part`) is composed of authority and path components. The authority component refers to the naming authority that governs the remainder of the URI (path, query, and fragment components). Some examples:

- `https://csarven.ca/#i` (WebID)
- `urn:isbn:9781584230700` (book identifier)
- `data:image/svg+xml;base64,PD94bWwg...` (encoded image)
- `file:///var/www/dokieli/index.html` (local file)
- `mailto:info@csarven.ca` (email address)
- `tel:+15551234567` (telephone number)

In 1996, Berners-Lee proposed a set of *axioms* for URIs as the backbone of what makes the Web a universal information space:

> **Axiom 0: Universality 1**
> Any resource anywhere can be given a URI
>
> **Axiom 0a: Universality 2**
> Any resource of significance should be given a URI.
>
> **Axiom 1: Global scope**
> It doesn't matter to whom or where you specify that URI, it will have the same meaning.
>
> **Axiom 2a: sameness**
> a URI will repeatably refer to "the same" thing
>
> **Axiom 2b: identity**
> the significance of identity for a given URI is determined by the person who owns the URI, who first determined what it points to.
>
> *Universal Resource Identifiers -- Axioms of Web Architecture* [35], Tim Berners-Lee, 1996

URIs make it possible for real or abstract *things* to be identified. When HTTP URIs are dereferenced different representations (data formats) are made available from corresponding *Uniform Resource Locators* [28] (URLs). While URIs makes it possible to identify (name) things on the Web, URLs make it possible for humans and machines to locate and to interact further with content.

The W3C's *URIs, URLs, and URNs: Clarifications and Recommendations 1.0* [36], 2001, distinguishes their purpose:

URI
Any type of identifier for a Web resource, eg. an HTTP URI.

URL
Any type of URI that can be resolved or dereferenced to locate a representation, eg. an HTTP URI is a URL.

URN
A specific type of URI that persistently identifies a resource. Once assigned to a resource, it cannot be reassigned.

A URI represents a resource as a conceptual entity and can have different electronic representations. The degree of genericity of a resource on the other hand is determined by the authority that allocates the URI. For example, a spectrum of URIs – generic to specific – can be assigned to a resource while conceptually being about the same thing. Time, language, content type, and the target medium are some of the common dimensions of genericity that Web software handles – *Generic Resources* [37].

On the Web, any entity is entitled to associate a URI with a resource in accordance with *URI ownership* [38]. The approach for the `http:` scheme and URI ownership is architecturally decentralised in that the location of a resource can physically exist anywhere in the world. However, architectural decentralisation is ultimately influenced by political or social centralisation at different levels. The *Domain Name System* [39] (DNS) is one such hierarchically decentralised structure for naming resources connected to the Internet. It is a kind of a social centralisation in that the *Internet Corporation for Assigned Names and Numbers* [40] (ICANN) coordinates the management of *Domain Names* [41], where top-level domains for like *countries, categories, multiorganization* [42], are then under the authority of specific organisations. Individuals then typically register fully qualified domain names (and map them to IP addresses) through accredited registries. Since DNS is a social construct, it can be brought down or controlled by an authority; states or people. As long as we continue to pay the required fees and manage the name server assignment, we can continue to use the URIs under that domain. Let us not forget that "you don't buy domain names, you rent them" – attributed to Ester Dyson, ICANN chair.

In response to emerging new identification mechanisms on the Web, the W3C TAG Finding, *URNs, Namespaces and Registries* [43], 2005, addresses the questions "When should URNs or URIs with novel URI schemes be used to name information resources for the Web?" and "Should registries be provided for such identifiers?" with respect to requiring persistence, standardisation within administrative units, protocol independence, location independence, structuring resource identifiers, uniform access to metadata, and flexibility authority. The finding states that the `http:` URI scheme can already achieve many of these requirements.

In this thesis, I focus mostly on URIs with the `http:` scheme because of its proven utility and wide adoption. As specified in *Axiom 2b: identity*, the owner of the HTTP URI defines the identity relationship which exists between a URI and the resources associated with it. Their reuse and persistence are defined by the information publisher – which may be also be the "owner" of the URI space. I expand on *Persistence and Preservation*.

2.2.1.2 Interaction

The Web architecture permits agents to communicate through different protocols eg. HTTP, FTP, SMTP are application layer protocols in the *Internet Protocol Suite* [44] (RFC 1123). In this thesis, the focus is on HTTP for data communication, ie. the structured requests and responses exchanged by Web agents. For example, when a *user-agent* (like a Web browser) initiates a request to access the contents of a resource identified by an HTTP URI (*dereferencing*), a server responds with a message body including a representation of the resource state, and data about the response. The content of the response URL (HTTP

URI) is a particular state in that it has a presentation and controls (affordances) that can be further engaged with, eg. hyperlinks to follow; data forms that can be submitted; location of alternate representations. One particular description of this interaction is given in 2000 by Fielding, in *Architectural Styles and the Design of Network-based Software Architectures* [45], commonly referred to as *Representational State Transfer* (REST): a set of architectural constraints for distributed hypermedia systems. There are diverse Web agents, eg. browsers, servers, autonomous scripts, with different capabilities, hence as per *HTTP/1.1 Message Syntax and Routing* [46] (RFC 7230), not all agents are expected to make the same interactions that are presented to them.

There are two categories of interactions: safe and unsafe. A *safe interaction* is one where the requested resource’s state does not change. When an agent makes a retrieval request (`HTTP GET`), it has no obligation beyond dereferencing the URL. An *unsafe interaction* on the other hand may cause a resource to change its state, and that the user may be held accountable, depending on the semantics of the interaction. For example, an agent sending an `HTTP POST` request with a payload can get a response indicating that a new resource was created, that the resource represented by the URI was deleted by an `HTTP DELETE` request, or that the agent is unauthorised to either create or delete the resource.

When a client requests a resource, the server may return a *particular* representation of a resource that is the result of a "negotiation" between them. For example, a client may indicate that it would accept the resource in HTML, Turtle, and JSON-LD data formats, and in English and Armenian languages. The server could provide a response body using HTML and English.

HTTP request and response functions can be expressed as follows:

$$\text{Request} = \text{Method} \times \text{IRI} \times \text{Header} \times \text{Body} \tag{1}$$

Method denotes the kind of HTTP request; *IRI* identifies the target resource for the request; *Header* provides context about the request and response eg. the `Content-Type` header indicates the format of the content in *Body*.

$$\text{Response} = \text{Status} \times \text{Header} \times \text{Body}$$

Status code denotes the result of the request.

Today, Web browsers provide a basis for many forms of human and machine communication; browsers are one of the most widely used platforms through which a wide range of Internet and Web standards used. The goal of the *Open Web Platform* [47] (OWP) is to advance such standards so that technologies like the Web browser can help us to effectively communicate.

2.2.1.3 Data Formats

The Web architecture has no constraints on the type of data formats that can be exchanged between clients and servers. A resource state may have different representations for the information it encodes. For example, HTML+RDFa, Turtle, JSON-LD, and RDF/XML are some RDF serialization formats – discussed in *RDF Syntaxes* – that can be used to express the same information (semantics) of a resource.

On designing computer systems and choosing languages, W3C TAG Finding, *The Rule of Least Power* [48], 2006, recommends that “when publishing on the Web, you should usually choose the least powerful or most easily analyzed language variant that’s suitable for the purpose”. Berners-Lee references RDF as a common language for the Web in *Principles of Design, Principle of Least Power* [49]: “the Semantic Web is an attempt, largely, to map large quantities of existing data onto a common language so that the data can be analyzed in ways never dreamed of by its creators”.

Berners-Lee states that a declarative language is preferable to a procedural language in terms of data reusability. HTML for example not only turned out to be simple for people to author, its modularity and extensibility paved the way for other standards. In this thesis, I use XHTML and HTML interchangeably (except when explicitly mentioning the lower-level differences in W3C specifications and tooling). Different kinds of *language mixing* – composition of data formats – emerged, including for example *Cascading Style Sheets* [50] (CSS), *Mathematical Markup Language* [51] (MathML), and *RDFa* [52], *Scalable Vector Graphics* [53] (SVG) in or alongside HTML. The extension mechanisms eg. in HTML and CSS, allowed new semantics or features to be declared in the content that can be processed by conforming Web agents. Web agents get to choose whether to ignore unrecognised content or try to understand and treat unrecognised patterns as errors. When a browser encounters a syntactical error in an XHTML document, it can stop processing the document or display an error to the user, also known as draconian error-handling. An HTML document on the other hand has a more forgiving error-handling model. An unrecognised component or non well-formed structure would not prevent processing and no alert needed to be passed to the user interface. To date, HTML acts as a unifying language for human interfaces to information systems.

2.2.2 Linked Data

The Web architecture initially used HTML as the language to represent and connect hypermedia documents with one another. HTML's success is indisputable today, and that's perhaps due to its simplicity to create and reuse; its flexibility, for partial understanding to proceed; and its extensibility to allow language mixing. While the HTML specification continues to evolve by reflecting on the needs of the Web, by design it is scoped to creating and connecting Web documents and applications.

Nevertheless, the potential *language* "to describe the interrelationships between things" as originally proposed in *Information Management: A Proposal* [54] was not fully realised until the *Resource Description Framework* [55] (RDF) came along. RDF is a constructed machine-interpretable language to describe and interlink arbitrary things at any level of abstraction on the Web, and reuses the existing architecture of the Web. In *Sense and Reference on the Web* [56], Halpin, 2009, states that in the context of RDF (and the "Principle of Linking"), "URIs are primarily referential and may not lead to access unlike links in traditional hypertext systems".

Conceptually, RDF was in the fabric of the Web from the start. By *naming* things with URIs, we can construct structured sentences with RDF. By comparison, HTML at its core is a terse approach to representing and linking knowledge. RDF on the other hand is intended to be a unifying language for machine interfaces to data systems. I will further discuss language mixing with RDF in HTML and XML-family languages via *RDFa*.

One conceptual view of the Web is that "anyone being (technically) allowed to say anything about anything" (AAA). One consequence of AAA is that while systems (and reasoners within) can operate under both the *open-world assumption* [57] (OWA) or the *closed-world assumption* [58] (CWA), the Web ultimately uses the OWA framework because new knowledge can always make its way into the system, and lack of knowledge does not imply falsity. RDF applies a particular restriction of the AAA principle in that nonsensical, inconsistent, or conflicting information can be created or inferred. It is up to processing applications to determine their confidence on the information. Hence, an important distinction: assertions made using the RDF language are claims, as opposed to facts.

Statements in RDF are expressed in the form of *triples* – subject, predicate, object:

> - the *subject*, which is an IRI or a blank node
> - the *predicate*, which is an IRI
> - the *object*, which is an IRI, a literal or a blank node
>
> *RDF 1.1 Concepts and Abstract Syntax, Triples* [59], W3C, 2014

The *International Resource Identifier* [60] (IRI) is a generalisation of URI (US-ASCII) in that a wider character set (Unicode/ISO 10646) can be used to identify resources on the Web. A Blank node (anonymous resource) is a local identifier scoped to the context it is used in, eg. file, RDF store, and they are not intended to be persistent or portable. Literals are used for values like strings, numbers, and dates.

A set of triples in RDF (an *RDF Graph*) can be represented as a directed edge-labelled graph: $G = (V, L, E)$, where V is the set of *vertices* (union of subject and object terms), L is a set of edge labels (predicate terms), and E contains triples from $V \times L \times V$. A subject may have several predicates, and predicates can also be mapped to subjects given that they have their own identity.

The structure of RDF statements is similar to:

- the *subject–verb–object* [61] sentence structure in *linguistic typology* [62] – used by nearly half of human languages
- the *entity–attribute–value* [63] (EAV) model – widely used for advanced (meta)data modeling

We can represent the sentence "Otlet influenced Berners-Lee" in one of the RDF syntaxes as follows:

```
<http://dbpedia.org/resource/Paul_Otlet>
  <http://dbpedia.org/ontology/influenced> <http://dbpedia.org/resource/Tim_Berners-Lee> .
```

As ad hoc exploration is one of the goals of the Web, the W3C TAG Finding, *The Self-Describing Web* [64], 2009, reports how HTTP and other Web technologies can be used to “create, deploy and access *self-describing* Web resource representations that can be correctly interpreted.” It is expressed that the RDF language can be used to integrate with the Semantic Web such that the information encoded in a representation explicitly provides interoperable means to relate Web resources.

RDF facilitates a uniform *follow your nose* [65] type of exploration by following the relations in statements in order to arrive at another unit of *self-describing* information. The uniformity in graph traversal enables applications to meaningfully interpret and integrate with disparate data without having any out-of-band knowledge, and retrieve more information about the terms as they need to from the source. As the content identified at `http://dbpedia.org/resource/Paul_Otlet` describes itself, in the same way, the relation `http://dbpedia.org/ontology/influenced` defines itself, and when we inspect a representation of `http://dbpedia.org/resource/Tim_Berners-Lee`, we will find that it describes itself.

We can describe Berners-Lee’s IRI (subject) by giving it a human-readable name (predicate):

```
<http://dbpedia.org/resource/Tim_Berners-Lee>
  <http://xmlns.com/foaf/0.1/name> "Tim Berners-Lee"@en .
```

> **Note**
> The terms IRI and URI are used interchangeably outside of technical specifications. While IRIs can be used to identify resources, retrieval mechanisms use URIs because the request URI in HTTP protocol is defined as a URI. Hence, the characters in IRI are first mapped to their URI counterparts before the request process.

In order to address different system and interface needs, different syntaxes for RDF emerged over time. *RDF/XML* [66] is the grammar to define the *XML* [67] syntax for RDF graphs; *Turtle* [68] is a human-readable syntax resembling (subject–verb–object) sentence structure, *RDFa* [52] expresses RDF statements in markup languages, *JSON-LD* [69] serialises RDF in *JSON* [70] (RFC 7159).

In order to facilitate querying and manipulating RDF graph content on the Web or in an RDF store, *SPARQL Protocol and RDF Query Language* [71] (SPARQL) 1.1 provides a set of recommendations. The specifications address *SPARQL 1.1 Query Language* [72] to match, filter, and construct graph patterns, as well as integration of disparate sources of information. *SPARQL 1.1 Federated Querying* [73] is an extension to executing queries distributed over different SPARQL endpoints. *SPARQL 1.1 Update* [74] to update, create, and remove RDF graphs in a store; a *SPARQL 1.1 Protocol* [75] to convey queries and updates over HTTP; *SPARQL 1.1 Service Description* [76] in order to discover SPARQL services, and a vocabulary for describing them. *SPARQL 1.1 Entailment Regimes* [77] retrieve solutions that implicitly follow from the queried graph; and an alternative interface to SPARQL 1.1 Update uses *SPARQL 1.1 Graph Store HTTP Protocol* [78] for operations on RDF graphs.

Using the example RDF data from earlier about Otlet's relation to Berners-Lee and his name in an RDF store, the following query returns the data in a tabular format for the names of objects that match the graph pattern where the subjects *influenced*.

SPARQL Query:

```
SELECT ?object ?name
WHERE {
  <http://dbpedia.org/resource/Paul_Otlet>
    <http://dbpedia.org/ontology/influenced> ?object .
  ?object
    <http://xmlns.com/foaf/0.1/name> ?name .
}
```

Resulting in extracted data:

Table 1. Example SPARQL Result

?object	**?name**
http://dbpedia.org/resource/Tim_Berners-Lee	"Tim Berners-Lee"@en

Figure 1. Example SPARQL Query and Result.

The SPARQL suite is a powerful set of tools for inspecting and manipulating information based on the graph nature of the data.

⁂

The "Linked Data" design principles build on the AWWW in order to identify, describe, and discover information, enabling a *Semantic Web* where humans and machines can explore the Web of data:

> 1. Use URIs as names for things
> 2. Use HTTP URIs so that people can look up those names.
> 3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
> 4. Include links to other URIs. so that they can discover more things.
>
> *Linked Data – Design Issues* [79] (revised in 2009), Tim Berners-Lee, 2006

> **Note**
> In the original version (earliest archived) of the Linked Data design principles, the third statement was "When someone looks up a URI, provide useful information." It was updated in 2009 (earliest archived).

In practice, what URIs identify, based on the resources fetched upon dereferencing, was ambiguous until the release of the famous W3C TAG Finding, httpRange-14 issue, 2007, as per *httpRange-14: What is the range of the HTTP dereference function?* [80]. The core issue had to do with machine agents having a deterministic mechanism to distinguish between accessible digital objects on the Web ("information resources") and references to a class of physical entities and abstract concepts ("non-information resources"), ie. things that are not technically "retrievable" over the wire on the Web, but may only be referred to. In *Sense and Reference on the Web*, Halpin posits that “the definition of information object and information realization can be thought of as the classic division in philosophy of mind between an object given on a level of abstraction and some concrete thing that realizes that abstraction, where a single abstraction may have multiple realizations.”

Having established an overview of the technical environment I will next discuss the *Web Science* framework, which situates the technical aspects alongside social processes.

2.2.3 Web Science

> Web science – what makes the Web what it is, how it evolves and will evolve, what are the scenarios that could kill it or change it in ways that would be detrimental to its use.
>
> *Still hates computers* [81], The Inquirer, Dame Wendy Hall, 2010

The openness of the Web and the ease of involvement for both humans and machines (through common protocols) helped its great expansion without a particular central point of failure or coordination. Scholarly communication could be exemplary of different kinds of agents interacting on the Web, whether they are sentient or something else. *Web Science – Creating a Science of the Web* [82] – is an interdisciplinary field of study concerned with understanding and developing sociotechnical systems like the Web alongside human society. In *A Framework for Web Science* [83], Berners-Lee, 2006, the authors state:

> the Web perhaps more than any other recent human construct carries with it any number of issues including privacy and protection, access and diversity, control and freedom. Structures that we design, engineer and research, and findings that emerge through analysis, will often have strong societal implications.
>
> *A Framework for Web Science* [84], Tim Berners-Lee, 2006

The challenges of Web Science comprise both social and technical aspects, from user interfaces to data, to information policy, resilience, access from diverse devices, collective creativity, and decentralisation. All of these areas are pertinent to the future of scholarly communication, and Web Science as a field provides a framework for ensuring we consider the various social and technical issues as part of an interconnected ecosystem.

In the same vein, furthering the study of *Web Science* itself can be done through advancements in how we communicate scholarly findings; through better enabling universal access and connections within our collective knowledge.

The Web not only expedited the flow of human knowledge with a friendly interface, it enabled social interactions to take place which would not have been possible otherwise due to physical or social constraints. The global adoption of the Web brought new forms of human association and social action. In essence, the Web creates one form of a “global village”, described by McLuhan, whereby there is greater requirement of individuals to participate than with earlier forms of electric media eg. telegraph, electric light, telephone, radio, television. The way society reacts to the Web as a medium is perhaps more important than the content that is on it.

In *Credibility of the Web: Why We Need Dialectical Reading* [85], Bruce, 2000 explores Kaufmann’s “modes of reading” from the 1977 essay *Art of Reading* (*The Future of the Humanities* [86]) in context of the characteristics of the Web:

- *exegetical*: author is authority, the reader is passive;
- *dogmatic*: scepticism on part of the reader;
- *agnostic*: continuous evaluation while acknowledging good and bad qualities;
- *dialectical*: possessing coding, semantic, and pragmatic competence

Bruce draws attention to the dialectical in that an observer enters into a “deep experience” of reading, engaging with the text with a critical eye, drawing advanced information and interpretations beyond what is merely presented, and actively seeking to understand the material in addition to political, social, and historical dimensions. Bruce states that under the dialectical view, the Web’s multimodal features allows new values and ways of making meaning as an holistic involvement for the consumer. This view is a useful exemplification of McLuhan’s notions on the effects of media on society.

Thus far I have covered the architectural foundations of the Web, and the study of the Web from a scientific point of view. I now look at how scholarship on the Web has developed over time.

2.3 A Brief History of Scholarship on the Web

> Choice is an illusion, created between those with power, and those without.
>
> *The Matrix Reloaded* [2], Merovingian, 2003

In *Mediums and Paradigms* I have discussed the wide reaching effects of communication mediums as well as scientific frameworks adopted by society. In this section I zoom in on the history of scholarship in order to contextualise the practices and their influences to date.

As the print age is an important historical backdrop of this thesis, I provide a brief *review of The Printing Press as an Agent of Change* to further contextualise social transformations and the systematisation of scientific exchange given the availability of a technology for the masses:

The Printing Press as an Agent of Change

The Printing Press as an Agent of Change [87], Eisenstein, 1979, traces the impact of the communications revolution on a variety of sociocultural movements that shaped the modern mind that started by the invention of the printing press developed by Johannes Gutenberg (c. 1439 CE). While Eisenstein’s account of the history is not regarded as doctrine, the work demonstrates that the printing press, as a technological innovation, was a potent force in the evolution of social systems, as well as a communication technology which intensified the evolutionary process in the storage and dissemination of information and data. Perhaps more generally, Eisenstein explains how “printing altered *written communications within the Commonwealth of Learning*”, p. xiv.

Republic of Letters: With the growing power of the printing press since the 15th century, and sufficiently large international readers across Europe, the mass production and distribution of printed material gave rise to the formation of intellectuals as a distinct and independent social class. The long-distance, international, and private correspondence network between the literary figures formed the basis for the concepts, *Commonwealth of Learning*; vernacular science-writing in Latin aimed at non-academic readers, as well as aiding the literary “underground” trade among researchers to propel theories and data collection, and the *Republic of Letters* [88] (ROL) refers to a metaphysical *literary* society that produced and had access to substantial (non)scientific knowledge in Europe and America in the 17th century, p. 136-139.

Royal Society: *The Royal Society* [89] (RS) effectively established authority for a metaphysical community for scientists and scholars. In essence, it acknowledged and legitimised the social activity in RoL, and acted as a way to filter and help disseminate the discoveries of the literati. *Journal des sçavans* [90] published in 1665 was the first academic journal in Europe, followed by *Philosophical Transactions of the Royal Society* [91] in the same year. The first journal that strongly focused on scientific communication came later with the *Mémoires de l'Académie des Sciences* [92] in 1666.

Due to societal circumstances, eg. sponsorship and censorship by authorities, the literary underground

> with the help of the printing press was well underway in the 16th century. By 1640 “science had risen” mostly in the underground, and pre-dated the discovery and improvement of the telescope p. 685; the RoL; and the establishment of the RS. With the help of the RS later on, new discoveries and theories were disseminated abroad, p. 664.
>
> Eisenstein postulates that the printed book played a central role in the “rapid dissemination of knowledge to whole new classes that created the modern new attitudes to both science and religion at the end of the fifteenth century.” The mere availability of prior knowledge to the public, stimulated scientific curiosity further, and eventually gave birth to the scientific revolution, p. 691. Eisenstein also contends that the communication shift in technical literature occurred even before astronomers changed their views about the structure of the universe, p. 685. Moreover, the origins of the scientific revolution was partly due to the rediscovery of classical scientific texts, and the effects of an internal strife between academic innovators and conservationists (“quarrels of learned men”), trying to control the field of knowledge outside of academia, p. 523-524, 570. Eisenstein emphasises on the “relevance of external social institutions to the internal, relatively autonomous, life of science”, p. 636, as well as the utilitarian “application of the mathematics to the problems of the natural world”, p. 683.

The publication and dissemination of information – in the most general sense of the words – played a key role in the advancement of societies’ knowledge for centuries. The first academic journal, *Journal des sçavans* [93] and shortly after the *Philosophical Transactions of the Royal Society* [94] in 1665, evolved out of the exchange of letters among scholars and the transcripts of the meetings in scholarly societies discussing experiments, publications and natural curiosities. With this in mind, the most widely accepted purpose of a scholarly article is a way to communicate research results and other intellectual work with the public; to have an accurate archival record of the research findings; to build on top of earlier knowledge; as well as to have it serve towards crediting authors for their work. Today’s scholarly journals, containing quality-controlled research reports, closely resemble the first journal. *Encyclopedia of International Media and Communications* [95], Scholarly Journals, Vol. 2, p. 688, 2003, states that due to the growth and demand for copies of the literature from non-members, mostly institutions, the larger journals began to sell subscriptions at relatively higher prices than members pay.

In *The Sociology of Science: Theoretical and Empirical Investigations* [96], Merton, 1973, discusses “four sets of institutional imperatives taken to comprise the ethos of modern science”, originally introduced in *The Normative Structure of Science*, Merton, 1942:

- *Communalism*: in order to promote collective collaboration, intellectual output should have common ownership.
- *Universalism*: scientific validity is kept independent of political and personal factors such that everyone has equal rights and possibility to obtain information.
- *Disinterestedness*: the practice of science should benefit the common scientific enterprise with no conflicts of individual interest or gain.
- *Organised skepticism*: scientific claims should be exposed to critical examination before being accepted to be part of common knowledge.

2.3.1 Open and Free

While there are many definitions for “open” and “free”, depending on the context assigned by different communities eg. *The Many Meanings of Open* [97], Berners-Lee, 2013, here I acknowledge some that are commonly used.

The definition given by *The Open Definition* [98] which is derived from the *Open Source Definition*, is as follows:

> Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.
>
> *The Open Definition 2.1* [99]

The definition of “free” (or “libre”) in context of software that is widely acknowledged is:

> “Free software” means software that respects users’ freedom and community. Roughly, it means that **the users have the freedom to run, copy, distribute, study, change and improve the software**. Thus, “free software” is a matter of liberty, not price. To understand the concept, you should think of “free” as in “free speech,” not as in “free beer”. We sometimes call it “libre software,” borrowing the French or Spanish word for “free” as in freedom, to show we do not mean the software is gratis.
>
> *What is free software?* [100], GNU Project – Free Software Foundation

In order for an item or piece of knowledge to be transferred in a way to foster rich culture, the *Free Cultural Works* [101] defines *freedom* to mean the freedom to use; to study; to make and redistribute copies; and to make changes and improvements. The definition can be used as a “tool to determine whether a work or license should be considered "free."”

The definition of *Open Content* [102] explains how users of works to have free and perpetual permission to engage in activities with the right to retain, reuse, revise, remix and redistribute content.

Creative Commons [103]’s (CC) *CC licenses* [104] is one of public copyright licenses that can be used on works which grant additional permissions for the purpose of free distribution of an otherwise copyrighted work. With the exception of CC0 1.0 (public domain) license, all CC licenses require users of the works to attribute the creators. CC0, CC BY (Attribution) and CC BY-SA (Attribution-ShareAlike) are compatible with the definition of Free Cultural works. CC licenses are open, flexible, interoperable, and has global uptake. The policy of CC license is also compatible with public laws in various jurisdictions.

I now discuss several movements which emerged since the 1990s towards fundamentally transforming scholarly communication using newly available digital network technologies.

2.3.2 Archives and Repositories

In 1991, recognising the need for scholars to expedite the exchange of scholarly information, Paul Ginsparg created the LANL preprint archive for “e-prints”. *Creating a global knowledge network* [105], Ginsparg, 2001, states that “self-archiving” provided a moderated central repository where scientists can deposit their articles to make them accessible to anyone, which was a cost-effective solution (estimates were less than 5 USD per submission for administration) for scholars to disseminate research results quickly. It initially started out as a central mailbox with an interface; it adopted FTP in 1991, Gopher in 1992, and finally the World Wide Web in 1994. The articles in the repository are not expected to be peer reviewed, but rather a place where versioned “preprints” – considered to precede publications that are peer reviewed – can be accessed for free. In 2001, the LANL preprint archive eventually became what is currently known as *arXiv* [106]. The arXiv license information permits different rights and licenses to be assigned to the scholarly records, ie. anything from public domain, Creative Commons licenses, or with copyright to the publisher or author, provided that arXiv is given sufficient rights to distribute. The repository accepts submissions in TeX, PDF, PostScript, and HTML. This sort of discipline-centric and institution-run preprint exchange mechanism had the technical ingredients to serve as a substitute for the traditional scholarly journal. From 1991 to 2017, arXiv had a *linear increase in submission rate* [107] (passing the 1 million mark in 2015). arXiv’s early success lead to the emergence of the *Open Archives Initiative* in 1999, and the *Open Access* movement in 2002.

2.3.3 Open Access

In 1994, Harnad made a proposal to appeal to the “esoteric”:

> It is applicable only to ESOTERIC (non-trade, no-market) scientific and scholarly publication … that body of work for which the author does not and never has expected to SELL the words. The scholarly author wants only to PUBLISH them, that is, to reach the eyes and minds of peers.
>
> *The Subversive Proposal* [108], p. 11, Stevan Harnad, 1994

After some years of prototyping and researcher-centric initiatives, the ideas eventually formulated into the *Budapest Open Access Initiative* [109] (BOAI) in 2002. BOAI is based on the principle that publicly funded research output, including peer reviewed journal articles as well as unreviewed preprints, should be made immediately available, and indefinitely, without access fees and free from most restrictions on copying and reuse, to increase visibility and uptake:

> By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself … the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.
>
> *Budapest Open Access Initiative* [110], 2002

In order to open up access to scholarly journal literature, BOAI recommends the adoption of two complementary strategies for open access (OA):

- Self-Archiving
- Open-access Journals

The term “self-archiving” in context of OA differs from that of the early practices around preprint “self-archiving” in that there were no constraints set for the status of the scholarly literature, ie. peer reviewed or not. For OA-centric scientific literature, quality-control and certification are integral.

The foundations of the OA approach were conceived at a time when the major revenue models to cover publishing costs for scholarly literature were through subsidies from institutions, journal subscriptions, or pay-per-view. In this payment model, consumers of the literature were charged either individually or through institutions. In contrast, the *article processing charge* (APC) model requires publication costs to be covered upfront, thereby waiving the costs for the consumers of the literature. Today, APC is the most widely used method to finance OA publishing, with varying levels of fees. The OpenAPC dataset includes publication fee spending of 238 institutions reveals an average of 1946 EUR per article based on “fully” and “hybrid” OA journals (average of 1558 EUR and 2504 EUR respectively). In this dataset, the institutional spending on article publishing in 2018 was at minimum 13 EUR, maximum 12000 EUR, and mean 1834 EUR.

The effect of open access and downloads ('hits') on citation impact: a bibliography of studies [111], OpCit, 2013, “is intended to describe progress in reporting these studies; it also lists the Web tools available to measure impact. It is a focused bibliography, on the relationship between impact and access.” Harnad argues that the essential message of OA is research that is not freely available online loses research impact – any reasons not to make research Open Access are merely excuses (see the “raincoat science” metaphor).

In *The deliverance of open access books* [112], Snijder, 2019, looks at the consequences of open access for books, and concludes that books that are freely accessible online attract more readers and are cited ten percent more than those that do not use open access platforms.

2.3.4 Deconstructing the Scholarly Journal

In *Modularity: the next form of scientific information presentation?* [113], Kircz, 1997, states that “the present-day linear (essay-type) scientific article is the result of a development over the centuries”, and that “different types of information, at present intermingled in the linear article, can be separated and stored in well-defined, cognitive, textual modules.” A natural consequence to having modular electronic articles, with their own unique characteristics, is that units are self-contained and interconnected to other units.

In *The deconstructed journal – a new model for academic publishing* [114], Smith, 1999, provides insights on networked-based scholarly publishing models:

1. “means/end” confusion, where journals continue to mimic or replicate existing mechanisms
2. modifications to the system should still match the needs of the traditional scholarly journal
3. cooperating agents can function without going through a central agency, eg. a “publisher”.

The role of the journal includes both “main” (editorial, quality control, recognition of work, marketing, and disseminating) and “hidden” (subject and community defining, and archiving) functions. The proposed *deconstructed journal* model is preferable to the “paper” influenced electronic publishing model in that both main and hidden roles can be accomplished in a distributed manner. Here, Smith emphasises allowing greater academic freedom and shifting of power and control from the monolithic third-parties to the knowledge producers, which could be realised in either an *institution-* or *researcher-centric* approach. In this process, journals with the refereeing process would link to the articles with full content hosted on the authors’ Web sites. To get there, however, Smith acknowledges that for such a new paradigm to be adopted, professional and funding bodies must transition themselves to accepting the decoupling and self-controlled literature publishing model as something equivalent (in terms of academic value) to the traditional model.

After two decades, the anecdotal evidence suggests that academia acknowledges the sort of equivalence that Smith describes at varying levels. For example, while literature quality-control can be generally anywhere on the Web, on the other hand, self-published literature remains to be treated as “grey literature”, that is, any material that is outside of the traditional scholarly distribution channels, especially when it comes to academic reward systems.

In *Decoupling the scholarly journal* [115], Priem, 2012, presents an analysis of the traditional functions of the scholarly journal – registration (recording, time-stamping), certification (quality-control, assessment), awareness (dissemination), and archiving (preservation) will be discussed in detail under *Forces and Functions*. The article echoes earlier work in that the scholarly journals bundle multiple functions, and that this tight coupling has hindered innovation, is inefficient, and makes it difficult to improve individual functions. The article also states that the current system, ie. the bundling of the closed publishing and certification models, is unlikely to change if the communication functions are not decoupled, regardless of all the other out-of-band activities – activism or innovation – to improve the publishing models or certification mechanisms. The article diverges from the traditional view of separating registration and archiving functions since the latter “necessitates both persistent storage and identification”. Hence, an archived “unit of communication” (*Rethinking Scholarly Communication*) has already fulfilled registration. The authors echo many of the ideas of Smith with respect to decoupling and decentralisation, however they only go as far as describing them within the context of scholarly functions operating alongside institutional or third-party online services. While the article does not exclude the possibility of units of information being self-hosted and individually-controlled by the researchers, it is not explored further. The article concludes that the nature of the communication revolution will be through the structure and organisation of the units in the system.

2.3.5 Open Archiving

The *First meeting of the Open Archives initiative* [116] in 1999 was held to agree on recommendations to address interoperability challenges, and outlined in *The UPS Prototype* [117], 2000 – such as search and linking – for data across publicly available distributed "self-archiving" solutions. To promote the global acceptance of "self-archiving", with interoperable systems being core to its successful adoption, the *Open Archives Initiative (OAI) Organization* [118] was established to take on the development of technical recommendations – for metadata standards, discovery, and retrieval. This lead to the evolution of recommendations such as *OAI-PMH* [119] "Protocol for Metadata Harvesting", and its successor *ResourceSync* [120] as a resource synchronization framework for servers; *OAI-ORE* [121] for "Object Reuse and Exchange"; and *Memento* [122] which specifies a time dimension to the Web architecture. *Reminiscing About 15 Years of Interoperability Efforts* [123], Van de Sompel, 2015, summarises the lessons learned on interoperability efforts by concluding that a shift from a *repository-centric* to a *Web-centric* perspective was essential in order have a "viable and sustainable" scholarly system.

2.3.6 Scholarly Declarations and Practices

While there is a plethora of overlapping guides, best practices, principles, manifestos, and declarations in the field of research, scholarly communication, data and publishing, here I highlight a few that are Web-centric.

Today, "Open Science" is a broadly understood notion that is intended to foster an accessible approach to conducting science and scholarship. While there is no single agreed definition of open science, the general consensus is about improving the practice of scientific communication in a transparent, traceable, verifiable, and reusable manner. One definition of "open science" is as follows:

> Open Science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.
>
> *Open Science Definition* [124], FOSTER

With the advent of the scholarly journal, the practice of open science have been continuously evolving. As the Web speeds up information exchange, there has been further interest in the systemisation and application of open science in research communities. Later on I will further discuss *Social Scholarly Web*.

Data on the Web Best Practices

The W3C Recommendation for *Data on the Web Best Practices* [125] (DWBP) provides a wide range of best practices, in order to benefit comprehension, processability, discoverability, reuse, trust, linkability, access, and interoperability. The DWBP can be used by "data publishers in order to help them and data consumers to overcome the different challenges faced when publishing and consuming data on the Web". DWBP applies to all data publishers but is pertinent for researchers.

FAIR Guiding Principles

The *FAIR Guiding Principles* [126], Wilkinson 2016, is a set of principles to facilitate knowledge discovery and reuse for machines and humans. The principles, *findable*, *accessible*, *interoperable*, *reusable* (FAIR) are intended to express policy goals, as opposed to a technical prescription to building a data infrastructure and exchange. The elements of the principles include:

Findable

For human and machine discovery of (meta)data and services by use of globally unique and persistent identifiers, and machine-readable way-finding.

Accessible

(Meta)data to be retrievable using open and free standard communication protocols via applicable authentication and authorization mechanisms.

Interoperable

(Meta)data using structured knowledge representations to be interpretable by autonomous agents.

Reusable

(Meta)data to optimise the reuse of data with domain-relevant, license, and provenance information.

Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud [127], Barend, 2017, posits that data are often open but not necessarily FAIR; data could be closed yet perfectly FAIR. FAIR does not directly proscribe data quality, trustworthiness, ethics or responsibility. Hence, conforming with the FAIR principles does not necessarily adhere to the scientific method.

Cost of not having FAIR research data [128], PwC EU Services, European Commission, 2018, provides a cost-benefit analysis for FAIR research data based on the following indicators: time spent, cost of storage, license costs, research retraction, double funding, cross-fertilization, potential economic growth (as % of GDP). The estimates reveal that the "annual cost of not having FAIR research data costs the European economy at least €10.2bn every year" and that the true cost cost to be much higher. To put €10.2 billion in perspective, the cost of not having FAIR is "~ 400%, of what the European Research Council and European research infrastructures receive combined." The study also highlights the consequences in the absence of FAIR data to include "impact on research quality, economic turnover, or machine readability of research data."

Data Processing, Privacy and Protection

In addition to institutional ethical principles and federal regulations, research can be subject to additional regulations. For instance, the *General Data Protection Regulation* [129] (GDPR) imposes obligations on organisations that process personal data of individuals in the EU and the EEA. Hence, GDPR is relevant for experiments and studies that include human subjects, personally identifiable information (PII), as well as acquired or inferred personal characteristics. In relation to scholarly communication, ie. research as a basis for data processing, GDPR is permissive and indicates exemptions on the "processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes should be subject to appropriate safeguards for the rights and freedoms of the data subject", as well as for scientific research purposes including "technological development and demonstration, fundamental research, applied research and privately funded research."

Digital Agenda

In order "to further open up the market for services based on public-sector information", the European Commission's *Revision of the PSI Directive* [130], 2012 — a digital agenda for Europe — as per *Office Journal C 240/2014* [131], recommends high-demand datasets from libraries and archives be "published in machine-readable and open formats (CSV, JSON, XML, RDF, etc.) to enhance accessibility", and "described in rich metadata formats and classified according to standard vocabularies", and "to facilitate re-use, public sector bodies should … make documents available … at the best level of precision and granularity, in a format that ensures interoperability".

As the importance of the new approaches are acknowledged by scholarly communities, there is an increasing social, as well as technical demands for their adoption.

2.3.7 Paper User Interface

> "Because paper enforces single sequence and there is no room for digression, it imposes a particular kind of order in the very nature of the structure. When I saw the computer, I said, 'at last we can escape from the prison of paper', and that was what my whole hypertext idea was about in 1960 and since. Contrarily, what did the other people do, they **imitated paper**, which to me seems totally insane."
>
> *Ted Nelson Demonstrates XanaduSpace (by Arthur Bullard)* [132], Ted Nelson, 2013

The Web goes beyond the confines of the physical paper. The user experiences between the Web and print-centric media differ greatly. This section briefly compares *Portable Document Format* [133] (PDF) and HTML.

Donald E. Knuth, the inventor of *TeX* [134], considers TeX to be a program written using the literate programming approach. The TeX source code can be imperatively programmed, then compiled by a processor to arrive at a view. *LaTeX* [135], which is widely used for scholarly articles, is a collection of TeX macros. In some scientific communities, PDFs are usually generated from LaTeX.

PDF (ISO 32000-2:2017) is intended for displaying and storing, and generally self-contained. However, it is not intended to be editable as the formatting instructions are no longer available. PDF is inherently layout oriented, and it is an optimal format for printing given its high-precision for typography. Third-party scholarly publishers tend to require a fixed layout with typographical guidelines for research documents, hence submissions are often required to be made as PDFs.

In order to facilitate metadata interchange, an *Extensible Metadata Platform* (XMP) package in the form of XML (most commonly serialized as RDF/XML) may be embedded in PDF as per *XMP Specification Part 1: Data Model, Serialization, and Core Properties* [136] (ISO 16684-1:2012). Otherwise, semantically useful information is not preserved when PDFs are generated, which makes it difficult to go back to source content format.

HTML is both a content format for encoding information, and a document format for storing information. HTML can reflow to fit different displays. HTML has ubiquitous support across devices.

If HTML is used as the original format to record the information, a PDF copy can be generated on demand, using the specifications of the user. With an accompanying CSS for the HTML, desired layout guidelines can be achieved. Converting from PDF to HTML+CSS is possible, but the precision rests on rendering the view, as opposed to creating a structurally and semantically meaningful document.

Formatting research articles as PDF comes at the cost of losing access to granular information. Thus, PDF's primary focus – the presentation layer – leads to legacy documents that are not reusable in the long run. A great amount of *reverse-engineering* is required to extract the underlying information (like content, tables, figures). The source format (like TeX, *JATS*, XML, HTML, or other) are often needed instead for content exchange and modifications.

TeX and HTML (as stacks) are compared in Table Comparison of TeX and HTML.

Table 2. Comparison of TeX and HTML

	TeX	**HTML**
System	Typesetting	Web
Programming paradigm	Imperative	Declarative
Device readiness	Moderate	Good
Applicable media	Screen, Print	Any*
Layout	Fixed	Reflowable
Modularity	Locked	Extensible
Immutability	Core	Feasible
Interactivity	Static	Dynamic
Accessibility	Varies	Core
Machine-readability	Low	High
Linkability	Basic	Rich
Reference granularity	Coarse	Fine

We consider TeX stack family to include DVI, XMP, LaTeX, PDF and ECMAScript, whereas the HTML stack family to include hypertext and semantic (W3C) technologies and JavaScript.

Any media refers to W3C CSS *Media Queries* [137], 2012, eg. braille, handheld, print, screen, speech.

Device readiness is an informal estimate on how likely a device will be able to open the stack, view, and interact.

Third-party research publishing services commonly impose constraints on file formats derived from print requirements. As a consequence, it regulates researchers’ choice of tools to prepare and share their work with their peers and the society. In essence, it impacts how researchers and scholars approach communicating their work based on the confines of the medium. While the choices for digital representation for research information may seem arbitrary, the effects of a print-centric scholarly publishing is analogous to McLuhan’s “we look at the present through a rear-view mirror”.

> I chose HTML not to be a programming language because I wanted different programs to do different things with it: present it differently, extract tables of contents, index it, and so on.
>
> *Principle of Least Power* [138], Tim Berners-Lee, 1998

2.3.8 Semantic Publishing

One of the shortcomings of the research publication system is that scientific literature in electronic formats resemble, mimic, and have similar affordances to printed papers. While the paper-centric approach has been adequate, although not necessary optimal, for human consumption and reuse, it is insufficient for machine-assisted analysis or reuse. The print-centric affordances lack structure that could enable better aggregation, discovery, processing, and integration – which machines excel at – and in turn help with human-led tasks. Representing and storing semantically structured knowledge on the Web enables automated agents to process information more intelligently.

The most wide use of the term “semantic publishing” in scientific literature is defined in *Semantic publishing: the coming revolution in scientific journal publishing* [139], Shotton, 2009. Here the term “semantic publishing” is constrained to “anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers”. In *Genuine Semantic Publishing* [140], Kuhn, 2017, posits that with respect to the Semantic Web vision, this definition is “too restrictive” in a sense that it is concerned mainly about narrated information, and “too inclusive” in that a shallow semantic publication would still qualify. The article argues that the definition for semantic publishing should be broader in that works without a narrative are accounted for, and narrower in the sense that semantic representations should be incorporated at the time of creating and publishing entities. The “genuine semantic publishing” concept prescribes the notion of publishing 1) *machine interpretable* formal representations, 2) having *essential coverage* of the claims, 3) *authenticated* by the contributors, 4) be a *primary* component of the work, and 5) be part of *fine-grained* and *light-weight* containers.

Analogous to the “FAIR principles”, data needs to "signal" potential opportunities for *reuse* as per “Research Objects” in *Why Linked Data is Not Enough for Scientists* [141], Bechhofer, 2013.

2.3.9 Arguments and Citations

A core component of the Web is hyperlinks between documents. Most, if not all, academic articles reference others in some way. For articles published using Web technologies for the content or even just the metadata, creating hyperlinks is an obvious way to enhance these references. Such links can create citation chains between articles at a high level, and graphs of claims and arguments at a more granular level.

The case is made for using a Web-friendly and machine-tractable model of scientific publications in *Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications* [142], Clark, 2014. The Micropublications model, formalised as an OWL ontology, can reduce the human-labour cost of checking the support of a claim, as it is capable of representing scientific controversy and evidentiary requirements in primary literature. Further, the Micropublications model can be applied to generate argument graphs representing the claim that is argued, and its support by other statements, references, and scientific evidence (data). Machine-navigable linkage at the level of assertions in scientific articles and data improves robustness of scientific citations and observations. *Argument graphs: Literature-Data Integration for Robust and Reproducible Science* [143], Clark, 2015, posts that creating data structures based on the relationships between arguments can be complementary to using entity recognition for mapping textual terms in articles to curated scientific databases.

Citations perform a variety of functions, and an author’s intention when referencing a particular source is not always evident from the prose, and certainly not from a machine-readability perspective. For example, does the author mean to point to a whole document (this is usually the level at which traditional citations operate) or are they only referring to a particular statement or section within it? Are they citing something because it supports their perspective, or because they disagree with it? This is particularly problematic in a system which simply counts the instances of a citation to determine impact or relevance. *Measuring academic influence: Not all citations are equal* [144], Turney, 2015, and *Why (almost) Everything We Know About Citations is Wrong: Evidence from Authors* [145], Teplitskiy, 2018, provide evidence indicating that authors with a means to express their citations and intentions in a structured way can improve the preservation of the connections between research assertions or units of information in general.

Fortunately there has already been work towards creating vocabularies around the relationships between articles and their citations. In *An annotation scheme for citation function* [146], Teufel, 2006, a classification system is proposed for relationship information between articles and their citations in order to improve automatic recognition to build better citation indexers, as well as to have more accurate interpretations of scientific arguments.

The Citation Typing Ontology (CiTO) model is equipped with common classifications eg. "cites for information", "disputes", "cites as evidence", "cites as potential solution", that can be used to create explicit relations between entities, as posited in *FaBiO and CiTO: ontologies for describing bibliographic resources and citations* [147], Peroni, 2012. It also has an extension mechanism to support customised properties in order to allow authors to express their intentions closer to their mental model.

In contrast to extracting rhetorical knowledge structured claimed by authors of scientific articles, *Corpora* *for the conceptualisation and zoning of scientific papers* [148], Liakata, 2010, describes how the structure of human-readable investigation in research articles can be retrieved based on a generic high-level annotation scheme for scientific concepts. It is shown that such complementary approach can assist in the generation of automatic summaries and intelligent querying.

Genuine Semantic Publishing [149], Kuhn, 2017, also contends that while narrative text remains an important part of research discourse, publishing the data using formal semantics without the need of a narrative is beneficial. Meanwhile, tooling for creating unique identifiers at granular levels, for clauses or statements within articles as well as pieces of supporting datasets, hinders researchers' in using structured mechanisms for describing their citations. Authors may be unable to obtain or generate a precise identifier, or lack an authoring or referencing environment which would allow them to assign identifiers appropriately. When an author's intentions are preserved only in prose, entity-recognition techniques are typically required to understand the original intentions of citations. This is generally a complicated task and error prone given that an algorithm attempts to reverse-engineer the purpose of the citation.

2.3.10 Registration of Identifiers

Identifying units on the (scholarly) Web is currently done with different kinds of social agreements and expectations. In *Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data* [150], McMurry, 2018, posits that actors should generally create their own identifiers for new knowledge, and to reuse existing identifiers when referring existing knowledge. Depending on the one or more roles an actor has in the creation, editing, and republishing of content, the decision to create a new or reusing identifier is left as a judgement call.

Here I discuss a few systems that pertain to registration of identifiers:

Digital Object Identifier [151] (DOI) is an international standard for document identification, and the services that are offered through `doi.org` is managed by the International DOI Foundation (at the time of this writing). The foundation permits *DOI Registration Agencies* [152] to manage DOI records. In order to have a resource assigned with a DOI, it has to be done by one of the DOI registration agencies which requires organisational membership, excluding individuals. For example, a typical journal publisher as a member of a DOI registration agency (eg. *CrossRef* [153]) submits mappings for objects of interest. Given the statement in AWWW, "URI persistence is a matter of policy and commitment on the part of the URI owner", DOI introduces an additional layer of social contract to govern the discovery of research artifacts. A DOI is aimed at being a persistent identifier for an object it is assigned to, and has the mechanism to refer to the URL of an object where it can be found. That is, a DOI may refer to a landing page about the actual object and not necessarily the object itself – this happens to be a common practice for literature that is not necessarily open and free. For literature that has restricted access, the resolved DOI is typically a webpage where an actor would need to go through the system's "paywall" or "access toll" in order to retrieve the content. From a machine's perspective, what the DOI identifies at the resolved location is not clear eg. company logo, tracking software, search form or the intended research object? For open literature, the resolved DOI would generally hyperlink to the actual scientific object in human-interpretable only fashion, or be part of a webpage accompanied by material unrelated to the scientific object itself, ie. information that the scientists neither assembled as part of their research nor intended for the consumer. If the actual URL of the object changes, the DOI's lookup table for the URL has to be updated. Otherwise, they'll point to dead links or unrelated information.

Open Researcher and Contributor ID [154] (ORCID) is a community-governed persistent identifier system for researchers and organisations. An ORCID identifier is specified as a URI. While anyone can create and update their ORCID profile, the identifier is ultimately owned and managed by the ORCID organisation and made available through `orcid.org`. Profiles are constrained to express to the UI's template and the kind of information ORCID is willing to store. For most intents and purpose, the free service offered by ORCID is sufficient for researchers, and provides a mechanism where researchers can aggregate information about their education, employment, funding, and works like academic publications and data. It also allows researchers to extend their profile descriptions, as well as linking to their alternative online profiles.

Persistent Uniform Resource Locator [155] (PURL) is a way to provide an HTTP URL a permanent redirect to another HTTP URL. PURLs can be registered and managed by anyone. The PURL service is currently administered by the *Internet Archive*.

Similar to PURLs, is the W3C *Permanent Identifier Community Group* [156]'s resulting *w3id.org* [157], where a group of entities that pledge to ensure the operation of the community-drive service. w3id.org provides a permanent URL re-direction service, and it can be done by making a request to register and update a URI path. There are shared second-level paths eg. for people, that can be registered the same way. The service is flexible in that different server rules for redirections can be set.

Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping [158] Van de Sompel, 2014, states that a *Persistent Identifier* (PID) is a "string that complies with a well-defined syntax, minted in a namespace controlled by a naming authority". PURL, w3id, ORCID, and DOI are PIDs. For digital objects accessible over the Internet, PIDs are intended to be used as a long-lasting reference to a unit of information, and be *actionable* in the sense that the environment can be directed to the identified source. For example, a DOI-based PID can be `10.2218/ijdc.v9i1.320`, and the organisation – in this case the International DOI Foundation – would be responsible to maintain an unambiguous mapping to an accessible resource. Typically once mapped, the process starts from a DOI URI `https://doi.org/10.2218/ijdc.v9i1.320` which (currently) resolves to the location `http://www.ijdc.net/article/view/9.1.331` where its contents can be retrieved. URN, International Standard Book Number (ISBN), Archival Resource Keys (ARK) are other examples of PID.

There are different forms of social contracts or *promises* made for long-term societal benefit in terms of archival and preservation of knowledge. They can be categorised as follows:

- any Web-wide publicly usable archival services, eg. *Internet Archive*, *archive.is* [159], *WebCite* [160], *Perma.cc, Webrecorder* [162] [161] [179];
- dedicated digital preservation organisations, eg. *Portico* [163];
- libraries, eg. [164];
- global archives preserving content on behalf of all libraries, eg. *CLOCKSS* [165];
- subscription based service for all kinds of libraries, federal institutions, state archives, NGOs, eg. *Archive-It* [166];
- state or federal archives, eg. *Swiss Federal Archives* [167], *Library and Archives Canada* [168];
- institutional-run digital archives, eg. *TIB* [169], *Zenodo* [170]

In 1996, the *Internet Archive* [171] begun to archive the publicly accessible resources on the Internet, and provides free public access to its digital library. The *Wayback Machine* [172] is a service that allows users to search and access archived Web documents in particular. The Wayback Machine can be used to navigate through different versions of URLs based on the date-time in which they were accessed and stored. For example, all of the captures of csarven.ca as well as the other articles of the site are available, along with a summary.

2.4 Social Scholarly Web

> Experience with the ªWorld-Wide Webº (W3) global information initiative suggests that the whole mechanism of academic research will change with new technology. However, when we try (dangerously) to see the shape of things to come, it seems that some old institutions may resurface but in a new form.
>
> *Physics World article for end March 1992* [173], Tim Berners-Lee, 1992

The scholarly ecosystem is inherently social. Researchers and scholars have been exchanging and collaborating on their findings and questioning, just as they were in the era of the *Republic of Letters*. The social dimension not only exerts a force within the research community, from the society, and eventually flowing back out to society. As the Web deeply impacts societies, in addition to using social media, researchers and academics have also been studying sociotechnical systems from their respective disciplines. We need to understand the effects of the tools we use on the ecosystem and on scientific and scholarly communication. There is a top-down social pressure for research openness and access to knowledge that is in the public interest, however, achieving it through the methods and ethics is challenging for societies. In this section, I will overview how researchers use social media, quality control and certification, and access to research knowledge in context of privilege and gatekeeping.

2.4.1 Social Media and Academia

Contemporary social media and networking sites has been facilitating academics to exchange information. More specifically, academics use social media for a variety of reasons, to name a few: to network and collaborate; disseminate and find research results, crowd-source and crowd-fund research challenges, engage with public and foster trust on research results by providing authoritative feedback.

There are online communities geared towards scientists, academics, librarians, and scholars with topical focus on various aspects of scholarly communication. In *Academics and their online networks: Exploring the role of academic social networking sites* [174], Jordan, 2015, discusses the proliferation of social networking sites predominantly operated by businesses independently of education institutions. Academics from different disciplines communicate to publish and review articles in different states (pre-prints to post-prints), pose and answer questions, construct identities and build profiles, exchange messages, as well as explore trends.

Online social media tools are affecting or influencing scholarly communication from a variety of fronts. For instance, with the advent of the Web and the notion academics having the right and means to exchange scholarly literature with their peers for free, there has been plethora of initiatives and developments over the course of three decades in academia, libraries, archives, as well as standards for creating, sharing and preserving (*Open Access Directory* [175]). Moreover, the call to take advantage of features offered by the new media, has opened new ways of publishing and disseminating scientific and scholarly knowledge globally, as well as institutional infrastructures being adapted to meet communication needs. For example, universities providing an online space for staff members and students to publish their profiles, as well as to share their research interests, academic output, and other educational material. Institutional repositories have been built to retain copies of scholarly work, as well as mechanisms to exchange data with other repositories.

In recent years, there has also been growing number of community-led initiatives for individuals to help researchers and scientists by performing *citizen science* [176] independently, from sharing photos of coral reefs, analysing radio signals, to solving protein structures.

In parallel to "internal" technical and social advancements, infrastructure towards open scholarly communication has been offered by commercial entities in various forms with different business models, from authoring, reviewing, messaging, organising, indexing, searching, to visualising scholarly information.

In essence, third-party controlled online social media exist simply because academics have needs to be fulfilled. *PASTEUR4OA Briefing Paper: Infrastructures for Open Scholarly Communication* [177], Moore, 2016, states that online platforms have become increasingly centralised locations for a range of user interactions, and they have a commonality in that, their primary “currency” is user’s data which is monetised in various ways; the platforms cater to different stakeholders while being “free”; the platforms are often funded by firms that offer venture-capital and are almost exclusively driven by profit-maximising; and in order to maximise the number of users and data generated through their interactions, the platforms are centralised and act as closed proprietary ecosystems (commonly referred to as “walled gardens”) where data access, portability, ownership and content structures are constrained in ways which makes it difficult for users’ to simply move to another system.

2.4.2 Quality Control and Certification

> **peer review**
> A method of quality control of scholarly articles, whereby each article submitted for publication in a journal is sent to one or more scholars working the same research field as the author and the scholar(s) assess(es) whether the article is of a high enough standard and appropriate for publication in that particular journal.
>
> *Encyclopedia of International Media and Communications* [178], Scholarly Journals, Vol. 2, p. 687, 2003

In *Free at Last: The Future of Peer-Reviewed Journals* [179], Harnad, 1999, states that “peer review is a quality-control and certification (QC/C) filter necessitated by the vast scale of learned research today.” In *Implementing Peer Review on the Net: Scientific Quality Control in Scholarly Electronic Journals* [180], Harnad, 1996, posits that electronic networks offer the possibility of efficient and equitable *peer review* on published and ongoing work, which can replicate the traditional paper form. As the medium revolutionised interactive publications, it was possible for peer review to be supplemented with online commentary as another form of quality control for scholarly literature at any state.

The scientific community has been aware of the crude and fallibility of traditional peer review, for example, in cases such as, crediting bad research, good research getting dismissed for political reasons, “predatory publishing” (quality of the research content being irrelevant), or even treating commentary or opinion pieces as equivalent to scientific evaluation. *Nevertheless*, the function of peer review in the scholarly system is a form of self-regulation and certification. Without having any maintenance for quality, distinguishing valid, reliable, useful and ethical work from those not, would be difficult to achieve systematically. This is particularly concerning where the general public can become aware of a piece of information, yet the validity of the research findings may be inaccessible or absent, and thus have major consequences to our society and the planet – as with cases where widely agreed scientific evidence and observable effects of global climate change can be easily "questioned" by propaganda spread through mass-media by elite forces with a personal or private agenda; the amplification of the flat Earth conspiracy, and so on.

What is open peer review? A systematic review [181], Ross-Hellauer, 2017, first categorises the problems in traditional peer review, second studies the diversity of definitions of “open peer review” (OPR) in literature, and sets out to resolve the ambiguity, and classifies the common traits. The articles argues that the OPR traits: “open identities”, “open reports”, “open participation”, “open interaction”, “open pre-review manuscripts”, “open final-version commenting”, and “open platforms (“decoupled review”)” can help to address common problems in the traditional peer review model categorised as: *unreliability and inconsistency*, *delay and expense*, *lack of accountability and risks of subversion*, *social and publication biases*, *lack of incentives*, *wastefulness*.

The peer review model towards quality control has received extensive criticism. These are studied in *A multi-disciplinary perspective on emergent and future innovations in peer review* [182], Tennant, 2017. The article examines different approaches to peer review models and systems. In context of “decoupled post-publication (annotation services)”, the article states that while there are online communities and services for open peer review, the systems are largely non-interoperable in that “most evaluations are difficult to discover, lost, or rarely available in an appropriate context or platform for re-use”.

Reproducible research findings refers to the notion that when the exact methods are used to replicate or repeat to reach the same findings. However, for some research, reproduction is not always possible, for example when data is sensitive, or influenced by time or space. In this case, replication can be carried out to answer the same research questions using similar or equivalent approaches. The reproduction of the original work can be either confirmatory, ie. able to reproduce, or; contradictory, unable to reproduce or inconclusive. Thus, in order to reconstruct how the data was generated and to make sense of it, data needs to be accompanied by enough detail about its provisional interpretations.

The preregistration revolution [183], Nosek, 2018, describes the uptake of the “preregistration” process as a way to “define the research questions and analysis plan before observing the research outcomes”. “Registered reports” was initially proposed as an open letter to the scientific community as a way to *Changing the culture of scientific publishing from within* [184], Chambers, 2012. Researchers would register a research plan which gets reviewed by their peers for significance, rationale, soundness and feasibility of the methodology, and replicability, it would be approved to carry out the research. The approach provides a conditional acceptance for the final manuscript.

> Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.
>
> *Goodhart’s law* [185]

The current scholarly system implements metrics for impact of various scholarly artifacts. *Citations impacts* [186] quantifies the importance of academic articles, journals, and authors. Article citations is used to indicate how often it is referenced. *Journal Impact Factor* [187] (JIF) is used to indicate the relative importance of an academic journal within its field. *h-index* [188] is one way to measure bibliometric impact of individual authors in their field. The validity of these metrics has received numerous criticisms.

Prestigious Science Journals Struggle to Reach Even Average Reliability [189], Brembs, 2018, examines accumulating data which suggests “methodological quality and, consequently, reliability of published research works in several fields may be decreasing with increasing journal rank.” As society becomes more literate and reliability of research becomes more important, the article posits that the scientific community should strive to abolish the pressure that JIF exerts towards measuring the productivity of scholars, in order to preserve society’s trust in the scientific endeavour. *Deep impact: unintended consequences of journal rank* [190], Brembs, 2013, posits that “fragmentation and the resulting lack of access and interoperability are among the main underlying reasons why journal rank has not yet been replaced by more scientific evaluation options, despite widespread access to article-level metrics today”.

Altmetrics [191] is “the creation and study of new metrics based on the social web for analyzing, and informing scholarship”. They are are complementary to traditional citation impact metrics by aggregating diverse online research output eg. citations, downloads, mentions, from public APIs (*The Altmetrics Collection* [192], Priem, 2012). It could be applied in researcher evaluations, as well as can pressure individuals to use commercially controlled social media.

2.4.3 Privilege and Gatekeeping

The scholarly communication system has been the subject of controversies internationally ever since the shift towards its systematisation and the effect of the forces in society, research market and knowledge industry. Controlling access to information or knowledge in all intents and purposes has been a race to controlling science (communication), financial and political power. Moreover, intelligence gleaned from users data has not only been profitable, but has also raised long-standing privacy issues.

2019 Big Deals Survey Report: An Updated Mapping of Major Scholarly Publishing Contracts in Europe

[193], European University Association, 2019, states that the consortia including universities, libraries, government representatives, and scientific organisations, “representing 30 European countries reported a total annual spend of €726,350,945 on periodicals Big Deals. The proportion of these costs covered by universities is about 72%, or approximately €519,973,578”. *SPARC Landscape Analysis* [194], SPARC, 2019, observes a company’s revenues per journal article in the region of 4100 USD. *Is the staggeringly profitable business of scientific publishing bad for science?* [195], The Guardian, 2017, reports that some publishing and data analytics businesses earn close to 40% profit margins meanwhile scientists get a “a bad deal”.

Inequality in Knowledge Production: The Integration of Academic Infrastructure by Big Publishers [196], Posada, 2018, commercial entities traditionally with the publisher role have evolved and increased their control of the research process, publishing process, as well as the research evaluation process.

The Oligopoly of Academic Publishers in the Digital Era [197], Larivière, 2015, posits that the value added by the publishers did not increase as their rise of ownership in the scientific publishing system over centuries. According to *Opening the Black Box of Scholarly Communication Funding: A Public Data Infrastructure for Financial Flows in Academic Publishing* [198], Lawson, 2015, the emergence and growing use of APC model provided a stable revenue stream to publishers, while the technical infrastructure remained relatively unchanged. In *Open Access, the Global South and the Politics of Knowledge Production and Circulation* [199], an Open Insights interview with Leslie Chan, 2018, states that “adding openness to an asymmetrical and highly unequal system simply amplifies the gap and empowers the already powerful.”

A corpus of research articles on the topic of “open” and “free” access to research is collected in *Group: Open Access Irony Award* [200], 2019, where individual articles at the time of their publication are themselves inaccessible, require subscription or payment. In *Practicing What You Preach: Evaluating Access of Open Access Research* [201], Schultz, 2018, sets to find out how many published research articles about OA fall into the category of publishing their work in paywalled journals and fail to make it open. One explanation of this phenomenon may be that while the research community has shown increasing interest in studying and pursuing open access to research (over several decades, if not more generally over centuries), as well as an overwhelming support for the idea of open access as shown in *Highlights from the SOAP project survey. What Scientists Think about Open Access Publishing* [202], Dallmeier-Tiessen, 2011, the actors in the scholarly system are constrained or tied to traditional models of publishing and distribution.

The *SPARC Landscape Analysis* [194], SPARC, 2019, reports that some of the major academic publishers are transitioning from providing content to data analytics business. The document provides broad-stroke strategies that higher institutes can consider eg. “revising existing data policies, establishing coordination mechanisms, negotiating to ensure institutional ownership of the data and infrastructure and establishing open terms and conditions … re-thinking the institution’s relationship to its data in terms of commercial exploitation, IP ownership, and research investment outcomes.”

Vertical Integration in Academic Publishing [203], Chen, 2019, confirms academic publishing industry’s expansion and control of “data analytics by building end-to-end infrastructure that span the entire knowledge production life cycle” through mergers and acquisitions. Authors state further vertical integration can potentially have an “exclusionary effect on less financially well-endowed journals and institutions, primarily those in the Global South, in their attempt to emulate the western modality of knowledge production”.

The article *Publisher, be damned! From price gouging to the open road* [204], Harvie, 2014, criticises the large profits made by for-profit publishers while taking advantage of academics’ labours. After internal disagreements (between the publisher and the editorial board), the article was published where the original version introduced a disclaimer warning by the publisher stating that “the accuracy of the content should not be relied upon.” Such unfortunate turn of events exemplifies how academic freedom can be jeopardised, subject to censorship and manipulation when external parties ultimately control the distribution of knowledge.

In response to paywalled articles with high costs, *Sci-Hub* [205] is a website that provides free access to research articles and books. The authentication credentials is illicitly obtained authentication credentials to commercial publisher’s online libraries. Sci-Hub has received a variety of responses including praise and criticism from different stakeholders including the research community and the commercial publishers.

> In the case of scholarly publishing, standards such as DOI, JIF, and even metadata standards, are not only technical decisions, but are also political decisions with public policy implications. This is because these standards are not neutral, but are designed to privilege certain types of knowledge or outputs, while rendering other invisible.
>
> *Open Access, the Global South and the Politics of Knowledge Production and Circulation* [206], an Open Insights interview with Leslie Chan, Leslie Chan, 2018

Some of the observable effects of commercial participants in the scholarly ecosystem on scholarly communication is that they have the means to monitor the usage and interaction of units of scholarly information, censor who or what information gets shared and at what cost, have the agency to control the flow of information, and influence the coupling of their notion of academic impact to a biased incentive system.

The current state of scholarly communication predominantly operates under the notion of requiring or relying on third-party publishers to make scholarly contributions available, as well as setting technical and ethical constraints on what can potentially be a scholarly contribution. Looking at this from another perspective, we can observe that actor-controlled participation is not possible towards “open” participation. I will highlight *some* of the technical practices (from a plethora of instances) put in place by third-party publishers (irrespective of genuine interest in adhering to the Open Science initiative) which raises a number of user experience (UX) issues. Some commonly observable UX includes:

- Session tracking.
- Requiring cookies to be enabled in order to access the core content.
- JavaScript-driven resources – forcing users to use certain software and enable JavaScript execution.
- Resources (eg. articles, annotations) including tracking software.
- Resources are actively blocked, rate limited, restrictive terms and conditions on reuse for content mining – preventing further research.
- Web crawlers are actively blocked – preventing archiving services to snapshot.
- Proprietary data models and access methods.
- Resources including advertisements.
- Resources including subscription forms to the journal.
- Resources are branded by the publisher.
- Resources change at the discretion of the publisher.
- Persistent identifiers eventually resolving to landing webpages with no deterministic way for machines to discover content.

2.5 Forces and Functions

The scientific communication *market* is an interplay between actors with an objective to exchange units of information along the lines of: academic literature, analysis, annotations (eg. assessments, bookmarks, comments), datasets, evidence, interactive representations, software, observations, provenance, and workflows. Further, researchers commonly want to expose their work beyond the scientific community to the wider world, seeking adoption or contradiction of new ideas and findings.

In the seminal work, *Forces and functions in scientific communication* [207], Roosendaal, 1997 provide a framework which has had historical influence on research and development pertaining to technical challenges on interoperability, workflows, service sharing, and information modeling in scholarly communication. Through this thesis, I will refer back to these concepts as a means to describe and compare various technologies and approaches to scientific publishing.

Roosendaal and Geurts disassemble the scientific communication market as an interplay between four forces and functions. The dynamics between the forces and functions is described as a tetrad:

Forces

The scientific communication market consists of four forces with each having complementary parts: *actors*, *accessibility*, *content*, and *applicability*.

Actor
Any kind of interactive agent that creates or uses information, eg. authors, readers, tool implementers, policy makers, application programs.

Accessibility
The availability and retrievability of information, content accessibility.

Content
The generated questions and answers by actors.

Applicability
Transfer of knowledge to science, technology, and society.

Functions

Scientific communication consists of four functions: *registration*, *awareness*, *certification*, and *archiving* (RACA).

Registration
Allows claims of precedence for a scientific activity.

Awareness
Remaining informed of the observations.

Certification
Establishes the validity of registered claims.

Archiving
Preservation of scholarly records over time.

The actor and content forces are generic and internal to the scholarly market, and are considered to be indispensable in that there would not be a scientific communication market if there are no researchers and the artifacts they exchange. The accessibility and applicability forces on the other hand are external forces because they shape, give purpose to, or influence the activities and content.

The registration and archiving are objective functions in that they are external to the research process. Information exchange cannot take place if there are no usable identifiers for the units of communication or if no knowledge preservation measures are put in place. Certification and awareness are subjective communication functions – internal to the research process. Registration and awareness are different aspects of making information findable, or *scientific observations*. Certification and archiving are aspects of *scientific judgement*, ie. they act as filtering mechanisms in the information space.

In today’s scholarly publishing, the journal article is considered to have four chief functions:

> - The dissemination of information
> - The archiving, or permanent preservation, of the information
> - The assignment of priority in the information
> - Quality control
>
> *Encyclopedia of International Media and Communications* [178], Scholarly Journals, Vol. 2, p. 690, 2003

In *Rethinking Scholarly Communication* [208], Van de Sompel, 2004, proposes to revise the notion of a unit of scholarly communication in two ways. In a technological sense, it suggests to recognise the compound and dynamic nature of a “unit of communication” that work on the Web. Content can be the result of a combination of other registered items; data may be available through interactions, or depend on acquiring externally registered items. In order to empower scholarly communities from a systems perspective, the article posits that there should be more flexibility in what constitutes a unit of scholarly communication in order to allow registration, as well as factoring timing and the quality of what is to be registered. The article also acknowledge the presence of an additional function in the market which reflects the academic credit and policy in the current system. A derived function, “rewarding” of actors “for their performance in the communication system based on metrics derived from that system”. For rewarding to take place, the information space needs to provide the ability to extract meaningful metrics.

In *A Perspective on Archiving the Scholarly Web* [209], Van de Sompel, 2014 characterises the ongoing evolution and transition of objects pertaining to scholarly communication. In particular, the communication is becoming more “Web-centric” ie. visible, continuous, informal, instant, content-driven, and the communicated objects vary in their type, have compound composition, diverse, networked, and open.

Given that the emergence of the Web influenced the communication market by transforming it from predominantly being medium-driven into a content-driven ecosystem, as well as the research process becoming more visible on the Web (per the “Open Science” movement), the article also posit that the aforementioned indicators impact the functions in scientific communication, and in particular have a significant impact on the future of archiving. The observation is based on the archival process held on the Web; unplanned and unconstrained archiving, organically occurring, by way of Web objects being deposited across distributed and dedicated archival services. They describe the effect of Web-based archiving with respect to the four functions:

- registration: a wide variety and versions of objects that are compound, dynamic, interrelated, and interdependent are available from different Web platforms;
- certification: a variety of mechanisms that are decoupled from registration;
- awareness: a variety of objects being communicated in real-time and social;
- archiving: “no constraints on the number or kinds of parties that can hold archived copies”, the need for “appropriate levels of cross-archive interoperability in order to support seamless, uniform access to archived objects”, that are temporally appropriate while context being preserved

The authors conclude that in order to address the needs of a distributed archival approach for scholarly Web objects, standards will play a central role for interoperable systems.

Given the dynamic nature of units of communication on the Web, vertical integration, stepping through the four functions in order (as per the traditional journal-centric ecosystem), turns out to be inadequate. Hence, an affirmation of the outlook as outlined in *Rethinking Scholarly Communication*, that the essential functions need to be “fulfilled in discreet, disaggregated, and distributed manners, and in which a variety of networked pathways interconnect the autonomous hubs that fulfill these functions”. The act of “recording” (or *registering* in the most general sense) a unit of communication is not a replacement for *archiving* or persistence, even if the freely available Web services are used towards this end. The authors distinguish recording and archiving in that the former is short-term, no guarantees provided, used for many read-write operations, and part of a scholarly process. Archiving serves a different function, in that it is intended for the long-term, and there is an attempt of a guarantee (or promise) and the archived artifact becomes part of the scholarly record.

The authors call for another archival paradigm that is based off the characteristics of the Web. That is, given the inherently interlinked and dynamic nature of Web objects, archived resources cannot be treated as atomic resources as they do not necessarily contain the complete content. For instance, a research document on the Web may refer to external media or interactive scripts, aggregated objects, and changing or ephemeral objects. This is in contrast to the traditional journal-centric approach to archival where all material – physical paper – can be eventually tracked to library stacks. Hence, the future of archiving needs to consider the nature of the object, including the degree to which it can be archived. The complexity and care that is required for the proper preservation of digital objects is further described in *Requirements for Digital Preservation Systems* [210], Rosenthal, 2005, provides a taxonomy of threat models: software, communication, network, economic, organisational, to name a few, which pose a particular risk to permanent information loss. Thus, it is important to bare these potential issues in mind when considering what scholarly units of information will contain and how they will be made available.

2.6 Effects and Artifacts

> All of man's artifacts – whether language, or laws, or ideas and hypotheses, or tools, or clothing, or computers – are extensions of the physical human body or the mind.
>
> *Laws of Media* [211], p. 93, Marshall McLuhan, 1989

In recent decades there has been a global shift in communication forms used for scientific activities. This perhaps demonstrates a desire for amplification of the senses that are used to consume and create knowledge. For instance, print-centric emphasises the linearity of scientific information representation and its communication, whereas a Web-centric approach makes multimodal interactions possible within the same domain. Scientists need no longer fit their understanding of the world within the confines of static and two-dimensional descriptions of phenomena. The need to travel great distances for first-hand observations and collaboration is also reduced. With electronic media, and more specifically the Web, we have the option to also watch, listen, and participate around measured observations, instantaneously as the events occur, or after the fact around persistent and detailed logs. Just as the invention of the printing press transformed the codification and mobility of knowledge, the Web transformed and extended human and machine communication on all fronts. What will be further explored in this thesis is the extent to which our current scholarly activities make use of the inherent qualities of the Web, as well as ways to further take advantage of these qualities.

Tetrad of Media Effects: The Web is an *artifact* of our society. In order to help us understand the *effects* of the Web on scholarly communication, we can *tetrad of media effects* from the *Laws of Media* [212], McLuhan, 1989, as a heuristic device; a set of four questions:

> - What does it enhance or intensify?
> - What does it render obsolete or displace?
> - What does it retrieve that was previously obsolesced?
> - What does it produce or become when pressed to an extreme?
>
> *Laws of Media* [211], p. 7, Marshall McLuhan, 1989

Effects of Web Centric Scholarly Communication: The questions are intended to be complimentary and studied simultaneously. They correspond to four effects: "enhances", "obsolesces", "retrieves", and "reverses into". I use this tetrad to situate the effects and artifacts of scholarly communication on the Web in relative to adjacent mediums, in particular, to print media:

Enhances
access to information, autonomy, social connections, decentralisation, distribution, immediacy, multi-modality

Obsolesces
centralisation, isolation, distance, memorisation, privacy, tactility, travel, linearity, typographical conventions

Retrieves
community, correspondence, direct representation, free expression, dialogue, involvement, public space, self-control

Reverses into
collectivism, degraded information quality, group-think, information overload, fabrication, lack of accountability, noise, third-party control

Throughout this thesis, I will refer back to these concepts.

The tetrad serves as a conceptual categorisation of impacts of a medium that a particular media affords. It allows us to outline and analyse what the medium enables or has implications for based on observations in "the real world", or in day-to-day academic practice. In particular, we can examine the Web’s effects and artifacts within with regards to composition of scholarly information, individual autonomy to create and consume knowledge, and sociality of scholarly communication. By examining these effects, we can hone in on understanding what the underlying *form* does to us, society at large, and scientific communication itself. It provides us with a foundation to go further in uncovering the nature of scholarly communication’s interplay with the Web.

How does scientific communication evolve through the mediums it operates from?

2.7 Understanding the Medium

In the 1967 book, *The Medium is the Massage* [213], McLuhan wrote: "at the high speeds of electric communication, purely visual means of apprehending the world are no longer possible; they are just too slow to be relevant or effective". While McLuhan's critique was about the effect of the print media in the electric age, it stands especially valid today in the age of interactive new media. For instance, the *World Wide Web* is a "cool media" because its access, multi-linearity, and interactive possibilities offer multi-sensory states for individuals. This is in contrast to *print*, considered by McLuhan to be "hot media", demanding primarily the attention of our vision. As the printed word's requirement for the other senses is minimal, it *spoon-feeds* the content to the reader in a linear fashion. The literate or print culture is passive and is detached from immediate involvement because the medium has no such requirements. The cooler media on the other hand, requires a greater level of engagement from the audience at a faster rate. This increased engagement allows us to detect change and better recognise patterns, which, according to *Gestalt laws of grouping* [214], makes the whole of the information easier to follow and understand.

The interaction of media with our senses is an inherent property of media. The media's built-in sensory bias affects and transforms human society and psyche. Modern media, like the Web, offer extensible interactions, creative participation, and social engagement. There is a natural human desire to connect and exchange ideas, behaviour, or style with others, as well as to co-create. Given these opportunities, the fundamental question to ask here is, why should research communication be limited to, evaluated by, and rewarded on methods and practices be based on print-centric media? Scholarly communication still uses methods which date back to the invention of the *mechanical movable type printing press* [215] (c. 1450 CE), even in the advent of the globally accessible and interactive new media that are at our fingertips today, and at much faster turnaround times from publication to consumption than ever before. The practices correspond to publication workflows that are in place for academic communication, and have remained largely unchanged for over a century. What does it mean to have an academic achievement through a "publication"? If we consider the Web as a potential supplement to or replacement for these older practices, and given that the Web can be purposed for greater possibilities than print alone, then we (members of the academic community) also need to address the question, *what constitutes a scholarly contribution?*

Today, it is evident that the typical academic publishing scenario requires adherence to the *printability* aspects of the information eg. page length, typographical guidelines. That is, all of the encompassing data or supporting artifacts for any given research output typically must be channelled through the characteristics of print media. In practice, this leads the research community to use tools that are tailored solely for printing or viewing (on screen). This (sticking to print) leads to unnecessary information loss or arbitrary filtering of the content, which might be otherwise preserved if a different medium was used. Since print is only one possible way to represent and communicate the information, it is worthwhile to investigate and employ alternative or complimentary representations that are more appropriate given the affordances of the Web. Traditional electronic document formats, on the hot end of the media spectrum, are trying remain relevant inside a cool medium.

This section served to acknowledge initiatives and developments that went into scholarly communication in the age of information and Web, against their historical backdrop. It also introduces some frameworks for description and analysis of technologies and processes particular to scholarly communication, and communication mediums in general, which I will continue to refer to through the rest of this thesis. Subsequent sections are dedicated to devising a scholarly space that is more fitting to the medium of the Web.

3. Structure of Scholarly Information

Scholarly Communication on the Web served to explain the state of sociotechnical affairs in scholarly communication. In this section, I focus on facets of research and scholarly information and its exchange on the Web. The *Linked Statistics* section describes my knowledge and artifact contributions for publishing and exchanging statistical Linked Data.

3.1 Information Spaces

Given that global strategic research is influenced by policies and industry, different *information spaces* can be created by an ensemble of units of information. Such information spaces would serve to facilitate knowledge exchange in the scientific communication market. In this section, I investigate existing approaches towards configuring different aspects of Web-centric information spaces.

Scholarly knowledge includes a range of research artifacts that needs to be described, uniquely identifiable and be discoverable on the Web. These include research articles, peer reviews, research data, social interactions like review requests and notifications in general, as well as different kinds of annotations on research objects. The current state of access and use of scholarly knowledge is insufficient for society at large. By enabling accessible “Scholarly Knowledge” (*Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web* [216]) graphs as well as applications which make use of it, we hope to enable universal access to previous research. By improving the availability through linked research, we can facilitate discovery and building on existing research. A fundamental first step is to investigate and develop effective ways to represent fine-grained information that is accessible, human- and machine-readable, and interconnected.

A typical high-level interplay between the *actors* and the *content* forces, and their functions (registration, awareness, certification, archive) in scholarly communication have authors reply to call for contributions by sharing their literature. Reviewers provide feedback on the research results; individuals annotate and re-share the literature; editors filter and assemble a collection of work; archives are notified about the research. While the document-centric Web was mostly for human-readability, Linked Data-based Web is oriented towards improving discoverability, readability, and reuse for machines, and in effect better helping humans. For example, if research articles capture their problem statements, motivation, hypothesis, arguments, workflow steps, methodology, design, results, evaluation, conclusions, future challenges, as well as all inline semantic citations (to name *a few*) where they are uniquely identified, related to other data, and discoverable, then specialised software can be used to verify the article’s well-formedness with respect to the domain. In essence, this materialises the possibility of articles being executable towards reproduction of computational results. Similarly, user interfaces can manifest environments where readers can rerun experiments or observe the effects of changing the input parameters of an algorithm. This has the affordance for a more involving environment for the user, improves learnability of material, and supersedes the passive mode of reading.

In a 1999 article, *Practical Knowledge Representation for the Web* [217], Van Harmelen, 1999, stated that “the lack of semantic markup is a major barrier to the development of more intelligent document processing on the Web”, and “meta-data annotation of Web sources is essential for applying AI techniques on a large and successful scale”.

In this section I examine some of the existing Web standards and practices for the structure and semantics of scholarly information with focus on narrative documents and observational data. Scholarly documents with the characteristics of a continuous prose, such as those of research articles typically includes factual information about an investigation with supporting provenance-level information and references to underlying data. Annotations in the general sense can be similar to articles, however they are generally intended to encapsulate an indirection where target resources are associated with annotation activities with motivations, eg. commenting, bookmarking, or classifying. Notifications for scholarly activities serve as a way to inform a target of interest about events in general. I also examine representation and publication of statistical, experimental, or observational data on the Web following the Linked Data design principles.

One goal here is to identify and apply patterns so that content creators can register different units of information at varying semantic granularity. Ultimately, I seek solutions that embrace interoperability and reusability, device and medium independence, as well as favourable for accessibility and archiving.

3.2 Structure and Semantics

The structure of a scholarly journal article has been mostly conserved during the evolution of scientific communication. A typical scholarly article may include a sequence of sections, usually an abstract; introduction and background relevant to earlier work; methods used to justify experiments; research results; discussion and conclusions; and a list of references. Different forms of research eg. original, scientific, artistic, or research in the humanities, may have their own discipline-centric organisational structures for the journal article. For example, *Introduction, Methods, Results, and Discussion* (IMRAD) is a widely adopted standardised structure in original research articles as per survey, *The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey* [218], Sollaci, 2004. In the medical domain, *structured abstracts* are used to help with rapid comprehension by having distinct labelled sections, eg. introduction, objects, methods, results, conclusions, within the abstract of the article.

This section covers semantic structure of the narrative aspects scholarly communication. This can enhance *both* human- and machine-readability of information, in particular, representations of units irrespective of how they are produced.

Non-narrative units can also be semantically represented to enable easier reuse and interpretation; this is covered in more detail in Linked Statistics and *Interactive Linked Statistics*.

3.2.1 Units of Communication

Researchers exchange “units of communication” like scientific literature, annotation, datasets, workflows, argumentation, provenance, citation, and software. Knowledge may be represented or presented dynamically, or as a compilation of various independent units (compound). Semantic structure may be embedded in narrative or prose-based scholarly communication as well as used to enhance non-narrative units like experimental results or datasets. Narrative and non-narrative units may in turn include or refer to each other.

While Web resources and their representations can be composed of different kinds of hypermedia, I focus on those that are generally referred to as “documents”. For instance, articles, annotations, notifications, and profiles when combined can cover a wide range of use cases with respect to units of communication. Each instantiation share common characteristics in that they can be both human- and machine-readable – forming a particular “knowledge graph”.

Article

An article in the most general sense is a domain-agnostic unit of information encoded in a document. Research contributions including, manuscripts, reports, assessments, technical specifications, news, social media posts and slideshows are some examples for different kinds of articles.

Annotation

Annotations include information that is generally about an association between resources with different intentions. For example, assessing, commenting, liking, bookmarking, describing, linking, are some of the motivations to annotate an article.

Notification

Notifications generally express actor activities. For example, announcements about a scientific article or parts within; a quality assessment of some literature; reports of observations; annotations or social activities.

Profiles

Actors have online profiles where they generally describe themselves, refer to their contacts (address books) or curriculum vitae.

3.2.2 Human- and Machine-Readable Information

A *human-readable format* (or medium) is a representation of information that enables humans to read, edit, or act upon. A *machine-readable format* entails that information can be effectively stored and processed by a machine. Both approaches are equipped with *affordances* that can be used by respective agents.

Declarative programming paradigm: Declarative programming is a style of building the structure and expressing the semantics of programs, ie. *what* it should accomplish. The imperative programming style on the other hand is about *how* the program should execute and change the program state. The declarative approach tends to be short, easy to understand, independent of implementation, less likely to contain errors as well as to easily correct, and tractable. For example, many domain-specific markup languages, such as HTML and XML-family (XSLT, SVG, MathML), and CSS are declarative. HTML simply tells the consuming agent – like a Web browser – what should appear as part of a Webpage. HTML's readability and ease of interpretation played a role in its adoption – anyone that wanted to express some information in a Webpage can "view source" to learn without any further steps.

RDF as the language: HTML is a prominent media form to publish on the Web. While this is sufficient to cover various use cases in a scholarly information space, it is limited in the sense that the granularity of machine-readable content is based on the classic hypertext model, ie. ultimately a relationship with loose semantics between documents or their parts. It is considered to be limited in terms of capturing domain-specific knowledge. This is in contrast to using RDF as the language to communicate knowledge about arbitrary things at different levels of abstraction. Atomic statements about scientific information and scholarly activities can be expressed, as well as each component of a statement being universally identifiable. The language enables the information to be serialised using different syntaxes eg. HTML+RDFa, Turtle. Perhaps most importantly, the underlying data is intended to be manipulated as a *graph* of things, where its syntactical representations remaining isomorphic across serialisations.

Human- and machine-readable units: As indicated earlier, the high-level units of communication that I am examining are primarily in the form of prose which may be accompanied with or linked to supplemental (meta)data eg. statistical Linked Data. With this premise, I emphasise on the point that the underlying information is intended for *both* humans and machines, where each can create and interact with the information through applicable and desired interfaces.

Content negotiation: The HTTP *Content Negotiation* [219] (RFC 7231) mechanism can be used to serve a possible representation of a resource that the client prefers from a URI. The decision algorithm may be based on different dimensions of content negotiation, eg. media type, character set, encoding, language, time. All things equal, here I consider having simplicity in the design as a desired quality for a server requirement, in order to make data available for the purpose of read-write operations that a client can perform. For instance, to what extent can we serve a resource representation "as is" without having to perform media type conversions in order to satisfy both human and machine consumers? By raising this, it is neither the case that a one-size-fits-all solution is required or desirable.

RDFa: From the available RDF syntaxes, this line of reasoning brings us to encapsulating information using *RDFa 1.1* [220] – attribute-level extension – inside host languages like (X)HTML, and various XML-family languages eg. SVG, MathML – language mixing. What makes *RDFa in HTML* [221], for instance, an attractive combination is that it maps information patterns in RDF which is ideal for enhanced machine processing, while retaining the same document that is interpretable by humans. Essentially, interoperable information exchange and reuse by machines is intended to work over the RDF graph that is expressed and encoded with RDFa. Here HTML merely acts as the container to encapsulate the underlying information and to offer presentations and interactions (in regardless of the RDF syntax embedded in HTML). RDFa specifies only a syntax and allows independently defined specifications (like vocabularies) to be used. RDFa defines a single, non-domain-specific syntax, used alongside a hosting language's syntax, so that fragments of information can be consistently interpretable. Existing HTTP servers can virtually serve HTML content without additional server or client-side modules or application logic. Ergo, RDFa in HTML manifests a low barrier to make human- and machine-readable information available from a single URL. The W3C TAG Finding, *The Self-Describing Web* [222], 2009, also posits RDFa in HTML as a good practice: "To integrate HTML information into the self-describing Semantic Web, use RDFa." Finally, W3C *RDFa Use Cases: Scenarios for Embedding RDF in HTML* [223] describes the "Augmented Browsing for Scientists" use case where actors can add RDFa to their articles to indicate the scientific components by reusing the existing vocabularies defined by the scientific community.

> **Note**
>
> The initial motivation and development on encoding "RDF in HTML" started in the late 1990s and continued into early 2000, and eventually became the *RDF/A Syntax* [224] – currently known as RDFa – in 2004. The goal was to express RDF by using markup languages' attribute mechanisms alongside existing content so that documents can be more machine-readable. There has been a number of *RDF in HTML: Approaches* [225] which helped shape the initial version of RDFa, the surviving approach i) retained the expressiveness of the RDF language ii) and extending host languages (like HTML, SVG, MathML) through common attributes as necessary, iii) was aligned with the original vision of the Web, iv) was built through consensus through an open Web standards body like the W3C.

Why RDFa: HTML's extensibility mechanism allows authors to embed "data blocks" using the `<script type="">` mechanism to include content in Turtle, JSON-LD. Information in `script` are hidden from human view in native HTML interfaces until supplementary processing takes place. For information extraction to take place, the consumer needs to 1) process the HTML, 2) select the `script` node, 3) parse the content in RDF. On the other hand, RDFa can be included on any HTML tag, and so all human-visible content in HTML can be complemented with machine-readable counterparts in the same context node. This is optimal in avoiding data duplication in the same document which happens to be the case with the other serialisations in RDF, as well as avoiding any further intervention to synchronise data across different nodes. Having a single, unambiguous, and authoritative representation of the information as human-visible and marked as RDFa conforms to the *Don't Repeat Yourself* [226] (DRY) principle. It is efficient in that it has no dependency on JavaScript, external or third-party applications to make the hidden machine-readable content be consumable from a human user interface – whereas HTML based user-agents can be expected to conform to *user interaction* [227] *loading Web pages* [228] specification. For example, a *text-based Web browsers* [229] can access and allow interactions with documents. RDFa in HTML is all around a simple but effective design pattern that is able to serve both humans and machines without additional parts or machinery. With all things equal, having all content in HTML is optimal for Web-based archives, as well as helps towards meeting the *archiving* and *awareness* functions of scholarly communication.

Content serialisations: Having the canonical representation for articles, annotations, and notifications in HTML+RDFa, still leaves the servers, if so desired, to provide alternative serialisations, eg. Turtle, and JSON-LD, depending on the content negotiation with clients. Articles are represented in HTML+RDFa so that information is usable by both humans and machine consumers while maintaining lowest requirements for publishing, eg. a single URL with full content in HTML+RDFa can be accessible from any HTTP server. No additional requirements are necessary from clients such as JavaScript support or servers with additional RDF based content negotiation.

Separation of concerns: By adopting the *progressive enhancement* [230] strategy as described in *Progressive Enhancement and the Future of Web Design* [231], Champeon and Nick, 2003, for the structural (HTML+RDFa), presentational (CSS), and behavioural (JavaScript) layers, we can allow content and base functionality to be accessible through different media and devices. Ultimately, the unit of communication represented in HTML+RDFa can be accessible – readable – for both humans and machines without requiring any CSS or JavaScript. This approach is considered to cover the base requirement of making the content available with lowest server requirements, ie. practically, today any HTTP server that can serve HTML and a client that can consume HTML.

Affordance: Hypertext has perceivable affordances in that both humans and machines can choose to act. For example, when a Web document is presented with a hyperlink, its users have the option to follow the link, if they so desire, by pressing a key on their keyboard, clicking on it with their pointing-device, or with voice activation and so on. Hovering a link with the cursor can trigger the Web browser to display the full target URL in the status bar of the application, and hence informing the user on what lies ahead before engaging further. If the user is already familiar with the link (visited earlier) or uninterested, they can skip. The interaction with the hypertext is in essence non-linear, ie. we can follow the links as far as we are able to, or interested, and then come back to where we started. Similarly, machines, like Web-crawlers are able to perform exactly the same process. Moreover, hyperlinks marked up with specific relations to their target reference can signal information that can effectively induce unique interactions. In that respect, RDFa in HTML for instance can help enrich and enable the structure for different interaction possibilities, for example though runtime JavaScript or the Web browser’s built-in understanding of the underlying actions. With respect to scholarly referencing, hypertext can not only reference other resources – as typical to print-based expressions – but it can also link.

Units with unique identifiers: Any *thing* that is deemed to be of importance can be identified on the Web with a URI as per *Axiom 0: Universality 1*. Having globally unique identifiers for all sorts of objects at fine granularity facilitates precision for interlinked knowledge. We have the means to guide both human and machine users to better discover and exchange scientific and scholarly objects. Thereby making it possible to fulfill the *registration* function for virtually anything.

Markup patterns: HTML has several *extensibility mechanisms* [232], eg. `class`, `data-*`, `rel` attributes, and `meta`, `script` elements to support vendor-neutral markup patterns. In addition to HTML’s *common infrastructure* [233], it is possible for documents to signal formal grammars, eg. HTML profiles and XML schemas, to announce the structure of documents such that the machine-processing instructions are well-defined; predetermined and fixed. In practice, grammars are used to help extract markup patterns according to specified definitions. Although profiles are no longer supported in HTML5, *DroppedAttributeProfile* [234], W3C, HTML Working Group, 2010, they were originally intended to assign additional meaning that would otherwise remain local (private) to the document. For instance, *Gleaning Resource Descriptions from Dialects of Languages* [235] (GRDDL) provided a way for custom *XSL Transformations* [236] (XSLT) to handle out-of-band transformation workflows, and to interpret the structure and semantics in the original document.

Formal grammars: While HTML markup with formal grammars has the quality of being efficient for cooperating systems, it comes at the cost of wider interoperability. That is, if an application is designed to *only* work with a particular flavour or subset of HTML, then it would not be able to work with arbitrary HTML documents on the Web. Hence, having profiles or grammars in and itself not an issue, however its *effects* are likely to be that they are less interoperable than they can be. This is in contrast to applications safely ignoring information patterns they do not understand or interested in using, even if they have their preferred internal grammars or HTML templates. From the consuming side, if the application’s information *reuse* is dependent on HTML a specific schema or tree structure, effectively locks the application into coping with those patterns, and ultimately less flexible about handling expressions that the application is unfamiliar with. The effect is a constrained or a closed information system.

3.2.3 Vocabularies

Here I describe a wide range of well-known Semantic Web vocabularies that can be used to model and describe units of communication. It is not intended to be an exhaustive list but to provide an overview on the kind of things that could be potentially described, anywhere from publishing articles, scholarly activities, to datasets holding scientific measurements.

General purpose: *DCMI Metadata Terms* [237] is used to describe digital and physical resources along with metadata on provenance and rights. *schema.org* [238] is composed of a common set of schemas that occurs in Web pages. These vocabularies are widely used on the Web and are generally domain agnostic, making them suitable to describe document-level concepts with respect to scholarly units.

Publishing and referencing: *Bibliographic Ontology Specification* [239] and *Semantic Publishing and Referencing Ontologies* [240] (SPAR) cover a wide range of concepts in the publishing domain while being agnostic about the types of content, eg. “document description, bibliographic resource identifiers, types of citations and related contexts, bibliographic references, document parts and status, agents’ roles and contributions, bibliometric data and workflow processes”. *The SPAR Ontologies* [241], Peroni, 2018, describes ontologies for bibliographic resources and their parts (*FaBiO* [242], *FRBR-DL* [243], *DoCO* [244], *DEO* [245], *DataCite* [246]); citations of scholarly resources (*CiTO* [247], *BiRO* [248], *C4O* [249]); publishing workflow (*PRO* [250], *PSO* [251], *PWO* [252], *SCoRO* [253], *FRAPO* [254]); metrics and statistics for bibliographic resources (*BiDO* [255], *FiveStars* [256]).

Knowledge Organisation: *Simple Knowledge Organization System* [257] (SKOS) defines a common data model for structures like thesauri, taxonomies, classification schemes and subject heading systems in library and information sciences. SKOS is suitable to bridge different scientific communities to organise concepts by providing definitions, links, and possible mappings across collections.

Participation: Actors are integral to social Web activities and the four functions of scholarly communication. Generally speaking, agent – human or machine – profiles can be described with the *Friend of a Friend* [258] (FOAF) vocabulary to have a combination of what or who they are, what they have created, who they know, their memberships, and so forth. *vCard Ontology - for describing People and Organizations* [259] describes the mapping of the vCard specification to RDF/OWL in order to allow compatible representations for personal data interchange. *SIOC Core Ontology Specification* [260] provides concepts to describe information from online communities like weblogs, message boards, wikis, etc., as well as connections between their content items. *Activity Vocabulary* [261] provides specific activity structures and types for objects and links (eg. article, event, place, mention), actors (eg. application, group, person, service) and for past, current or future social activities (eg. announce, create, like, invite, question). While the core vocabulary covers a wide range of activities, it can be extended to cover domain-specific scientific activities. The *Web Annotation Vocabulary* [262] underpins the annotation model and is used to express information about a set of connected resources, typically conveying some sort of a relationship where the “body” of a resource is about the “target” resource. *Embedding Web Annotations in HTML* [263] describes and illustrates potential approaches for including annotations within HTML documents using current specifications like RDFa or JSON-LD. *Selectors and States* [264] describes the use of annotation Selectors as URI fragment identifiers relying on the formal specification and the semantics in the data model. Any resources can advertise a location where it can receive *Linked Data Notifications* [265] for social activities eg. invitation to review an article, annotation being created. I describe LDN in detail later serving as one of the foundational components towards a decentralised and social Web. In order to indicate an agent’s cognitive pattern within contexts, their temporal dynamics and their origins, *The Cognitive Characteristics Ontology* [266] can be used eg. a human language competence.

Rights and Responsibilities: *Creative Commons Rights Expression Language* [267] lets licenses to be described – jurisdiction, permissions, requirements, license and work properties – and having them attached to digital works. The *Creative Commons licenses* [268] is intended to enable sharing and reuse of creative works and knowledge, by specifying how something – like scientific and scholarly resources – may be copied, distributed, edited, remixed, and built upon. The *Open Digital Rights Language* [269] (ODRL) makes it possible to express permissions and obligations about the usage of content and services. *ODRL Vocabulary & Expression* [270] describes how to encode such policies.

Access Control and Certificates: The *Web Access Control* [271] (WAC) mechanism with the *Access Control List* [272] (ACL) vocabulary describes authorization policies in particular to access and operations – read, write, append, control – that can be done on Web resources by agents or groups. *The Cert Ontology 1.0* [273] can be used to indicate digital certificate information for agents.

Provenance: *The PROV Ontology* [274] (PROV-O) can be used "to represent and interchange provenance information generated in different systems and under different contexts". It can be also be "specialized for modeling application-specific provenance details in a variety of domains", like for instance *The OPMW-PROV Ontology* [275] to describe workflow traces and their templates, which also extends the *P-Plan Ontology* [276] that is designed to represent scientific processes. The *Wf4Ever Research Object Model* [277] is used to describe workflow-centric *Research Objects*: aggregations of resources along with annotations on those resources relating to scientific workflows. *Verifiable Claims Data Model 1.0* [278] is aimed at expressing a claim; statement made by an entity about a subject, on the Web in a way that is cryptographically secure, privacy respecting, and automatically verifiable. *The Memento terms vocabulary* [279] contains a set of terms for the *Memento Framework* to facilitate obtaining representation of resource states. *Link Relation Types for Simple Version Navigation between Web Resources* [280] (RFC 5829) defines link relation types to related to navigating between versioned Web resources, such as latest and predecessor versions and working copies.

Design Lifecycle: *Design Intent Ontology* [281] can be used to "to capture the knowledge generated during various phases of the overall design lifecycle" like for instance the artifacts about software or technical specification requirements, issues, solutions, justifications and evidence.

Datasets: For multi-dimensional statistical data, the *RDF Data Cube vocabulary* [282] can be used to both model the structure of the hypercube as well as aggregate data which uses its structure definition. The structure is typically an arbitrary number of components: dimensions that helps to identify an observation, and one or more measurements about the phenomenon being observed. The *DDI-RDF Discovery Vocabulary* [283] (Disco) enables publishing and discovery of metadata about datasets like research and survey data. Disco is generally concerned with microdata descriptions that is on a lower level of aggregation than the RDF Data Cube. It is not intended to represent the data itself, but only its structure, where the record-level raw data in its original format is only referenced. Disco also enables a way to describe the aggregation methods that was used to collect the data. The *Semantic Sensor Network Ontology* [284] (SSN) can be used for "describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators". *Describing Linked Datasets with the VoID Vocabulary* [285] (VoID) is concerned with metadata about RDF datasets to help with deployment, discovery of data and services, cataloguing and archiving of RDF datasets, as well as linksets between datasets. *Evaluation and Report Language* [286] (EARL) can be used to describe the test results and facilitate their exchange between applications. It provides reusable terms for generic quality assurance and validation purposes.

Scientific expressions: *Semanticscience Integrated Ontology* [287] can be used to describe scientific experiments, for example including their procedure, hypothesis, objectives, study design, analysis, and observations. *STATO: the statistical methods ontology* [288] covers statistical methods, tests, conditions of application, results, as well as aspects of experimental design and descriptions. *Nanopublication Guidelines* [289] helps to publish scientific assertions that can be uniquely identified with associated context, attributed to their author, as well as help to preserve associated provenance. *Micropublications* [290] can be used to formalise the arguments and evidence in scientific publications.

Software Projects: *Description of a Project* [291] (DOAP) can be used to describe (open source) software projects like their repositories, bug databases, maintainers, technical specifications they implement, programming languages they use or the platform they run on.

3.2.4 Accessibility, Usability, and Inclusion

We acknowledge the diversity of people using the Web, not only the actors in scholarly communication, but anyone that may create or reuse information. Our aim is to have inclusive designs for wide range of people and their abilities. I outline and borrow key initiatives and solutions on content accessibility, accessible applications, authoring tools, and internationalisation. I refer to “accessibility” (a11y) in the widest sense that any agent; human, machine, or other can effectively access information and participate.

Web Content Accessibility: The accessibility of units of communications can also be seen as an aim for all-inclusive design that is usable by humans with widest possible range of abilities and situations. For this, I defer to W3C’s *Web Content Accessibility Guidelines* [292] (WCAG) to cover an array of recommendations to make content accessible to a wider range of people regardless of any disability, limitation, or sensitivity, through different media and interfaces. The goal is to provide a reliably mutually consistent expression of the content. For human-centric scholarly communication to thrive, units of communication should be targeted to aim to meet highest level of conformance criteria. W3C WCAG 2.1 provides a range of guidelines that can be adopted for better content accessibility:

> **Perceivable**
> Information and user interface components must be presentable to users in ways they can perceive.
>
> **Operable**
> User interface components and navigation must be operable.
>
> **Understandable**
> Information and the operation of user interface must be understandable.
>
> **Robust**
> Content must be robust enough that it can be interpreted by a wide variety of user agents, including assistive technologies.
>
> *Web Content Accessibility Guidelines (WCAG) 2.1* [293], W3C, 2018

Accessible Applications: The W3C *Accessible Rich Internet Applications* [294] (WAI-ARIA) recommendation provides an ontology to help assistive technologies to provide a consistent user interface and understanding of the objects. Host languages like HTML and SVG can include the ARIA ontology: “roles” to alert the purpose of an element; “properties” to indicate an elements relationship to other things; “states” to indicate what the element is doing, as well as alerting users about changes in state. ARIA helps to inform consuming assistive technologies such as screen magnifiers, screen readers, text-to-speech software, speech recognition software, alternate input technologies, and alternate pointing devices to create a particular accessibility tree, and to adapt the user interface to a form that works for the person, eg. a screen-reader can read the menu items, or selected option is receiving keyboard focus. WAI-ARIA is extensible and has a number of accessible API mappings. *Digital Publishing WAI-ARIA Module 1.0* [295] (DPUB-ARIA) specialises WAI-ARIA’s ontology to enable semantic navigation, styling and interactive features in context of digital publishing. *WAI-ARIA Graphics Module* [296] is another extension of WAI-ARIA aimed to support structured graphics such as charts, graphs, technical drawings and scientific diagrams, to assistive technologies in order improve accessibility of graphics or diagrams through detailed annotations.

Authoring Tool Accessibility: In order to contribute to the proliferation of Web content that is accessible to a broad range of people, authoring tools would need to be accessible as well. The W3C *Authoring Tool Accessibility Guidelines* [297] (ATAG) is a recommendation to assist with the design of authoring tools. The guidelines have a success criteria covering two areas:

> - Make the authoring tool user interface accessible
> - Support the production of accessible content
>
> *ATAG 2.0 Guidelines* [298], W3C, 2015

User Agent Accessibility Guidelines: To support the general principles for the development of accessible user agents, ie. “any software that retrieves, renders and facilitates end-user interaction with web content”, the W3C *User Agent Accessibility Guidelines* [299] (UUAG) specifies three layers of guidance for developers to integrate:

- overall principles: perceivable, operable, understandable, programmatic access, specifications and conventions;
- general guidelines to provide a framework to make user agents more accessible to users with disabilities;
- and success criteria in order to test conformance

Internationalization and Localization: Adaptability of content and software to the needs of target audiences helps towards accessibility. For example, the mechanisms to cater information and interfaces so that people from any culture, region, or language preference can participate better. The W3C *internationalization* [300] (i18n) and *localization* [301] (l10n) initiatives and best practices helps towards this end. Internationalization refers to the design and development of mechanisms so that adaptable localization can take place in users’ environment.

Internationalization of Web content and technologies is intended to make it possible for people to use them with different languages, scripts, and cultures. For example, content authors can:

- include links to navigate to different languages of the content;
- declare the base language of a document, indicate multiple languages and their directional flow – to help with translations;
- use Unicode character encoding, eg. UTF-8, in data forms and text to ensure correct effects;
- check and minimise inappropriate cultural bias, and improve translatability;
- restrict markup use to structure and semantics.

The localization of content would mean that user’s preferred (or acceptable) visual design can be presented. For instance, human-readable numeric, date and time formats can be adapted to what we are familiar with; symbols, icons, and colours can be anywhere from what’s culturally acceptable, familiar or comfortable for us to use; text and graphics can be normalised or transformed to minimise misinterpretation.

3.2.5 Archivability

Archivability of Web resources “refers to the ease with which the content, structure, functionality, and front-end presentation(s) of a website can be preserved and later re-presented, using contemporary web archiving tools”.

In *CLEAR: a credible method to evaluate website archivability* [302], Banos, 2013 posit that Web archivability can be measured by five facets: “accessibility” as network access to content; “standards compliance” in terms of information using common open formats and specifications; “cohesion” as information being independent of external support; the level of “metadata” that is available alongside the content; and, server’s “performance”. Such facets can be used to quantify website archivability.

In the Web development community, a rule of thumb to improve archivability of Web content is to aim for general standards compliance and content accessibility. In the case of standards compliance, well-formed and valid markup ensures internal integrity of Web documents. As for better accessibility, incorporating WCAG where applicable can help user-agents, including text-only Web browsers, Web crawlers, and other agents with minimal capabilities to parse and render content. With respect to traditional Web documents, the information that is visible, observable, or human-readable from HTML is considered to set the lowest barrier to obtain content. As CSS and JavaScript are concerned with presentation and behavioural layers respectively, they play a secondary role towards archiving of "content".

The impact of JavaScript on archivability [303], Brunelle, 2015, study the quality of archived Web resources (based on URIs from Twitter and Archive-It) and report that JavaScript-driven content played a significant role in the reduction of preservation and recall of content. For example, resources referring to external scripts may load unsuccessfully given that their own independent availability. Authors refer to the notion of “deferred representations” where the final state of a resource is when all code and events have finished executing. Hence, representations rely on JavaScript to fully render do not serve well for archival crawlers that are unequipped to handle JavaScript. On the other hand, certain archival services include a feature known as *headless browser* that is capable of processing the resource similar to common GUI-based Web browsers with JavaScript enabled, and thus capturing the final state of the representation once all scripts are executed. This too however may be problematic for content that dynamically changes based on the instructions of the script. For example, a script that presents temporal content may be different at each time the archived resource is recalled.

While evaluation services like *ArchiveReady* [304] check for website archivability based on the five metrics, it is not able to determine the *accuracy* of intended information. As the content in the HTML representation and the state of rendered content (after script execution) may differ, their delta can be compared to what is intended for archiving. In *Significance is in the Eye of the Stakeholder* [305], Dappert, 2009, contend that a stakeholder can specify the characteristics of resource representations and the conditions for preservation. For human-readable documents, it can be reasoned that the rendered content is what is intended for archiving by the publisher.

3.3 Existing Markup Patterns

There are plethora of approaches and developments to representing scholarly information with markup languages. Here I focus on a few and describe their characteristics.

Journal Article Tag Suite [306] (JATS) is an XML format used to describe content and metadata of research and non-research articles eg. including reviews, editorials, instructions to authors. It can be used by publishers and archives to interchange scholarly content in a uniform way. The intent of its Tag Suite is to prescribe and to preserve the structure and semantics of the original content independent of the presentation. JATS includes three Tag Sets: journal archive and interchange, journal publishing, and article authoring. In the case of peer reviewed scholarly articles that are intended to be archived, the publication may also include information about the conference the research was presented at, data about the publication and journal, document history, license, funding, and various kinds of annotations. Given that JATS is XML-based, different manifestations of the original content are generated through transformations eg. as a print-version, HTML, as well as RDF as described in *From Markup to Linked Data: Mapping NISO JATS v1.0 to RDF using the SPAR (Semantic Publishing and Referencing) Ontologies* [307].

Text Encoding Initiative [308] (TEI) has guidelines to “define and document a markup language for representing the structural, renditional, and conceptual features of texts.” TEI is similar to JATS in that it aims to improve understanding of text made explicit with XML, as well as for information interchange, integration, and preservation. Its encoding methods are mainly aimed at machine-readable texts in academic, linguistic, literary and technical documents. Its document structure is similar to JATS in that it has a front (header), body (article), and back pattern, and including a generalistic approach to coping with text with any size, complexity, writing system, language, date, or media. For instance, the specification helps with information identification (eg. page numbers), structural divisions, pictures and diagrams with captions; different writing modes (eg. prose, verse); each with formal structural units (eg. paragraphs, lists, stanzas); with textual distinctions (eg. titles, names, quotations); as well as metatextual indications (eg. corrections, annotations, revisions).

Electronic Publication [309] (EPUB) is implemented as a distribution and interchange format based on XML and Web Standards. EPUB is delivered as an archive file (ZIP) packaging XHTML, CSS SVG, along with supporting media files. The *EPUB Content Documents* [310] defines profiles – inheriting XHTML and extensions to the underlying document model – for the content to be used in the context of EPUB Publications, and *EPUB Open Container Format* defines the file format and processing model for the single-file that encapsulates all material. Software on devices complying with EPUB can read and interact with its content. It retains the reflowable features of XHTML and CSS in that content presentation can be adapted by a consuming device. The specification allows embedding of RDFa attributes.

Research Articles in Simplified HTML [311] (RASH) is a format accompanied with the RASH Framework "for writing HTML-based scholarly papers". RASH aims at sharing scholarly documents through the Web while working within the existing publishing workflow, and is expected to be produced from MS Word, ODT and LaTeX sources. RASH has been developed as a formal grammar using RelaxNG which includes a subset of elements – 32 – from HTML5. With the exception of the `script` element, it is compiled of printable elements. Its grammar facilitates RASH documents to run against a markup validator, to convert between sources, rendering for Web and print media, and to extract additional semantics. Its rationale for representing human-visible and machine-readable information in HTML varies eg. title, authors, keywords of an article are used as hidden metadata, and require additional processing to make it human-visible in the Web browser, as such a dependency on JavaScript support and being enabled. The resulting grammar favours simplicity towards authoring over expressing accurate or appropriate semantics eg. h1 for all heading levels is used, as opposed to HTML5's required heaving levels (hX). The format allows DPUB-ARIA and RDFa annotations, as well as embedding JSON-LD, Turtle, and RDF/XML data islands with the `script` element.

W3C Scholarly HTML Community Group [312] has its mission to build a common, open format for the exchange of scholarly information. Its specification on *Scholarly HTML* [313] proposes a vernacular document format aimed at encoding *scholarly articles* built on open standards, as well as having them compatible with off-the-self Web browsers. One of its high-level goals is to enable structured metadata as well as semantic enrichments, accessibility, internationalisation. It applies the notion of "semantic overlays" for its markup patterns with focus on role-based semantics as defined in WAI-ARIA, DPUB-ARIA, and semantic representations using RDFa. The – work-in-progress – specification explains how concepts for people and organisation, article semantics, schema roles, actions, citations, document rights, and so forth can be marked using appropriate vocabularies, using the schema.org vocabulary as base. In order to simplify authoring, where applicable, it is encouraged to use RDFa markup patterns where the depth (level of nesting in the DOM tree) for the RDF statements would be relatively flat. One consequences of this rule of thumb may be that there can be duplicate information or parts of the information being human-visible whereas some others would be intended only for machine use.

Summary of Characteristics: For historical and technical reasons, here I summarise some of their (shared) characteristics:

- With the exception of *Scholarly HTML*, the markup methods first cater to printability (as opposed to being Web-centric) or require transformations to other formats in order to be viewed and interacted further.
- Data representations are prescriptive in that they are coupled with specific applications or workflows.
- The generation or use of markup is dependent on specific environments for use or interchange, media or devices.
- The structure of the information is scoped to general-purpose scholarly documents, as opposed to for example encapsulating social (scholarly) Web activities.
- The structure and semantics of information is fundamentally based on HTML or XML (as opposed to the RDF language).
- In some cases, there is information duplicity as human and machine consumers are served with different data.

- There are post-processing dependencies (eg. via JavaScript execution) in order to turn hidden metadata into visible and readable by humans.

3.4 Linked Statistics

Research data in the form of observational or experimental data follows a different structure than narrative content – which is a common characteristic of a research article, annotation, notification or a profile. Structured data may be (semi-)automatically aggregated through instruments, stored, and viewed in different ways.

As pointed out in *Statistical Linked Dataspaces* [314], 2012, Capadisli, what linked statistics provide, and in fact enable, are queries across datasets: Given that the dimension concepts are interlinked, one can learn from a certain observation's dimension value, and enable the automation of cross-dataset queries. Moreover, domain-specific user interfaces can be built based on federated queries as discussed in *Linked Statistical Data Analysis* in *Sparqlines: SPARQL to Sparkline*.

3.4.1 Linked SDMX Data

As statistical data is inherently highly structured and comes with rich metadata (in form of code lists, data cubes etc.), the application of the Linked Data design patterns is inherently suitable for representation and reuse on the Web. There exists no simple, standardised or (semi)automated way to transform statistical data into Linked Data since the raw data comes in different shapes and forms.

While access to statistical data in the public sector has increased in recent years, a range of technical challenges makes it difficult for data consumers to take advantage at ease. These are particularly related to the following two areas:

- Automation of data transformation of data from high profile statistical organizations.
- Minimisation of third-party interpretation of the source data and metadata and lossless transformations.

Statistical Data and Metadata eXchange [315] (SDMX) is an ISO standard which provides the possibility to consistently carry out data flows between publishers and consumers. SDMX-ML (using XML syntax) is considered to be the gold standard for expressing statistical data. It has a highly structured mechanism to represent statistical observations, classifications, and data structures. Given that SDMX is arguably the most widely used standard for statistical data exchange (among statistical agencies and others), a great amount of statistical data about our societies is yet to be discoverable and identifiable through the Open Web Platform. Development teams often face low-level repetitive data management tasks to deal with someone else's data. Within the context of Linked Data, one aspect is to transform this raw statistical data eg. SDMX-ML, into an RDF representation in order to be able to use publicly available data in a uniform way.

Linked SDMX Data [316], Capadisli, 2013, ie. my colleagues and I, contribute as follows:

- an approach for transforming SDMX-ML based on XSLT 2.0 templates and the implementation which transforms SDMX-ML data to RDF/XML
- The extract, transform, load ETL process: a) retrieval of SDMX-ML data from APIs, b) their transformations to RDF/XML, and c) enrichment, d) storage in an RDF data store, and e) publicly accessible Linked Data publication based on the statistical data from the following agencies:
  - *Australian Bureau of Statistics* [317] (ABS)
  - *Swiss Federal Statistical Office* [318] (BFS)
  - *Bank for International Settlements* [319] (BIS)
  - *European Central Bank* [320] (ECB)
  - *Food and Agriculture Organization of the United Nations* [321] (FAO)
  - *Federal Reserve Board* [322] (FRB)

- *International Monetary Fund* [323] (IMF)
- *Organisation for Economic Co-operation and Development* [324] (OECD)
- *UNESCO Institute for Statistics (UIS)* [325]

Each Linked Statistical Data (LSD) endpoint:

- modeled using RDF, RDFS, XSD, OWL, XSD, DC Terms for general purpose descriptions; the RDF Data Cube vocabulary to describe multi-dimensional statistical data; PROV-O is used for provenance coverage; SKOS and XKOS to cover concepts, concept schemes and their relationships to one another;
- cross dataset and concept scheme interlinking;
- uniquely identifiable provenance level information at retrieval, transformation, and post-processing phases;
- licensed under CC0 1.0 Universal Public Domain Dedication;
- follows a URI template, where the URIs are versioned with respect to the original data counterparts
- data dumps;
- queryable SPARQL endpoint;

The results are part of the *270a.info* [326] LSD Cloud. The LSD Cloud is also part of the broader *Linked Open Data Cloud* [327] (LOD) Cloud.

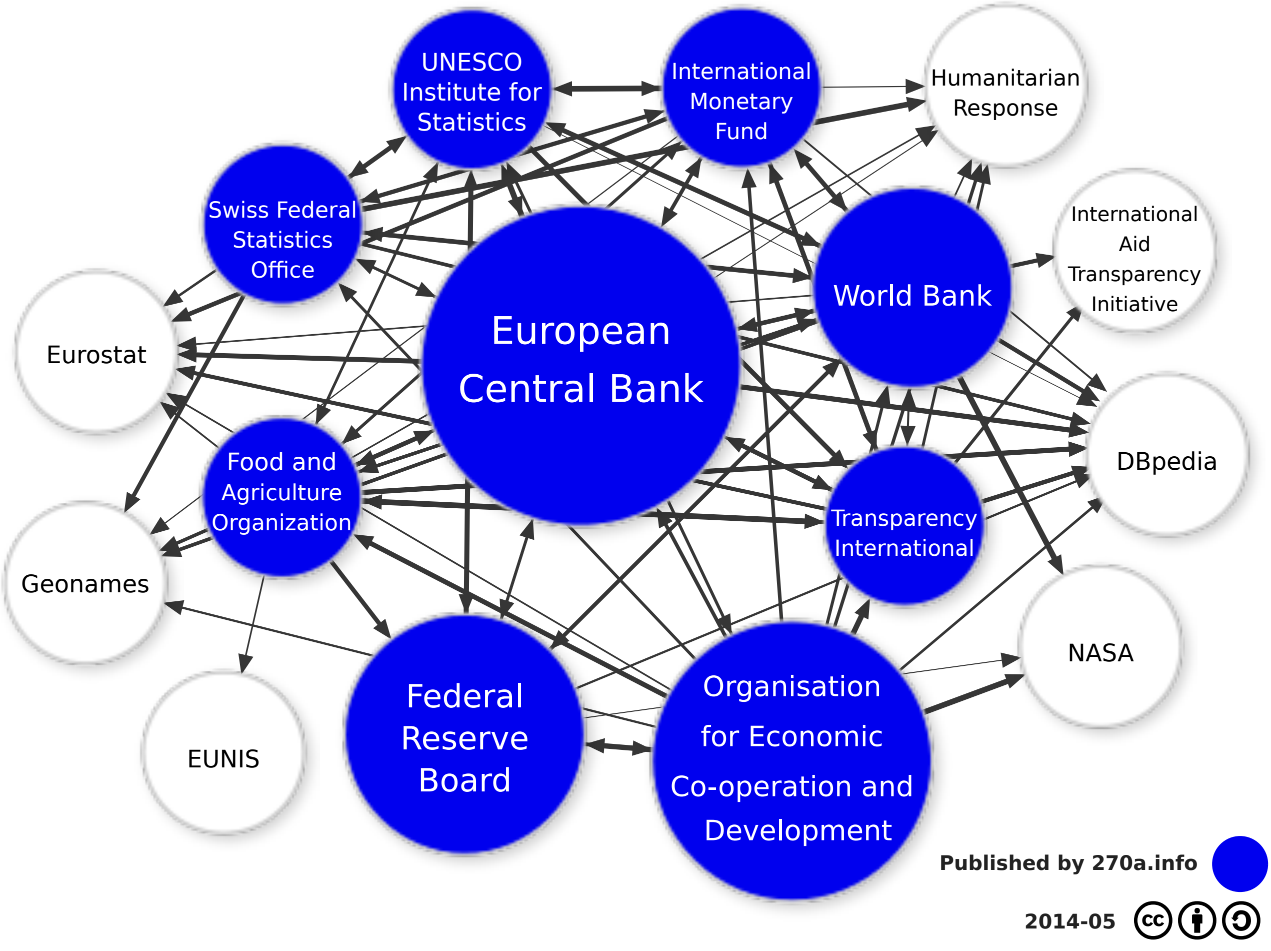

UNESCO Institute for Statistics
International Monetary Fund
Humanitarian Response
Swiss Federal Statistics Office
World Bank
International Aid Transparency Initiative
Eurostat
European Central Bank
DBpedia
Food and Agriculture Organization
Transparency International
Geonames
NASA
EUNIS
Federal Reserve Board
Organisation for Economic Co-operation and Development
Published by 270a.info
2014-05

Figure 2. Interlinks of Linked Statistical Data endpoints at 270a.info.

What the LSD Cloud enables is that highly structured and interlinked statistical data and classifications that can be queried and used by decentralised applications.

3.4.2 Linked Statistical Data Analysis

In *Linked Statistical Data Analysis* [328], Capadisli, 2013, puts forward an approach based on decentralised (and federated) structured queries to retrieve statistical data from various SPARQL endpoints, conducting various data analyses (eg. regression analysis), and providing the results of the analysis as Linked Data back to the user. The system demonstrating the mechanism stores the analysis in RDF to enable future discovery and reuse eg. researchers looking up statistically significant results based on a set of indicators for a specific reference area. As a result, distributed linked statistics with accompanying provenance data can be more easily explored and analysed by interested parties.

The approach expects that the data is modelled using the RDF Data Cube vocabulary and has *Well-formed cubes* [329]. Essential checks for integrity constraints include: 1) a unique data structure definition (DSD) is used for a dataset, 2) the DSD includes a measure (value of each observation), 3) concept dimensions have code lists, and 4) codes are from the code lists.

In order to compare variables in observations across statistical datasets, there needs to be an agreement on the concepts that are being matched for in respective observations. In the case of regression analysis, the primary concern is about reference areas (ie. locations), and making sure that the comparison made for the observations from dataset_x (independent variable) and dataset_y (dependent variable) are using concepts that are interlinked (using the property `skos:exactMatch`). Practically, a concept eg. Switzerland, from at least one of the dataset's code lists should have a relation to the other dataset's concept. It ensures that there is a reliable degree of confidence that the particular concept is interchangeable or the degree in which the concepts being comparable. Hence, the measure corresponding to the phenomenon being observed, is about the same location in both datasets. To this end, concepts in the datasets were interlinked.

In order to foster trust and confidence for data consumers (human and machine), the analysis is accompanied with provenance data. The analysis includes a *Oh yeah?* [330] reference (in HTML as well as in RDF serializations) intended to guide the data consumer to a resource about the provenance activity – using the PROV-O vocabulary – about the performed analysis. These previously generated provenance activities provide links to all data sources which were used for the analysis, query construct for data aggregation, as well as metadata about the used tools, assigned license, production timestamps, and responsible agents for the generated analysis. Thus, in addition to analysis metadata, the user is able to track the data all the way back to its origins (at the statistical agencies), and reproduce or compare their results.

The proposed approach focuses on the scalability of the system with minimal human intervention. That is, new statistical dataset can be independently published while being expressed with relevant statistical vocabularies, and sufficient interlinks between concepts, then applications can discover them and perform queries in a uniform way. Hence, applications *only* need to be made aware of the location of new datasets.

⁂

Enabling Scientific Data on the Web [331], Miłowski, 2014, focuses on “enabling scientific data to exist on the Web in such a way that it can be processed both as viewable content and consumed data”. Miłowski posits that in order to enable user interactions through the Open Web Platform while having the same information machine-readable, “data must to be partitioned into “Web-sized” representations”. This can be contrasted with the approaches mentioned from earlier where the data is treated as "as is" without being structured in any particular way for applications to consume.

In this thesis I describe and exemplify multiple implementation approaches. For example, the tabular data in *Characteristics of Specifications*, *Comparison of Notification Mechanisms*, *Linked Data Notifications Implementations*, and in *Decentralised Linked Research Application*, as well as the data in *Linked Specifications, Test Suites, and Implementation Reports* are all human- and machine-readable. Whereas in *Interactive Linked Statistics*, I will describe *Sparqlines: SPARQL to Sparkline* where SPARQL queries can be executed by client-side applications in an article authoring environment to fetch live data from remote endpoints, and be represented in SVG+RDFa.

4. Decentralising Scholarly Communication

In this section, I discuss the design of socially-aware decentralised systems. The sections *Degree of Control*, *Forces and Functions in Specifications*, *The Effects and Artifacts of Autonomous Engagement* are my knowledge contributions.

4.1 Control Yourself

Throughout history, the complex connections between communication technologies and society transformed control structures of civilisations. Harold Innis and Marshall McLuhan argued that electronic media, unlike any other, compressed time and space, creating a “global village” as well as facilitating the influence and potential of centralised control.

While the Internet reinforces democratic (re)distribution of communicative power, and the Web is inherently a decentralised social system, centralisation of all forms of information exchange and various forms of censorship – state or self-imposed – still occurs. In response to the *Telecommunications Act of 1996*, John Perry Barlow voiced a fundamental societal concern in *A Declaration of the Independence of Cyberspace* [332] given the attempts of the “governments of the industrial world” to exert control over it, the builders of the cyberspace; a self-sovereign online society that is cross- border, culture, with diverse economies is ultimately an opposing force.

Over the years, large-scale centralised systems were built to collect, organise and share data in diverse sectors. The domination of centralisation is essentially due to economic and state incentives, and it has shaped the technical infrastructure: it is currently much easier and more efficient to author, manage, publish, read, and search large amounts of content using centralised services or proprietary platforms. As such the convenience that comes with centralisation is at the user’s expense where ownership and access control is compromised, and ultimately the user’s privacy and security. Moreover, from the service provider’s perspective, there is no particular incentive to build interoperable systems as that would entail losing control over data. Since such services hinder information flow; information in one service is not necessarily usable by other applications and services and vice versa, they effectively form information silos – *Decentralization: The Future of Online Social Networking* [333], 2008.

In addition to state controlled network infrastructures, information empires like *Google*, *Facebook*, *Amazon*, *Elsevier*, as well as a long tail of (social media) services, exert a great deal of influence on how information can be accessed, expressed and reused. This influence impacts society, affecting the cost of accessing information, the effort required to produce and distribute, long-standing privacy concerns (in exchange of some convenience for users), as well as how and the amount of information that is aggregated, shared and filtered.

Independence from centralised services and platforms is a necessity for ownership of published ideas, and to establish authenticity and trust. For example, *Facebook* has been accused of data scandals, bias, false information, and censorship – but rather than blaming this on any particular service or platform, we identify it as an unavoidable result of centralisation. After all, there is a continued tension between unrestricted publication rights on the one hand, and a guarantee of balanced, verified information on the other. In a fully decentralised setting, each source is filterless and responsible for its own quality and reputation, while consumers are free to selectively (dis-)trust certain sources using any mechanism they desire, for example through trusted parties. This can have both favourable and unfavourable consequences depending on actual implementations. *Google Search* or *DuckDuckGo* [334] for instance apply different filtering algorithms based on their Web crawls. Whether consciously or not, why we use a particular search engine naturally correlates with the amount of use or trust we have on that system. Similarly, as the academic peer review system acts as a layer to regulate credibility of academic contributions, we can approach the research results provided that there is sufficient provenance, accountability of the quality control processes, and able to distinguish different levels of certification given or lack of from the scholarly community.

With the presumption that the major external forces – governments, companies – are at play, the situation for individual-controlled communication is complicated and ever challenging. Given that academic research is a global activity, it is important to acknowledge that the Web enables an environment in which actors in the space can both autonomously conduct their endeavours and participate in knowledge exchange. Hence, there is a need to *re-decentralise the Web* so that it benefits individuals, the public, as well as the scholarly communication market as a whole. We need to improve decentralisation and take further steps to apply it to research activities and scholarly publishing.

The information needs of individuals, communities and societies vary on personal, local and global levels. In *Understanding Knowledge as a Commons* [335], 2011, Hess and Ostrom, and contributing authors state that scientific communities can make their discoveries and digital resources available to all which can be shared without having a subtractable characteristic like that of print. Through electronic knowledge representations, researchers can take part in reproducing and replicating others' findings; be involved in quality-control and certification processes, as well as incorporate interactions among researchers, examiners, journalists, policy makers, and the general public. Given access to the Internet and the Web, academic researchers can register their URIs, and control how they express their ideas and what they refer to, establish data storage and choose which applications to use for their social and scholarly activity.

⁂

So far I have used the term "decentralisation" liberally, but in fact it can mean different things depending on the context, community, and collection of standards and technologies that are used. From here on end, I work with the the following definition in context of the Web architecture:

> **Decentralized system**
>
> A distributed system in which multiple authorities control different components and no single authority is fully trusted by all others.
>
> *Systematizing Decentralization and Privacy: Lessons from 15 Years of Research and Deployments* [336], Troncoso, 2017

Messages and operations in centralised or distributed systems "may be managed by a single root of trust or authority". From this position "decentralized systems are a subset of distributed systems" where they "embody a complex set of relationships of trust between parties managing different aspects of the system". These systems are conceived of as a graph where the nodes correspond to the components of the system, and the edges correspond to connections between them. From a global perspective, the Web is a decentralised system, however, it can have components which are disjoint from the rest of the graph. Figure *Coupling of Centralised and Decentralised Architectures* depicts the contrast between typical centralised and decentralised architectures.

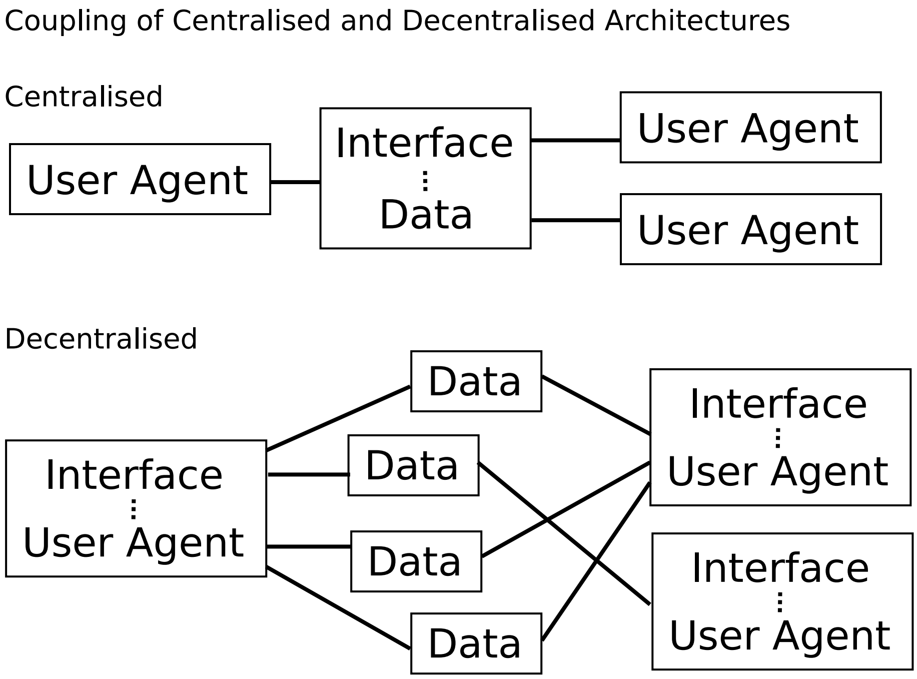


Figure 3. Typical centralised and decentralised architectures.

Troncoso et al, consider *security* and *privacy* to be integral aspects of systems. Security includes “confidentiality, integrity, authentication … availability, accountability, authorization, non-repudiation or non-equivocation”, and privacy considers “protections of users’ related data (identities, actions, etc.)”, and “formalized in terms of privacy properties (anonymity, pseudonymity, unlinkability, unobservability).”

How do we control ourselves on the modern, overly-centralised Web? I focus on the feasibility of using a set of open Web standards and technologies to address certain kinds of challenges in scholarly communication. I conclude with a focus on three areas of control (*Degree of Control*) that are at the forefront of the way (scholarly) communication is conducted on the Web and its ramifications on the rest of the sociotechnical system.

4.2 Decentralised Dataspaces

In this section I discuss existing research involving the decentralisation of dataspaces on the Web.

In *From Databases to Dataspaces: A New Abstraction for Information Management* [337], Franklin, 2005, raises a challenge of organisations “relying on large number of diverse, interrelated data sources, but having no way to manage their *dataspaces* in a convenient, or principles fashion.”. The article brings forth the notion of dataspaces as a “data co-existence” approach, with the goal of providing functionality over all data source. In such systems, there is a set of participants and relationships, while remaining sensitive to their requirements of autonomy. The authors propose the development of different “DataSpace Support Platform”s (DSSP) where interrelated services would allow data to be managed by participant systems.

In *Principles of Dataspace Systems* [338], Halevy, 2006, explores challenges that are involved in realizing DSSPs based on motivating applications eg. personal information management (PIM), scientific data management, structured queries and content on the Web. Authors conclude that “interoperability and interchangeability of multiple components is key to the success of DSSPs.”

In *Web-scale Data Integration: You can only afford to Pay As You Go* [339], Madhavan, 2007, propose the notion of a PAYGO data integration architecture with the premise that it is impossible to fully integrate vast heterogeneous collections of structured data that exists on the Web. The PAYGO principle states that a system needs to be able to incrementally evolve in its understanding of underlying data's structure, semantics, and relationships between sources.

In *A decentralized architecture for consolidating personal information ecosystems: The WebBox* [340], Van Kleek, 2012, proposes a realisation of Berners-Lee's socially aware cloud storage, in which "Web applications running on a user's devices gain privileged access a unified, user-controlled data space", where Webbox applications are decoupled from personal data spaces.

In *An Architecture of a Distributed Semantic Social Network* [341], Tramp, 2012, and *Distributed Semantic Social Networks: Architecture, Protocols and Applications* [342], Tramp, 2014, posits that the landscape of Social Web is increasingly losing its distributed nature. Tramp highlights that there is an increased use of centralised social platforms, where users are locked in to their respective platforms with not much opportunity to communicate easily with users on other platforms, as well as difficulty to relocate their social graphs and personal data. Interoperability between platforms being largely limited to proprietary APIs, and changes being at the discretion of the service provider. Users of the online services are being required to keep their – often overlapping – data individually at each location up to date in order to minimise divergent information. Tramp argues that technical solutions should empower users to regain control and ownership over their data and its use, ensure privacy policies and rules that's user-centric (as opposed to driven by commercial interest), increased data security, information extensibility based on user's needs, reliability through distribution, and freedom of communication without centralised control. The output of Tramp's research is that Semantic Web technologies can be deployed to some extant to support the structure, maintenance, and usage of federated and distributed social networking on the Web. The evaluation is based on integration use case tests, eg. *SWAT0* [343], *SWAT1* [344], formulated by the W3C *Federated Social Web Incubator Group* [345].

In *Data ownership and interoperability for a decentralized social semantic web* [346], Sambra, 2013, acknowledges the same core challenges of the modern Web as Tramp, and sets out to identify key components that would help achieve data ownership and interoperability. Sambra demonstrates how a stack of interoperable Web technologies around WebID and Linked Data can be used to build class of social Web applications, where users without centralised intermediaries can authenticate themselves, and participate in creating and exchanging data with servers that are equipped with social access controls.

In *Amber: Decoupling User Data from Web Applications* [347], Chajed, 2015, posit that "users control their own data" choosing which applications to manipulate their data as well as their ability to share data between applications and with other users. *Amber* is a proposed architecture that looks into overcoming challenges on inexpensive querying, an expressive authentication and access control system, users trusting their platform, being able to handle large volumes of data and query results, offline capabilities, and a sensible economic compatibility among its users, platform, and application developers. The results from Amber show that the architecture is most useful when data is associated with users, and operations on the data are carried out in Web browsers.

In *A Demonstration of the Solid Platform for Social Web Applications* [348], Mansour, 2016, raise the concern that well-known Social Web applications are essentially "data silos" involving design patterns where each application is custom built to store its own data, along with custom authentication and access control mechanisms. Hence, users of such systems often cannot easily switch personal data storage services, or reuse their data with similar applications. Authors demonstrate the *Solid platform* to address the challenges in decentralised authentication, data management, as well as the development of interoperable and portable Social Web applications that interact. A number of servers and client applications are implemented where it is possible to create, store, serve RDF and non-RDF based resources, as well as permitting operations via SPARQL queries for complex data retrieval and link-following. The client-side applications cover common "day-to-day" tasks eg. contacts management, event organisation, collaborative authoring, annotating, and social notifications, all meanwhile users are able to use their WebIDs and switch between applications with similar functions.

4.3 Read-Write Linked Data

Standards-based Scholarly Communication: In order to manifest a Web-centric interoperable information space, I investigate existing research, Web specifications and practices that can fulfill the forces and functions in scholarly communication. This is in contrast to an arbitrary ensemble of implementations which are not guaranteed to be interoperable with each other, alternatively may only after post-facto interoperability initiatives eg. the survey on *400+ Tools and innovations in scholarly communication* [349] for the most part observes applications implementing propriety designs. The following are a selection of open Web specifications; standards and practices that are designed with the notion of decentralisation, interoperability, and extensibility. They are based on use cases that facilitate discovery, read-write operations of interlinked Web resources.

Both Tramp and Sambra approach from the perspective that the Web has proven itself architecturally and universality, and argue to extend the decentralised architecture for information reuse and user interaction. The core approaches use the Linked Data technology stack for publishing, retrieval, and integration; decoupling services and applications from the users data so that content creators are owners and have desired rights on their data and its use; and fostering data extensibility and distribution, as well as privacy.

The foundational communication and transfer protocols of the Internet and the Web are designed to operate in a decentralised and distributed manner. In terms of global functionality, removing a single node from the network – a machine with an IP address – does not halt the whole network. Similarly, when an HTTP URL is no longer available (eg. typical "404 Not Found", "410 Gone" status messages), the other URLs are not directly affected. This particular behaviour, where the Web being functional without requiring bi-directional hyperlinks is considered to be a "feature" (rather than a "bug") of the overall system. In the meantime, bi-directional links can still be established when so desired eg. as in the case with notifications.

The AWWW enables the identification, discovery and description of information for the notion of a “Read-Write Web” where servers and clients interact using Web standards.

In *Socially Aware Cloud Storage* [350] and *Read-Write Linked Data* [351], Berners-Lee discusses where the Web architecture together with existing or future communication protocols and data standards can be used to materialise a “socially-aware decentralized access control of reading and of writing to linked data, and of notification of changes.” The overarching goal is to enable people and software to co-create information, and as well as to interact with information using social applications. This is all meanwhile technically and legally enabling participants to retain their autonomy, identity and storage, access and rights over their contributions. In order to achieve this, one of the goals is to make an infrastructure where data and applications are decoupled by design. That is, an ability to work with Linked Data without constraints on which applications can be used, as long as they follow consensus-based open protocols and data models. At the core of this initiative is where users, groups, and applications use global identifiers (URIs), and access control is applied using policies assigned to those identifiers. This also leads to the commodification of read-write data storage; the storage location on the Web can be decided by its owners and contributors, and it remains independent from the applications interacting with it.

⁂

Linked Data Platform [352] (LDP) specifies a “RESTful” read-write protocol and vocabulary for Web resources. LDP has distinct notions for “resource” and “container”; where the container construct can group and manage resources. LDP uses the RDF data model to describe the state of RDF and non-RDF resources. *Linked Data Platform Use Cases and Requirements* [353] outlines a wide-range of user stories, use cases, scenarios, and requirements which was used as the bases for the LDP specification. LDP can be used to mimic a file system abstraction interacting with Web resources over HTTP. The *LDP Paging* [354] mechanism can be used by clients and servers to efficiently request and serve a resource description, eg. resources in a container, in multiple parts. All LDP servers and clients can interact with each other using the same HTTP interface in context of Linked Data operations. An LDP server can facilitate researchers to register and store their resources, as well as share their content by applying different access controls.

Hydra [355] is a vocabulary to enable the creation of hypermedia-driven Web APIs by enabling a server to advertise affordances of its resources – machine-readable valid state transitions – to a client. A client uses this information to construct HTTP requests in order to perform possible read-write operations. Hydra enables generic client applications to be built without hardcoding knowledge about the available operations against Web APIs. With Hydra, applications can automate some of their tasks and only request researchers' engagement when need to.

Linked Data Fragments [356] (LDF) is the concept of having a uniform view on Linked Data interfaces that can be used towards reliable Web querying. To that end, a client asks a server about the kinds of Linked Data fragments that are available and then dynamically adapts its query plan for the server to execute. A Linked Data fragment is defined by three characteristics: data (what triples does it contain?), metadata (what do we know abut it?), controls (how to access more data?). LDF-based interactions can help researchers to inspect available scientific data through their distinct characteristics.

Triple Pattern Fragments [357] (TPF) is one possible way for a server to define a Linked Data fragment that enable live querying over the dataset on the client-side. TPF's hypermedia controls are expressed using the Hydra vocabulary.

Linked Data Templates [358] (LDT) provides the means to define read-write Linked Data APIs declaratively using SPARQL and specify a uniform interaction protocol for them. An LDT based server applies re-usable RDF ontologies to define application structure declaratively as a set of instructions for resource representation processing. To this end, templating is where an HTTP operation is mapped to a certain URI pattern that a server executes a SPARQL command in order to drive RDF processing. When a client triggers a particular request pattern, the server gives a suitable response based on available templates that are identified with URIs. LDT allows arbitrary academic and research applications to be built while exposing a uniform API that can be used in decentralised environments.

RDF/POST Encoding for RDF [359] uses the W3C HTML 4.01 Recommendation to specify read-write operations of RDF data through form-encoded RDF serialisation. Read operations are done by having an RDF/POST document encoded in HTTP GET URL, and write operations are sent via HTTP POST with form-urlencoded media type. Operations similar to LDP and LDT can be achieved with RDF/POST.

Fedora API Specification [360] refines the semantics and interaction patterns of the LDP specification in order to address the needs of repositories for durable access to digital data. Its goal is to facilitate interoperability with client applications. While the interactions patterns provides a mechanism for different client applications to communicate with servers, servers can be expected to vary in the kinds of services and affordances they offer for their resources. Fedora uses LDP as a foundation and defines the version identification and navigation scheme with respect to the *Memento specification*, integrates resource authorization and *Access Control List* rules (both described below), provides a design for the publication of event notifications, and interaction patterns to support binary resource fixity verification. The Fedora API can fulfill the registration, awareness and archiving functions, while enabling different actors to cooperate across instances on each others resources through independently built applications.

HTTP Framework for Time-Based Access to Resource States -- Memento [361] (RFC 7089): introduces a uniform, datetime-based, version access capability that integrates present and past Web resources. As representations of resources change over time, there is a need to preserve (archive and version) earlier representations. One common resource versioning pattern on the Web consists of generic URIs referring to the latest version of an accessible resource, as well as having a dedicated version URI for each resource. The Memento framework facilitates the discovery and retrieval of distributed versioned resources with datetime negotiation, a variant of content-negotiation; and TimeMaps an index of URIs referring to the prior states of a resource. The framework also helps to recognise frozen states of resources – a promise that the resource state has not and will not change. The Memento protocol can assist scientific applications to discover and reveal the variations of scholarly artifacts.

Web Annotation [362] (WA) is set of specifications for an interoperable, sharable, distributed Web

Annotation architecture. The annotations convey information about a resource or associations between resources to meet different motivations and purposes, eg. assessing, replying, describing, bookmarking, as well as “linking arbitrary content to a particular data point or to segments of timed multimedia resources”. *Web Annotation Protocol* [363] (WAP) is the HTTP API for publishing, syndicating, and distributing Web Annotations. Much of the protocol is based on using and extending LDP and REST best practices. *Web Annotation Data Model* [364] describes the underlying annotation *abstract* data model as well as a JSON-LD serialization. The WAP uses an LDP container to manage annotations with some constraints derived from the WA Model. Scholars can use servers and applications that support Web Annotation in order to annotate scholarly literature to meet the certification function.

Activity Streams 2.0 [365] (AS2) detail a model for representing potential and completed activities with the intention of using specific types of activity vocabularies defined elsewhere. The *Activity Vocabulary* [261] provides a foundational vocabulary to describe past, present, and future activities eg. announcing, creating, following, offering, and about objects eg. actors, media. Various academic and research activities can be recorded and brought to the attention of actors to determine their applicability and allow management of scholarly knowledge.

Linked Data Notifications [265] (LDN) is a resource-centric communication protocol for applications to generate notifications about activities, interactions, and new information, which may be presented to the user or processed further. It allows any resource (target) to advertise a receiving endpoint (inbox) for the messages anywhere on the Web. The server (receiver) hosting the inbox can have messages pushed to them by applications (sender), as well as how other applications (consumer) may retrieve those messages. Each notification in an inbox is an identifiable and reusable unit (URI), and can contain any data using any vocabulary. LDN can be combined with Web Annotation and Activity Streams to support applications to deliver and consume notifications about research and scholarly activities. I will revisit *LDN* and discuss its role in research communication where it supports the accessibility, content forces, and how it can perform the registration and awareness functions.

ActivityPub [366] (AP) is a decentralised social networking protocol based upon the Activity Streams 2.0 data format. It provides a client to server API, as well as a federated server to server API for delivering notifications and content. User accounts on servers have an inbox (to receive messages from the world), and an outbox (to send messages to others). AP’s inbox property is the same as LDN’s, and the targeting and delivery mechanism can be interoperably combined. AP can facilitate distributed scholarly interactions by allowing actors to disseminate their operations on their own, as well as others’ data. It serves to fulfill various social scholarly Web functions.

The design pattern for these read-write Linked Data centric protocols and architectures can be categorised as follows:

- The ability to perform HTTP interactions against Web resources.
- Decoupling software from domain or application-specific operations and data.
- Descriptive, extensible, and interconnected units of information.
- Declarative and machine-readable affordances for valid read-write operations.

4.4 Universal Identity for the Web

One of the functions of an online identity is to a identify a persisting entity at a particular time and context. The construction of an online identity of an entity eg. individual, organisation, community, state, can be complex in that it is interconnected with social and technical systems, and it can be issued – declared to exist, named, or be referenceable – in different ways. Entities being universally and uniquely identifiable on or off the Web entails that they can be attributed for their contributions, develop their reputations, be accountable for their actions and research output, as well as play an integral role towards building trust. An actor is an indispensable force in research communication, and enabling identity agency is both relevant and foundational.

In *Social Personal Data Stores: the Nuclei of Decentralised Social Machines* [367], Van Kleek, 2015, considers decentralised social applications from the user’s perspective based on the shortcomings and dangers of de facto practices where user data and interactions being controlled by centralised service providers. The work states that existing personal data stores for the most part do not “account for the need for multiple identities, effective separation of roles and anonymity, and to prevent unwanted tracking and clickstream profiling”, and that future work should integrate people’s need for privacy, creation, management, and switching of separate identities, pseudonyms, personas in order to be used in different contexts, eg. personal, professional.

In *The Presentation of Self on a Decentralised Web* [368], Guy, 2017, the “What is a profile?” study describes “affordances of systems which integrate online profiles in a social capacity and raises five features of systems with regards to their representations of users: flexibility, access control, prominence, portability, representation.” Furthermore, Guy posits that “online self-presentation is both constituted and affected by *who* sees a representation of an individual, and *what* it is they see, both of which are encompassed by the situation *whereby* it is seen”. The results of the studies are summarised as a conceptual framework as the “5Cs” for online profiles and self presentation: “context, control, customisability, connectivity, cascade”.

Third-party controlled identity: Community or state controlled identifiers for identities are a class of identifiers that are ultimately third-party controlled in that their ownership, governance, description is beyond an individual’s jurisdiction. The creation and control of identifiers by a group has the quality of long-term promise for persistence. Typically there are policies in place which promise their longevity, and procedures in the event that the group ceases to manage the identifiers. For example, ORCID URIs are intended to be part of the scholarly commons where they managed by the *ORCID community* [369]. The ORCID URI and profile description is useful to researchers, organisations, as well as to interlink researcher objects, however it is ultimately intended to serve the research community, and the systems that use it. One unintended consequence of this kind of centralisation in scholarly communication is that a growing number of institutions and publishers require researchers to participate with their ORCID identifiers, as opposed to any URI of a researcher. The resulting effects of *such initiatives* [370] are counter to autonomous participation in that a researcher has to conform to the terms and policies of multiple third-parties. On the other hand, the administration of the URI by the ORCID organisation removes the burden of maintaining the URI as well as the profile URL. In this case, a "trusted" third-party takes care of the identifier, whereas researchers are responsible of updating the content.

While some scholarly societies adopt identifiers like ORCID given their agreed potential benefits, it can also be harmful in the case of tracking researchers; where they work, who they work for or collaborate with, what they work on. Such ease of identification opens up the possibility to target and surveil researchers, as well as to discriminate or exile them if their findings act counter to others’ interests. *[GOAL] eLife collects ORCIDs from authors of accepted papers at proofing* [371], Morrison, 2017, argues that “it is not clear that facilitating researcher identification is in the best interests of academic freedom … participation in a service like ORCID should be optional”. *User Tracking on Academic Publisher Platforms* [372], Hanson, 2019, studied 15 different publisher sites and found that on average, each had 18 third-party assets loaded, with a total of 139 distinct third-parties across the platforms. The loaded JavaScript and cookies tracked user behaviour, interests, and research when combined with user identity information or fragmentary identity. The collected user data is shared with ad networks and data brokers. Unauthorised third-party tracking of users online activities is particularly concerning even in the case of *public sector* websites, which are expected to be more privacy respecting than "ordinary" websites (as per GDPR in the EU). For example, *Ad Tech Surveillance on the Public Sector Web* [373], Cookiebot, 2019, reports that advertising technology (ad tech) companies are extensively tracking EU citizens using government and public health service websites – “a total of 112 companies were identified using trackers that send data to a total of 131 third party tracking domains”. Thus, behavioural data combined with third-party identifiers in any shape or form (including syncing across services) is in breach of individual’s privacy in many jurisdictions.

Self-controlled identity: The concept of a personal self-sovereign identity entails that entities can register, describe, and manage their own identities of their choice. Controlling our online identities is important because it reflects our individuality and privacy. When identities are verifiable, we can also distinctly identify applications and their (human or machine) users, as well as switch between them depending on the context of use.

Different types of strings, eg. a username, an email address, an RSA public key, a URI, can be used to directly or indirectly identify agents. While each can be useful within their own context, Web-based identifiers (HTTP URI) make it possible to provide machine-readable descriptions when dereferenced, descriptions being extensible by their owners, as well as being interoperable global identifiers that can be interlinked unambiguously with other things.

A server can collect information about a client based on its device, machine, or browser fingerprint, even if the user did not directly disclose information about themselves. On the other hand, global identifiers like government-issued identifiers, credit card numbers, and full names, can be used to determine, track, and correlate an entity.

The *W3C Workshop on the Future of Social Networking Report* [374], 2009, concluded that distributed social networking is a possibility on the Web, given available data interoperability technologies, as well as further support of open source implementation of decentralised architectures. At the crux of the initiatives was to: foster preservation of privacy best practices for users; and their ability to trust claims; deepening user contexts and roles in social networking; enabling protocols for exchanging goods and services within communities, and; creating adapted user experienced with improved accessibility and mobility by closing the gap between implementations of social networks and device capabilities.

One of the conclusions of the workshop was that many of the existing technologies needed to create decentralised social networks already existed, eg. FOAF, OpenID, XMPP. One of contributions to the workshop was presented by Henry Story based on the original proposal for *FOAF+SSL* [375] (combining *RDFAuth* [376] by Henry Story and *sketch of a simple authentication protocol* [377] by Toby Inkster in 2008) for the notion of using Semantic Web vocabularies such as FOAF with SSL certificate exchange mechanism permits distributed and interlinked social networks to exist. The protocol was later posited in *FOAF+SSL: RESTful Authentication for the Social Web* [378]. Its goal was set to protecting and controlling access to personal information distributed on the Web where identification and privacy is at its centre. The work on WebID and related extensions described below builds on this work.

In order to enable people, group and software to act as autonomous agents that can cooperate on the Web, we need methods for distributed and decentralised identity, secure authentication, and access control. The collection of *WebID* [379] specifications outlines the approaches to meet those needs.

Web Identity and Discovery [380] (WebID) is a specification that outlines a distributed and extensible universal identification mechanism on the Web. A WebID is an HTTP URI denoting an agent, for example a person, organisation, or software. It can be used towards the declaration of an agent’s existence; discovery, describing or reference; authentication; and authorization. WebID is distributed and decentralised in that any agent can be *named* anywhere on the Web, and the server hosting the URI space controls the identity. It is extensible in that it can be used in conjunction with secure authentication and access control mechanism, as well as having dereferenceable machine-readable descriptions which are themselves extensible. WebID can improve privacy, security, and control be in the hands of users. WebID fulfils registration of identities of academics and researchers, helps them to be associated with scholarly artifacts they interact with eg. authoring an article, having access to an annotation.

WebID Profile [381] is an RDF document which uniquely describes an agents denoted by a WebID. The profile description is a way for researchers to *extend* the information about themselves; specify their preferences for applications, the location of their storage, and scholarly artifacts; how they can be contacted; their associations with other researchers, as well as other identities, and so forth. It serves to make identity claims which can be coupled with verification systems.

I have authority of the domain name `csarven.ca` where I publish my *WebID Profile* document https://csarven.ca/ describing my *WebID* `https://csarven.ca/#i` referring to me (in real life).

https://csarven.ca/ #i

```
<https://csarven.ca/>
  a foaf:PersonalProfileDocument ;
  foaf:primaryTopic <https://csarven.ca/#i> .

<https://csarven.ca/#i>
  a foaf:Person ;
  foaf:name "Sarven Capadisli"@en ;
  foaf:mbox <mailto:info@csarven.ca> ;
  foaf:img <https://csarven.ca/media/images/sarven-capadisli.jpg> ;
  foaf:interest <https://en.wikipedia.org/wiki/Media_theory> ;
  cert:key <https://csarven.ca/#cert> ;
  foaf:account <irc://irc.freenode.net/csarven,isnick> ;
  foaf:knows <https://www.w3.org/People/Berners-Lee/card#i> ;
  pim:storage <https://csarven.ca/> ;
  ldp:inbox <https://csarven.ca/inbox/> ;
  as:outbox <https://csarven.ca/outbox/> ;
  foaf:made <https://dokie.li/> ;
  rdfs:seeAlso <https://csarven.ca/cv> .
```

Listing 1. A WebID Profile describing a WebID.

Multiple identities: Individuals may want to be associated with multiple identities to fulfill diverse needs and social expectations. For instance, academics, politicians, employees, and artists, to name a few, may want to perform their online activities using public identifiers, so that they can be discovered, attributed, rewarded, trusted etc. There can also be quasi-public identities issued by a government or a bank to connect individuals to the physical world. Furthermore, individuals may wish to have several personal – public or private – identities and self representations, eg. on social networks, to address different needs. At any event, identities and corresponding identifiers can be registered and managed by different entities, and may be interlinked.

For instance, while ORCID profile description has constraints, it allows extensibility in the form of asserting a "same as" relation through its UI, so that the ORCID URI can be linked to other WebIDs. This essentially makes a connection to another *self-describing document* and facilitates *follow your nose* discovery. My personal WebID which is under my control is https://csarven.ca/#i. I also have an ORCID identifier that also serves as a WebID: https://orcid.org/0000-0002-0880-9125 which also has an RDF representation. As ORCID's user interface allows people to extend their descriptions with certain fields, it is possible to add *same as* or *preferred URI* relations:

```
<https://orcid.org/0000-0002-0880-9125> ;
  owl:sameAs <https://csarven.ca/#i> ;
  contact:preferredURI "https://csarven.ca/#i" .
```

Listing 2. WebID with same as and preferred URI relations.

The `owl:sameAs` and `contact:preferredURI` information provides an opportunity for applications consuming the ORCID WebID to potentially discover more information about other entities. This is particularly useful as the ORCID WebID Profile is not completely open to add arbitrary information. For example, specifying a person’s picture or their contacts is not currently possible with ORCID, however, if the alternative profiles (https://csarven.ca/#i) contain such information, they can be purposed to know more about the entity.

Social Graph: Identity profiles can further describe their relationships to other identities and hence form a network of self-described profiles. This is useful in that relationships between profiles of people, groups, organisations, and even software agents can manifest an information space that at its core has social affordances. That is, different kinds of relationships between entities can be expressed for different use cases, where each identity is independently controlled by its owners. For example:

- identifying a layer of trusted (or acknowledged) network of people
- sending notifications to the members of one’s personal social graph
- discovering annotations made by a particular social graph

Both the entities (nodes) and the relations (edges) of a social graph may have public and private parts. An example of a social graph where someone knows someone that knows someone, is as follows:

```
<https://www.w3.org/People/Berners-Lee/card#i>
  foaf:knows <https://csarven.ca/#i> .

<https://csarven.ca/#i>
  foaf:knows <https://bblfish.net/people/henry/card#me> .
```

Listing 3. Tim knows Sarven, and Sarven knows Henry.

4.5 Authentication and Authorization

Trust is ultimately tied to verifiable claims about identity and their provenance. While the creation and protection of third-party controlled online identities have received much focus, claiming and using personal identities online is still a challenge. Today’s Web applications predominantly make use of online identities that are essentially tied to and controlled by authorities in which users are merely borrowing an identifier and its management. It is a common practice to where users are required to create accounts at each online service they wish to use, eg. social media sites. Moreover, the trust model of such design dictates claims and usage of online identities is not ultimately in the hands of the users, and can be taken away from the users by the owners of the URIs and service at any time and reason. Even in cases where third-party issues identities can be used across different services, it is subject to a privacy violation as the identity provider essentially gets to know which services the user’s online identity is used with and for what purpose. Consequently such services can gradually build up information about users’ actions as well as behavioural patterns on the Web. Lastly, portability and interoperability of the identities is generally under developed as there is no particular incentive for an identity provider to make it so – after all, providing “free” services for online identities, and then locking identities to well-defined circle of services is taken advantage of by large corporations given their business models. Perhaps most importantly, such online services (as well as applications) do not recognise the identity of users as a distinct notion from the applications and services they use towards authentication.

While there are different authentication mechanisms, here I limit our coverage based on open standards and decoupling of identity, authentication and authorization process, as well as the resource server that the accounts are available from. Hence I categorically omit service-based authentication workflows which are known to require an account name and password, as well as for instance *The 'Basic' HTTP*

Authentication Scheme [382] (RFC 7617).

Web of Trust [383] (WoT) is a concept for a decentralised trust model used in systems like *Pretty Good Privacy* [384] (PGP) to establish the authenticity of the binding between a public key and its owner. A public key cryptographic system uses a pair of keys: public keys (disseminated openly) and private keys (withheld by the owner). It has two main functions: authentication and encryption. For authentication, public key verifies that a holder of the paired private key sent the message. In the case of encryption, for example, when Guinan wants to send Picard a message, Guinan takes Picard's public key to encrypt a message where only Picard can decrypt the message with his private key. WoT essentially leaves trust decisions to the users, as opposed to a centralised authority. Thus, WoT uses self-signed certificates and third-party attestation of those certificates, eg. at *key signing parties* [385]. Self-signing certificates removes the dependency on hierarchical certificate authorities to assert identity.

X.509 [386] is a standard for public key certificates that specifies a structure for a digital certificate, including a public key associated with one or more identities, as well as extensions for new information. X.509 essentially serves as a identity card that can be used to store identity claims. X.509 certificates can be used to authenticate both servers and clients. *HTTP over TLS* [387] (RFC 2818) – an extension of HTTP for secure communication (HTTPS) – uses the *Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile* [388] (RFC 5280). All combined, HTTPS ensures privacy, integrity, and identification.

The Transport Layer Security Protocol Version 1.2 [389] (RFC 5246) is a cryptographic protocol aimed at providing secure connections between applications. When applications eg. a server and a client (Web browser), communicate using the TLS protocol, it is private (secure) in that data is encrypted, identity of either party can be authenticated using public key cryptography (like PGP), and integrity of the exchanged data can be assured during transit.

WebID-TLS [390] defines how a server can authenticate a user with a client application holding the user's public key certificate, and serves towards identity claim verification. As the extension mechanism in X.509 allows additional statements to be specified under the `Subject Alternative Name` (SAN) field, it is used to declare the user's WebID for the `URI` field. When a client application running on the Web browser requests a resource from the server, the TLS connection gets established between them. Once TLS has been set up, the HTTP application protocol exchange takes place (eg. `HTTP GET`, `HTTP POST`). If the requested resource requires WebID authentication the server can request the client to authenticate itself. When the Web browser encounters this certificate request, it may prompt its user to choose a certificate (from its certificate manager) – thereby signing a token with its private key – to let the client send it to the server. The verification agent extracts the public key as well as the WebID from the certificate. As the WebID enables a global dereferencing mechanism for finding a key, the server uses can decide on dereferencing the WebID Profile and verify the claims about the WebID. Once the identity verification is established, the server can look up its access control rules for the requested resource to determine if and how the request can be fulfilled. Researchers can use servers and applications implementing the WebID-TLS workflow towards securely identifying profile claims and to have access to resources may be otherwise protected.

The Cert Ontology 1.0 [273] can be used in WebID Profile documents to specify an agent's certificate information. For example, an agent can indicate their public and private keys. The fields and corresponding values of `Public Key Algorithm`, `Modulus`, and `Exponent` in the X.509 certificate are used in the WebID Profile as part of `cert:RSAPublicKey`, `cert:exponent`, and `cert:modulus`. URI under SAN is our WebID. An example where information about an RSA Public Key is associated with a WebID:

```
<https://csarven.ca/#i>
  cert:key <https://csarven.ca/#cert> .

<https://csarven.ca/#cert>
  a cert:RSAPublicKey ;
  cert:exponent "65537"^^xsd:nonNegativeInteger ;
  cert:modulus
"caf6a78d16e80f930337453d84f790764c56ea58acbeda5de3d17b6e673569ef581b896b74466655fb3
da2f9a961c6d46d99e77053a0aaf83fac4eab8b9198f20930672de22cb1b22f0ab85a66c95a6830eaa7b
e1e28ec197ff7a4a448243af2f206d2be458bbc71f32aa073315e22c9b2668fe15c732a33edcfd9fa39d
04706191044f33e580f954e6a1d9c4bf08c820c4666f5ef8554cf1430a33b630c80d11f7d309cb641b21
3fcbfc9a5b699af625fe23627c403ff12d79514178728537aa173aba5bc5aeb87569eba1be3274aa53d9
b2f8fe72f13ba8125e734e8e31128303c76594327c4c2e431e38f2b326cc244080e597a4d27a98ae3fe1
edb5de785"^^xsd:hexBinary .
```

Listing 4. RSA Public Key information associated with a WebID URI.

WebID Profiles using the Cert Ontology can help authenticating servers to discover researchers public keys which can be used towards verifying identity claims.

WebID-TLS-Delegation is an an extension to WebID-TLS that enables users to delegate other agents to act on their behalf. In *Extending the WebID Protocol with Access Delegation* [391], Tramp, 2012, states how a “principal” agent can delegate authentication and access to a trusted “secretary” agent – typically a third-party software – who can act *on behalf of* of the principal agent to perform (asynchronous) requests. A secretary with its own WebID acting on behalf of a user can potentially have the same (or required) privileges while still being distinguishable eg. via different public key, and capabilities. An agent can declare a delegation as follows:

```
<https://csarven.ca/#i>
  acl:delegates <https://example.org/application#i> .
```

Listing 5. An agent delegating another agent to act on its behalf.

whereas the agent making requests on behalf of another agent would include the HTTP `On-Behalf-Of` header when making requests.

OpenID Connect [392] (OIDC) is an authentication extension to *The OAuth 2.0 Authorization Framework* [393] (RFC 6749). OAuth 2.0 provides an authorization layer that separates the role of a client from that of the resource owner. Clients obtain an access token issued by an authorization server with the approval of a resource owner, and clients can use that delegation-specific credentials (access token) to obtain and use limited access to HTTP resources on the server. In addition to the authentication process, OIDC enables authorization servers to obtain profile information about the user.

WebID-OIDC [394] is an authentication delegation protocol resulting in OIDC’s verified ID Token once a client authenticated, and deriving the WebID from the ID Token. The complete WebID-OIDC workflow as follows: an initial request is made to access a protected resource; whereby the user specifies their identity provider to the authorization server; the user authenticates themselves at the provider and gets redirected back to the resource they were trying to access and provides the signed ID Token; the server controlling the resource validates the token and extracts the WebID; and finally the confirmation from the resource server. For researchers, WebID-OIDC is equally applicable as WebID-TLS.

Access Control List [272] (ACL) ontology, enables a server to provide four modes of access: “Read”, “Write”, “Append”, and “Control”. As implied by the name, *read* allows a server’s resources to be accessible and interpretable. *Write* allows resources to be modified or deleted. *Append* allows a specific kind of Write in that while adding of information is permitted, removal is not. *Control* permits full read and write access to an ACL resource, which is typically given only to the "owners" (or administrators).

The following code snippets of authorization policies are used to specify the ACL information (eg. `resource.acl`) associated with a resource (eg. `resource`) on a server. For example, setting an authorization policy where an agent (`https://csarven.ca/#i`) has write access to resources, as well as control access to their respective ACLs:

```
<#write-control>
  a acl:Authorization ;
  acl:agent <https://csarven.ca/#i> ;
  acl:accessTo <./> ;
  acl:mode acl:Write, acl:Control .
```

Listing 6. An authorization policy that assigns write and control access for an agent.

Setting a class of agents to perform read and append operations on a resource eg. an inbox:

```
<#read-append>
  a acl:Authorization ;
  acl:agentClass foaf:Agent ;
  acl:accessTo <./> ;
  acl:mode acl:Read , acl:Append .
```

Listing 7. An authorization policy that gives append access to a class of agents for a resource.

ACL plays a role in determining which authorization policies to apply when a resource is requested eg. raw sensitive data, private social graph, initial versions of an article.

Prohibiting Delegation [395], Miller, 1999, provides an analysis on four delegation problems: “perimeter security, confinement distributed confinement, confused deputy, communicating conspirators”, and examines the basic differences between Capabilities and ACLs. The article posits that users and programmers can be lead to “make decisions under a *false sense of security* about what others can be prevented from doing, ACLs seduce them into actions that compromise their own security.”

4.6 Persistence and Preservation

As knowledge builds on knowledge, it is vital to preserve the connection between units of information in scholarly communication. There are a number of challenges in this respect. In this section I focus on the persistence of the identifiers for units, as well as content integrity and preservation of context. In the most general sense, trust and accessibility are integral to preservation of content, integrity, and the context in which they are used. This is so that knowledge can be scrutinised with the help of transparent trail of reproducibility and replicability of research results. I look at existing research and approaches in this space.

Domain name registration and maintenance is one of the key factors for long-term reliable persistence as per *URI ownership*. Berners-Lee outlines two issues for the persistence of HTTP URIs:

> 1. The persistence of the opaque string which follows the domain name, and
> 2. the persistence of the domain name itself.
>
> *Persistent Domains* [396], Tim Berners-Lee, 2000

In *Cool URIs don't change* [397], Berners-Lee, 1998, discusses some of the approaches that can be taken towards usefulness and longevity of URIs. The article focuses on practices that a publisher making a commitment to persistence by designing and managing the URI path and the content it resolves to, as well as the domain name it uses. The owner of a domain name has the obligation to define what the things mean for its URIs. This is also a form of social contract made by the authority that names and defines a URI to anyone that's using it – see also *Philosophical Engineering and Ownerhip of URIs* [398].

Persistence policies can come in different forms. For example, W3C's *URI Persistence Policy* [399] is a document making a pledge about how some of the resources under its domain will persist throughout the lifetime of the Consortium; any changes to persistent resources will be archived; and in case that the organisation is disbanded, its resources can be made available under the same rights and license. These human-readable statements are useful institutional commitment to persistence. The ODRL vocabulary can be used in a similar way to provide a machine-readable policy about resources.

From the archiving perspective, Van de Sompel came to the conclusion that in a long enough timeline, HTTP URIs are not inherently persistent but persistable. The units of information that are registered using URIs are more of a promise made by its original or current authority. Hence, along with the examples from earlier, URI registration is ultimately a social agreement. URI owners declare a policy eg. implicit, written, verbal. If a policy is announced for a collection of URIs eg. what happens in 1000 years, then that says something about its intentions and expected level of availability. From this perspective, as discussed earlier in the *registration of identifiers with social contracts*, PIDs such as DOI, PURL, w3id, and ORCID can help to prolong such promises and to extend the lifetime of accessibility of units of scholarly information.

Decentralized Identifiers [400] (DID) are identifiers for verifiable "self-sovereign" digital identity, where they are "under the control of the subject, independent from any centralized registry, identity provider, or certificate authority." In a way, DIDs go around the shortcomings of the domain name system where they can be created and managed without the authority of the registrar.

In *Analyzing the Persistence of Referenced Web Resources with Memento* [401], Sanderson, 2011, presents the results of a study on the persistence and availability of Web resources cited from research articles in two scholarly repositories. The results show that within a few years of the URL being cited, 45% of the URLs referenced from arXiv still exist but are not preserved, and 28% of the resources referenced by articles in the UNT digital library have been lost. In order to address this commonly known as URIs ceasing to exist (link rot), authors suggest that repositories expose the links in the articles through an API so that Web crawlers can be used to archive. With the help of archives supporting the Memento protocol, the original context of the citation can still be reconstructed.

Given the dynamic and ephemeral nature of the Web, and in particular management of URIs and corresponding representations at URLs, it poses a threat to integrity of Web-based scholarly content, and the consistency of scholarly records, as well as everywhere else. One special area is about the formal citation of scholarly resources eg. DOI, HTTP-DOI-URI, and informal referencing of other resources on the Web ie. any HTTP URI. The *Hiberlink* [402] project investigates "reference rot" in Web-based scholarly communication, and introduces the term to denote two problems in using URI references:

> **Link rot**
> The resource identified by a URI may cease to exist and hence a URI reference to that resource will no longer provide access to referenced content.
>
> **Content drift**
> The resource identified by a URI may change over time and hence, the content at the end of the URI may evolve, even to such an extent that it ceases to be representative of the content that was originally referenced.
>
> *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot* [403], Martin Klein, 2014

In *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot* [404], Klein, 2014, acknowledge extensively studied phenomenon on link rot and content drift, and themselves "investigate the extent to which *reference rot* impacts the ability to revisit the web context that surrounds STM articles some time after their publication". The results show that significant amount of HTTP URIs cited in STM articles are no longer responsive or adequate archived snapshots available. Authors state that it is impossible to adequately recreate the temporal context of the scholarly discourse, hence suggest that robust solutions are needed to combat the problem of reference rot. Authors can take practical steps to remedy some of these issues eg. using archives that support on-demand snapshots, embedding the archived URI with datetime information alongside the reference to the original resource – I discuss this further in *Robust Links*.

In *Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content* [405], Jones, 2016, reuse the same dataset from Klein study, to investigate to what extent the textual content remained stable since the publication of the referencing article based on various well-established similarity measures based on the comparison of the representative Memento and the live resource. They "find that for over 75% of references the content has drifted away from what it was when referenced." The authors support the idea that in order to partly work around this issue, authors should pro-actively create snapshots of the referenced resources at Web archives, and referencing them in their scholarly literature. However, the authors also state that such robust embedding in the infrastructure of the existing authoring, reviewing, and publishing workflow is still an open challenge. To that end, applying the *Robust Links* approach can help. While the DOI-paradigm for scholarly units help to improve the link rot scenario when the custodians of the domains or the URLs of the scholarly resources relocate, the resources on the Web at large remain to be a problem given that their incentives towards longevity and access differ.

From the point of persistence of the domain name (losing ownership) as Berners-Lee describes, one kind of content drift would be if the content published at https://csarven.ca/ today may be different tomorrow if another authority gets to own csarven.ca and defines what goes there. Alternatively, it may be that the content at that location is dynamic, and could differ from one request to another. In both cases, content drift creates a situation where if the originally referenced resource is still the same. From the perspective of Web-based scholarly publications, changes to content – accidental or intentional – can impact the degree of reproducibility, replicability, comparability of research results, and maintaining a reliable scholarly record.

In *Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping* [158] Van de Sompel, 2014, posit that while PIDs are assigned to resources outside of information access protocols (like HTTP), there is a need to unambiguously bridge the discovery of the Web-oriented resource eg. from PID to HTTP URI, in a way that is machine-actionable. For example, the PID paradigm has the following discovery path:

1. PID is the resource identifier eg. `10.2218/ijdc.v9i1.320`
2. HTTP-URI-PID is the resolving URI eg. `https://doi.org/10.2218/ijdc.v9i1.320`
3. HTTP-URI-LAND is the redirect URI (landing page) eg. `http://www.ijdc.net/article/view/9.1.331`
4. HTTP-URI-LOC is the location URI of the content eg. `http://www.ijdc.net/article/download`

`/9.1.331/362/`

Authors propose that using existing standards and practice, the essential ingredients for such a mapping is as follows. A PID has a Web equivalent HTTP-URI-PID, which is a requisite – minted by the naming authority. The HTTP-URI-PID can be content-negotiated to result in a) a human-readable representation at HTTP-URI-LAND or b) machine-readable representations with distinct HTTP-URI-MACH. The HTTP-URI-LAND remains the same for discovery, however, HTTP-URI-MACH uses an RDF-based approach to describe the aggregations of scholarly assets based on the OAI-ORE specification.

It is worth briefly revisiting the notion of social agreements around Web resources. Conceptually, the agreement that I make with you about the persistence of my website's resources is in essence the same as a naming authority controlling a PID, as well as all of the nodes in between HTTP-URI-LOC. The kind of mapping between a domain name and the IP address it points to is similar to a PID being mapped to a HTTP-URI-PID. Hence, *URI Ownership* essentially involves two possibilities: either I "own" and control a URI space or someone else does.

In *Persistent URIs Must Be Used To Be Persistent* [406], Van de Sompel, 2016, reveal the results of a study where authors do not use persistent URIs like DOIs even when available, and instead use the location URIs. In order to alleviate this issue, authors propose that an HTTP `Link` header is used at the location URI to announce the identifying HTTP-URI-PID. The current proposal is to use *cite-as: A Link Relation to Convey a Preferred URI for Referencing* [407], Van de Sompel, 2019, and a number of related patterns are outlined at *Signposting the Scholarly Web* [408].

Robust Links [409], Van de Sompel, 2015, is based on the idea that in a long enough time line, content-drift and link-rot pose challenges to maintain context and integrity between resources. In order to preserve the context of linked resources at the time of linking, a *link robustness* strategy is applied by way of creating a snapshot of the original resource and making it available to data consumers. The *Link Decoration* approach decorates links in hypertext documents in order to be human- and machine-actionable. By additionally associating the datetime information when linking helps to retain such context. The datetime information can be accompanied with a snapshot of the target URI made at a Web archive or a versioning system when linking, and using either the original or the versioned resource as the primary reference. Including datetime information in the source document can make it possible to interpret the context of each of the links in the document and to discover relevant snapshots. In the event that a linked resource or a snapshot becomes inaccessible, the datetime information helps to discover temporally close alternative snapshots.

Trusty URI [410] is a technique to include cryptographic hash values in URIs to uniquely associate them with an artifact. In *Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data* [411], Kuhn, 2014, outline how specific resources and their entire reference trees can be verifiable. If the trusty URI of an artifact is known, it can be used to verify if the content of an artifact corresponds to what it is suppose to represent. This is useful to determine if the content is corrupted or manipulated. It follows that trusty URI artifacts are immutable as each version of a content generates a unique trusty URI. Trusty URI artifacts are considered to be permanent in that once archived or cached, the artifact can still be verified even if the original location is no longer available. Applications can use trusty URI to encode their own references, as well as compute an artifacts' trusty URI and verify before using it. Trusty URIs can be used to represent byte-level file content and for RDF graphs, and is compatible with named information URIs – *Naming Things with Hashes* [412] (RFC 6920). An example from the *Trusty URI Specification – Version 1* [413]: given resource `http://example.org/r1`, its trusty URI would be `http://example.org/r1.RAcbjcRIQozo2wBMq4WcCYkFAjRz0AX-Ux3PquZZrC68s`, where RA identifies the module as an RDF graph (independent of its serialization) of the resource, and the remaining characters signify the computed hash of the content.

The Anatomy of a Nanopublication [414], Groth, 2010, propose to improve the efficiency in finding, connecting, and curating core scientific statements with associated context, with an annotation model and a format based on the RDF language realized with Named Graphs. The *Nanopublication Guidelines* [289], 2015, specify how to denote unique RDF graphs for *assertions*, *provenance*, and *publication information*, which make up the body of a nanopublication that can be used as a single publishable and citable entity. Nanopublications can be independently used to expose and disseminate individual quantitative and qualitative structured scientific data, without being accompanied with narrative research articles. For example, hypothesis, claims, and negative results can exist on their own, be identifiable, and reused in different places, as well as embedded in articles.

In *Decentralized provenance-aware publishing with nanopublications* [415], Kuhn, 2016, argue that due to publication and archival of scientific results is still based on print-centric workflows and commonly considered to be a responsibility of third-party publishers, there is currently no efficient, reliable, and agreed-upon Web-centric methods for publishing scientific datasets, and therefore a bottom-up process is necessary. To this end, the authors present a decentralised server network with a REST API to store, archive, find, and serve data in the form of nanopublications, where the identifiers for the units of information are based on the trusty URI method. The authors argue that the underlying architecture can serve as a reliable and *trustworthy* low-level semantic publishing, archiving, and data sharing layer that can be used by different knowledge domains.

Signing HTTP Messages [416], 2019, “describes a way for servers and clients to simultaneously add authentication and message integrity to HTTP messages by using a digital signature.” The signing is made by signing the HTTP message body and some HTTP headers at the time of making the HTTP request. As the signature is carried in the HTTP headers, the HTTP message is not altered, and thereby does not require a normalisation algorithm to be used on the message structure.

Linked Data Proofs [417] is a mechanism to ensure authenticity and integrity of Linked Data documents through the use of public/private key cryptography. The digital signature is comprised of information about the i) signature suite that was used to create the signature, ii) parameters required to verify it, and iii) the signature value generated by the signature algorithm. The signature typically accompanies the Linked Data document so that the receiver can verify its authenticity using the available information. In contrast to the *Signing HTTP Messages* method, *Linked Data Proofs* requires a normalisation algorithm.

Verifiable Claims Data Model 1.0 [278] is a mechanism to express credentials on the Web that is cryptographically secure, privacy respecting, and machine-verifiable. The core model outlines the concepts for claims, credentials, and presentations that can be used by issuers, holders, verifiers, and verifiable data registries.

4.7 Decentralised Storage and Interoperable Applications

The Presentation of Self on a Decentralised Web [418], Guy, 2017, describes various types of decentralised social systems. As per the client–server model, their arrangement generally falls in the spectrum of tightly or loosely coupled architectures. In these systems, actors control client applications typically through user-agents. The degree of coupling that is exerted on the system varies depending on the class of components involved. More specifically, *Why is the Web Loosely Coupled? A Multi-Faceted Metric for Service Design* [419], Pautasso, 2009, proposes “discovery, identification, binding, platform, interaction, interface orientation, model, granularity, state, evolution, generated code, conversation” as metrics for the degree of coupling in service-oriented systems.

Loose coupling reduces the number of assumptions each party makes about the other when exchanging information. For instance, AWWW includes various protocols for interaction, URIs to identify, and standard data formats to communicate the content. The Linked Data design principles then sets further constraints within this framework by favouring HTTP URIs, and the RDF language. A Linked Data driven system using uniform interaction and information exchange can be viewed to be more tightly coupled in comparison to an arbitrary system that is less specified or unspecified. Corollary, the loosely coupled systems set a higher requirement for using Web specifications emerging through open consensus as well as being widely deployed in order for the ecosystem to function. Otherwise, it would be impossible to communicate if there is no higher-order agreement. The loose coupling of interoperable systems based on open standards however can go beyond the arrangement of clients and servers. For instance, the notion of WebID enables loose coupling of identity, identification, authentication, authorization, and storage.

Decentralised datastore (or dataspaces) under the custodian of different actors implicates organic growth and operation as a by product of autonomous operation. In contrast to third-party controlled centralised systems, decentralised systems can have persistent availability (globally), resistance to external censorship, monopoly disruption, increased interest in privacy and security, and enable free expression with respect to information structure and semantics. In order for decentralised datastores to be effective, both servers and client application have an increased requirement on implementing consensus-based open standards in order to cultivate an interoperable ecosystem.

Shared hosting refers to a kind of *web hosting service* [420] where the server's resources are shared among its users. While the hosting has usage limits, it is considered to be economical and have reliability features put in place by the host's system administrators. Any website with its own domain name can act as an independent node in a decentralised system. However, given that multiple websites reside on the same server, system-wide configuration changes or disruptions on the network usually effects all of the websites. Moreover, there are privacy and security concerns if the information on or transferring through individual dataspaces are not encrypted. Hence, the user places their trust on the hosting service's administrators.

Personal data service (or store) (PDS), *personal online datastore* (POD) or "pod" as a noun, loosely refer to an individual-operated Web space with mechanisms for data storage and management that is capable of providing access-controlled read-write operations for independently built interoperable applications. A generic PDS can be virtually realised with any HTTP server with varying configurations and capabilities offering client-side applications to communicate through open protocols and data formats. Web resources that are only intended for public read entails that the custodian allows resources to be available in response to client requests. Servers would need to factor in an authentication mechanism and authorization rules to determine who to serve the resources to and under which access constraints. A client-server interaction based on a read-write Linked Data architecture has affordances for distributed and semantically interconnected Web resources.

The concepts above generally fall into the spectrum of *self-hosting* [421] for Web space management. An individual responsible for a PDS entails that they are responsible as well as accountable of its operation, abilities, policies and practice decisions on data, persistence strategies and so forth. PDS makes it possible for actors to safeguard their data by setting authentication and authorization rules for any or class of agents, being able to act as the source of truth for WebIDs and the descriptions in the profiles. Running a PDS fundamentally requires the system to fulfill the registration function for the content it makes available. As archiving is an external function, we expect other services to meet that requirement. The certification function can be fulfilled in a similar manner to registration in that the actors that are involved in the quality-control process can register their own units of information in their PDS in response to target units.

4.8 Design Decisions for Decentralisation

I use the same set of questions raised in *Systematizing Decentralization and Privacy* to study the design choices for an actor-controlled system based in context of the *Architecture of the Web*:

How is the system decentralized?

The nodes in the system are the Web resources on servers that make up information such as profiles, articles, annotations and notifications, as well archived snapshots. The advantage of this infrastructure is that once a resource is registered (named and stored), they do not require users for its availability or functionality – a reduced attack surface. Users retain their privacy in that, access and use of resources are encapsulated, and such information about identity or actions taken is not required to spread to the rest of the network. In the case of compound resources including information from distributed resources, like with transclusions, or applications dynamically fetching information from different context, nodes may get to be aware of cross-origin requests. The *network topology* is by default *point-to-point*, ie. connections between two endpoints. However, in the case of federated servers, information can be routed and disseminated to the whole network, for example, an actor sends out a notification about their activity. The *authority topology* is that nodes (client-server or server-server) *interact directly* with each other, without necessarily requiring third-party nodes to participate. In order to establish trust, a certain amount of transparency can be used to make an authority *accountable*. In federated designs, users place their trust on their provider nodes who act as authorities on their behalf with other providers (server-server).

What advantages do we get from decentralizing?

There is no single point of failure, has high availability, system performance, and generally *survivable*. However, the system's usefulness may degrade if certain resources in the network disappear or become inaccessible. For example, if a notification about an annotation is no longer available, then the annotation may need to be discovered through other routes. There is a flexible trust model in that, there is *no natural central authority*. Multiple independent authorities exist, and each verify claims based on their own (or shared) decisions eg. self-signed certificates as part of web of trust, or in the case of *anyone can say anything about anything*, consumers of such information verify the statements based on their needs. There is a *distributed allocation of resources assists ease of deployment*, and independent extensibility. As with the open-world assumption, nodes are not expected to have complete knowledge, and conflicting knowledge can enter the system. Decentralised systems while benefit from network *location diversity*, they may also be subject to censored both on the network and legal levels. The *development and operation is decoupled* in that open software can be publicly visible and auditable before being used by authorities. Original resources as well as versioned or archived copies can have *publicly verifiable integrity*. Decentralised system are relatively secure in that compromising or weakening the system requires attackers to engage with different services, or code bases. *Public accountability* can be devised to detect and exclude misinformation, as well as misbehaving or compromised actors. A decentralised system can incorporate trusted centralised nodes eg. community agreed scholarly reviews, archives, caching servers, proxies, and so forth, while if compromised by adversaries or disappearance is temporarily kept local – possibly a performance or discovery problem.

How does decentralization support privacy?

Due to the way resources are distributed in decentralised systems, they offer greater user data privacy with respect to their communication, identities, and actions than centralised systems in that there is no single authority that can observe all data and interactions. In essence, it makes it challenging for an adversary to aggregate and connect all information in a way that can undeniably hold a participant in the system accountable. These systems can allow different degrees of confidential participation in that agents (human or machine) can remain anonymous or use their pseudonymous identities. For instance, sensitive research data can be kept confidential. Access controlled resources can be accessible by trusted parties. Actions of an actor can potentially be unlinkable and so rendering a global search of information useless. The communication between parties can be undetectable and unobservable by unauthorised third-parties.

What are the disadvantages of decentralizing?

Decentralised systems inherently increases the number of attack vectors that an adversary can use due to number of software implementations and configurations. A compromised node in the network can potentially monitor activities, corrupt information exchange, insert malicious code, distribute data to interested parties while undetected, reduce or expose anonymity, provide inconsistent views, and so forth. As there is no single authority to globally manage the nodes in the network, the quality and availability of the services will vary and becomes challenging for each node to actively and effectively protect themselves. For example, an article, its inbox, and the annotations referred from each notification may all reside at different network locations, and thus data consumers would need to factor in potential issues arising from each node by defending themselves or routing around disturbances. Decentralisation also poses an obstacle for accountability and reputation in that while actors may control a node, handling the ramifications of a misbehaved node becomes difficult in the broader ecosystem. Finding, discovering, and interactions on fine-grained resources are typically less efficient than centralised systems.

What implicit centralized assumptions remain?

Decentralised systems include social and technical aspects of centralisations. The DNS is the most prominent example of centralisation in that domain names are administered by an entity other than the actors interacting in the information space. The name which resolves to a node is generally governed by a single authority solely determining the level of privacy the node offers as well as its availability. For instance, since registration of units of communication are mere records of content, their persistence and preservation are at the discretion of the owners of the URI space. Thus, archiving by trusted parties is necessary to take into account – which essentially introduces additional centralised entities. The use of certain certification authorities also introduces hierarchically structured centralised governance into the system, hence fundamentally allowing single entities to revoke authenticate credentials, or eventually influencing abuse-prevention, reputation, and trust. Along these lines, carrying out access policy management and enforcement by independent nodes introduces central locations for potential attacks. Web standards are a form of social centralisation in that they reflect the participants’ norms and agenda. Similarly, content delivery networks, for wide reuse introduces a single point of failure for large number of services – a potential risk even for open-source software.

4.9 Degree of Control

We can observe different degrees of control; individual and third-party actors at the opposite ends of the control spectrum for identifiers, data, and applications:

Identifiers

Rules and governance of naming things.

Data

Information that can be shaped and used with intended semantics.

Applications

Computer software that performs functions on data on behalf of a user.

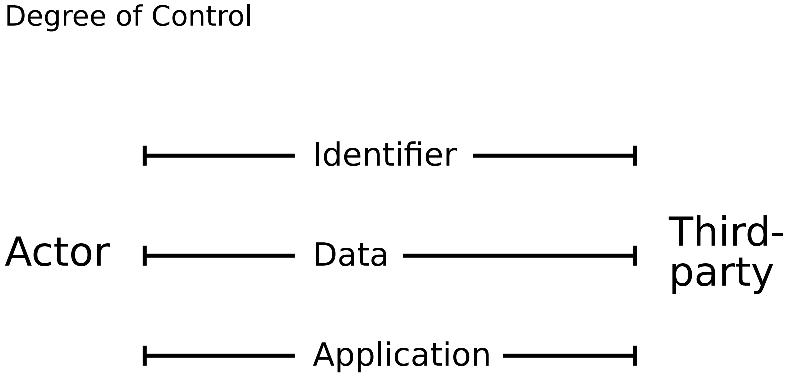


Figure 4. Degree of control (actor or third-party) for identifiers, data, and applications.

On one end of the spectrum, an actor eg. an individual, registers their own domain name, hosts their personal storage, and uses applications of their choosing. With that, they have highest degree of autonomy. On the other end of the spectrum, an individual's identifier as well as a range of expressible identities, the data location and conditions of use, and assigned applications by third-party services.

I focus on interactions based on the *client–server model* [422] under the *Architecture of the Web*, identifiers and data are conceptually decoupled components. In the case of identifiers, I focus mainly on HTTP-based URIs, and the social conventions around *URI Ownership*. Finally, the applications comprise software that actors can use.

While centralisation of data and applications are two distinct areas of concern, some of the challenges we face in scholarly communication make them appear as one. When we discuss challenges around access to knowledge, the location and quality of the data is often conflated with the tools or services which are required to access the data. Being required to use particular software to create, format or share research articles, to create accounts on particular services, or agree to certain terms and conditions impose a set of characteristics on research output or data itself.

Data and application appear to be intertwined because the workflows that the researchers are required to use tend to be unique and proprietary and predominantly enforced by third-parties. While some of the differences in designs may be due to historical reasons they also ensure various forms of vendor lock-in on both the research outputs and the applications to generate, share and reuse data. The use of non-interoperable (or even post-facto) methods to create and exchange data on the Web leads to the fragmentation of data through silo-services, and along with it the dependency on applications having their interaction mechanisms established via out-of-band knowledge. By strictly focusing on the current quality of machine-readable scholarly information on the Web, the technical obstacles are evident and ample. The applications need to be continuously hard-wired to know where and what to look for; what to do once something of interest is found; or where to go next in the discovery phase. Similarly, automating discovery of fine-grained information without human interference is still central to the limitations.

Different degrees of control would influence the forces and functions in scholarly communication. For instance, while ORCID is community controlled, it is possible for individuals to customise their profiles and to link to their other identities. The profile description can indicate the location of researcher's personal online storage, inbox, as well as outbox, which can be controlled by the individual. The DOI system on the other hand is governed by an organisation where the mapping between an identifier and the location of content they refer to is managed by select few organisations (as opposed to individuals).

Desired qualities: Based on aforementioned research, standards, and practices in the field towards Web-based personal identifiers and representations, I infer a set of qualities that can support individuals when participating in a decentralised scholarly system:

- Self-hosted and individually-controlled personal identifiers denoted with HTTP URIs
- Profile customisability and connectivity
- Linkable multiple identities
- Identity privacy and control

4.10 Forces and Functions in Specifications

Standards-based Forces and Functions: While every system of scholarly communication can have its forces and functions arbitrarily configured irrespective of their actual implementation, I have focused on a subset arrangement that is based on existing Web specifications facilitating the manifestation of interoperable information spaces. The model can be extended to include other or future specification with respect to forces and functions in scientific communication.

Table 3. Characteristics of Web specifications for forces and functions in scientific communication

	Forces				Functions			
Specification	**Actor**	**Accessibility**	**Content**	**Applicability**	**Registration**	**Awareness**	**Certification**	**Archiving**
Protocol								
Linked Data Platform	⌙	✔	✔	⌙	✔	✔	⌙	⌙
Linked Data Templates	⌙	✔	✔	⌙	✔	✔	⌙	⌙
Linked Data Fragments	⌙	✔	✔	⌙	⌙	✔	⌙	⌙
Fedora API	✔	✔	✔	⌙	✔	✔	⌙	✔
Memento	⌙	✔	⌙	⌙	✔	✔	⌙	✔
Web Annotation	⌙	✔	✔	⌙	✔	✔	✔	⌙
ActivityPub	✔	✔	✔	⌙	✔	✔	✔	⌙
Linked Data Notifications	✔	✔	✔	⌙	✔	✔	⌙	⌙
WebID	✔	✔	✔	⌙	✔	⌙	⌙	⌙
WebID-TLS	✔	✔	✔	⌙	⌙	⌙	⌙	⌙
WebID-OIDC	✔	✔	✔	⌙	⌙	⌙	⌙	⌙
Model								
Web Access Control	✔	✔	⌙	⌙	⌙	⌙	⌙	⌙
Hydra Core	⌙	✔	✔	⌙	⌙	✔	⌙	⌙
Activity Streams	✔	⌙	✔	✔	✔	✔	✔	⌙
Robust Links	⌙	⌙	✔	⌙	⌙	✔	⌙	✔

✔ Passed
✗ Failed
? Cannot tell
⌙ Inapplicable
○ Untested

The data in this table was manually determined, where a specification whether intended to and can fulfill a force or a function.

The suite of Web Annotation standards are grouped. Triple Pattern Fragments is grouped into Linked Data Fragments. WebID incorporates identification systems that can be used for actors (any HTTP, PURL, w3id, ORCID), as well as the WebID Profile that describes it.

4.11 The Effects and Artifacts of Autonomous Engagement

The Web’s design stands out because of its absence of centralised control, both for technical reasons of scalability and resilience as well as a societal need to exercise creative and effective communication. A challenge in such large-scale decentralised networks is how related publications can be semantically interlinked, even if they are authored and published by different parties. Centralising their publications is practiced by the majority of authoring networks today, demanding authors to give up some or all of their control in exchange for technical simplicity.

We can approach an alternative to this centralisation by adopting Web-centric workflows that are inherently interoperable by design, but most importantly by enabling the creators and the users of data and applications to be autonomous entities, working towards the democratisation of scholarly knowledge for humans and machines. One way to achieve this is by decentralisation and decoupling of data and applications. By *decentralisation*, I mean data and applications are loosely coupled, and users can choose where their data is stored. I focus on Web-based decentralisation, where content is transported over HTTP, and resources are identified with URIs.

The future of self-registration of scholarly units depends in large part on how well its practitioners, advocates, educators alike, are able to reconcile the ideals that the medium offers with the realities of the commercial third-party systems within which scholarly communication operates and the homogenizing influences of scholarly socialisation.

While these challenges are technically surmountable, they need to be taken seriously if, the extreme case where *anyone can say anything about anything* is subject to be treated on equal grounds as a piece of information that is well certified. One contemporary and popular example of this scenario is where nearly all scientists being in agreement on the presence of global warming, climate change, and its consequences. In order to honour a commitment to public service, scientific communication – independent to self-registration – must remain grounded on reproducible and replicable facts and explanations, as well as pertain attributable and accountable participants. The output affects society, from journalism, advancement of scientific knowledge, policy making, as well public opinion and perception.

Decentralised authoring, publication, and annotation furthermore have the potential to impact areas in which centralised services currently determine the pace of evolution. Scientific publishing, for instance, is often bound to centralised review and dissemination processes. Instead, rigorous scientific discourse can still be realised with an open, decentralised environment for the annotation of manuscripts, which can potentially reach interested parties in a timely manner. Trust then no longer stems from a finite process with limited transparency, but is rather continuously assessed by repeated independent validation. Publication thereby becomes the starting point rather than the end point of the scholarly communication process. An application built with interoperable open Web standards can potentially communicate with services that conform to the same protocols and data models.

Non-Functional Requirements: While there are plethora of *architecturally significant requirements* such as institutional or orientations in society and industry, I focus on non-functional and functional requirements which enable personal and academic freedom, in particular to creating affordances towards autonomy and interoperable participation on the open Web. I derive non-functional requirements for systems to aim at addressing the forces and functions in scientific communication through the following:

- *Interoperability* to improve discovery, accessibility, integrability, and reusability.
- *Security and privacy* to honour desire to remain safe and information not to be shared without consent.
- *Modularity and extensibility* to enable evolvable and reconfigurable systems.
- *Data integrity, transparency, and reliability* towards building trust.
- *Persistence and preservation* to ensure that units of scholarly information preserves into the future.

In order to support publishers and consumers (in the most general sense of *actor* roles), it is important to seek system designs that give control to its users at various fronts. This is to foster a scholarly ecosystem where users having the means to participate:

- by creating and using their preferred identifiers for identities and data;
- by deciding what can access data and under which conditions;
- by using different applications on the same data for different purposes.

5. Linked Data Notifications

> **JANEWAY**
> To be honest, we're having a little trouble understanding your technology.
>
> **ABARCA**
> The problem is your technology. Interference from your antiquated scanning devices shut down our propulsion system.
>
> **TINCOO**
> Forgive us, Captain. We live in a closed system. We are not accustomed to interacting with other species, especially inferior ones.
>
> **JANEWAY**
> Well then, I guess we'll leave the repairs in your superior hands.
>
> *Virtuoso, Star Trek: Voyager* [423], 2000

In this section I describe my knowledge and artifact contributions of a decentralised notification protocol. Parts of this section are adapted from work published as *Linked Data Notifications: a resource-centric communication protocol* [424], Capadisli, 2017, *Extended Semantic Web Conference* [425] (Best Student Paper Award). Additionally some parts are adapted from *Linked Specifications, Test Suites, and Implementation Reports* [426], Capadisli, 2018, TheWebConf Developers Track. Finally, parts of the W3C Recommendation *Linked Data Notifications* [427], editors Capadisli and Guy, 2017, have been integrated as well.

5.1 Fulfilling the Awareness Function

As scholarly content is created and registered, becoming globally identifiable on the Web, the next challenge is to ensure it is *findable* and *discoverable* within a decentralised system.

I have reviewed a number of Web standards and specifications pertaining to read-write Linked Data, vocabularies about publishing resources and social participation, and explained how they can be combined in various ways to realise a wide array of scholarly and social activities in a decentralised fashion. This section describes the design and standardisation of a decentralised notification system that works on the HTTP application layer.

Notifications are sent over the Web for a variety of purposes, including social applications: “You have been invited to a vegan lunch!”, “Guinan annotated your article”, “Eunice tagged you in a photo”. The notification data may be displayed to a human to acknowledge, or used to trigger some other application-specific process (or both). Such notifications are also useful as part of a decentralised scholarly communications ecosystem, primarily for fulfilling the *awareness* function.

Notifications are a common part of many centralised Web applications, and variety of uses are in social media. In this case, notifications are structured arbitrarily and are typically only usable by the sanctioned applications or those that generated them in the first place. So far, notifications are *ephemeral* resources which may disappear after delivery, and thus are not designed for reuse. In such systems, communication and semantic interoperability across independently built applications is not expected. This design however places limitations on possible interactions around a notification. For example, current major social media services on the Web act as disjoint closed ecosystems, commonly known as “walled gardens” in contrast to open platforms with more possibilities for extensibility and reuse of applications and content. In essence, notification data is locked into particular systems, and users are restricted to using tightly coupled applications. Such software and systems are only capable of exchanging notifications if they have prior knowledge about the actors involved, location of the data, data formats, and user interfaces. Customarily, switching applications entails switching social providers, and having to create new profiles, and yet again reconnect or build one’s social network in the new system from scratch. This scenario is analogous to using an email application that’s customised to each service provider, and only being able to send and receive emails from people that use the same provider. In order for information to flow between systems customised communication interfaces or APIs need to be created after the fact, thus imposing additional maintenance tax for systems to cooperate.

In a decentralised architecture, notifications can be a key element for federation of information, and application integration. Current efforts towards *re-decentralising* the Web are moving towards architectures in which data storage is decoupled from application logic, freeing end users to switch between applications, or to let multiple applications operate over the same data.

Similarly for a decentralised scholarly Web, notification data should not be locked into particular systems. Data generated from independently built applications should be discoverable and increase their chance for reusability by other applications. Users can choose where to store their notifications as well as using their preferred applications to create, discover, and consume the data.

This section addresses the following research question for this thesis:

> What technical mechanisms, standards or protocols can be employed for decentralised information exchange on the Web?
>
> *Mechanisms*

The following are core requirements of a notifications protocol in context of *Degree of Control*:

- Actors can use their preferred applications to discover, reuse, and send notifications.
- Actors store incoming notifications where they prefer.
- Actors switching between applications without having to move their data.
- Actors change the location of their data without having to change their application.

In the following subsections I review existing decentralised protocols for notifications, and then outline use cases for notifications in scholarly communication. These use cases yield key design considerations, which are taken into account for the design of a new protocol – Linked Data Notifications (LDN) – which can be implemented to meet the awareness function for scholarly communication in decentralised systems, as well as meet the degree of control requirements listed above. LDN also supports systems with respect to the registration, certification and archiving functions.

I developed LDN and served as the specification editor as part of the W3C Social Web Working Group (SWWG), and LDN became a W3C Recommendation in May 2017. This section outlines key features and discusses design decisions made in the context of scholarly communication; for reference the specification in its entirety can be found at https://www.w3.org/TR/ldn/.

5.2 Overview of Web Notification Systems

Many systems which make use of notifications operate either in a completely centralised way, or are decentralised only in the sense that different instances of the *same* codebase need to interoperate; this review is restricted to mechanisms which do not expect the notification to be received or used only by the same software or platform which sent it.

Existing decentralised notification mechanisms can be grouped according to the notification contents. Notification contents are either: 1) One or more URLs, indicating relations between Web resources (aka a ‘thin ping’), or 2) a ‘fat ping’ containing a blob of information. Semantic Pingback, Webmention, and Provenance Pingback follow the first form, and are also known as *linkbacks*, the suite of protocols that essentially allows Web documents to automatically reciprocate hyperlinks. This has the advantage that a verification mechanism can be tightly specified (the URL of the target must appear in the content of the source), but the disadvantage that notifications are generally suitable for website publishing. Fat pings, in contrast, are flexible and extensible in that payloads can be in any content type, carry any information, and be used for various use cases.

Semantic Pingback [428] and *Webmention* [429] are protocols to send and receive notifications when a relationship is created, updated, or deleted between two URLs. They both update the original *Pingback* [430] mechanism by replacing the XML-RPC transport mechanism by a `x-www-form-urlencoded` request with two parameters (`source` and `target`). Resources which are the target for a notification advertise the respective receiving service or endpoint via a `Link` relation, either in HTTP headers or HTML. Semantic Pingback additionally enables discovery of the Pingback service where target description is available in RDF. While the content at source may indicate (in any convention or serialisation format) the type of relation between the source and target URLs, this information about the relation is not transmitted to the receiver's endpoint; only the source and target URLs are sent. As such, there is also no way to deterministically distinguish between multiple mentions of the target at the source based on the information in the mention. Hence, each receiver implementation is responsible for interpreting the underlying information at the source, thereby potentially inconsistent across implementations. In contrast, Semantic Pingback does encourage generation of additional semantics about the relation(s) between the source and the target by processing the source as RDF if possible, and also defines specific ways for a receiving server to handle incoming pingback data in order to add the source data to an RDF knowledge base. Beyond verifying that the source contains the URL of the target, Webmention does not specify any further requirements of the receiving server; nor is it expected that "mentions" are retrievable once they have been sent.

A *Provenance Pingback* [431] endpoint is also advertised via the HTTP `Link` header; it accepts a list of URIs for provenance records describing uses of the resource. Provenance Pingback does not specify any further behaviour by the receiving server, but the contents at the URIs listed in the notification body must be semantic data.

Other notification mechanisms send more information than just URLs in the notification body in the form of fat pings. With the exception of WebSub, due to each mechanism's focused use case, the payload of notifications is restricted to a particular vocabulary.

WebSub [432], previously known as *PubSubHubbub* [433] (PuSH), is an implementation of a publish-subscribe messaging pattern for content changes. In WebSub, content publishers delegate subscriptions to and distribution of new and updated content to a hub, which acts as a broker service. Subscribers send a request with a topic of interest – the URL of a resource – to a hub and specify a location where content can be delivered to. When a publisher informs a hub about content changes, a hub then handles the distribution by sending a copy of the content that a publisher makes available to the subscriber's preferred location. The distributed content may be any arbitrary content type. Subscribers only indicate that they have received hub's message or to retry later.

DSNotify [434] is a centralised service which crawls datasets and observes changes to links with the specific use case of preserving link integrity between "Linked Open Data" resources. Third-party applications can register with the sending service to receive notifications of changes in the form of a specific XML payload.

With the *sparqlPuSH* [435] service, users may input a SPARQL query, the results of which are the specific updates they are interested in. The query is run periodically by the service, and the results are converted to RSS and Atom feeds, which is sent to a PuSH hub to which the user can subscribe.

The *ResourceSync Change Notification* [436] specification also sends update notifications via a PuSH hub, with an XML payload based on the Sitemap format.

Each of these mechanisms are triggered by subscription requests. That is, a user must actively solicit messages from a particular service, rather than having a way for a service to select a notification target and autonomously discover where to send notifications to.

We present and discuss a *Comparison of Notification Mechanisms* to better understand current limitations, and how the proposed protocol addresses them and at what cost.

5.3 Use Cases for Decentralised Notifications

In this section I refer and outline commonly known user stories as the basis to developing social and Linked Data based protocols. I also describe common scenarios in scholarly communication.

In 2015, the *W3C Social Web Working Group* [437] documented *user stories* [438] in online social media in the process of producing technical specifications to enable decentralised and federated social communication. The user stories were derived from existing open and proprietary platforms on the Web, including notifications about events, certifications, annotations, requests, and announcements. For example, the following user stories concern notifications: “user posts a note”, “reading a user’s recent posts”, “following a person”, “adding comments to bespoke software”, “direct messaging”.

The *Linked Data Platform Use Cases and Requirements* [353] (LDP-UCR) documents “Aggregation and Mashups of Infrastructure Data” as one of the user stories and “Aggregate resources” as the derived use case.

In addition to general social Web activities, researchers and scholars, as well as software on their behalf (actors) perform discipline-specific activities to send (registration), discover (awareness), and reuse public and private notifications (accessibility, content). I describe some scenarios in scholarly communication which are facilitated by notifications:

Call for contributions
Scenario: A conference’s organising committee would like to make a public announcement that they are now accepting research contributions that can be peer reviewed.

Publicising an article
Scenario: One of the authors of an article would like to let their colleagues know about their publication.

Request for review
Scenario: The authors of an article requests their work to be reviewed by the members of their community.

Annotating artifacts
Scenario: A reviewer or commenter wants to let the authors of an article know about their feedback.

Authors notify referenced article
Scenario: In order to promote the relevancy of their article and its discovery, authors would like to announce that their article cited another article.

Social actions and reactions
Scenario: Researchers perform public and private actions on resources they come across, including bookmarking, re-sharing with others, reactions such as “like“, as well as following another researchers’ activities.

Providing research information
Scenario: A researcher wants to inform the scientific community that the research methods and data sources of their work is publicly accessible.

State change
Scenario: As automated software detects changes in datasets, new version of an application, anomalies in sensor readings or experimental observations, it informs the laboratory.

Status update
Scenario: Researchers would like to inform indexing and archiving systems about new or updates to existing artifacts so that they can perform their specialised operations.

Artifacts of interest
Scenario: A recommender system, upon coming across articles and annotations that may be of interest to a researcher, wants to inform them based on their listed interests (research results, disputes over data analysis, new relations between research objects).

Requesting access
Scenario: When one of the authors does not have write-access to an article, they want to make a request to their peers to have the required permissions.

A useful decentralised notification protocol would be able to realise some of these use cases for scholars. In the next section I describe specific design considerations derived from these uses cases, which can be used to inform the design of the protocol.

5.4 Design Considerations

This section details design considerations for a decentralised notification protocol are based on the *Use Cases for Decentralised Notifications*. I will take conformance to the Linked Data design principles into account, as well as Web application best practices. I use these considerations to establish the concrete requirements and the points of implementation-specific flexibility for the protocol.

The non-functional features include: modularity, reusability, persistence and retrievability, adaptability. The functional features include: target representation, notification body, notification verification, subscribing.

Modularity (R1): To encourage modularity of applications, one should differentiate between different classes of implementation of the protocol. Two parties are involved in the creation of a notification: a *sender*, generating the notification data, and a *receiver*, storing the created resource. We also have the role of a *consumer*, which reads the notification data made available by a receiver, and repurposes it in some way. A software implementation can of course play two or all three of these roles; the important part is that it need not. A consuming application can read and use notification data without being concerned about ever sending or storing notifications.

Reusability (R2): The relationship between the *consumer* and *receiver* roles is key to notifications being reusable. A consumer must be able to autonomously find the location of notifications for or about the particular resource it is interested in. To achieve this, we place a requirement on the receiver to expose notifications it has been sent in such away to permit consumer applications to access them; and specify how any resource can advertise its receiving endpoint for consumers to discover. To promote fair use or remixing of notification contents, applications can incorporate rights and licensing information into the data. Similarly, applications may include additional information on licensing resources that the notification refers to. The presence of this type of information is important for consumers to assess the (re)usability of data.

Persistence and Retrievability (R3): Given the traditionally ephemeral nature of notifications, here we refer to persistence as units of information that can be registered. Applications may benefit from referring to or reusing notifications if the notifications are known to be retrievable in the long term, or indicate their expected lifespan. This enforces the notion that notifications are considered resources in their own right, with their own dereferencable URIs.

Adaptability (R4): A notification protocol should be adaptable for different domains, but that there is no need to create multiple domain-specific notification protocols; the fundamental mechanics are the same given the three main architectural layers of the Web – identification, interaction, data formats.

Target representation (R4-A): Any resource may be the *target* of a notification. By target, we mean a notification may be addressed *to* the resource, be *about* the resource, or for a sender to otherwise decide that it is appropriate to draw the attention of the resource (or resource owner) to the information in the notification body. As such, any Web resource must be able to advertise an endpoint to which it can receive notifications. Resources can be RDF or non-RDF (such as an image, or CSV dataset), and may be informational (a blog post, a user profile) or non-informational (a person).

Notification body (R4-B): We consider the *contents* of a notification to be application specific because different domains have different needs. From a sender's perspective, we permit a notification to contain *any data*, and a notification can use *any vocabulary*. From a consumer's perspective, interoperability between different applications occurs through vocabulary reuse, and shared understanding of terms. This is in accordance with Linked Data principles in general. The practical upshot of this is that a calendar application which consumes event invitations using the *RDF Calendar* [439] vocabulary is likely to completely ignore notifications containing the *PROV Ontology* [440], even if it finds them all stored in the same place. For two independent applications operating in the *same* domain, a shared understanding of appropriate vocabulary terms is assumed.

Notification verification (R4-C): From a receiver's perspective, exposing itself to receive any blobs of RDF data from unknown senders may be problematic. Thus, it should be possible for the receiver to enforce restrictions and accept only notifications that are acceptable according to its own criteria (deemed by eg. user configuration; domain-specific receivers). This can be used as an anti-spam measure, a security protection, or for attaining application and data integrity.

Subscribing (R5): In general, applications may require that new notifications are pushed to them in real-time, or to request them at appropriate intervals. To take this into account, we expand our definition of senders, receivers and consumers with the following interaction expectations: notifications are *pushed* from senders to receivers; and *pulled* from receivers by consumers.

The table below shows the design considerations as applicable to each use case.

Table 4. Design considerations derived from use cases for decentralised notifications

	R1	R2	R3	R4-A	R4-B	R4-C	R5
Call for contributions	✔	✔	✔	✔	✔		
Publicising an article	✔	✔	✔	✔	✔	✔	
Request for review	✔	✔	✔	✔	✔		
Annotating artifacts	✔	✔	✔	✔	✔	✔	
Authors notify referenced article	✔	✔	✔	✔	✔	✔	
Social actions and reactions	✔	✔	✔	✔	✔	✔	✔
Providing research information	✔	✔	✔	✔	✔	✔	
State change	✔	✔	✔	✔	✔	✔	✔
Status update	✔	✔	✔	✔	✔	✔	
Artifacts of interest	✔	✔	✔	✔	✔	✔	✔
Requesting access	✔	✔	✔	✔	✔	✔	

5.5 A Decentralised Notifications Protocol

Linked Data Notifications (LDN) is an HTTP-based protocol to facilitate exchanging messages between applications. LDN describes how servers (receivers) can receive messages pushed to them by applications (senders), as well as how other applications (consumers) may retrieve those messages. Any resource (target) can advertise a receiving endpoint (Inbox) for notification messages. Messages are expressed in RDF, and can contain arbitrary data. LDN is a *W3C Recommendation* [441].

LDN is defined using the following basic principles:

- The protocol is developed within the framework laid out by the AWWW.
- Interactions are designed to take place over HTTP.
- Any Web resource may be the target of a notification.
- Notifications may contain any data and use any Linked Data vocabulary.

Inbox
An inbox is a container or directory (attached to a Web resource) which is used to store and serve a collection of notifications.

Notification
A notification is a retrievable resource in an inbox which returns content expressed in RDF. The contents of notifications are intended to describe a change in state of some other resource, or contain new information for the attention of a user or process, and may be subject to constraints of the Inbox it is contained in.

We designed the LDN protocol to support sharing and reuse of notifications *across* applications, regardless of how they were generated or what their contents are. I describe how the principles of identification, addressability and semantic representation can be applied to notifications on the Web. Specifying LDN as a formal protocol allows independently implemented, heterogeneous applications which generate and use notifications, to seamlessly work together. Thus, LDN supports the decentralisation of the Web as well as encourages the generation and consumption of Linked Data.

LDN is built on existing W3C standards and Linked Data principles. In particular, the storage of notifications is compatible with the *Linked Data Platform* specification; notifications are identified by HTTP URIs; and notification contents are available as JSON-LD (or other RDF syntaxes as per content-negotiation). A key architectural decision is the separation of concerns between *senders*, *receivers*, and *consumers* of notifications. Implementations of the protocol can play one or more of these roles, and interoperate successfully with implementations playing the complementary roles. This means that notifications generated by one application can be reused by a completely different application, accessed via the store where the notification data resides, through shared Linked Data vocabularies. LDN also pushes the decentralised approach further by allowing any *target* resource to advertise its Inbox anywhere on the Web; that is, targets do not need to be coupled with or controlled by a receiver operating from the same server, and can make use of a third-party *inbox-as-a-service*.

Beyond a generic understanding of hypermedia, LDN senders and consumers need little to no prior knowledge about how to interact with receivers. Furthermore, senders and consumers determine a target’s inbox location dynamically. This design approach is aligned with the notion of self-describing resources in that applications are only required to identify a particular relation indicating the inbox location, and perform a *follow your nose* type of exploration. Given that a target’s preferred inbox location may change at their discretion, applications can be adaptive in inbox discovery.

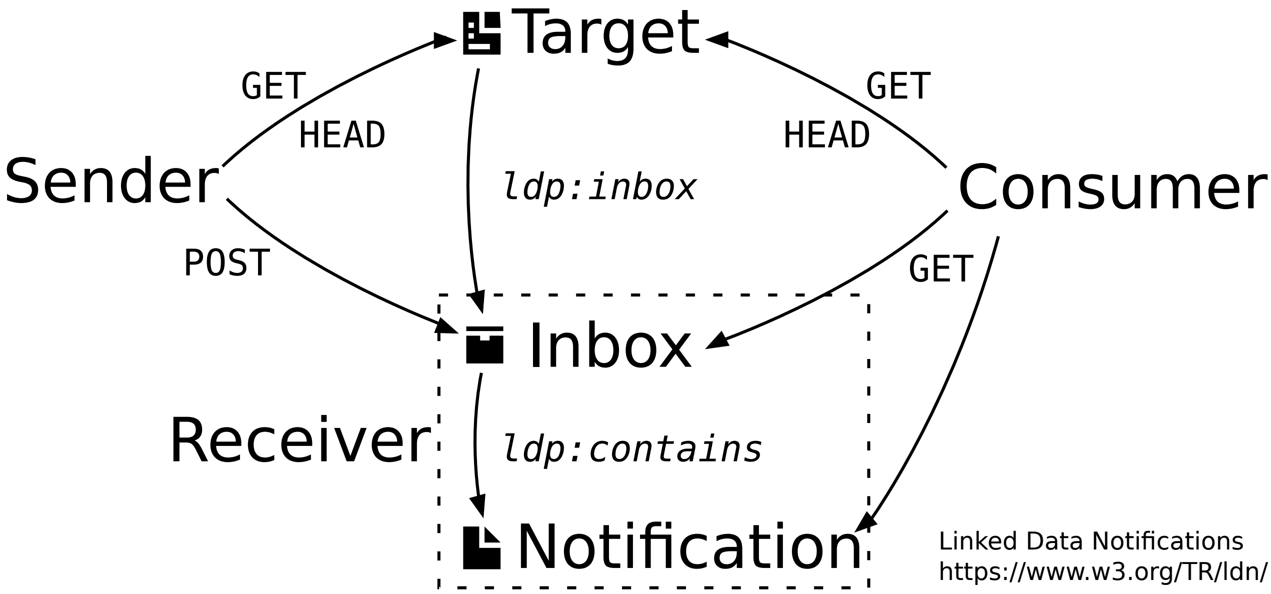


Figure 5. *Overview of Linked Data Notifications* [442].

5.5.1 Application Interactions

Building on the notion of modularity, we consider that *interoperability* occurs between two classes of interactions, senders and receivers, or between consumers and receivers, when the sender/consumer makes a request to the receiver, and the receiver sends the expected response.

5.5.1.1 Sender to Receiver interactions

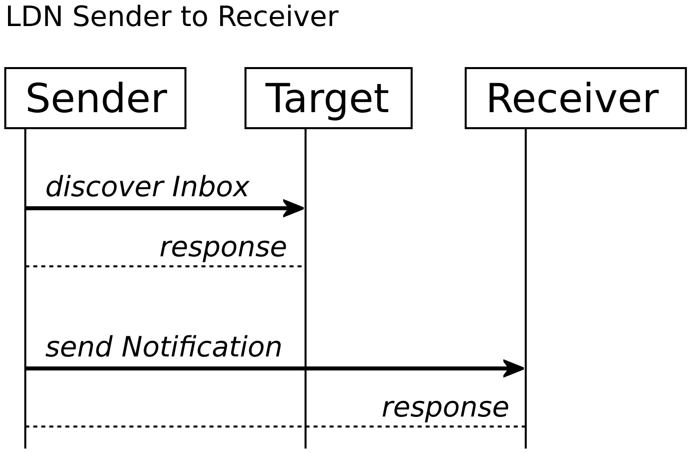


Figure 6. Linked Data Notifications Sender to Receiver interaction.

The following steps (in order without skipping) describe the interaction between sender and receiver:

1. A sender is triggered, either by a human or an automatic process, to deliver a notification;
2. The sender chooses a target resource to send notifications to;
3. The sender discovers the location of the target's *Inbox* through the `ldp:inbox` relation in the HTTP `Link` header or RDF body of the target resource;
4. The sender creates the body of the notification according to the needs of application;
5. The sender makes a `POST` to the Inbox URL, containing the body in JSON-LD or in another serialisation acceptable by the server;
6. The receiver optionally applies filtering rules, and sends the appropriate HTTP response code to accept or reject the notification;

7. The receiver exposes the notification data (according to appropriate access control) for use by consumers.

5.5.1.2 Consumer to Receiver interactions

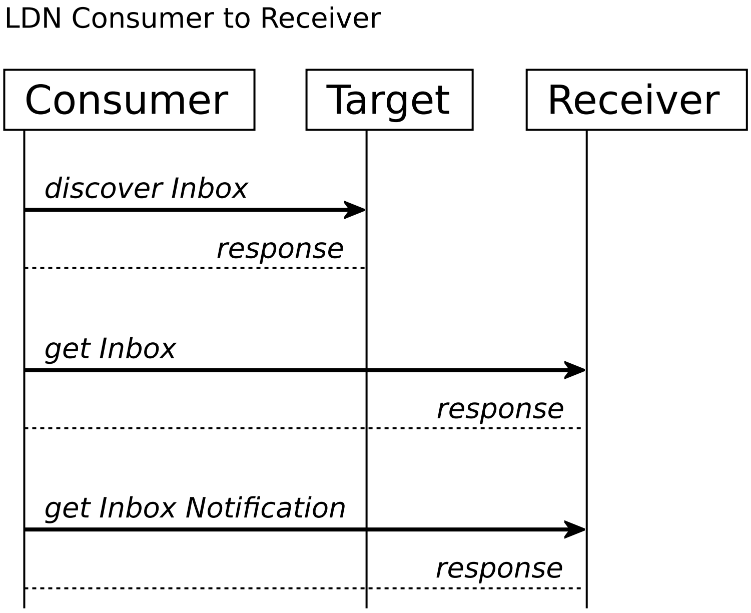


Figure 7. Linked Data Notifications Consumer to Receiver interaction.

The following steps (in order without skipping) describe the interaction between consumer and receiver:

1. A consumer selects a target and discovers the location of its Inbox in the same way as the sender;
2. A receiver responds to HTTP `GET` requests made to the Inbox URL with a listing of the URLs of notifications that have previously been accepted, linked to the Inbox with the `ldp:contains` predicate;
3. The receiver responds to HTTP `GET` requests made to the individual notification URLs with JSON-LD (or optionally other serialisations);
4. Following the retrieval of notification listings or individual notifications, the consumer may perform further processing, combine with some other data, or simply present the results in a suitable human-readable way.

5.5.2 Data Formats and Content Negotiation

The data formats which are used by a notification specification must be chosen to maximise interoperability between implementations. *Structure of Scholarly Information* makes the argument for using RDF as the language to express human- and machine-readable information on the Web. Allowing LDN actors – senders, receivers, and consumers – to handle data irrespective of the particular RDF serialisation permits flexibility, however it can also be costly to support. We take into account:

- *Application interoperability* eg. should applications support all current RDF formats, as well as new versions or formats in the future?
- *Maintenance of RDF parsers and serialisation libraries* eg. will the RDF libraries that applications use be kept up to date?
- *Complexity of their inclusion in applications* eg. what are the costs or concerns of using all RDF formats?
- *Run-time efficiency* eg. how well will applications perform going forward?

Why JSON-LD: To address these issues, choosing a single RDF serialisation to *require* is necessary for consistent interoperability, as well as keeping processing requirements or external code dependencies minimal. While any RDF notation would satisfy that need, we decided that LDN requires all applications to create and understand the JSON-LD syntax, both for the contents of Inbox as well as for individual notifications.

To some extent, JSON-LD is compatible with existing JSON libraries or in some cases native programming language data structures. For example, Web browsers have a built-in interpreter for JavaScript that can handle JSON objects, that can take advantage of native object operations. Same data processing and manipulation operations are also available in a Node.js environment. Having said that, JSON-LD is a format that serves to be an RDF notation, and can also be processed as plain JSON. Hence, JSON-LD is just one convenience to certain kinds of environments that the application is confined to. JSON-LD is also advantageous in being familiar for developers who are *used to JSON-based APIs but not RDF* [443].

Content negotiation: Optionally, applications may attempt to exchange different RDF serialisations by performing **content negotiation**. As in compliance with HTTP principles, a sender may make an HTTP `OPTIONS` request to the receiver to determine the RDF content types accepted by the server. A receiver can expose `Accept-Post` headers for senders, and so the sender can serialize the notification in the request body accordingly. Consumers can send `Accept` headers to receivers to signal which RDF content types they prefer and capable of processing.

By mandating one RDF syntax (JSON-LD) for applications to communicate with, as well as permitting other syntaxes to discover and exchange notifications, compatibility across applications is assured.

5.5.3 Security, Privacy and Content Considerations

Target ownership: As per AWWW's *URI Ownership*, the Inbox is ultimately controlled by the owner of LDN Receiver, and subject to third-party access to HTTP headers and content. Hence, publishers of the resources advertising an Inbox (target) are expected to be aware of using servers they trust or control for the Inbox location.

Constraints: One way to filter unwarranted notifications from being created on the server and exposed is where Inbox URLs announce their own constraints eg. *Shapes Constraint Language (SHACL)* [444], *ShEx* [445], Web Annotation Protocol, via an HTTP `Link` header or body of the resource with a `rel` value of `ldp:constrainedBy`. For example, an LDN receiver may want to allow notifications with a certain shape; a model specifying allowed and required values in a notification. This is so that senders can comply with the advertised constraint specification or the receiver may reject their notification. Rejecting notifications which do not match a specific pattern in their contents, or the *shape* of the data, is one way to filter. For example, if the Inbox owner knows that they will only ever use a consuming application which processes friend requests, they can configure their receiver to filter out anything that does not match the pattern for a friend request, helping their consumer to be more efficient. If the notification constraints are also advertised by the receiving service as structured descriptions, generation and consumption of the notifications can be further automated. Constraints are particularly useful towards providing notifications that meets a certain criteria so that it can be used effectively by consuming applications.

Authenticated inboxes: The requirement for reusable notifications could be seen as a potential risk for privacy. The LDN Receivers are expected to consider implementing access control on the Inbox URL as well as the individual notification URL and may restrict reading and writing to a whitelist of trusted senders. Various authentication methods which could be used alongside LDN with the idea that notifications are not necessarily public but only visible or reusable by intended parties. The receivers decide (based on their use case) which consumers (based on any criteria) can reuse. They achieve this by setting authentication and authorization settings on the notifications. As different authentication mechanisms are appropriate for different applications, the notification protocol should ideally be usable alongside various methods such as clientside certificates, eg. WebID+TLS, token-based, eg. OAuth 2.0, or digital signatures.

Personally identifiable information: Notification payloads may contain any data including that which identifies the sender or the receiver. As access to the Inbox and notification data is under the control of the receiver, they ultimately determine what information may be exposed to the world. Once fetched by a third-party, any piece of information could potentially be subject to further (unauthorised) distribution and reuse. Hence, this is orthogonal to whether the notifications are ephemeral or persistent.

Security and Privacy Review: The LDN specification includes a *Security and Privacy Review* [446] covering threat models categorised as: “Passive Network Attackers, Active Network Attackers, Same-Origin Policy Violations, Third-Party Tracking” as outlined in W3C *Self-Review Questionnaire: Security and Privacy* [447].

5.5.4 Protocol Interaction and Content

This section demonstrates some interactions between a server with a target resource, and among sender, receiver, and consumer implementations.

Discovering an inbox:

```
GET / HTTP/1.1
Host: csarven.ca
Accept: application/ld+json

HTTP/1.1 200 OK
Content-Type: application/ld+json

{
  "@context": "http://www.w3.org/ns/ldp",
  "@id": "https://csarven.ca/#i",
  "inbox": "https://csarven.ca/inbox/"
}
```

Listing 8. Discovering an Inbox with a HTTP GET request to retrieve JSON-LD. Response in JSON-LD compact form.

Consumer requesting the inbox and receiver’s response:

```
GET /inbox/ HTTP/1.1
Host: example.org
Accept: application/ld+json
Accept-Language: en-GB,en;q=0.8, en-US;q=0.6

HTTP/1.1 200 OK
Content-Type: application/ld+json
Content-Language: en

{
  "@context": "http://www.w3.org/ns/ldp",
  "@id": "http://example.org/inbox/",
  "contains": [
    "http://example.org/inbox/5c6ca040",
    "http://example.org/inbox/92d72f00"
  ]
}
```

Listing 9. Receiver responding to a HTTP GET request on the Inbox with a listing of notifications.

Consumer getting a notification:

```
GET /inbox/14a792f0 HTTP/1.1
Host: example.org
Accept: application/ld+json, text/turtle, application/xhtml+xml, text/html
Accept-Language: en-GB,en;q=0.8, en-US;q=0.6

HTTP/1.1 200 OK
Content-Type: application/ld+json;profile="https://www.w3.org/ns/activitystreams"
Content-Language: en

{
  "@context": [
    "https://www.w3.org/ns/activitystreams",
    { "@language": "en" }
  ],
  "@id": "http://example.org/inbox/14a792f0",
  "@type": "Announce",
  "actor": {
    "@id": "https://csarven.ca/#i",
    "name": "Sarven Capadisli"
  },
  "object": {
    "@context": "http://www.w3.org/ns/anno.jsonld",
    "@id": "http://example.net/note",
    "@type": "Annotation",
    "motivation": "http://www.w3.org/ns/oa#assessing",
    "rights": "http://creativecommons.org/licenses/by/4.0/"
  },
  "target": "http://example.org/article",
  "updated": {
    "@type": "http://www.w3.org/2001/XMLSchema#dateTime",
    "@value": "2016-06-28T19:56:20.114Z"
  }
}
```

Listing 10. Result of a GET request on an individual notification discovered in an Inbox.

Sending a notification:

```
POST /inbox/ HTTP/1.1
Host: example.org
Content-Type: application/ld+json;profile="https://www.w3.org/ns/activitystreams"
Content-Language: en

{
  "@context": "https://www.w3.org/ns/activitystreams",
  "@id": "",
  "@type": "Announce",
  "actor": "https://rhiaro.co.uk/#me",
  "object": "http://example.net/note",
  "target": "http://example.org/article",
  "updated": "2016-06-28T19:56:20.114Z"
}
```

Listing 11. Example request to send to an Inbox.

Response to send request:

```
HTTP/1.1 201 Created
Location: http://example.org/inbox/5c6ca040
```

Listing 12. Example response to a POST request on an Inbox.

Sender initiates content-negotiation and sends notifications:

```
OPTIONS /inbox/ HTTP/1.1
Host: example.org

HTTP/1.1 200 OK
Allow: GET, HEAD, OPTIONS, POST
Accept-Post: application/ld+json, text/turtle

POST /inbox/ HTTP/1.1
Host: example.org
Content-Type: text/turtle
Content-Language: en

@prefix as: <https://www.w3.org/ns/activitystreams#> .
@prefix cito: <http://purl.org/spar/cito/> .
<> a as:Announce
  as:object <https://linkedresearch.org/resources#r-903b83> ;
  as:target <https://csarven.ca/dokieli-rww#architecture-and-technologies> .
<https://linkedresearch.org/resources#r-903b83>
  cito:citesAsPotentialReading
  <https://csarven.ca/linked-data-notifications#protocol> .
```

Listing 13. Sender initiated content-negotiation, followed with an announcement of a specific citation relation between two entities (Turtle).

5.6 Test Suite and Implementation Reports

The W3C “Technical Reports” are intended to help different (current or future) implementations to have common core functionality and quality, comply with accessibility and internationalisation guidelines, and take security and privacy considerations into account. When an application, for instance, implements a specification, it can be checked against that specification’s conformance criteria for normative requirements. Specifications are typically accompanied with test suites to assist implementations to identify their conformance level as well as areas for improvement. Similarly, reports and feedback help specifications to improve and advance towards publication. So, specifications and conforming implementations are integral to ensuring interoperable applications.

As a concrete example of the benefits of resource-centric notifications, I took the liberty to both use the W3C process of creating a test suite for LDN, and the submission of reports about implementations to exemplify the LDN protocol itself, as well as to generate discoverable Linked Data about the specification and its implementations. The LDN *test suite* [448] is itself an LDN implementation for the purpose of automating the collection and aggregation of implementation reports which were used directly towards the formal standardisation process.

The LDN test suite covers each requirement of the specification with individual tests, and the test suite documentation is semantically linked with the specification itself accordingly. Once the tests have been run, the test suite generates a similarly linked implementation report. This report is submitted using LDN itself as the notification mechanism. The listing of implementation reports (software conforming to the specification) is automatically updated and are accessible as their own registered units of information.

The LDN test suite puts the LDN protocol into practice by acting as an LDN receiver implementation (based on *mayktso* [449]). It also acts as a sender and consumer LDN implementation. Each part of the test suite (for Senders, Receivers, and Consumers) advertise an `ldp:inbox`. Upon completion of a run of the tests, the system generates the report data and sends an LDN notification to the Inbox. The payload of the notification is the full report as RDF.

As an LDN Consumer, the test suite generates the summary of the reports by fetching and processing Inbox contents. The notifications are aggregated automatically, and the semantics of the submitted reports are retained.

The services are decoupled; that is, an implementer may generate their report independently of the test suite, and submit it vial the standard LDN protocol. Furthermore, projects can implement their own consumers and reuse the report data generated by the test suite directly, for example to demonstrate to potential users their conformance to the LDN specification.

All reports have their own URLs, and a human-readable representation in HTML+RDFa, in addition to JSON-LD or other RDF syntax per `HTTP OPTIONS` and `Accept-Post` negotiation at this particular test server. See for example *dokieli* [450]’s implementation report and test results as a sender. This provides the human-visible information, eg. what was tested and the results also in machine-readable form. The report can be seen as a dataset composed of observations based on the structure that was specified in the specification. Hence, each test report is a `qb:DataSet` (and generally equivalent in `as:Object`) where its `qb:structure` refers to `https://www.w3.org/TR/ldn/#data-structure-definition`. The dataset has `as:published` and `as:actor` for the agent that initiated the test and generated the report. The report may be accompanied with an additional `as:summary`. An example report at https://linkedresearch.org/ldn/tests/reports/2c5af2f0-f832-11e6-a642-0dd857219753 has the following core information:

```
<>
  a qb:DataSet ;
  qb:structure ldn:data-structure-definition .

<https://linkedresearch.org/ldn/tests/reports/2c5af2f0-f832-11e6-
a642-0dd857219753#test-sender-header-post-content-type-json-ld>
  a qb:Observation, earl:Assertion ;
  qb:dataSet <> ;
  earl:subject <https://dokie.li/> ;
  earl:test ldn:test-sender-header-post-content-type-json-ld ;
  earl:mode earl:automatic ;
  earl:result [
    a earl:TestResult ;
    earl:outcome earl:passed ;
    earl:info "<code>Content-Type: application/ld+json; profile=&quot;
http://www.w3.org/ns/anno.jsonld&quot;</code> received."^^rdf:HTML ] .
```

Listing 14. A snippet of a test result dataset and an observation in Turtle syntax.

The test results are provided in an HTML table, where each test is expressed as an `qb:Observation` (and equivalent `earl:Assertion`) in RDFa containing:

- a `earl:subject` that refers to the URI of the application, eg. dokieli, a `doap:Project` as an LDN Sender.
- a `earl:test` with the range being one of the requirements (concepts) from the specification.
- a `earl:mode` referring to one of the EARL test modes that were carried out: automatic, manual, semi-automatic, undisclosed, unknown.
- and a `earl:result` that gives information on the test `earl:outcome`: passed, failed, inapplicable, cannot tell, untested, as well as detailed `earl:info` about the particular experiment.

The implementation test report has some basic information linking to the `doap:Project` with a `doap:name`, and its `doap:maintainer`.

All of the sender, receiver, and consumer reports are available in separate aggregate tables in *LDN Tests Summary* [451]. The summary is a `void:Dataset` where each report is linked as a `void:subset`. This makes individual reports alternatively findable if the exploration starts from the summary of all test results.

The test suite software fulfills the base required functionality of the LDN protocol. Next, I take a closer look at the LDN implementations. Later I describe further details on how the LDN specification and implementation reports are interlinked and retrievable to facilitate their automated discovery and reuse in *Linked Research*.

5.7 Linked Data Notifications Implementations

In order for the LDN specification to advance to W3C Proposed Recommendation status, an *Exit Criteria* [452] outlined the requirements for implementations. At least two independent, interoperable implementations of each feature had to be fulfilled. Each feature was implemented by a different set of products. There was no requirement that all features needed to be implemented by a single product.

In order to verify the conformance and support of the LDN features in implementations, the *LDN Test Suite* [453] was development. Individual *LDN Test Reports* [454] as well as the *LDN Test Reports and Summary* [455] of all reports were made public using CC BY license. *LDN Test Reports and Summary* [456] is documented and readable by human and machine consumers. The *Receiver*, *Consumer*, and *Sender* reports listed below are snapshots of the tests summary data as of 2019-01-05.

Once an implementation has passed the relevant tests, an implementation report was submitted to help enable the specification progress to W3C Recommendation. Implementations were also submitted after the Recommendation, and remain open to accept new implementation reports. Running the tests generated a checklist, filled in where possible, to indicate the outcome of the tests with additional information, and thus indicating the features implemented according to the specification.

Test outcomes

It returns a passed/failed response for individual requirements of the LDN specification. It also tests some optional features; you will get an inapplicable response if you do not implement them, rather than a fail. Some of the test outcomes will require manual checking, hence they will be marked with cannot tell. If a test was skipped or no value provided, it will be marked as untested.

So far we have seen decentralised notifications applied in social networking scenarios, as well as for archival activities and scientific experiments through monitoring the state of online resources, datasets and files, or sensor outputs, and sending notifications when changes occur.

5.7.1 Receiver reports

Derived From

https://linkedresearch.org/ldn/tests/summary#ldn-report-receiver

Derived On

2019-01-05

The receiver implementations are either LDP-based, extension of LDP, generally LD-based platforms, stand-alone libraries, or integrated into existing domain specific systems, eg. personal websites. For instance, implementations range from minimum proof of concepts like DIY Inbox to enterprise-ready platforms like Apache Marmotta or Virtuoso Universal Server. *Sloph* [457], *The Presentation of Self on a Decentralised Web*, Guy, 2017, and IndieAnndroid are part of personal publishing and quantified self platforms geared around social media-like interactions. The *Scholastic Commentaries and Texts Archive* [458] (SCTA) inbox is used for annotating scholarly manuscripts, and serving notifications about the annotations. ldn-streams is a specialised receiver that is capable of accepting and serving notifications as RDF streams. distbin is a general purpose “pastebin” to store data. The rest of the implementations are non-domain specific servers that can be used for different purposes.

Table 5. Receiver tests summary

	Receiver tests														
	Required for interop								**Optional**						
Implementations	**PR**	**PC**	**PL**	**PLP**	**GR**	**GLC**	**GNJL**	**GNRS**	**OR**	**OAP**	**OAPCJL**	**PRCU**	**GNL**	**GLCR**	**GLCB**
SCTA inbox receiver (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	⌙
Sloph (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	✔
Linked Data server for Go (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	⌙	✔	⌙
ldn-streams (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	⌙
Carbon LDP (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙
solid-server in Node (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	⌙
pyldn (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	⌙
Virtuoso Universal Server (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	✔	✔	⌙
maytkso (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	✔	✔	⌙
DIY Inbox (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	⌙
Trellis LDP (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	⌙
Apache Marmotta (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	✔
IndieAnndroid (report)	✔	✔	✔	✔	✔	✔	✔	✔	⌙	○	○	✔	✔	⌙	⌙
LDP-CoAP (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	✔	✔	✔
distbin.com (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙	⌙	✔	⌙
Fedora Repository (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	⌙

Number of implementation reports 16

✔ Passed
✗ Failed
? Cannot tell
⌙ Inapplicable
○ Untested

PR: Accepts POST requests. [source]
PC: Responds to POST requests with Content-Type: application/ld+json with status code 201 Created or 202 Accepted. [source]
PL: Returns a Location header in response to successful POST requests. [source]
PLP: Succeeds when the content type includes a profile parameter. [source]
GR: Returns JSON-LD on GET requests. [source]
GLC: Lists notification URIs with ldp:contains. [source]
GNJL: Notifications are available as JSON-LD. [source]
GNRS: When requested with no Accept header or */*, notifications are still returned as RDF. [source]
OR: Accepts OPTIONS requests. [source]
OAP: Advertises acceptable content types with Accept-Post in response to OPTIONS request. [source]
OAPCJL: Accept-Post includes application/ld+json. [source]
PRCU: Fails to process notifications if implementation-specific constraints are not met. [source]
GNL: Restricts list of notification URIs (eg. according to access control). [source]
GLCR: Inbox has type ldp:Container. [source]
GLCB: Advertises constraints with ldp:constrainedBy. [source]

5.7.2 Consumer reports

Derived From
https://linkedresearch.org/ldn/tests/summary#ldn-report-consumer

Derived On
2019-01-05

The consumer implementations so far include streaming-capable implementations, stand-alone libraries, and domain specific applications. As part of the *International Image Interoperability Framework* [459] (IIIF) Discovery Support Technical Specifications Group, an LDN implementation aggregates metadata to discover resources made available by a IIIF service. Later in this thesis, I describe how LDN is implemented in context of dokieli, a clientside authoring and publishing application that can consume notifications to discover independently published annotations on the Web.

Table 6. Consumer tests summary

	Consumer tests									
	Required for interop				**Optional**					
Implementations	**HDC**	**BDC**	**LJLC**	**LJLE**	**NAE**	**NCG**	**NCN**	**NAG**	**NCT**	**NR**
dokieli (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
LDP-CoAP (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
ldn-streams (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Linked Data Notifications for aggregation of IIIF Services (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
boa (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
LDN Test Suite (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
py-ldnlib (report)	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔

Number of implementation reports 7

✔ Passed
✗ Failed
? Cannot tell
⌙ Inapplicable
○ Untested

HDC: Inbox discovery (Link header). [source]
BDC: Inbox discovery (RDF body). [source]
LJLC: Notification discovery from Inbox using JSON-LD compacted form. [source]
LJLE: Notification discovery from Inbox using JSON-LD expanded form. [source]

NAE: Contents of the `announce` notification. [source]
NCG: Contents of the `changelog` notification. [source]
NCN: Contents of the `citation` notification. [source]
NAG: Contents of the `assessing` notification. [source]
NCT: Contents of the `comment` notification. [source]
NR: Contents of the `rsvp` notifications. [source]

5.7.3 Sender reports

Derived From
https://linkedresearch.org/ldn/tests/summary#ldn-report-sender

Derived On
2019-01-05

Some of the sender implementations include streaming-capable implementations of LDN, stand-alone libraries, and domain specific applications. Some of the sender implementations also fulfill the role of a consumer. For instance, Linked Edit Rules checks the consistency of statistical datasets against structured constraints, and delivers the consistency report as a notification to the user. py-ldnlib is a library that can be reused by other applications to send LDN notifications.

Table 7. Sender tests summary

	Sender tests							
	Required for interop							
Implementations	**HDS**	**HPR**	**HPCTJL**	**HPVJL**	**BDS**	**BPR**	**BPCTJL**	**BPVJL**
dokieli (report)	✔	✔	✔	✔	✔	✔	✔	✔
ldn-scta-sender (report)	✔	✔	✔	✔	✔	✔	✔	✔
LDP-CoAP (report)	✔	✔	✔	✔	✔	✔	✔	✔
py-ldnlib (report)	✔	✔	✔	✔	✔	✔	✔	✔
Linked Edit Rules (report)	✔	✔	✔	✔	✔	✔	✔	✔
mayktso (report)	✔	✔	✔	✔	✔	✔	✔	✔
ldn-streams (report)	✔	✔	✔	✔	✔	✔	✔	✔
distbin.com (report)	✔	✔	✔	✔	✔	✔	✔	✔

Number of implementation reports 8

✔ Passed
✗ Failed
? Cannot tell
⌙ Inapplicable
○ Untested

HDS: Inbox discovery (`Link` header). [source]
HPR: Makes `POST` requests (`Link` header). [source]
HPCTJL: `POST` includes `Content-Type: application/ld+json` (`Link` header). [source]
HPVJL: `POST` payload is JSON-LD (`Link` header). [source]

BDS: Inbox discovery (RDF body). [source]
BPR: Makes `POST` requests (RDF body). [source]
BPCTJL: `POST` includes `Content-Type: application/ld+json` (RDF body). [source]
BPVJL: `POST` payload is JSON-LD (RDF body). [source]

5.8 Forces and Functions in Linked Data Notifications

This section goes into detail about how LDN fulfills or supports some of the forces and functions in scientific communication. LDN can facilitate the realisation of interconnecting scholarly communication systems as posited in *Rethinking Scholarly Communication*.

Assuming multiple interoperating implementations of LDN, I observe the *forces*:

Actor

The actors involved in executing the LDN protocol can be categorised into documents (or things, or entities) and software applications. An LDN target (like a document or a researcher profile) with an inbox is the actor that affords the start of any communication. Applications consist of LDN implementations with one or more roles: sender, receiver, and consumer, with the purpose of delivering information to or about a target that can be used by interested parties.

Content

Target resources, inboxes, and notifications collectively comprise the content in LDN’s flow. That is, any target resource advertising an inbox relation; an inbox description with a collection of notifications (with optional data about creating and using notifications), and individual notifications with any RDF contents. A notification's content can include or refer to any unit of information or communication in the scientific and scholarly system.

Accessibility

An inbox and its notifications are made *available* to users and applications via HTTP over a network connection. When an HTTP URL resolves, its contents are *retrievable*, subject to authentication and authorization. Content can be created by and made available to different kinds of consumers: humans or machines.

Applicability

Applicability of scholarly communication comes from research which responds to problems or answers questions, which can be used or built upon by others. Notifications enable communication between different parties (individuals, universities, research institutions, industry, etc.) and LDN enables more direct, non-centralised control over this communication.

The *functions*:

Registration

Prior to LDN, notifications generally would not be considered as persistent units of information. Notifications being registered as their own entities as HTTP URIs facilitates information discovery and reuse. Registration of scholarly artifacts are not addressed by LDN, however, notifications which reference scholarly artifacts are registered.

Awareness

The LDN protocol provides a way for content to be disseminated, as well as read by different consumers. LDN can be used to inform systems that a registration of a unit of information took place; a request for quality-control and verification is sought-after, or has occurred; and for instance, an archiving service is informed to place a request for preserving the units.

Certification

LDN can be implemented to support scholarly certification mechanisms. Notifications can be sent to request review, as well to return the results of a certification process such as a registered unit of information's significance and soundness. Units of information at any level of granularity can be referenced in notifications, from entire articles to individual data points. Similarly, information about a revocation of a unit of information can be circulated.

Archiving

LDN can be implemented to support archival processes. When new units of information are registered, or existing units of information are updated, archives can be notified in order to store a persistent copy of the resources.

The information herein is factored into table *Characteristics of Specifications*.

5.9 Contextualising LDN

I have discussed how notifications may be purposed in a number of ways in a decentralised scholarly communication ecosystem. In addition, outlined specific use cases, and used these to derive design considerations. An overview of existing protocols for decentralised notifications indicates their respective strengths and shortcomings, and a new protocol is designed to meet our needs in light of this. The Linked Data Notifications protocol separates the concepts of *senders*, *receivers* and *consumers* of notifications for modularity and leverages Linked Data concepts of shared vocabularies and URIs, thus providing a building block for notifications between diverse decentralised and loosely coupled Web applications. The three roles can be implemented independently from each other or all together in one system. This permits end users more freedom to switch between the online tools they use, as well as generating greater value when notifications from different sources can be shared between applications and used in combination.

The utility and flexibility of LDN is demonstrated to some extent by the various implementations reported as part of the standardisation process. In this final concluding section, I evaluate LDN by comparing it along various axes with the potential alternatives. I also explain its relationship with other Web standards in the Linked Data or decentralisation space, and describe how LDN addresses users' degree of control over their data and applications. Finally, I discuss additional considerations that were omitted from LDN at the time of standardisation but may be important to take into account for certain implementations.

5.9.1 Comparison of Notification Mechanisms

This section covers a direct comparison between existing notification mechanisms from the *Overview of Web Notification Systems* with Linked Data Notifications, followed by a discussion of the trade-offs that were necessary to realize certain benefits of LDN. The comparison criteria include our *design considerations* (*Rx*) along with additional technical information which helps to capture some design differences (*Tx*).

The figure *Comparison of notification mechanisms* is derived from https://csarven.ca/linked-data-notifications#comparison-of-notification-mechanisms and modified to include *WebSub*.

Table 8. Comparison of notification mechanisms

Mechanism	T1	T2	T3	R1	R2	R3	R4-A	R4-B	R4-C^{p}	R4-C^{v}	R4-C^{o}	R5
Semantic Pingback	Linkback	POST	RDF	S R	/	/	Any[r]	form urlencoded[k]	!	! parse source	Any[r]	X
Webmention	Linkback	POST	HTML	S R	–	–	Any[h]	form urlencoded[k]	!	! parse source	Any[h]	X
Provenance Pingback	Linkback	POST	RDF	S R	/	/	/	URI list	/	/	RDF[q]	X
WebSub	Fat ping	POST	Varies	S R U	-	-	Any	form urlencoded	/	+ app	-	!
DSNotify	Fat ping	POST, PUT	XML, PuSH	S U	/	–	–	XML	/	–	RDF[t]	!
sparqlPuSH	Fat ping	POST	XML, SPARQL, PuSH	S U	–	–	–	XML[ra]	/	–	RDF[t]	!
ResourceSync	Fat ping	POST	XML, PuSH	S U	/	–	–	XML[s]	/	–	?	!
Linked Data Notifications	Fat ping	POST	JSON-LD	S R C	!	! URI	Any	JSON-LD[j]	+ app	+ app	–	O app

T1: Notification type
T2: Delivery method
T3: Dependencies
R1: Modularity (application classes: S Sender, R Receiver, C Consumer, U Subscriber)
R2: Reusability
R3: Persistence – required? how?

R4-A: Target representation
R4-B: Notification body
R4-C^{p}: Payload processing required?
R4-C^{v}: Verification – required? how?
R4-C^{o}: Requirements for referenced resources?
R5: Subscription

⁂

–: not applicable, out of scope
/: not specified, in scope
X: explicitly disallowed
app: application specific decision

!: required (*MUST*)
+: recommended (*SHOULD*)
O: optional (*MAY*)
PuSH: PubSubHubbub

⁂

[h] HTML recommended
[j] Alternate RDF formats can be negotiated
[k] `source` and `target` key–value pairs is required
[q] Provenance records with PROV Ontology

[r] RDF representation recommended
[ra] SPARQL results transformed to RSS/Atom
[s] Sitemaps
[t] Described in an RDF store or dataset

Table is derived from *comparison of notification mechanisms* [460] originally published in *Linked Data Notifications: a resource-centric communication protocol*, Capadisli, 2016, CC BY.

Figure 8. Comparison of Notification Mechanisms.

All of the notification mechanisms specify multiple roles, with the Sender (S) being common across all of them. LDN is the only mechanism which describes the Consumer (C) role; this explicitly highlights how notifications can be useful to applications other than those concerned with sending and receiving. For mechanisms which include a Subscriber (U) role, the Receiver (R) is essentially also the Consumer, so Receivers are less generic and require specialised payloads. In this respect, LDN is a more modular specification. LDN Receivers need not be concerned with additional actions beyond storing and exposing notifications; complex and domain-specific uses of the notifications are left to Consumers, resulting in a lot more flexibility in terms of lightweight end-user applications.

WebSub differs in terms of the number of roles, but not the actual functions of the specified roles. The hub is an intermediate party which plays the role of a Receiver and a Sender, relaying notifications from the original Sender (or publisher) to the Subscriber (or ultimate Receiver), but nonetheless only makes notifications available to the Subscriber and no other applications.

LDN enforces that notifications should be reusable, and specifies how Receivers should make notifications available for Consumers. The other mechanisms either consider notification reuse as inapplicable or leave it optional and unspecified.

This is also tied to to persistence and retrievability in that LDN requires notifications to be identified by dereferenceable HTTP URIs. The other notification mechanisms treat notifications as ephemeral resources which only exist until they are delivered to a Receiver.

The mechanisms with the *fat ping* approach require structured data in XML or JSON-LD. Only LDN requires the use of RDF through JSON-LD, and optional content negotiation for other RDF syntaxes. Thus, notification payloads can include references to other retrievable resources, or can embed all relevant data directly in the payload.

Approaches which send only URLs (as with linkback) rely on the receiver interpreting a third-party resource, which may or may not contain structured content or be under the control of the sender. Approaches which offer additional guidance to aid the receiver in interpreting the source document(s) nonetheless still restrict the sender. LDN therefore offers flexibility to senders, increasing the potential uses for the notification mechanism. LDN compensates for increased complexity on the receiver’s end by recommending filtering mechanisms, and moving most of the burden of understanding notification contents to the consumer role. In placing no constraints on the notification payload, LDN enables a sender to be precise and lossless with the data it is transmitting. As such LDN can cover a broader variety of use cases.

LDN recommends that Receiver applications perform some processing of the payload. This is to enable users to configure their Receivers to filter out spam or customise them for particular domains, but is not required; such filtering could also be left to or supplemented by Consumers. In contrast, Semantic Pingback and Webmention require receivers to parse the form-encoded payload and then to perform fetching of the URLs within and additional parsing of the documents those URLs identify. This is required in order to verify the message (by finding the target URL). Such processing in LDN is left at the discretion of the Receiver; linkback style verification is not required, but also not forbidden.

As WebSub requires subscribing as part of its core design, hubs are expected to perform specific actions on receipt of a *subscription* request (where the payload is tightly constrained), but not necessarily to process the contents of a notification sent by a *publisher* (where the payload can be anything; how they handle this depends on the content-type).

Though LDN does not specify a subscription mechanism in the same vein as WebSub, thanks to the modularity of specification conformance classes and the flexible nature of the notification contents, it is possible for applications to perform subscription-like interaction by taking on multiple roles.

LDN requires that Senders and Consumers are equipped to discover Inboxes (receiving endpoints) through both HTTP headers and the body of an RDF resource. This slightly increased complexity around Inbox discovery for Senders and Consumers is a worthwhile tradeoff to lower the bar for publishers of target resources; Inboxes can be advertised from *any* document on the Web (through HTTP headers), and attached to even non-informational resources (through RDF statements). This increases the potential use cases for LDN.

It is difficult to objectively measure performance metrics of different mechanisms as the various implementations use different servers and programming languages whose efficiency do not necessarily have any bearing on the protocols themselves. Mechanisms which require large payloads may increase the time for HTTP requests to be processed, and mechanisms that require verification add an additional network request to complete the process.

5.9.2 Interoperability Across Specifications

There are several technical specifications that LDN may be interoperably combined. Here I discuss a few.

Relationship with Linked Data Platform: An LDN Receiver is not dependent on a complete implementation of the LDP specification, but comprises an easy-to-implement subset. An LDN Inbox is comparable to an LDP *BasicContainer* [461]. LDP does not define how clients discover LDP Containers. Through LDN's inbox discovery, LDP Containers for the purpose of holding notifications can be discovered. Thus, the core features necessary to exchange notifications between LDN applications is effectively possible with existing, as well as future implementations of LDP.

Relationship with ActivityPub: ActivityPub uses LDN's targeting and delivery mechanism with some specific constraints. Notifications must use a single AS2 Activity in compact JSON-LD syntax. Receivers are required to authenticate requests made by Senders, as well as verify the existence of an object (that the activity is about) that is mentioned in the notification by fetching its source from the origin server. It is possible for LDN Senders to deliver notifications to AP servers. It is also possible for AP clients to deliver messages to LDN Receivers with some bridging.

Relationship with Fedora API: The Fedora API Specification is in the process of being formalised (as an extension of LDP) by the Fedora community. As LDN can be used to support external integrations, Fedora API's repository event stream draws upon the LDN specification, allowing LDN consumers and senders to react asynchronously to repository events. Fedora implementations were included in the LDN implementation reports.

Linked Data vocabularies: As LDN is agnostic about the contents of a notification, any Linked Data vocabulary can be used, including the *Vocabularies* I have previously discussed in *Structure of Scholarly Information*.

Relationship with other federation protocols: Further similarities and differences between LDN and other federation protocols and social APIs are described in W3C Note *Social Web Protocols* [462]. For example, a bridging code can be applied by applications implementing different specifications to achieve further interoperability between systems.

5.9.3 Degree of Control in LDN

At the beginning of this section, four requirements relating to *Degree of Control* were listed. LDN addresses each of these as follows:

Actors can use their preferred applications to discover, reuse, and send notifications

LDN was developed as a W3C standard, with strict criteria for interoperability. Applications which implement any of the roles of the protocol can confirm their conformance with the test suite, and are thus expected to be able to interoperate with other conformant applications. Actors should be able to choose any available conformant application and maintain expected notification-related functionality.

Users store incoming notifications where they prefer

Any conforming LDN Receiver (server) under a particular user's control – which they trust, and are authorised to read – can be chosen to store notifications.

Users can switch between applications without having to move their data

LDN specifies the mechanism for information exchange, so it is possible to switch between Receiver applications at the protocol level without any further action (implementation-specific details like database storage mechanism may require additional work for portability of data already stored there; this is out of scope of the LDN specification). A user's Receiver, where notification data is stored, has no bearing on their ability to alternate between sending and consuming applications as needed.

Users change the location of their data without having to change their application

Users can serve their inbox and the notifications from different HTTP URIs provided that they do not mind persistence. In this case, the target resource's inbox location would simply point to the new location. As inbox discovery starts from looking up a target resource's description, the new inbox location will be discovered as before.

5.9.4 Additional Considerations

The LDN specification itself covers privacy and security considerations on *authenticated inboxes* and *personally identifiable information*. As for persistence and retrievability, any consumer application can potentially benefit from being able to reuse notifications if the owners of a receiver makes the necessary commitments. There are further considerations which may be of interest to implementers, or may constitute future work around extending the core protocol, which are worth outlining here.

Subscribing to Notifications: The interaction between consumers and receivers describes a *pull* mechanism. A *subscribing* mechanism in which consumers *request* that receivers *push* content changes to them is left out of scope. Much of the related work *requires* notifications to be explicitly solicited to trigger sending. Since in a decentralised model, receivers may not be aware of possible sources for notifications, our sender-receiver relationship depends on the sender's autonomy to make such decisions by itself. This does not preclude the scenario in which a receiver may wish to solicit notifications from a particular sender, but as there are already *subscription* mechanisms in wide use on the Web eg. ActivityPub, WebSub, *The WebSocket Protocol* [463] (RFC 6455), *Generic Event Delivery Using HTTP Push* [464] (RFC 8030), and can be interoperably combined with LDN, we do not need to specify it as part of our protocol. A push based interaction can still be arranged between LDN applications, where it would be implementation specific. For example, in a notification, a sender specifies the location of an inbox where it can receive updates about content changes. Then, an application that consumes the notifications manages its own subscriber list, and takes the role of sender in order to push content to the inbox as specified in the notification. Another specification-level integration of LDN pertaining to subscribing is that ActivityPub's *Follow Activity* [465] is a notification requesting to be notified of target's activities when they are created. Once the request is approved, the requesting actor is added to target actor's *Followers Collection* [466] – a list of actors (with their own inboxes). The server then dispatches notifications to the inboxes of these actors.

Semantic Organisation of Notifications: While LDN facilitates a decentralised architecture for notification exchange, it does not prescribe the structure and semantics of the notifications themselves. The composition of the notifications are unspecified to foster different kinds of applications to be built to communicate information in any domain. Consequently, organisation of notification models materialises through the fact that domain-centric applications create and consume what is meaningful to them in order to coordinate among themselves. So while the notification system is globally decentralised, it enables communication that can be specialised and be only useful and meaningful to applications that understands a particular message's semantics.

Inbox Paging: We need to consider both the needs of software systems and humans when large amounts of notification data are being generated and shared between diverse applications which may be operating without knowledge of each other. To organise and manage large amount of notifications over time, mechanisms should be in place to break representations of collections of notifications into multiple paged responses that may be easier to consume by applications.

Updating and Deleting Notifications: Receivers may want to carry out resource management or garbage collection, or permit consumers or other applications to do so. For example, an application to consume messages might let an authenticated and authorised user ‘mark as read’ by adding a triple to the inbox or notification contents.

Social Implications: In an ecosystem where anyone can (technically) say anything about anything, we can acknowledge the social challenges surrounding that in context of notifications. As the LDN protocol facilitates distribution and discovery of information, it can amplify misinformation and (community-centric) inappropriate content, as well as online and offline social dangers effectively. That is, consumers of notifications can be exposed to such information if an inbox is not managed properly. LDN may offer one way to mitigate this problem in that, the actors that control an inbox can decide on which notifications can persist and which to be removed, thereby one way to eliminate references to undesirable content created by other parties. At the same time, the same privilege can also be used to filter information and have bias towards certain parties. This is a form of decentralised curation in that while no single inbox can oversee the discovery of a resource (as there are alternative ways), it enables the owners of receivers to exert their authority on the notifications that makes it into their system. It is a middle-ground in that, on one hand, individuals can voice themselves by registering and making their resources accessible, owners of inboxes can choose not to provide pointers to it. I will further revisit some of these concerns in *Addressing Social Implications*.

As we have seen in *Scholarly Communication on the Web*, research communication is continuously transitioning from print-centric to Web-centric. LDN being rooted to Web-centric protocols and formats builds on a reliable and relevant foundation for it to be used as part of the future scholarly communications ecosystem.

There are many diverse use cases for notifications within scholarly communication now and likely to be more in the future, so designing the protocol to be flexible permits a common thread of data sharing between these different cases, as well as accommodating unexpected use. Specialisation may be necessary for notifications to be particularly useful in a lot of scenarios, however the protocol is not the level to do this. Thus, LDN leaves it up to applications to specialise their functionality for particular domains or user needs by using semantically *self-descriptive* notifications.

In the next section, I present LDN as one part of the broader scholarly communication ecosystem through an implementation which ties together many relevant protocols, and encompasses article authoring, annotations, and social interactions.

6. Decentralised Linked Research Application

> Information systems start small and grow. They also start isolated and then merge. A new system must allow existing systems to be linked together without requiring any central control or coordination.
>
> Non-Centralisation, CERN Requirements, *Information Management: A Proposal* [467], Tim Berners-Lee, 1989

Academic articles are commonly represented in packages, like binary file formats (PDF, Word) merging structural, presentational, and behavioural layers. Such forms are often disconnected from other knowledge, in that navigation from one unit of information to another is non-continuous or not uniformly accessible and usable by both human and machine users. This is generally due to historical policy, cost, and social or technical reasons intertwined with the systems around units of scholarly communication. One side-effect of scholarly publication still being primarily driven by the constraints of print media in this way is that it is challenging for efficient machine-readable coordination on the Web, as well as for continuous advancement of creating, communicating, and consuming information as part of the Open Web Platform.

I have covered existing initiatives and progress in research, standards, and practices towards different aspects of realising a decentralised information space. Each dataspace can be individually controlled with access controls for participating agents. Different applications can be implemented to operate on the same data, as well as send activities about the interactions with the data.

This section addresses the following *research question* for this thesis:

> How can Web technologies be employed to fulfill the core functions of scholarly communication in an open and interoperable way?
>
> *Artifacts*

To this end, I will discuss a set of requirements for software implementations which combine the various technologies and concepts into a complete package which enables decentralised publishing of articles and globally identifiable units of information, annotations and social interactions on the Web, as well as their linking and dissemination. I note the overarching influence of the concept of *Degree of Control*. Such software must enable:

- Creating human- and machine-readable Web resources.
- Socially-aware operations using Web Annotations, Linked Data Notifications, and ActivityPub.
- Actors to use their preferred WebID in clientside applications.
- Actors to use their preferred clientside application to perform HTTP-based read-write operations on resources they are authorised to.
- Actors to switch between applications without having to alter the schema of their data.
- Actors to move data between servers or datastores without having to change their preferred application.
- Querying and visualising [statistical] data from clientside applications.

These requirements are met by *dokieli* [450], a domain-independent clientside authoring and publishing application that is loosely coupled with a data server. This section outlines key features and discusses design decisions made for this implementation, and extends *Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli* [468], Capadisli, 2017. I will focus on various aspects of the application, with some use cases specific to scholarly communication, and examine the *Forces and Functions in dokieli*.

The implementation of dokieli is a core contribution of this thesis, and serves as a culmination of the various background research, Web standards work, and advocacy I have done so far. It is an implementation which showcases various disparate parts – Web standards, technologies, and social ideals – coming together in a cohesive way to serve a future scholarly communication ecosystem which takes full advantage of the capabilities of the Web.

dokieli is not intended to be *the* solution for decentralised academic publishing and communication. My contribution is not the implementation itself, but rather the deep exploration and demonstration of how the, often complex and controversial, moving parts of the scholarly communications ecosystem can come together. What is presented here is a description of the state of the work at the time of writing. It is intended as a conceptual guide or model, and a minimum-viable baseline to be bettered by future developers and researchers. Certain development decisions were made based on existing literature, development, and observations in the relevant fields, which are discussed throughout the prior sections of this thesis. The major undertaking here was to develop dokieli with a thorough analysis of the technical possibilities of the Web *as well as* the social context in which academic researchers currently find themselves. This is not a quick-fix or a patch, but an ongoing and lengthy iterative process. It is my hope that future tooling built to support the forces and functions in decentralised scholarly communication can build upon the rationale for the ways in which dokieli is designed, and learn from the things that work and the things that do not.

While I am the original developer of dokieli, the code is open source, maintained in public, and accepts attributed community contributions. As a direct result of my work on dokieli I have also contributed to and been influenced by the following:

- *Embedding Web Annotations in HTML* [263], a Note output of the W3C Web Annotation Working Group. The Note includes my contributions to representing annotations in HTML+RDFa derived from dokieli's implementation of the WA Data Model and Vocabulary for HTML+RDFa representations.
- The MIT *Social Linked Data* [469] (Solid) project repositories, eg. *node-solid-server* [470], "a proposed set of conventions and tools for building decentralized social applications based on Linked Data principles."
- The *ORCID source* [471], to enable existing ORCID identities to take advantage of extensible profile descriptions in RDF by making it possible for ORCID users to declare the location of their preferred Personal Data Storage, Linked Data Notifications inbox, and ActivityPub outbox.

6.1 Linking the Decentralised Information Space

By using different arrangements of LDN, WA, AP, LDP mechanisms and as well as relevant Linked Data vocabularies, scholarly and social activities can be systematically created, linked, discovered, and reused for a variety of use cases. In this section, I present an overview of *a* decentralised information space including actor profiles, notifications, and annotations. As units of information are structured and semantically interconnected, it forms a knowledge graph with the following key components with respect to technical specifications and practices:

- Assertion of preferred service for maintaining annotations about the subject resource (via `oa:annotationService`)
- Assertion of a service for receiving and serving notifications about the subject resource (via `ldp:inbox`)
- Assertion of preferred storage location (via `pim:storage`)
- Assertion of location of an actor's activities (via `as:outbox`)
- Assertion of type index registrations (via `solid:publicTypeIndex`)
- Members of an actor's social graph (via eg. `foaf:knows`) asserting their preferred storage location
- Members of an actor's social graph asserting the location of their activities
- Members of an actor's social graph asserting their type index registration (via `solid:publicTypeIndex`)
- Inbox or annotation service locations which are known prior to interaction eg. trusted or used by a community.
- Inbox or annotation service locations which are manually input to the system eg. not publicised or private.

When discussing discovery throughout this section, this means the *follow your nose* approach of using the kinds of links listed above to find relevant services and other resources. Applications can incrementally arrange and implement a combination of these components to deliver and discover distributed activities on the Web – in a similar fashion to the *paygo* approach. For example, the preferred locations for data storage, annotations and notifications for resources can be used as input to create new resources eg. assigning an inbox for an annotation where the inbox location is derived from a profile's preferred location for a class of information. Thereby a chain of operations can be orchestrated for decentralised units of information. This matches the notion of decoupled functions and reconstructing information flows for a unit of communication as posited in *Rethinking Scholarly Communication*.

The following figure depicts a generalisation of a decentralised information space with emphasis on personal profiles, articles, annotations, and notifications.

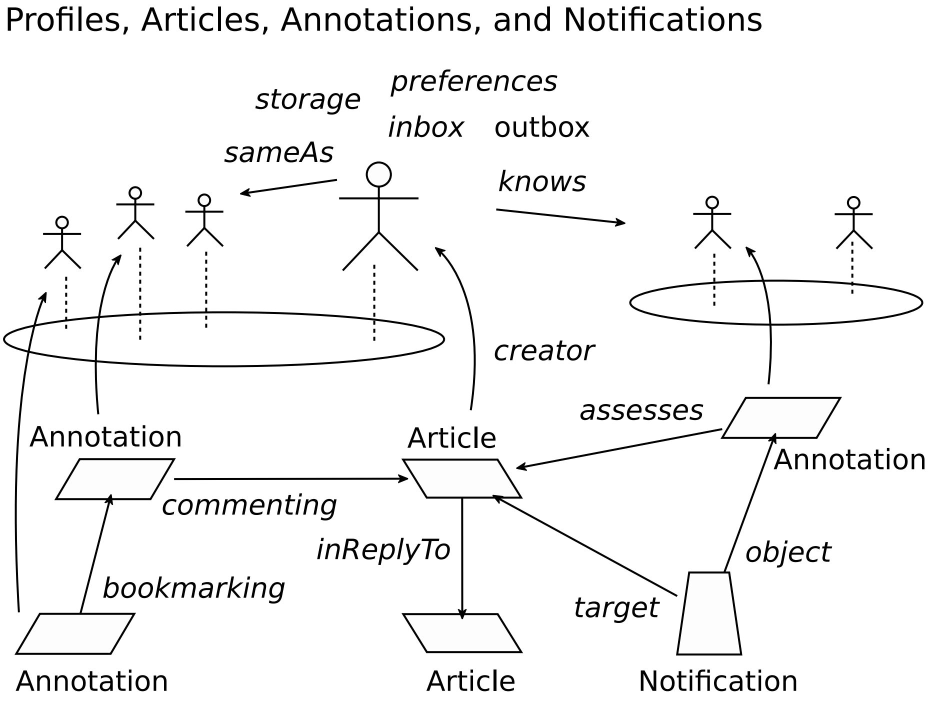


Figure 9. Interplay of Profiles, Articles, Annotations, and Notifications.

6.2 Implementing a Read-Write Application

In this section I present how a collection of open standards can be integrated into an application that can be used autonomously by researchers to participate in a scholarly communication ecosystem that puts the principles of *interoperability* and *composability* at the forefront. *dokieli* implements the use cases pertaining to *Linking the Decentralised Information Space*, as well *scholarly communication notification use cases*.

Project website
https://dokie.li/

Source code
Git repository

Code license
Apache License, Version 2.0

Content license
Creative Commons Attribution 4.0 International

Documentation
https://dokie.li/docs

Community
Public chat

dokieli is a general-purpose clientside editor for decentralised article publishing, annotations and social interactions. While it implements use cases specific to research or scholarly communication, it is not constrained to any knowledge domain or field of work. For instance, dokieli is capable of rendering the following use cases from the W3C *RDFa Use Cases: Scenarios for Embedding RDF in HTML* [223]: “Basic Structured Blogging”, “Publishing an Event - Overriding Some of the Rendered Data”, “Content Management Metadata”, “Web Clipboard”, “Advanced Data Structures”, and “Augmented Browsing for Scientists”, “Publishing a RDF Vocabulary”. The functional and non-functional requirements from the *Linked Data Platform Use Cases and Requirements* [353] (LDP-UCR) are also met from the point of expectations of a client application for authoring, publishing, and sharing content with LDP servers.

dokieli uses *LDN* to send and consume notifications. For example, when a reader comments on a fragment of text in an article, the application discovers the article’s Inbox and sends a notification about the annotation. dokieli also consumes notifications from this Inbox to fetch and display the annotation as marginalia (figure dokieli Web Annotation), or indicates the back-references as in the case with citations. A reader can share a dokieli-enabled article with their contacts; dokieli discovers each contact’s Inbox and sends a notification there (figure dokieli Share). When editing an article, the author can add a citation. If an Inbox is discovered in the cited article, dokieli sends a notification there to indicate what part of the article was cited by whom and where. dokieli-enabled articles also consume citation notifications to display these metrics for the author and other readers (figure dokieli Citations and Notifications).

Next I describe dokieli’s architecture according to the following capabilities:

- Actors use their own WebIDs to identify themselves with, and use their profile information to adapt the user interface.
- Actors publish and consume human- and machine-readable Linked Data, and decide where they are made available from and the conditions for access and reuse.
- Actors trigger the notification system to announce, as well as discover relevant information.

6.2.1 Core Techniques

I summarise the key protocols and data models that dokieli uses in this section. Later, I will describe important use cases in more detail.

6.2.1.1 Core Concepts

Separation of Layers: dokieli adopts *progressive enhancement* strategy for the structural, presentational, and behavioural layers to allow content and base functionality to be accessible through different media and devices. The content representation in HTML+RDFa that dokieli produces is accessible (readable by human and machine consumers) without requiring CSS or JavaScript, ie. text-browser safe, as well as backwards-compatible with the *WorldWideWeb* [472] browser.

Human- and Machine-Interpretable: *Why RDFa* explained having all human-visible content in HTML, and all structured statements be made in context of their content in RDFa, thereby a close association between structured data and visual locality. That is, machine-readable semantics are embedded directly into human-visible prose through RDFa. dokieli applies this technique to avoid data duplication in the same document – adhering to the DRY principle. The approach also helps to avoid the creation and usage of multiple data islands, ie. separately for humans and machines. Consequently, it has the advantage of not having any dependency on JavaScript in order to make the hidden machine-readable content be consumable from a human user interface. One exception to this rule is where dokieli allows authors to embed Turtle, JSON-LD, or TriG data blocks (useful for Nanopublications) in HTML typically for the purpose of including complementary non-prose structured information. We cannot know in advance what new kinds of information or data can be integrated into articles, so to facilitate extensibility, dokieli sets no constraints on what information can be embedded nor how it can be be expressed. Hence, information can be expressed using different vocabularies, enabling reuse of existing data models and the addition of new semantic expressions.

Composability: Resources generated by dokieli are composable on the data layer as there are semantic relations between units of information, both internally and externally identifiable units. All articles, annotations, and notifications can be assembled in various combinations since they are ultimately expressed using the RDF language, which acts as a layer to describe them at any level of abstraction. All units of information in articles, annotations, and notifications are self-contained, context neutral, globally identifiable and accessible without additional processing. Resources can be versioned, marked as immutable, and archived, thereby treating them as stateless entities on the Web. Another way to reuse the resources is by dynamically embedding them in articles while communicating the original context eg. link to the source, parameters in which an annotation was created. Notifications can serve to establish bi-directional linking eg. in citations. Articles and annotations can embed media objects eg. images, videos, audio, or documents, as reusable resources.

Presentation: The appearance of HTML in the Web browser is handled with CSS3. Different stylesheets can be applied to the same HTML structure so that a document can be presented flexibly, in the most appropriate way for a particular circumstance, eg. a theme resembling a blog post or a scholarly article. Stylesheets can be switched from either dokieli’s menu or through Web browser with native controls. External stylsheets can also be applied dynamically.

Behaviour: When JavaScript is enabled on the clientside, dokieli provides a rich editing interface which includes visual and structural formatting of text as well as embedding machine-readable semantics, media, dynamic citations, and inclusion of statistical charts from live endpoints. An Internet connection or a personal data store are not needed at this stage as modifications to a document made in the browser this way can be persisted to a local filesystem using the dokieli menu *Export* function; the Web browser’s *Save Page As* function, or through the Web browser’s local storage.

6.2.1.2 Core Mechanisms

Read-Write: dokieli operates on a “thin server architecture” in that it has no expectations from a server other than to provide a representation of a resource (based on content-negotiation). Reading primary resource that dokieli is activated on (same-origin) does not require a JavaScript capable user agent. Dynamically fetching additional resources from the same or cross-origin requires JavaScript. Writing to same or cross-origin requires JavaScript.

Identifiers: The following kinds of identifiers can be registered (generated or requested by the application and fulfilled by a service):

- Resource-level (articles, activities, annotations, notifications)
- Versioned resources
- Immutable resources
- Archiving
- Contents, list of tables, figures and abbreviations
- Sectioning content
- Paragraphs
- Selections (arbitrary)
- References and citations
- Definitions
- Figures, tables, code listings
- Code line numbers
- Access-control policy
- Canonical identifier (annotation)
- Embedded data resources

Content-Negotiated Serialisation: dokieli can serialise articles, annotations, and notifications in HTML+RDFa, Turtle, and JSON-LD, depending on server content-negotiation. Articles are represented in HTML+RDFa so that information is usable by both humans and machine consumers while maintaining lowest requirements for publishing, eg. a single URL with full payload in HTML+RDFa can be accessible from any HTTP server. No additional requirements are necessary from user-agents (eg. JavaScript support), or servers (eg. content-negotiation). For annotations and notifications, dokieli first meets interoperability requirements (for protocols and vocabularies), and remains flexible about the serialisation that servers prefer. Similarly for consuming content, it can work with any of the serialisations. All HTML serialisations use HTML5 *Polyglot Markup* [473].

Notification Inbox: An article, or any unit of information within the article, may be associated with a notification inbox using LDN's `ldp:inbox` property. dokieli detects inbox relations and sends notifications when annotations are created. While orthogonal, articles and inboxes can also preserve context in that a unique inbox can be associated with a unique version of an article, so that annotations will refer to the specific version of an article, as well as being specifically recalled through the unique inbox. dokieli also enables users to assign an inbox for the annotation that they are creating. This has the advantage that the annotation has its own inbox, which means it can stand alone and receive replies, or enable threaded conversation on the article itself. The authors of articles or annotations choose their preferred storage location for notifications and assign their inbox accordingly, based on Solid's Type Index Registration approach. An inbox may also be discovered through an actor's profile and used to send notifications eg. to announce the availability of a new resource.

Annotation Service: Articles can be associated with an annotation service located anywhere on the Web with WA's `oa:annotationService` property. dokieli can detect the annotation service of an article and offer optionally to send annotations there using the WA Protocol.

Provenance and Memento: dokieli can create resources that are derived from other resources, such that provenance data is retained eg. using `prov:wasDerivedFrom`, `prov:generatedAtTime`, in the derived copy. Versioned and immutable resources can be created and include information on their history eg. `rel:latest-version`, `rel:predecessor-version`. Memento TimeMap is also created to keep references to memento versions of the created resources (using `mem:memento`), as well as resources linking to their TimeMap with `mem:timemap`.

Resource archiving: dokieli uses Internet Archive's API to create archived resources and to cite the snapshots with Web Annotations. Actors are able to visit the snapshot URI or reuse by including it back in the document as Robust Links. As per *Archivability*, as all human- and machine-readable (meta)data can be encoded in a static standards-compliant HTML document, they are archivable by on-demand Web archiving services.

Robust Links Link Decoration: In order to work around content-drift and link-rot, authors of articles can *decorate links* to include *Robust Links*, eg. the URI of a snapshot from the *Internet Archive*, datetime stamp at the time of linking, and other context. The application UI shows the versioned URI, as well as the *TimeTravel* [474] URI that redirects to a Memento of a given resource. The documents also include common date information to provide context for external references eg. through created, published, modified datetime stamps.

Internationalization and Localization: UTF-8 content encoding is used in HTML and RDF serializations, as well as in CSS. Language tags are used at the document level as well as parts within, and languages can be mixed. Users can individually tag the language of their articles and annotations from the UI. Date and time are expressed according to *ISO 8601* [475].

License: Actors can assign any license and rights to their documents or any unit of information within. All of the Creative Commons licenses are available from the dokieli UI for actors to use for articles and annotations. Notifications are licensed with Creative Commons CC0 1.0 Universal (public domain dedication).

6.2.1.3 Core Artifacts

dokieli applies the notion of articles, annotations, notifications, and profiles as described in *Units of Communication*. dokieli's user interface is used to author articles, annotations, and notifications.

Articles: An article – in the most general sense of the term – may contain RDF classes like `schema:CreativeWork`, `prov:Entity`, `schema:Article`, `as:Article`. While the body of annotations may be also considered as articles in their own right, I reserve the usage of the term articles in particular to primary (research) documents typically containing structured prose. Actors can assign unique identifiers to any fragment of information.

Annotations: dokieli uses the Web Annotation *Data Model* and *Vocabulary* to express the annotations it creates, as well as implementing the Web Annotation *Protocol* in order to communicate with conforming annotation servers. dokieli initially prepares the annotation in RDFa using the *Embedding Web Annotations in HTML*, where it may be re-serialised to another RDF syntax depending on content-negotiation with the server. An annotation has the `oa:Annotation` RDF class, contains a link to its body content and the target resource in which it is associated with. dokieli also uses the WA *Selectors and States* for selecting part of a resource and generating a HTTP URI. Annotations can be done at the document as well as sentence level.

Notifications: dokieli implements the LDN protocol to discover social and scholarly activities, such as annotations with different motivations, as well as sending contextual notifications to recipients. A notification may have one of the following classes: `as:Announce`, `as:Like`, `as:Dislike`, `as:Relationship`.

Profiles: Actors can use their preferred *WebIDs*, and depending on the information available in their *WebID Profile* descriptions, the user interface automatically adapts to accommodate their preferences while interacting with Web resources. For instance, the profile description is used to discover an actor's preferred storage locations when creating annotations, as well as discovering their social network in order to send notifications to their inbox.

6.2.2 Resource Interaction

dokieli implements the following client operations to communicate through LDP, WAP, AP.

Table 9. HTTP Interactions

Method	Resource	Intent
`HTTP GET`	Any resource	Read
`HTTP HEAD`	Any resource	Read
`HTTP OPTIONS`	Any resource (collection)	Read
`HTTP POST`	Annotation, notification	Create
`HTTP PUT`	Any resource	Create, Update
`HTTP PATCH`	Memento TimeMap, ACL	Create, Update, Delete
`HTTP DELETE`	Any resource	Delete

The term "collection" (as in WA and AP) is used synonymously with "container" (as in LDP, LDN).

Any resource pertains to any kind of object identified with HTTP URI eg an article, annotation, inbox and notification, profile, media and scripts, ACL, SPARQL or other API endpoints. Its application depends on context.

Retrieve and read HTTP message and body, parse and use relevant information.

Before dokieli makes a write request to a server, it triggers content-negotiation through `HTTP OPTIONS` and checks for `Allow` and `Accept-Post` HTTP headers in order to determine server's preferences. In the absence of successful negotiation, dokieli proceeds to make the request as per the required parameters of LDN, WAP and AP. For example, `HTTP POST` with `application/ld+json` content type would be used for annotations and notifications, if a server's preference is not determined.

Articles are by default use HTML with embedded RDFa. Where a server implements `HTTP GET` with `text/html`, and allows `HTTP PUT` for writing, the assumption is that it can allow payloads with `text/html` content type. HTML+RDFa is considered to be an important default for articles for being human- and machine-readable. Hence, if a server is capable of storing and serving resources in HTML, it will receive articles, annotations, and notifications in that form. Otherwise, dokieli will re-serialize the content into an alternative RDF syntax that the server prefers or is capable of handling per content-negotiation.

> As of this writing, the reasons to use `HTTP POST` and `HTTP PUT` over `HTTP PATCH`, for documents like articles, are as follows:
>
> - Servers with *XML Patch* [476] (RFC 5261) and mediatype `application/xml-patch+xml` – *A Media Type for XML Patch Operations* [477] (RFC 7351) – capability in the Linked Data ecosystem are not well supported.
> - Servers implementing `HTTP PATCH` with *SPARQL Update* [478] (using mediatype `application/sparql-update`) would also need to re-serialize the content eventually as HTML+RDFa.
> - `HTTP PATCH` can help to optimise HTTP requests provided that the server processes the SPARQL query and eventually publishes the final state of the article. dokieli's use of `HTTP POST` and `HTTP PUT` on the other hand does not particularly expect or impose operations other than to eventually store and serve the resource.
> - `HTTP POST` and `HTTP PUT` are functionally valid. `HTTP PATCH` can still be used in the future if server implementations can better handle updates to documents encoded in HTML+RDFa.

Due to *Mixed Content* [479] implementations in Web browsers, ie. "fetching of content over unencrypted or unauthenticated connections in the context of an encrypted and authenticated document", is subject to being blocked by the Web browser. Hence, a document available from `https:` will not be able to use the contents of a document available from `http:`. As a workaround, dokieli implements the notion of using a *preferred proxy* that a user-agent can use in order to access the contents of an `http:` resource. A document on `https:` fetching an `https:` resource will not use the proxy.

An actor may require to be *authenticated* and *authorized* with their *WebID* to create Web resources. The table on *unit registration and content in dokieli* includes the functions in dokieli that requests write-operations to a server for articles, annotations, and notifications. These client-triggered mechanisms and methods applied to creating resources are done with HTTP URIs, and helps fulfill the registration function of scholarly communication.

Table 10. Unit registration and content in dokieli

Action	Method	Resource	Notification
New	HTTP PUT	Actors can create new resources (articles) by inputting the location where they are authorized to. New resources include a HTML+RDFa document (a basic template) with dokieli's JavaScript and CSS. Uses `HTTP PUT`.	None
Save	HTTP PUT	Overwrites the current resource with the current state of the resource in user-agent. Uses `HTTP PUT`.	None
Save As	HTTP PUT	Actors creates a derived copy of the current resource (with its scripts and media) including provenance information at preferred location. A unique notification inbox and annotation service can be assigned to the derived resource. Uses `HTTP PUT`.	None
Reply	HTTP POST	A note reply modelled around `oa:Annotation` and other data motivated by `oa:replying`.	Sends and consumes notification using the Activity vocabulary's Announce.
Annotation	HTTP POST	A note modelled around `oa:Annotation` and other data.	Sends and consumes notifications for activities `as:Add`, motivated by `oa:bookmarking`; `as:Announce`, motivated by `oa:replying`, `oa:questioning`; `as:Like` and `as:Dislike` motivated by `oa:assessing`. Can also consume notifications using `as:Relationship`.
Activity	HTTP POST	Sends activities about the annotations and encapsulates with Activity vocabulary's Create.	None
Share	None	None	Sends notification using the Activity vocabulary's Announce about the target resource.
Citation	HTTP GET	Sends an archival request to the Internet Archive.	Notifies cited entity about the citing entity's reference and the author's intention for the citation. Consumes notifications about citations.
Archive	HTTP GET	Sends an archival request to the Internet Archive.	None

Articles by default are generally modelled to express `schema:CreativeWork`, `prov:Entity`, `as:Article`.

Notifications include information about its actor, notification date, rights and license, as well as information pertaining to the activity that it is in reference to eg. an annotation motivated by replying, and the target resource. All notifications are created through a `HTTP POST` request.

The Web Annotation Data Model is used for annotations. An annotation includes information about its author, publication date, rights and license, system that was used to render the target resource (in this case dokieli) for the annotation. All annotations are created through a `HTTP POST` request.

Actions such as *New*, *Save*, *Save As*, *Reply*, *Version* are intended to create or update mutable resources, whereas *Immutable* is intended to create an immutable resource. The server ultimately decides whether to promise immutability or Memento based discovery.

The 'Archive" action uses `HTTP GET` according to the requirements of the Internet Archive's API.

Action	Method	Resource	Notification
Version	HTTP PUT	Generates a derived copy of the current document as a mutable resource including provenance.	None
Immutable	HTTP PUT	Generates a derived copy of the current document as an immutable or frozen Memento resource including provenance. Creates or updates a Memento TimeMap including a reference to the immutable resource.	None
Export	None	Downloads the document to local system.	None

Articles by default are generally modelled to express `schema:CreativeWork`, `prov:Entity`, `as:Article`.

Notifications include information about its actor, notification date, rights and license, as well as information pertaining to the activity that it is in reference to eg. an annotation motivated by replying, and the target resource. All notifications are created through a `HTTP POST` request.

The Web Annotation Data Model is used for annotations. An annotation includes information about its author, publication date, rights and license, system that was used to render the target resource (in this case dokieli) for the annotation. All annotations are created through a `HTTP POST` request.

Actions such as *New*, *Save*, *Save As*, *Reply*, *Version* are intended to create or update mutable resources, whereas *Immutable* is intended to create an immutable resource. The server ultimately decides whether to promise immutability or Memento based discovery.

The "Archive action uses `HTTP GET` according to the requirements of the Internet Archive's API.

These client-triggered mechanisms and methods applied to creating resources helps fulfill the registration function. All created resources are considered to be archive-friendly ie. through the application of *Separation of Concerns* and keeping *Human- and Machine-Readable Information* where possible.

6.2.3 Personal Identities

Authentication: Actors can authenticate themselves using *WebID-TLS* and *WebID-OIDC* mechanisms and then the server determines their level of authorization to read, write, and control resources. The workflow for authentication starts by a user inputting their WebID (HTTP URI) or their Identity Provider URL as a shorthand for their WebID into the dokieli UI. The Web resources with which the user wishes to interact may be under any domain, whether self-hosted elsewhere, so long as they are hosted by a server application conforming to the same standards.

Profile Description: In order to adopt dokieli's user-interface and the kinds of interactions that a user can perform, the user's *WebID* (HTTP URI) is dereferenced to a *WebID Profile* document for the following kinds of information:

- Name (eg. `foaf:name` and equivalents)
- Image (eg. `foaf:image` and equivalents)
- Homepage (eg. `foaf:homepage` and equivalents)
- Social graph (`foaf:knows`)
- Storage (`pim:storage`)
- Inbox (`ldp:inbox`, and alias `as:inbox`)
- Outbox (`as:outbox`)
- Preferences (`pim:preferencesFile`, `solid:publicTypeIndex`, `solid:privateTypeIndex`)
- Preferred proxy (`solid:preferredProxy`)
- Multiple identities (`owl:sameAs`)
- Additional information (`rdfs:seeAlso`)

- Delegated agent (`acl:delegates`)

When dokieli encounters `owl:sameAs` and `rdfs:seeAlso` relations in WebID Profiles, it traverses the unique connections continuously to mine for relevant information. The reason for this is both to discover a graph that can be used to adapt the interface and the possible interactions. For example, driven by *follow your nose* type of exploration, a human-readable name, an inbox relation, list of contacts, and other information can be put together by looking at the interconnected identity graph that is semantically associated.

6.2.4 Annotations

Annotation of artifacts facilitates a social behaviour in that it is a collaborative effort: some actors create primary resources while others create associations between those resources, add new information, or extend existing information. Consequently, such activities help actors to learn, discover or explore new knowledge. Annotations of analogue as well as digital resources have been historically essential to sociality, education, stronger engagement, as well as being a valid form of user-generated content as per background provided in *Social Annotations in Digital Library Collections* [480], Gazan, 2008.

dokieli implements the following use cases from the W3C Interest Group Note, *Digital Publishing Annotation Use Cases* [481]:

- 2.1.1 Comment on Publication Title
- 2.1.2 Tagging a Publication
- 2.1.3 Structured Review of a Publication
- 2.1.6 Annotated Resource is an Annotation
- 2.1.7 Annotation Metadata
- 2.1.8 Annotation has Multiple, Independent Comments and/or Tags
- 2.2.1 Bookmarking Current Reading Position
- 2.2.2 Highlighting a Span of Text
- 2.2.3 Commenting on a Span of Text
- 2.3.2 Cross Version Annotations
- 2.4.2 Persistence of Annotations
- 2.4.4 Annotation (or Part) is not Published Openly
- 2.4.5 Publication (or Part) is not Published Openly

A Web Annotation can be created to associate a node or a selection with a particular user motivation. The annotation can be registered and stored as a unit of information at actor’s *storage*, *type index location*, *outbox*, or article’s *annotation service*. The users can provide a text note for their annotations, assign a *license*, as well as designate an *inbox* for their annotation, so that it can receive its own notifications. The Table on WA motivations includes the method applied, location of the annotation, and associated notification.

Table 11. Implementation of Web Annotation motivations and notifications in dokieli

Motivation	Method	Storage	Notification
assessing	Assessments in the form of approval and disapproval with comments and license that can be created from and displayed in the UI. User mode: social.	Actor’s preferred storage, outbox, or annotation service.	Sends notification to target’s inbox about the annotation as a `as:Like` activity.
bookmarking	Node or text level targeting for an annotation including a description and tags. User mode: social.	Actor’s preferred storage, outbox, or annotation service.	None
classifying			
commenting	Node or text level targeting for an annotation including a description and tags. User mode: author.	Integrated into the target resource.	None
describing	A note to indicate a footnote. Used as part of purpose for bookmarking and commenting. User mode: social, author.	Actor’s preferred storage, outbox, annotation service, or integrated into the target resource.	None
editing			
highlighting	Node or text level targeting that gives the user a URI based on Web Annotation Selector. User mode: social.	URI selectable from the user agent or interface.	None
identifying			
linking			
moderating			
questioning	A note to indicate a request for specificity (like “citation needed”) including a license that can be created from and displayed in the UI. User mode: social.	Actor’s preferred storage, outbox, or annotation service.	Sends notification to target’s inbox about the annotation as a `as:Create` activity.
replying	A note to indicate a reply including a license that can be created from and displayed in the UI. User mode: social.	Actor’s preferred storage, outbox, or annotation service.	Sends notification to target’s inbox about the annotation as a `as:Announce` activity.
tagging	Used as part of purpose for bookmarking and commenting. User mode: author, social.	Actor’s preferred storage, outbox, annotation service, or integrated into the target resource.	None

Modified 2019-01-23

Motivation Web Annotation Data Model’s motivations.
Method User actions, data creating and reuse. "author" user mode refers to document structuring. "social" refers to common online interactions.
Storage The location of the annotation: actor’s preferred storage, outbox, or integrated into the target resource.
Notification Using the Linked Data Notifications protocol with notifications using any vocabulary. Notifications include Creative Commons license.

An brief example of a Web Annotation with the reply motivation:

```
<http://example.org/annotation>
  a oa:Annotation ;
  oa:motivatedBy oa:replying ;
  oa:hasTarget <http://example.net/article> ;
  oa:hasBody <http://example.info/note> .
```

Listing 15. Example annotation that is motivated by replying.

The Figure on private annotations serves to explain how each CERN requirement is realised with a standards-compliant Web application:

Sarven Capadisli replies

Authors

Sarven Capadisli

Published

2019-02-18 00:30:11

Rights

CC BY 4.0

Canonical

urn:uuid:d82804f8-e0ef-4a1e-95cd-4b2c8383eaf0

In reply to (part of)

> **Fragment selector conforms to**
>
> RFC 3987
>
> **Refined by**
>
> One must be able to add one’s own private links to and from public information. One must also be able to annotate links, as well as nodes, privately.

Rendered via

dokieli

Note

Language

English

Rights

CC BY 4.0

This annotation is independently created by an actor with a WebID and stored at their preferred location.

Figure 10. Annotation rendered.

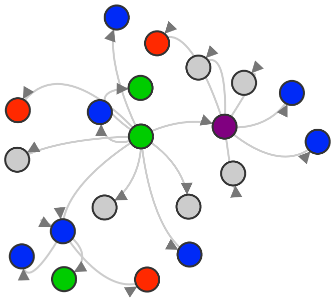


Figure 11. Annotation graph.

```
<article about="#a112cb4b-6814-4687-933d-8c25c917c746" id="a112cb4b-6814-4687-933d-
8c25c917c746" typeof="oa:Annotation" prefix="rdf: http://www.w3.org/1999/02/22-rdf-
syntax-ns# schema: http://schema.org/ dcterms: http://purl.org/dc/terms/ oa:
http://www.w3.org/ns/oa# as: https://www.w3.org/ns/activitystreams# ldp:
http://www.w3.org/ns/ldp#">
  <h3 property="schema:name">Sarven Capadisli <span rel="oa:motivatedBy"
resource="oa:replying">replies</span></h3>
  <dl class="author-name"><dt>Authors</dt><dd><span rel="schema:creator"><span
about="https://csarven.ca/#i" typeof="schema:Person"><img alt="" height="48"
rel="schema:image" src="https://csarven.ca/media/images/sarven-capadisli.jpg"
width="48" /> <a href="https://csarven.ca/#i"><span about="https://csarven.ca/#i"
property="schema:name">Sarven Capadisli</span></a></span></span></dd></dl>
  <dl class="published"><dt>Published</dt><dd><a href="#a112cb4b-6814-4687-933d-
8c25c917c746"><time datetime="2019-02-18T00:30:11.101Z" datatype="xsd:dateTime"
property="schema:datePublished" content="2019-02-18T00:30:11.101Z">2019-02-18
00:30:11</time></a></dd></dl>
  <dl class="rights"><dt>Rights</dt><dd><a href="https://creativecommons.org
/licenses/by/4.0/" rel="dcterms:rights" title="Creative Commons Attribution 4.0
International">CC BY 4.0</a></dd></dl>
  <dl class="canonical"><dt>Canonical</dt><dd rel="oa:canonical"
resource="urn:uuid:d82804f8-e0ef-4a1e-95cd-4b2c8383eaf0">urn:uuid:d82804f8-
e0ef-4a1e-95cd-4b2c8383eaf0</dd></dl>
  <dl class="target"><dt><a href="https://www.w3.org/History/1989/proposal.html"
rel="oa:hasTarget">In reply to</a> (<a about="https://www.w3.org/History
/1989/proposal.html" href="https://www.w3.org/History/1989/proposal.html"
rel="oa:hasSource" typeof="oa:SpecificResource">part of</a>)</dt><dd><blockquote
about="https://www.w3.org/History/1989/proposal.html" cite="https://www.w3.org
/History/1989/proposal.html"><div rel="oa:hasSelector" resource="#a112cb4b-
6814-4687-933d-8c25c917c746-fragment-selector" typeof="oa:FragmentSelector"><dl
class="conformsto"><dt>Fragment selector conforms to</dt><dd><a content="" lang=""
property="rdf:value" rel="dcterms:conformsTo" href="https://tools.ietf.org
/html/rfc3987" xml:lang="">RFC 3987</a></dd></dl><dl rel="oa:refinedBy"
resource="#a112cb4b-6814-4687-933d-8c25c917c746-text-quote-selector"
typeof="oa:TextQuoteSelector"><dt>Refined by</dt><dd><span lang="en"
property="oa:prefix" xml:lang="en"></span><mark lang="en" property="oa:exact"
xml:lang="en">One must be able to add one’s own private links to and from public
information. One must also be able to annotate links, as well as nodes,
privately.</mark><span lang="en" property="oa:suffix" xml:lang="en"></span>
</dd></dl></div></blockquote></dd></dl><dl class="renderedvia"><dt>Rendered via</dt>
<dd><a about="https://www.w3.org/History/1989/proposal.html"
href="https://dokie.li/" rel="oa:renderedVia">dokieli</a></dd></dl>
  <section id="note-20338295" rel="oa:hasBody" resource="#a112cb4b-6814-4687-933d-
8c25c917c746-note-20338295"><h4 property="schema:name">Note</h4><dl
class="language"><dt>Language</dt><dd><span content="en" lang=""
property="dcterms:language" xml:lang="">English</span></dd></dl><dl class="rights">
<dt>Rights</dt><dd><a href="https://creativecommons.org/licenses/by/4.0/"
rel="dcterms:rights" title="Creative Commons Attribution 4.0 International">CC BY
4.0</a></dd></dl><div datatype="rdf:HTML" lang="en" property="rdf:value
schema:description" resource="#a112cb4b-6814-4687-933d-8c25c917c746-note-20338295"
typeof="oa:TextualBody" xml:lang="en">This annotation is independently created by an
actor with a WebID and stored at their preferred location.</div></section>
</article>
```

Listing 16. Annotation HTML+RDFa.

Once an annotation is created and has a dereferenceable HTTP URI (subject to access control rules), a notification can be sent to the inbox of the target resource about the annotation if any part of an article advertises an LDN inbox. Depending on the motivation of the annotation, the notification is modelled with the *Activity vocabulary* eg. the Create activity (`as:Create`), to indicate that the actor has created the object.

In order to discover and view previously created annotations associated with an article, dokieli follows LDN's discovery mechanism for notifications about annotations, and then subsequently fetches the annotations in order to display them in context of their target selector.

In 2016, the *Annotating All Knowledge Coalition* [482] derived user stories based on the notion of what open annotation can enable. Below is a description of how dokieli realises the user stories derived from *A Coalition for Scholarly Annotation* [483]. The term "publisher" that is used in these user stories usually refer to an entity that is other than the creators of the content. Whereas in dokieli, a "publisher" is *any* actor capable of triggering the registration of resources and having their content potentially accessible to anyone.

As an author of a paper, I want to invite a small group of colleagues to provide feedback.
Implemented with the *Share* feature using the actor's social graph data, and LDN for dissemination.

As a pre-print publisher, I want to create and manage an overlay journal.
It is possible to open and render a collection of activities eg. published articles, which can be treated as a journal volume.

As a journal publisher, I want to implement a peer review system that brings authors and reviewers together in role-defined groups.
Actors can notify each other and annotate resources. However, there is no particular system to manage it as a whole.

As a university librarian, I want students' dissertations to be annotated openly so they carry forward an expectation that things should work that way.
Annotations with various motivations are implemented.

As a biomedical researcher, I want to use human curation of the literature to develop examples we can use to train our text mining algorithms.
Actors can select, add, and edit semantic markup (HTML+RDFa) as well embed structured data blocks for Nanotations using TriG, as well as JSON-LD and Turtle. Common document formatting features are supported.

As a journal publisher, I want to create safe spaces for commentary so that authors and expert readers can participate comfortably.
Actors with the required access control privileges to an inbox manages the information flow. There is no overarching system to create or manage *a space* as each resource is individually controlled by their owners. It is possible for LDN inbox owners to filter notifications which refer to unwanted annotations.

As a vision-impaired user, I want annotation tools to be accessible to me.
WCAG are applied, and ARIA is used where applicable. The application of ATAG is in progress.

As an annotator, I want to be able to assign DOIs to individual contributions, or to sets of them.
DOI generation or DOI registration is out of scope. Moreover, current DOI registration systems do not have authentication mechanisms that allows annotators to identify themselves with their WebID in order to assign a DOI.

As an author, I want annotation to integrate with my preferred writing tool so that annotations can flow back into the text.

Annotations that are generated use the Web Annotation Data Model. Web Annotation Selectors and States is used to create selectors that an actor can use.

As a scientist, I want to participate in a journal club that travels around the web, annotating papers in many different publishing systems.

Any accessible URL rendered as HTML (including different publishing systems) can be annotated, and the annotation can be stored at actor's preferred location eg. community's annotation service.

As an author, I want to annotate my own published work, noting errata and supplying new context.

Authors can annotate with various motivations and include "codetags".

As an editor, I want to document key editorial decisions and expose them to the public.

Editors can annotate with various motivations.

As a scientist I want to annotate everywhere using my ORCID identity.

Scientists using their ORCID URI (as their WebID) including the storage location in their profile can annotate anywhere.

As a researcher, I want control over the annotations I make in various contexts so I can assemble this work into a coherent form and claim credit for it.

Researchers with access controls to their annotation storage can.

As a publisher, I want to commission expert annotators whose commentary adds value to my content.

Publishers of articles and annotations can assign different inboxes to receive notifications about commentaries.

As a user, I want to be able to upvote or downvote annotations so I can help improve the quality of discussion.

Annotations can be created with the "assessing" motivation, as well as indicate their approval, disapproval, questioning with comments.

As a funder, I want to bring grant reviewers together in private annotation spaces where discussion is tightly coupled to the documents they are reviewing.

Reviewers can annotate with various motivations, including approval, disapproval, questioning with comments.

As publisher, I want to be able to review and moderate annotations in my own authoritative layer.

Actors can create their reviews where they prefer and have access to. Moderating annotations owned by other actors can only be done if the publisher has access to.

As an author, I want to be able to annotate and interconnect the references I cite.

Citing and annotations with various motivations is implemented.

As a publisher, I understand that readers may view my content through a variety of lenses, but I want to present my own annotation layer as the authoritative one.

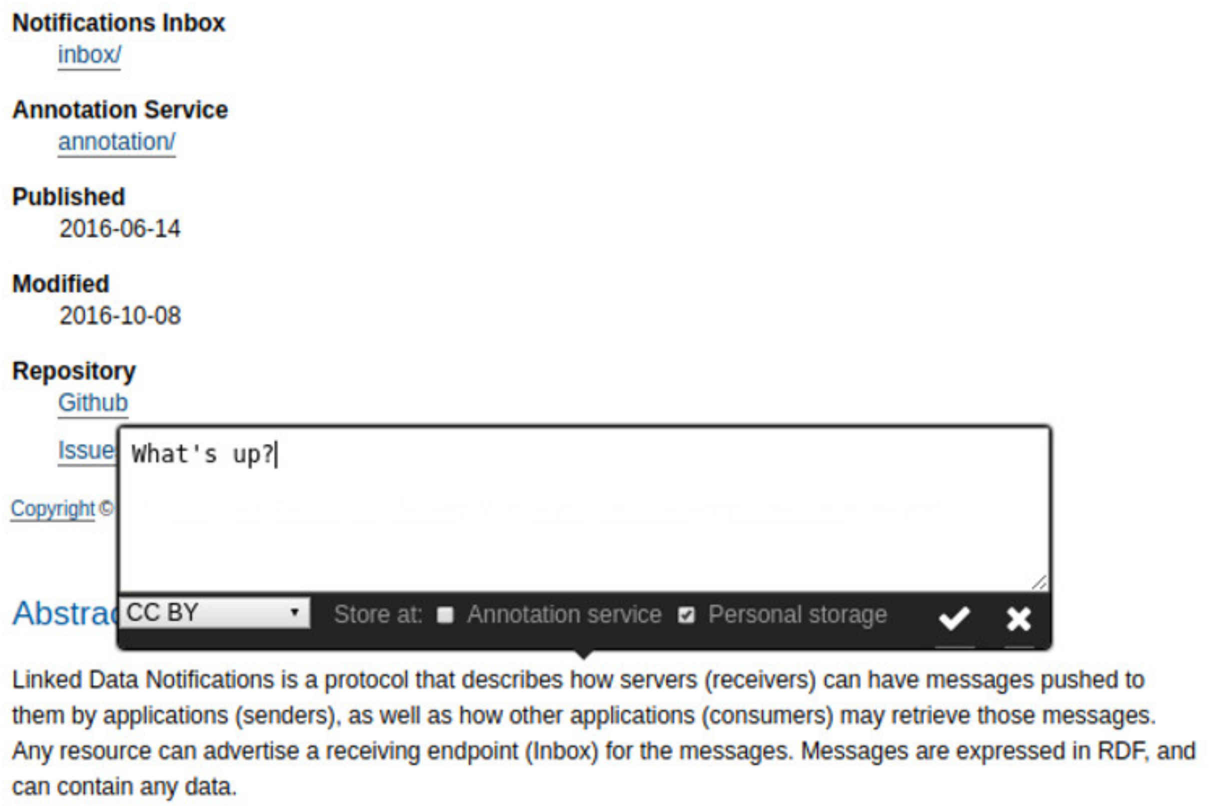
Annotations can be created with the "commenting" motivation, which gets embedded into the document (achieved by those with write access).

As the developer of a journal publishing system, I want to deliver a bundled annotation service that integrates with my identity system.

Using open standards for profiles, authentication and annotations means that anyone wishing to provide a hosted service need only adopt the same standards, eg. WebID, OIDC and WA.

As a publisher, I want to project a curated view of a paper's peer review process onto the published paper.
Memento and provenance information is generated and consumed in articles. All annotations and unique inboxes can be used. Robust Links are also shown.

A screencast of an annotation in dokieli is shown below.



Notifications Inbox
inbox/

Annotation Service
annotation/

Published
2016-06-14

Modified
2016-10-08

Repository
Github

Issue

Copyright ©

Abstra

What's up?

CC BY

Store at: ■ Annotation service ☑ Personal storage ✓ ✗

Linked Data Notifications is a protocol that describes how servers (receivers) can have messages pushed to them by applications (senders), as well as how other applications (consumers) may retrieve those messages. Any resource can advertise a receiving endpoint (Inbox) for the messages. Messages are expressed in RDF, and can contain any data.

Figure 12. Video of dokieli Web Annotation [31s, WebM].

6.2.5 Activities

The mechanism for personal identities, social graphs, and self-publishing articles and annotations enables a new layer of decentralised interactions. Here I describe how actors can discover units of information created or annotated in response to Web resources (like articles) by their social network without intermediaries (which includes the authors of the article or the features supported by the server it is published at). This scenario expands on *identity A knows identity B*, and *identity B has a storage or outbox* as declared in Linking the Decentralised Information Space.

As each unit of information can be independently published and made accessible by an actor, one way to discover the information is by determining the relation between their WebID and the location of their collection of artifacts eg. under their personal storage (`pim:storage`) or activity outbox (`as:outbox`). The second connection between the actor that is interested in discovering such information and the actors which created them (`foaf:knows`). Hence, we are able to discover the reactions (like replies, comments, approvals, questioning, or annotations in general) made by the members of an actor's social graph.

In order to realise this scenario, the required components are as follows:

- Actor A wants to read annotations or activities around a unit of information.

- Actor A knows actor B.
- Actor B has a collection of artifacts.
- Actor A has read access to Actor B’s activity collection.
- Actor A’s application associates the unit of interest with actor B’s annotation or activity.
- Actor A’s application displays the annotation in context of the unit of information.

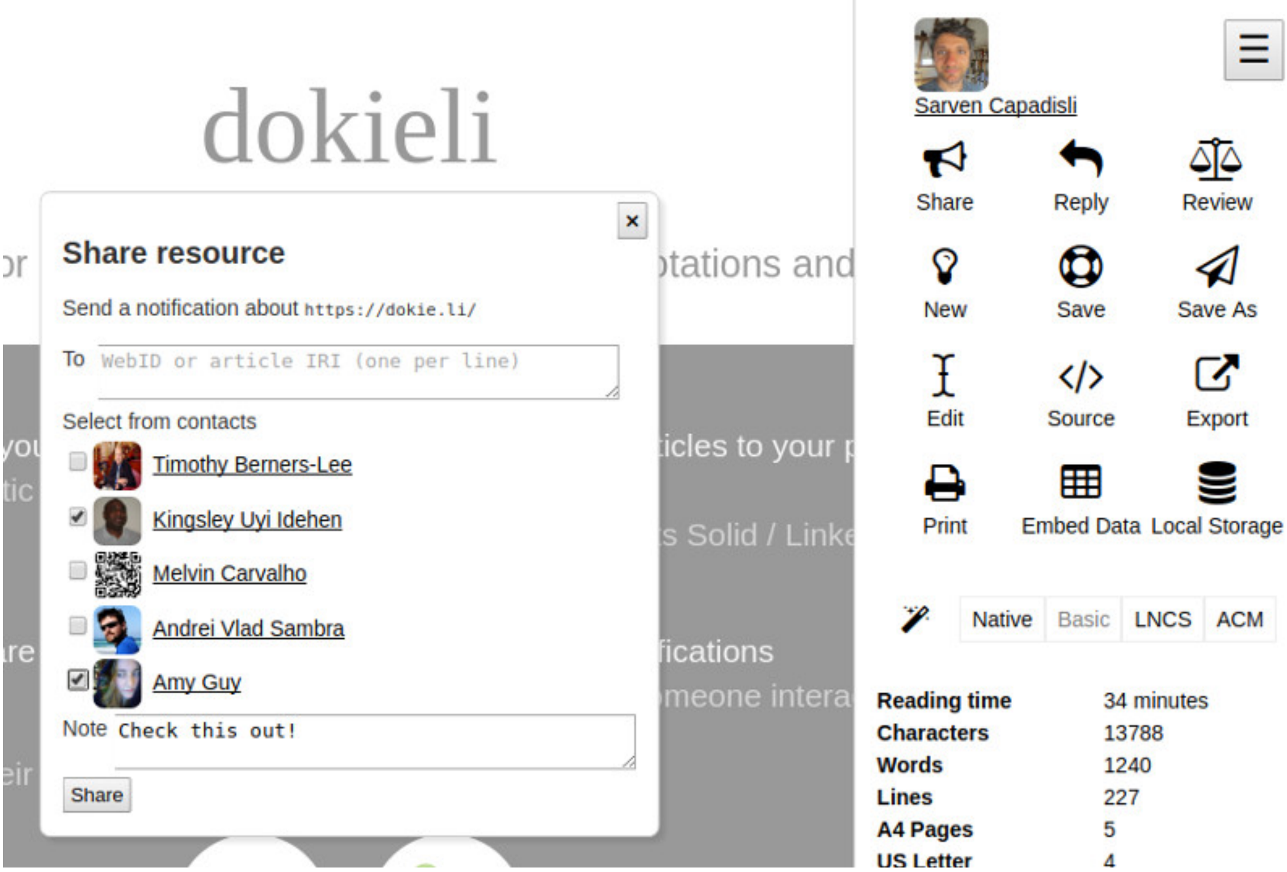
As these components are solely based on interoperable standards, any conforming application is able to interact with them, and perhaps more importantly, is not subject to *vendor lock-in* [484].

This scenario is realised in dokieli, and works as a *single-page application* or via the *Web browser extension*. The Web Extension enables a new dimension to discovering and consuming independently published activities in that, a user can navigate to any HTTP URL and potentially discover and interact with annotations made by the members of their network. Thereby, a layer of decentralised and linked research space can exist independently, where any Web resource, perhaps typically an article, can be incorporated into the social or research discourse.

As each annotation (or unit of information) can have its own inbox, subsequent interactions with the annotation can be both stored at the annotator’s storage and notified at the advertised inbox. This further decouples the components required to publish and consume content on the Web in an interoperable way. The decentralised activities in dokieli demonstrates the utility of independent publishing, semantic relations, as well as extensibility of structured Web objects.

6.2.6 Sharing Resources

A key aspect of the Social Web is sharing our creations and interests with the others. After (optionally) authenticating with a WebID, dokieli documents can be shared with contacts, which are discovered from the user’s WebID profile. Contacts whose profiles advertise an LDN Inbox will receive a notification of the share. The notification is modelled with the Activity vocabulary, eg. Announce, to indicate that the actor is calling the target’s attention to the object. Recipients can use any LDN-compatible application to view the notification, without needing to have ever used dokieli before.



Share resource

Send a notification about `https://dokie.li/`

To `WebID or article IRI (one per line)`

Select from contacts

- □ Timothy Berners-Lee
- ☑ Kingsley Uyi Idehen
- □ Melvin Carvalho
- □ Andrei Vlad Sambra
- ☑ Amy Guy

Note `Check this out!`

Share

Reading time	34 minutes
Characters	13788
Words	1240
Lines	227
A4 Pages	5
US Letter	4

Figure 13. Video of dokieli Share [36s, WebM].

6.2.7 Semantic Citations

Publishing and Referencing and *Scientific Expressions* described some of the existing specifications to identify, integrate, and link assertions in research. The registration of units of information with URIs enables us to create RDF relations between citing and cited entities. For example, actors can link to specific statements within and across resources, as well as annotate their intentions while preserving the context when they are created.

Contextual Citations: One way to create a citation in dokieli is by authors selecting a node or a text fragment of interest in their document, and then inserting the URI of the resource to be linked to, as well as a specifying the type of semantic link between them in a specific or general way using the classifications available from the CiTO ontology eg. using “agrees with”, “confirms”, “cites as evidence”. The input for the cited entity can be any HTTP URI, DOI string, or ISSN string, and when resolved from a corresponding HTTP URL, the content in RDF is used towards constructing the citation. dokieli takes the input target URI, and retrieves data (like title, authors) expressed in RDF in the linked document, integrates a statement expressed in RDF eg. figure cites as evidence, in context of the content, the citation intention as human-readable text, and adds a scientific endnote reference including the retrieved information. As a citation results in the assigning an HTTP URI for the citing entity, it can be further semantically annotated by the actor.

```
<http://example.org/article#argument>
  cito:citesAsEvidence <http://example.net/observation#results> .
```

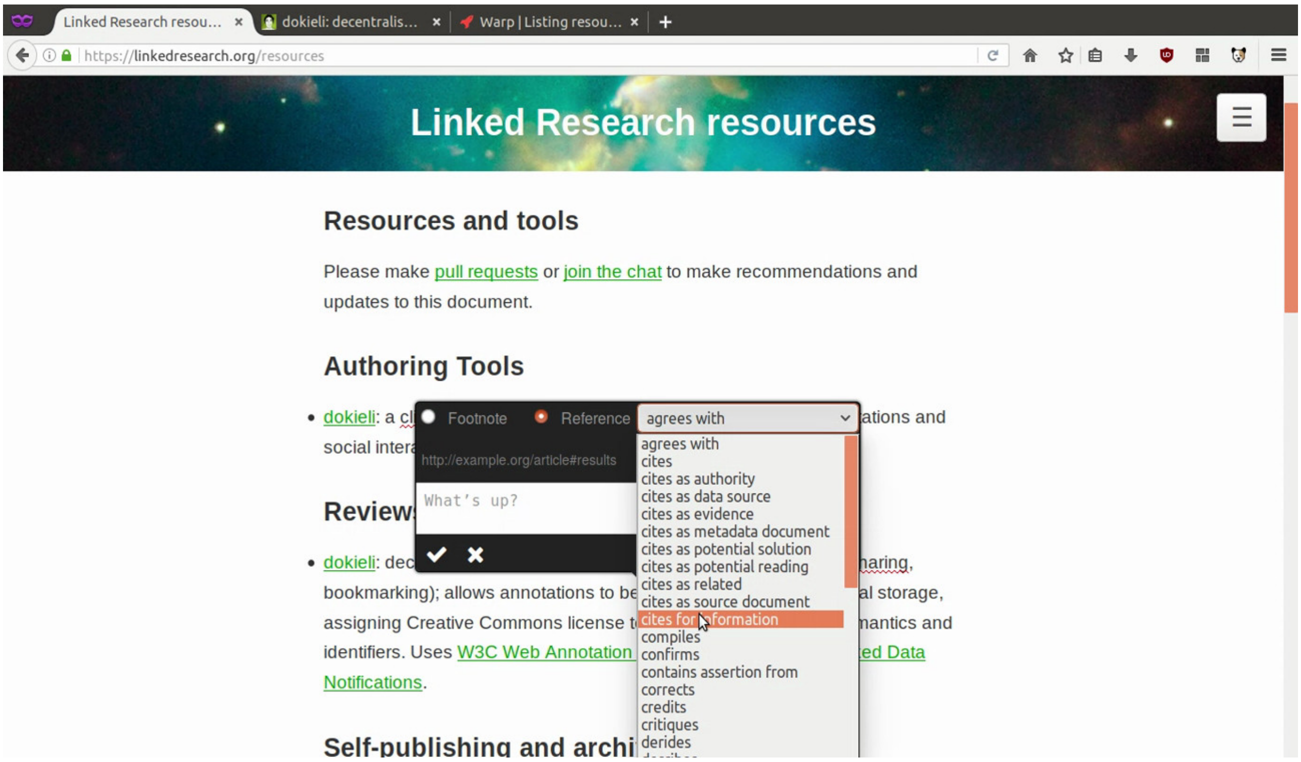
Listing 17. The citing entity (an argument) cites the cited entity (observation results) as source of factual evidence for statements it contains.

Robust Citations: In order to preserve the context of the citation as well as taking content-drift and link-rot into account, first a snapshot of the cited entity's URI is created at the *Internet Archive*, and then both the archived URI and its datetime stamp is incorporated to the referencing hyperlink – the original citation resource.

Notifying Cited Entity: Once a citation is made through the dokieli user interface, dokieli proceeds to check if the cited resource advertises its own LDN inbox, and if so, sends a notification indicating the citation function. This notification serves as a back-reference announcement, ie. being informed that a citation was made and thus creating discoverable *bi-directional linking* between the units of information. With the help of an inbox consumer, like dokieli, the cited article can also inform the reader about the back-references out there on the Web.

Registrations: The citation activity results in the registering HTTP URIs for the citing resource, archived snapshot, and the notification resource.

The following screenscast illustrates a clientside interaction in dokieli where the user cites a unit of information with a particular relationship, and a notification is sent to the cited entity's inbox.



Linked Research resources

Resources and tools

Please make pull requests or join the chat to make recommendations and updates to this document.

Authoring Tools

- dokieli: a cl… ations and social intera…

Review…

- dokieli: dec… haring, bookmarking); allows annotations to be… al storage, assigning Creative Commons license t… mantics and identifiers. Uses W3C Web Annotation… ed Data Notifications.

Self-publishing and archi…

Figure 14. Video of semantic inline citations and notification in dokieli [43s, WebM].

6.2.8 Social Reviews

It is possible to use dokieli's create operations as a composite to fulfill social and scholarly activities. One common scenario includes soliciting reviews, reviewing, announcement of the review, and final decisions made by community members. Core mechanisms, annotations and sharing resources in dokieli is as follows:

- Actor A publishes their article at preferred location.
- Actor A sends Actor B a notification requesting a review of their article.
- Actor B creates a derived copy of Actor A's article at their preferred location. The derived resource optionally includes its own inbox and annotation service.
- Actor B reviews and annotates the derived resource by storing their annotations at preferred location. Notifications are sent to the inbox about the annotation.

- Actor B archives their derived copy with annotations.
- Actor B sends Actor A a notification sharing the location of the annotated resource.
- Actor A visits Actor B’s derived resource and discovers the review and annotations.
- Actor B sends a notification to a community inbox that is dedicated to monitor scholarly activities about their review and annotations so that it can be discovered by other actors.
- Actor C discovers Actor B’s review through the community inbox.

This particular scenario shows that it is possible for each actor to create and control resources, including articles, annotations, and notifications at their PDS or preferred locations. It is one way to combine standards-compliant mechanisms where actors can shape their content, self-publish, as well as share their works with other actors and services. In essence, the registration, certification, awareness, and archiving functions in scholarly communication are fulfilled in this scenario without a central authority.

6.2.9 Privacy Considerations

Creating awareness and enabling actors to protect their data privacy is essential to fostering user’s trust in the systems they use. The adoption of standards and practices with different degree of control can yield different privacy measures. In the decentralised and interoperable environment that dokieli operates in, there are a number of challenges and approaches. I discuss user privacy, data privacy, preferred proxy, and private activities below. The terms “anonymity”, “unlinkability”, “undetectability”, “unobservability”, and “pseudonymity” follow the definitions from *A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management* [485], Pfitzmann, 2010.

User Privacy: dokieli acts as a conduit to perform read-write operations against PDS. As the rules for read-write operations are set by servers, dokieli does not require the creation of user accounts at any server, in order to interact with the server’s resources. Where applicable, actors can participate by maintaining their *anonymity*, *pseudonymity*, or can choose to identify themselves to dokieli by using their preferred personal identities, without necessarily disclosing "real" information. Actors authenticate against the servers they prefer in order to perform access-controlled actions. It is possible for the environment in which dokieli operates in to transmit device, machine, or browser fingerprint, thereby potentially risking *unlinkability* – however, they are outside the control and scope of dokieli.

Data Privacy: With respect to data privacy, actors can exercise their autonomy through the use of PDS for their profiles, articles, annotations, and notifications. As a PDS can provide us with the means to safeguard data, dokieli as a clientside application enables actors to perform read-write Linked Data operations where they prefer under their own conditions. With the help of access controls, a PDS that dokieli communicates with can potentially fulfill all minimizations – “anonymity”, “unlinkability”, “undetectability”, “unobservability”, “pseudonymity” – possibly with the exception of the domain name and IP that the PDS operates from.

Preferred Proxy: A proxy acts as an intermediary between the clientside application and the destination server. This is useful in a number of scenarios for dokieli, such as getting around Mixed Content issues, absence of *Cross-Origin Resource Sharing* [486] (CORS) support, as well as concealing the identity of the actor, fulfilling *anonymity*, *unlinkability*, *pseudonymity*, when making requests to a target server. An actor’s preferred proxy can be determined by checking their WebID Profile description for the `solid:preferredProxy` relation, and if found, dokieli can use it to route requests when needed. Alternatively, a built-in (but user configurable) proxy location in dokieli is used.

Private Activities: The ability for actors to create and discover activities made within their social network without going through intermediaries, makes it possible to isolate units of information that is both access-controlled and invisible (or be unknown) to third-parties. These class of interactions in dokieli is aligned with the notion of *undetectability* and *unobservability*.

6.2.10 Interactive Linked Statistics

dokieli implements *sparqlines*: statistical observations fetched from SPARQL endpoints and displayed as inline-charts. An inline-chart, also known as a *sparkline* [487], is concise, and located where it is discussed in the text, complementing the supporting text without breaking the reader’s flow. For example, the GDP per capita growth (annual %) [Canada] claimed by the World Bank Linked Dataspace. dokieli demonstrates how scientists or authors can better support their argumentation with sparklines generated from their own or public statistical linked datasets.

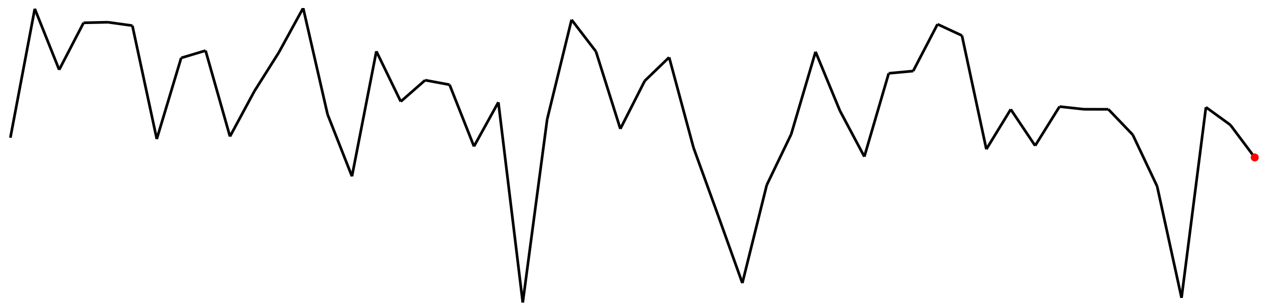


Figure 15. Video of Sparqlines interaction in dokieli [26s, WebM].

6.2.11 Deployment

dokieli is a clientside application that uses the user-agent to interact with server applications. It has two deployment approaches including the same set of features: *single-page application* [488] and *Web browser extension* [489].

dokieli’s presentational and behavioural code layers can be included in Web "pages" in order to trigger them as active single-page applications. It is a smart client that allows different kinds of articles eg. academic, blog posts, news, to be authored and annotated from within Web browsers, without necessarily having them deployed from a server, ie. it can be used offline or on localhost. dokieli internally handles its content and structural and semantic representation based on user’s actions and available information. Articles, profiles and their contact information, notifications, annotations with different motivations, for instance, can be read and written ubiquitously to any Web space with supporting HTTP methods and access control mechanisms.

The Web browser extension is a thin wrapper around dokieli’s core code in order to embed itself in any HTML-based Web page on the Web. It inherits all of the features of a single-page application. While HTML based documents on the Web vary in their quality, dokieli’s write operations generate HTML+RDFa. One of the utilities for the extension is to have a consistent interface for annotating node or text selections on a Web resource, building structured citations, creating derivations of Web resources as well as sharing document fragments with ones contacts via notifications, without having a service dependency or being limited by the Web page’s UI.

6.3 Semantic Content Authoring Quality Attributes

In *User Interfaces for Semantic Authoring of Textual Content: A Systematic Literature Review* [490], Khalili, 2011, provide an analysis on approaches for “Semantic Content Authoring” (SCA), which is defined as “a tool-supported manual composition process aiming at the creation of semantic documents. With an ontology and a user interface appropriate for the type of content, semantic authoring can be easier than traditional composition of content and the resulting content can be of higher quality.” The article proposes a set of “quality attributes” – as non-functional requirements – to evaluate the performance of SCA systems: usability, customizability, generalizability, collaboration, portability, accessibility, proactivity, automation, evolvability, interoperability, and scalability. Based on the starting point of the authoring process, the authors distinguish the approaches, and generally categorised as “bottom-up” (lifting unstructured content to a semantic level) or “top-down” (semantic representations from the beginning requiring users with the knowledge of the domain), and sometimes a hybrid of the two. A survey of tools are presented (in article’s Table 3) covering different types of users, expected domain knowledge, and authoring approach. The article also discusses *Research and Technology Challenges* and identifies a gap in addressing “accessibility”, “handling complexity in UIs”, “formal and systematic methods for user interface evaluation”, “heuristic evaluation”, “support of crowdsourcing”, UIs for ubiquitous devices.

The scope of SCA is relevant because it covers some of the aspects of dokieli that can be examined. Hence, I use the structure of the table of results for SCA systems and include a dataset about *dokieli’s quality attributes*:

Table 12. Quality attributes of dokieli

Quality Attribute	**UI Feature**	**User Type**		**Authoring Approach**		**Domain**	
		Expert	**Non-expert**	**Top-Down**	**Bottom-Up**	**Domain-independent**	**Domain-specific**
Usability	**Single Point of Entry Interface**	✔	✔		✔	✔	
	Faceted Browsing						
	Faceted Viewing	✔	✔		✔	✔	
	Inline Editing and View Editing	✔	✔		✔	✔	
Customizability	**Living UIs**	✔	✔		✔	✔	
	Providing Different Semantic Views	✔	✔		✔	✔	
Generalizability	**Supporting Multiple Ontologies**	✔	✔		✔	✔	
	Supporting Ontology Modification						
	Supporting Heterogeneous Document/Content Formats						
Collaboration	**Access Control**	✔	✔		✔	✔	
	Support of Standard Formats	✔	✔		✔	✔	
	UIs for Social Collaboration	✔	✔		✔	✔	
Portability	**Cross-browser Compatibility**	✔	✔		✔	✔	
	UIs for Mobile Devices	✔	✔		✔	✔	
Accessibility	**Accessible UIs**	✔	✔		✔	✔	
Proactivity	**Resource Suggestion**	✔	✔		✔	✔	
	Real-time Semantic Tagging						
	Concept Reuse	✔	✔		✔	✔	
	Real-time Validation						
Automation	**Automatic Annotation**						
Evolvability	**Resource Consistency**	✔	✔		✔	✔	
	Document and Annotation Consistency	✔	✔		✔	✔	
	Versioning and Change Tracking	✔	✔		✔	✔	
Interoperability	**Support of Standard**	✔	✔		✔	✔	

Extension of *Table 6: Overview of results* [491] in *User Interfaces for Semantic Authoring of Textual Content: A Systematic Literature Review* including dokieli.

		User Type		Authoring Approach		Domain	
Quality Attribute	**UI Feature**	**Expert**	**Non-expert**	**Top-Down**	**Bottom-Up**	**Domain-independent**	**Domain-specific**
	Formats						
	Semantic Syndication	✔	✔		✔	✔	
Scalability	**Support of Caching**	✔	✔		✔	✔	
	Suitable Storage Strategies	✔	✔		✔	✔	

Extension of *Table 6: Overview of results* [491] in *User Interfaces for Semantic Authoring of Textual Content: A Systematic Literature Review* including dokieli.

6.4 FAIR Metrics

As the *FAIR guiding principles* has been increasingly accepted and endorsed by research communities, I apply the corresponding *FAIR Metrics* [492] as outlined in *A design framework and exemplar metrics for FAIRness* [493], Wilkinson, 2018, on dokieli. The FAIR metrics is intended to help assess the degree in which the resources issued through dokieli meets the FAIR principles.

6.4.1 FAIR Metrics Data Structure Definition

Dimensions

Subject - the test subject of an assertion

Test - the test criterion of an assertion

Mode - the mode in which the test was performed

Measures

Result - the result of an assertion (Outcome, Info)

6.4.2 FAIR Metrics Dataset

Identifier

a2ebd5be-b9ca-11e8-bdaa-338eda30385e

Published

2018-09-16

Creator

https://csarven.ca/#i

Structure

FAIR Metrics

Table 13. FAIR Metrics results

Subject	Test	Mode	Outcome	Info
dokieli	FM-F1A	semiAuto	✔	Initiates the registration of HTTP URIs. Creates local identifiers (including Web Annotation Selectors and States) in documents. Uses HTTP URIs, DOI and ISSN names for citations.
dokieli	FM-F1B	semiAuto	✔	Policy about identifier persistence can be created and retrieved.
dokieli	FM-F2	semiAuto	✔	All information of significance is machine-readable based on the RDF language, encoded using RDFa in HTML and SVG, JSON-LD, Turtle.
dokieli	FM-F3	semiAuto	✔	No distinction made for "data" and "metadata" as all documents are human- and machine-readable, and can refer to other resources.
dokieli	FM-F4	semiAuto	✔	Units of information is findable from search engines, Web archives, *Linked Open Data Cloud*.
dokieli	FM-A1.1	semiAuto	✔	Uses WebID-TLS and WebID-OIDC to access HTTP URIs.
dokieli	FM-A1.2	semiAuto	✔	HTTP URIs requiring authentication relies on an actor’s use of WebID-TLS or WebID-OIDC, and the authorization by the server will be handled by its own access policies eg. ACL, communicated back to the client. Actors can publish ACL rules for HTTP URIs.
dokieli	FM-A2	semiAuto	✔	Policy about metadata plan can be created and retrieved.
dokieli	FM-I1	semiAuto	✔	All units of significance can be expressed through accessible Linked Data vocabularies.
dokieli	FM-I2	semiAuto	✔	All units of significance can be expressed through accessible Linked Data vocabularies.
dokieli	FM-I3	semiAuto	✔	All units of significance can be expressed through accessible Linked Data vocabularies. Generated HTML, SVG, MathML are the host language for RDFa. JSON-LD and Turtle are also used to create and consume Web resources.
dokieli	FM-R1.1	semiAuto	✔	Uses Linked Data vocabularies to express statements about particular rights and licenses in resources eg. Creative Commons. Information pertaining to rights and license can be created and retrieved.
dokieli	FM-R1.2	semiAuto	✔	Information pertaining to provenance is generated and can be explicitly included by actors eg. can express "was derived from", "cites as evidence" to any unit of information.
dokieli	FM-R1.3	semiAuto	✔	Resources can be shared with the *Linked Open Data Cloud*, as well as with *Linked Open Research Cloud* using the *Linked Data Notifications* protocol. Any consumer conforming to various Read-Write Linked Data protocols and Linked Data vocabularies can reuse.

✔	Passed
✗	Failed
?	Cannot tell
⌙	Inapplicable
○	Untested

FM-F1A: Identifier uniqueness [source]
FM-F1B: Identifier persistence [source]
FM-F2: Machine-readability of metadata [source]
FM-F3: Resource Identifier in Metadata [source]
FM-F4: Indexed in a searchable resource [source]
FM-A1.1: Access Protocol [source]
FM-A1.2: Access authorization [source]
FM-A2: Metadata Longevity [source]
FM-I1: Use a Knowledge Representation Language [source]
FM-I2: Use FAIR Vocabularies [source]
FM-I3: Use Qualified References [source]

Subject	Test	Mode	Outcome	Info
FM-R1.1: Accessible Usage License [source] **FM-R1.2:** Detailed Provenance [source] **FM-R1.3:** Meets Community Standards [source]				
The generated and consumed data reflects dokieli's capabilities with respect to potentially interacting with servers adhering to common open Web standards and practices.				

This table shows that dokieli fares well with the FAIR principles. The technical standards and practices that are applied to dokieli fulfill each metric. One observation here is that the application of Linked Data principles “out-of-the-box” consequently scores high against the FAIR Metrics.

6.5 Forces and Functions in dokieli

This section goes into detail about how dokieli fulfills or supports some of the forces and functions in scientific communication.

As dokieli is employed, I observe the *forces*:

Actor

Any *web agent* [494] — person or software — can use WebIDs for their identities to read and write resources. Actors can contribute anonymously, be pseudonymous, or unlinkable. The application adapts to the information that is available from the profile description. An actor's social graph can be discovered and used to communicate.

Content

Actors can author, publish, and reuse Web resources, including articles, annotations, notifications, as well as query statistical data and lookup citation references. Any fragment of *resource of significance* can be assigned a URI, defined or described, as well as interlinked with other URIs. Information is semantically expressed with Linked Data *Vocabularies* by the application and actively by the actor.

Accessibility

Actors can access publicly available HTTP URIs directly or authenticate themselves in order to retrieve representations of access controlled Web resources – including articles, annotations, and notifications – provided that they are authorized to. Actors can set access rules on resources they control.

Applicability

Applicability of scholarly communication comes from research which responds to problems or answers questions, which can be used or built upon by others. The application makes it possible to both express and share units of information pertaining to applicability.

The *functions*:

Registration

Articles, annotations, and notifications can be registered as their own entities by assigning HTTP URIs at any location the actor has ownership or access to, facilitating information discovery and reuse. Registration of units of information are triggered by the application and then realised by servers. Registration is also triggered by requesting a snapshot from archiving services.

Awareness

The LDN protocol is used to disseminate notifications to document, profile, or community managed inboxes. The kind of notifications range from information sharing, request for feedback, replies, assessments, or annotations of resources, to bringing attention to citation back-references.

Certification

Units of information at any level of granularity can be annotated or replied to, from entire articles to individual data points as part of quality-control. All artifacts for the purpose of quality-control are registered units of information that can be dereferenced, and can be announced to individuals or groups.

Archiving

dokieli can inform the APIs of archiving systems about the existence and state of units of information that can be crawled and preserved. Actors can reuse the archived URIs. Robust citations are created by registering an archived version of the cited entity and the datetime stamp is preserved at the time of linking.

6.6 Contextualising dokieli

dokieli is composed of use cases that are realised through a collection of open Web standards. I contend that dokieli is a *socially-aware read-write Linked Data* application. It demonstrates how decentralised units of communication can be co-created and used by actors and applications.

6.6.1 Web Publishing Systems

Throughout this article I have highlighted the environment that the AWWW enables with the goal of fostering a universally accessible and interoperable information space. I have also looked at the ongoing initiatives towards the decoupling of functions in scientific communication. Here I reuse some of these notions to situate dokieli with some existing work that can be (subjectively) *considered as representative* for different areas in “Web publishing”. There is plethora of examples for centralised and decentralised systems on the Web. Some of the authoring and publishing systems already fulfill some aspects of decentralised systems which operate over HTTP. I focus on degree of control with respect to applications, data and identity, privacy, accessibility, and interoperability.

Currently, *Google Docs* [495], *Medium* [496], *Twitter* [497], *Facebook* [498] are examples of well-known Web applications for collaborative creation and publication of content. Users create accounts on these systems to perform access controlled read-write operations on data stored on the same server. These kind of systems allow users to authenticate with their identity provider from preconfigured authorities, as well providing their own accounts on the system. Each service generally has a custom API and vocabulary that a trusted client application (normally hosted on the same server) uses to interact with the server in order to perform CRUD operations through the user interface. They are categorically a client-server that is *tightly coupled* on identity, identification, authentication, authorization, and storage.

WordPress [499] is a free and open-source client-server application for article publication which can be self-hosted on a server controlled by the user. Users may optionally sign-in with their WordPress accounts to leave comments on others’ articles where they are stored by the hosting site.

MediaWiki [500] is a wiki software for collaborative content creation, structuring, and modification, most well known for its use by *Wikipedia* [501]. It runs as a service with a server and a frontend client application in the Web browser. Users create accounts per service in order to further control the frontend to perform CRUD operations. Content can be licensed, and changes are versioned with HTTP URIs. Human-readable content is available in HTML, and considered to be accessible and archive-friendly. Data into the system can flow in from *Wikidata* [502], a storage for structured data – expressed and dereferenceable as RDF with a SPARQL query service – used by Wikimedia projects.

Hypothes.is [503] has a server and client application that makes it possible for users to leave annotations on different types of documents on the Web using either a browser extension, proxy service, or including its JavaScript library in HTML documents. Annotations may be private, public, or for a group, and can be threaded to form conversations around a piece of content. Each instance of a client application is preconfigured to use its own API endpoint against a particular server to perform CRUD operations, such as creating and finding annotations and related data. Actors use the client application to control the user accounts at a server or preconfigured identity providers eg. hypothes.is, orcid.org, acknowledged by a server. Annotations can be self-hosted, however different servers do not federate. The HTML representations of annotations requires JavaScript to be available and enabled on the user-agent in order to look up and fetch the annotation's contents in JSON from its API and update the DOM. The annotation URLs, proxy, and the browser extension uses the *Google Analytics* tracking service per active URL. On active URLs, the client application by default makes API calls to the Hypothes.is server to check for annotations. On-demand archiving services like the Internet Archive are currently unable to properly archive an annotation's contents. It is possible to fetch an annotation that is represented with the WA Data Model from a server, however the annotation is not machine-discoverable ie. other applications will need to be uniquely configured with Hypothes.is's annotation URL template in order to acquire a JSON-LD representation - which is semantically a subset of the information provided in the HTML and JSON representations.

One common architectural aspect to these systems is that while they have some characteristics of a loosely coupled system, the server is expected to work with a designated client application. For instance, they require account creation and data storage with respect to an appointed centralised service. In essence, the user interface is restricted to communicate with a particular service endpoint, as well as an authority that governs all actor identities and the data they have produced. They allow multiple participants to annotate and hold discussions around the primary content; users must access their accounts to be notified of updates to conversations, and data from both the main content and related discussion is confined to the service which was used to create it. The data formats and schemas are consistent across storage instances in that it is possible to migrate data from one location to another using the same server application, storage mechanism and data schema. Archivability of Web resources varies and their access being at the discretion of server owners.

LibreOffice Online [504]'s *Writer*, for example, allows collaborative editing of office documents from the Web browser. Content can be stored under different CMSs in the cloud. The document's interface consists of image tiles which are sent from the server and rendered in the browser. However, it hardly provides accessibility, rich interlinking and annotations.

> I have saved discussing the landmark to last, Amaya and Annotea:

Amaya [505] is a desktop Web browser and editor that was developed by W3C (1996-2012) to provide a framework to experiment and validate Web specifications. It allows documents to be created and updated with remote access features.

As per *Annotea: An Open RDF Infrastructure for Shared Web Annotations* [506], Kahan, 2001, the *Annotea* [507] system was the earliest approach to creating and sharing Web-based annotations built on the RDF infrastructure. The client application was integrated into Amaya where it was able to store and recall annotations locally or from a remote annotation server, with the expectation of an access controlled server using the standard HTTP authentication mechanisms. Annotea annotations are independent Web resources. Annotea has a *protocol* [508] that specifies the CRUD operations for servers and clients, and enables different applications to be built. Annotea in Amaya for instance allowed the user to specify the URI of an annotation server. In order to reference target resources, annotations used *XML Pointer Language* [509] (XPointer) to specify the part of HTML or XML-based documents that the annotation was in context of. Annotea laid out a path for subsequent systems, influencing *Open Annotation Data Model* [510] which later evolved into the WA Data Model.

Amaya and Annotea emerged at a time when publishing and consumption standards on the Open Web Platform were still relatively in the early phases. The most apparent conceptual similarities between the Amaya and Annotea combination with dokieli is that they treat the hosting document as one of the environments to perform read and write operations from.

⁂

A tentative view of the interoperability between Hypothes.is and dokieli is as follows. The client applications Hypothes.is and dokieli are non-conflicting in that when used in the same content or context, each can still be used on its own. A Hypothes.is client application can search and fetch annotations from an associated server using the Hypothes.is API, and render them on the target resource. dokieli can discover the WA annotation service of a Web resource, fetch annotations from servers conforming to WAP and LDP specifications, and render them on the target resource. Additional data about the annotation can be created and reused in both applications. A Hypothes.is client application can authenticate with a particular server using OAuth. Actors with WebID in dokieli can authenticate with servers supporting WebID-TLS and WebID-OIDC. Both can create annotations given authorization. Hypothes.is application can extend and direct annotations to a community group using the same server. dokieli triggers content negotiation with servers to determine preferred RDF serialization of Web Annotation and profile. The Hypothes.is client can only fetch annotations from a corresponding Hypothes.is server. dokieli can only fetch annotations from servers implementing LDP, WAP or AP. Putting the method of fetching an annotation resource aside, it is possible for either client application to render an annotation serialized as JSON-LD that is created by the other application. The notion of "private" content in Hypothes.is is tied to the annotation server that client application is associated with, whereas "private" content in dokieli is associated to actor’s preferred storage location and access control policies based on individual, class of, or group WebIDs.

Existing server and client applications in particular to authoring and publishing are complex systems in that while they adhere or adopt some aspects of open Web specifications, they also include proprietary designs. Such systems also have varying characteristics of centralisation and decentralisation, hence, while some aspects are interoperable, other areas require post facto efforts to interop.

The tools which provide good collaborative editing user-interfaces appear to do so at the expense of data ownership and interoperability; those which promote creation and publication of data in open reusable formats are lacking facilities for linking discourse and conversation to concepts published. Decentralised creations also mean that each author can choose their own semantics eg. their own vocabulary to annotate using the RDF language, and then such decentralised documents can link to each other and their schemas can also be mapped to each other, whereas in centralised systems this is (if they support structured semantics at all) often prescribed, either technically enforced, or encouraged by social convention.

6.6.2 Adoption of dokieli

dokieli is first and foremost intended to demonstrate the potential of an application that is built entirely on open Web standards, meanwhile enabling actors to control their own data and respecting their privacy. During the process of exemplifying its core principles as an implementation, it has raised awareness and had some success in its adoption. Here I list some of those:

The W3C Working Group Note *Embedding Web Annotations in HTML* [511] includes examples from dokieli’s publishing and consuming of the Web Annotation Data *Model* [512] and *Vocabulary* [513] with motivations based on general use cases, for example: personal annotating, offline-first annotations, lightweight, decentralised annotation tools, collaborative annotating, and wholly internal annotations.

The W3C *Linked Data Notifications* [514] specification uses dokieli’s HTML+RDFa template, and the *Editor’s Draft* [515] showcase dokieli as a consumer of LDN and Web Annotations, as well as allowing the user to create new annotations and send notifications. The LDN *Test Suite* [516] also uses dokieli’s templates and stylesheets.

The academic workshop *Semantic Statistics* [517] series use dokieli in its website templates, including the call for contributions.

CEUR-WS.org [518], an “Online Proceedings for Scientific Conferences and Workshops” offers the tooling *ceur-make* [519] to help organisers generate proceedings using dokieli's HTML+RDFa template, eg. *SemStats 2016 Proceedings* [520].

A community of researcher who self-publish their articles and thesis using dokieli with different stylesheets and derived scripts are listed in dokieli's *examples in the wild* [521].

The conference series: WWW (eg. LDOW and WOW workshops), ISWC, and ESWC propose dokieli as one tooling in which authors can use to make their contributions to the calls with.

Solid Panes [522] is a set of core Solid-compatible applications includes to create human-readable and Linked Data documents. A pane displays the dokieli document using a part of its window.

The *Linked Research* [523] website uses dokieli in its templates on the site as well as workshop proposals and call for contributions.

My personal website at *https://csarven.ca/* [524] uses dokieli, where some articles – like this one you are reading – offer pointers to a public annotation service in which users can optionally send their annotations to, in addition to their own storage. Some articles – like this one you are reading – have an inbox associated with the article to send notifications to eg. replies, reviews, citations.

6.6.3 Degree of Control in dokieli

At the beginning of this section, requirements for *decentralised Web publishing* relating to *Degree of Control* were listed. dokieli fulfills variety of these use cases pertaining to decentralised Web publishing by meeting these requirements. dokieli demonstrates one configuration of open Web standards that can operate in a loosely coupled environment.

Future work can examine how additional features can be realised on top of existing Web standards, or where more development is required. Real-time collaborative editing is often realised with centralised communication (even though some peer to peer alternatives exist). Services like top-down annotations or automated entity marking can improve the discoverability of a publication, yet the question of how to offer these without being tied to certain servers as well as Web identities served from fixed servers needs further work.

Actors can read-write human- and machine-readable Web resources

dokieli uses HTML+RDFa, JSON-LD, and Turtle to express human- and machine-readable units of communication, which fulfills both internal and external requirements from relevant Web specifications. All units of significance can be expressed through Linked Data vocabularies, in particular, the common kinds of resources such as articles, annotations, and notifications.

Application can read-write Linked Data Notifications and Web Annotations

The use cases pertaining to annotations with different motivations, social activities, sharing resources, semantic citations, social reviews, as well as privacy can be addressed.

Actors use their preferred Web identity

It is possible for actors to use any WebID in order to adapt dokieli's user interface and possible actions based on the available (extended) profile description.

Actors use their preferred clientside application on access-controlled resources.

WebID-based access-controlled resources can be read and manipulated by actors where they authenticate through WebID-TLS and WebID-OIDC.

Actors switch between applications without having to move or alter the schema of their data.

The data and information that dokieli generates is expressed with Linked Data vocabularies. Any other application with the ability to understand or work with the same Linked Data vocabularies can reuse the dokieli generated data.

Actors relocating data without having to change their preferred application.

All units of communication are decentralised in that no central server acts as root or starting point in discovery. dokieli's awareness of the actor's profile starts by where they input their WebID, which can be at any location. Similarly, inboxes of various objects and annotations services are dynamically discovered.

Querying, visualising, and creating statistical data from clientside applications.

Statistical Linked Data can be looked up based on actor's actions, visualised and integrated with their document, and finally the document include the Linked Data can be saved.

6.6.4 Effects and Artifacts of dokieli

dokieli shows that it is possible for people to interact with each other without central coordination. Users can choose storage space for their content independently of the applications with which they edit and view that content. Documents are connected statically through links and dynamically through Linked Data Notifications. This is a proof for the viability of a social and decentralised authoring and annotation environment based on Web standards. dokieli happens to be one application that demonstrates the potentials of decentralised notifications and annotations. The information it creates does not result in a vendor lock-in, as other applications are able to reuse them – given agreement on using open Web standards.

On the other hand, dokieli's use of standards shows that dokieli itself is only one means to an end: once the document has been created, it lives on independently. The *social machine* [525] consists of people and documents, connected by Web standards, with dokieli acting as just one possible catalyst. Different Web applications can incorporate any of dokieli's functions and implement the principles to varying extent. Since the data is decoupled from the application, we avoid the *walled garden* problem of current social platforms today, as well as various forms of vendor lock-in with some applications, even if they are are operating in a loosely coupled environment. In essence, it is possible to swap dokieli with another client application and still communicate with any standards-compliant server. This is one of the key steps to liberating actors to create and share without setting burdens on URI, data and application control.

A couple of important sociotechnical challenges remain, and implications on autonomy. Resources might want to indicate in a granular way which actions they support or encourage, such as assessments, bookmarking, or sharing, and perhaps conditions about which notifications should be sent when any of these events take place. In order to encourage positive behaviour, we might want ways to provide moderation, and solutions to prevent harassment and abuse. Closely related is the issue of identity, pseudonymity and anonymity, and its relation with trust and verification. There are also social implications of dokieli's Web browser extension and social graph. For instance, the ability of annotating and sharing information even if the primary article does not have an inbox or an annotation service. There is no administrator that oversees everything, but that each actor has control privileges to what they are given. Having the ability to annotate anything and stored at a preferred location enables actor autonomy. This can be contrasted with annotation systems that may restrict annotations from getting created with particular targets or annotations having particular information in its body.

7. Linked Research

> You never change things by fighting the existing reality. To change something, build a new model that makes the existing model obsolete.
>
> Attributed to *Buckminster Fuller* [526].

The *Motivation* of this thesis along with the *Research Goals* and *Research Questions* can be used as a lens to configuring *a* scholarly communication ecosystem. I generalise the configuration to formulate the *Linked Research* [527] (LR) notion and initiative.

Next, I will formulate a set of *Design Principles* by generalising the core qualities of the concepts and implementations exhibited in this thesis.

7.1 Design Principles

Based on the observations and patterns involving the *Knowledge Goals* and *Technical Research Problems*, I state the *Linked Research design principles*:

- Autonomy
- Universal access

With respect to the *Effects and Artifacts*, the design principles are intended to “enhance” the physical, social, psychological, and intellectual functions of an *actor* in context of the Web and scholarly communication. Furthermore, they primarily “obsolesces” centralisation and linearity of information; “retrieves” communal involvement; and when pressed to an extreme, they can “reverse into” information overload, collectivism, and third-party control.

7.1.1 Autonomy

Today it would be difficult to think of a Web without social functions. Perhaps the Web’s support and ability to facilitate social interactions – above anything else – contributed to the transformation of societies in the last quarter century. The Web, as a medium, expedited all patterns of social interrelationships in a manner that can be more efficiently documented, discovered and reused. It necessitates that system designs should be inclusive and facilitate *involvement* within the range of an actor’s autonomy. Furthermore, actors should be able to use socially-aware decentralised applications that can perform the core functions in scholarly communications – registration, awareness, certification, and archiving.

The systems that are designed to fulfill the functions in scholarly communication should operate with the expectation of autonomous participation. One value that can be upheld through the principle of autonomy is to not have any constraints on where units of communication can exist on the Web. For instance, the actors of a system can have the freedom to designate identifiers to their artifacts: profile(s), articles, annotations, notifications, data, and so forth, as well as where they can be retrieved from. Similarly, the actor’s autonomy to use preferred applications to interact with data. The *decentralisation* of the information space is *a* means to gain autonomy on the Web. With respect to *delocalisation*, units of information, eg. personal Web identifiers (WebID) controlled by their owners, can be used across different systems as they need not be tied to any particular URI space or system, and are free from external influences. Ultimately, autonomy is the right to freely choose and maintain sovereignty over URI ownership and data spaces, and the ability to use or switch between applications that can interact with them. Likewise, actors’ autonomy would not obstruct them to use shared URI spaces if they so desire.

Any type of object, eg. scientific literature, research objects, personal websites, profiles and social graphs, can be made available on the Web provided that the *actors* have sufficient control over their URIs, or delegate to entities they trust. Through URI ownership, deconstructing and decoupling of resources can take place.

The contents of HTTP URLs may be logically centralised or decentralised. For example, the media representations found at the URL can embed compound or dynamic components that are dependent on external (remote) URLs. In this particular case, the persistence of the external URLs, as well as whether their contents are immutable is a design concern.

> **Article 19.**
>
> Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.
>
> *Universal Declaration of Human Rights* [528], United Nations

Autonomy is an important ingredient to fully exercise *freedom of expression*, which is a civil liberty. If a URI is governed by a third-party entity, ie. other than the researcher who shares the contents of a scientific resource, then it is possible that the URLs may no longer resolve, and the content may be subject to change in ways beyond researcher's control. Whereas without intermediaries, the researcher has sufficient access and control as to what happens at their URI space. This gives actors the ability to exercise *self-registration* with respect to the registration function. Trust in this case may be placed on the Web space that is closer to the owner and publisher of the URI.

The right to have *freedom of expression* necessitates that the articulation of scholarly contributions are not subject to arbitrary interference from third-parties, and allow actors to seek the most appropriate ways to co-create and co-participate. We inherit:

Freedom of expression in context of academic communication does not entail that any arbitrary artifact eg. self-published articles, passes quality-control and certification from the community, or ethically appropriate. By free expression, I mean that actors should not be subject to *technical constraints* on how they communicate their discoveries and knowledge with the rest of the world. That is, to have the means to *shape* machine-readable structure and semantics, presentations, and interactions. Enrolling in the certification process of some literature can for instance mean that reviewers may engage in any aspect and granularity of the contribution itself, eg. a weakness for an argument in one literature can be contrasted (and linked) with a scientific result presented in another literature. Academic freedom to express entails that researchers caring out their work comply with discipline-specific methods and meet standards related to accuracy, precision, completeness, reliability, relevance, and so forth. Furthermore, as with human rights, the ability to express does not grant unethical practices at the expense of any person or group. Technical expressions that are harmful or discriminatory, eg. inaccessible content, constrained presentations, or undesirable user interactions, should be avoided.

The concerns about authenticity, persistence and permanence of data and knowledge is constant regardless of where the resources reside on the Web. Hence, strictly trusting scholarly knowledge based on its registered location is ultimately fragile in that it is unavoidable to prevent all sorts of threats to preservation in a long enough timeline. However, since the registration function is decoupled from archiving, we are able to delegate that trust to a third-party that is best fit to handle the preservation and to some extent guarantee the authenticity of the units of communication. This is particularly important as archival and trust are part of the global scholarly commons.

In context of the Web, technical or social decentralisation at all levels is unattainable. Even open Web standards are based on social and technical decisions, and can be seen as a form of centralisation in that conforming systems cluster and coordinate among themselves. Similarly, centralisation naturally occurs in social systems like concepts or language, eg. machine-readable vocabularies restrict the set of uses of a system. The alternative pushed to the extreme where each unit is expressed in their own language and without any language mapping, puts us in a situation similar to that of the *Tower of Babel* [529], ie. being incapable of understanding each other. Hence, autonomy in a system where researchers are isolated or information is not interpretable renders freedom of expression meaningless.

Self-publishing: In context of *Degree of Control*, *self publishing* entails that an actor can register identifiers for their content, shape and store data where they have access to, set access control policies, and use preferred applications to achieve them.

7.1.2 Universal Access

The desire to have scientific knowledge accessible and serve humanity for good was expressed through the concept of *memex* – to supplement collective memory – *As We May Think* [530], Vannevar Bush, 1945.

A globally democratic scientific and scholarly inquiry and progress fundamentally depends upon a well-informed citizenry. For societies to be well informed, the people must have ready access to information about the knowledge output of the scholarly community, the actions of the public officials based on available knowledge, and the development of public policies. The philosophical underpinnings of the *Freedom of Information* [531] (FOI) social movement trace back to the Age of Enlightenment. Over the centuries, governments have incorporated constitutional, legal, and historical expressions of the people's right to obtain public information. A global freedom of information movement can be traced to the *World Press Conference* in 1893. The movement revolved around the need for unrestricted "access to news at its source and free transmission" of the news – *Freedom of Information Laws*, p. 53-54, urn:isbn:0123876729. One parallel movement in scholarly communication today is under the umbrella of *Open Science*.

Universal access to information can be attributed to two distinct but related areas:

- freedom to inquire; the right and means to access information
- accessibility; the quality of accessible information

Human Rights: On the actors' rights and fostering an ecosystem for reuse, we borrow from the *Universal Declaration of Human Rights*:

> **Article 27.**
>
> (1) Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.
>
> (2) Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.
>
> *Universal Declaration of Human Rights* [528], United Nations

Free Culture: In order to disseminate scientific information on the Web with minimal friction, usage rights and responsibilities should permit anyone to share, reuse, and remix information under various conditions. We align this with existing movements like Open Science, and the *free-culture movement* [532] so that both humans and machines can participate by modifying and (re)distributing units of communication without delay in the form of "free content and open content" that is network-accessible through open communication protocols. Policies about the permissions and obligations should be human- and machine-readable to facilitate accurate, fair, secure, and privacy preserving use.

Accessibility: With the motivation to create and encourage inclusiveness and diversity towards accessible and usable communication for a Web that works for everyone, we refer to W3C's accessibility fundamentals:

> **Accessibility**: addresses discriminatory aspects related to equivalent user experience for people with disabilities.
>
> **Usability**: is about designing products to be effective, efficient, and satisfying.
>
> **Inclusion**: is about diversity, and ensuring involvement of everyone to the greatest extent possible.
>
> *Accessibility, Usability, and Inclusion* [533], W3C, 2016

The AWWW states that agents, *interpret* available “representation according to the data format specification that corresponds to the representation’s Internet Media Type.” Here we can use the term *interpretation* in the general sense for both human and machine consumers of representations. An interpretation entails that a given content can be identified or discovered; comprehended in context of where and how it is used; as well being to put it to further use. Scientific and scholarly knowledge representations using information models that are based on Web standards brought through open consensus benefit from wide reuse.

Interoperability: In order to achieve cooperative systems, interoperable mechanisms are essential. One widely successful approach to have interoperable systems is through the adoption of open standards for protocols and languages that are arrived through open consensus. As per Linked Data design principles, this is accomplished with HTTP for communication, URI for identification, and RDF as language. While a decentralised system can enable each entity to be autonomous, some level of agreement is needed on the network for communication to take place. Systems that are interoperable by design have the benefit of interacting with other systems without having prior or out-of-band knowledge about them. This kind of interoperability can be contrasted with *post-facto* interoperability that is typically due to propriety designs at the discretion of each party, and then bridging the systems with application adapters and semantic mappings. In such cases, given two systems, one system adopts the others’ protocol, API, or format, or they build specialised mechanisms in order to communicate. Systems that do not use open standards or have specialised technological extensions that are only useful within their own environment are generally considered to be vendor lock-in solutions. Here we seek agreements on Web standards established through open consensus, as opposed to independent and individual design choices made in application implementations.

Systems vary on their level of interoperability. For instance, a service may require their own user account or authentication mechanism to be used when interacting with data on the Web; users may have to use a particular application to access, create, share or reuse information; applications may need to be reconfigured to reuse the information that is created elsewhere. In the *Acid Test* section I propose criteria to self-assess accessibility and interoperability of artifacts in scholarly communication.

There is a balance between free expression of structure and semantics (through autonomy) and the degree of a communication (interoperability) that different parties can achieve. At the same time, as concentration or centralisation of ideas is ultimately inescapable, interoperability influences or limits the choices a system can take. Hence, we embrace the notion of interoperability as a trade-off, ie. it is better to seek and adapt design patterns that are expected or intended to be useful for the ecosystem. Hence, interoperability is a social construct with the intention of balancing autonomy and cooperative systems.

In order to have optimal means of discovery, exchanging, and reusing information, systems and data should opt for open standards and widely acknowledged best practices wherever possible. Linked Data technologies include some of the essentials to achieve this, and are well-suited to fulfill the four functions of scholarly communication on the Web.

Thus we refer to “universal access” in the widest sense that any agent; human, machine, or other can effectively access information and participate in interacting with different systems.

7.2 Call for Linked Research

In this section I highlight research community adoption and a range of practices under the notion of Linked Research.

Call for Linked Research [534], Capadisli, 2014, expresses the shortcomings of the traditional methods of research communication in the Semantic Web community, and proposes approaches that can be adopted to conduct information exchange more effectively. The initiative aims to:

- Enable and encourage researchers to use personal Web identities and decentralised Web publishing.
- Create and exchange human- and machine-readable units of communication by way of “Linked Open Data”.
- Provide suitable user interfaces and interactions to communicate information effectively.
- Participation in open discourse.

Some aspects of *Linked Research* has been implemented by *individual researchers* [535], teams, as well as academic communities:

- Public Call for Contributions: *Call for (Enabling) Linked Research* [536]
- Self-publishing: *Examples in the Wild* [537]
- Public Awareness: *Linked Open Research Cloud* [538] at *Extended Semantic Web Conference* [539]
- Proceedings Publishing/Archiving: *CEUR-WS.org* [518]
- Conferences: *Extended Semantic Web Conference* [539], *International Semantic Web Conference* [540]
- Workshops: *Decentralizing the Semantic Web* [541], *Enabling Decentralised Scholarly Communication* [542], *Enabling Open Semantic Science* [543], *Linked Data on the Web* [544], *Research Objects* [545], *Researcher-Centric Scholarly Communication* [546], *SAVE-SD* [547], *Semantic Statistics* [517], *Web Observatories, Social Machines and Decentralisation Workshop* [548]
- Tutorials: *Authoring, annotations, and notifications in a decentralised Web* [549] (Innovations in Scholarly Communication, OAI), *Authoring, Annotations, and Notifications in a decentralised Web with Dokieli* [550] (Semantic Web in Libraries)

7.2.1 Acid Test

An *acid test* based on the *Design Principles* can be used to evaluate systems’ openness, accessibility, decentralisation, interoperability in scholarly communication on the Web. Such test does not mandate a specific technology, therefore design challenges can be met by different solutions. The following assumptions can be adopted when addressing technical design problems in scholarly communication.

Assumptions

- All interactions conform with open standards, with 1) no dependency on proprietary APIs, protocols, or formats, and 2) no commercial dependency or priori relationship between the groups using the workflows and tools involved.
- All mechanisms are independently implemented by at least two interoperable tool stacks.
- Information and interactions are available for free, accessible, and reusable for any agent; human, machine, or other.
- Information is discoverable and interpretable for any agent; human, machine, or other.
- All interactions are possible without prior out-of-band knowledge of the user’s environment or configuration.

7.2.2 Linked Open Research Cloud

The *Linked Open Research Cloud* [551] (LORC) project aims to increase the awareness, discovery, and reuse of resources about scholarly communication on the Web in the form of publicly accessible Linked Data. It is intended to fulfill the *awareness* function of scientific communication.

It pursues this by accepting notifications *about* scholarly activities on the Web, making them available for reuse, and generates an interactive visualisation for additional exploration. Consequently, LORC can assist researchers with their applications to discover pertinent scholarly information on the Web, and connect their knowledge with the global knowledge graph. For the wider scientific community, LORC can facilitate scientometrics studies and decision making based on the scholarly commons ecosystem.

To this end, LORC, as a unit of communication, advertises its *inbox* to receive *Linked Data Notifications* [265] about scholarly activities, eg. publication of scholarly articles, *Web Annotations* [552] (eg. assessments, replies), citations, call for contributions, proceedings, scientific observations and workflows, arguments, funding, decisions, and so forth. The notifications are licensed with *Creative Commons CC0 1.0 Universal* [553].

The notifications serve as (pre-)registrations of research and scholarly activities, hence play a role in fulfilling the certification function. The inbox and the notifications are archivable. Furthermore, LDN can be used to notify institutional repositories, Web crawling agents, or archives about the existence of or updates to Web resources.

More generally, the LORC model to accepting notifications about Web resources can be used to update *Linked Open Data Cloud* [554]'s dataset metadata as well as in automating the generation of an accurate diagram; as initially outlined in *Re: Please update your resource in the LOD Cloud Diagram* [555], Capadisli, 2017. Based on the idea of the LORC, *Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud* [556], Debattista, 2019, proposes an architecture for the LOD Cloud service and data pipeline which strives to improve the survival of the LOD Cloud.

7.2.3 Linked Specifications, Test Suites, and Implementation Reports

To demonstrate the extensibility of units of communication within the frame of Linked Research, I extend the LDN *Test Suite and Implementation Reports* and *Implementations* and describe the development of an interlinked, human- and machine-readable W3C Recommendation, its test suite, and implementation reports as a whole.

The missing connection among existing specifications and test reports is that a uniform resource discovery is not possible between the test reports and the individual conformance criteria in the specifications, where a given information at a particular URL is both human- and machine-processable (see *Linked Specifications, Test Suites, and Implementation Reports Related Work* [557].)

Having the specifications and implementation reports interlinked and retrievable can facilitate their automated discovery and reuse. One use case is to be able to find applications that match a certain conformance criteria, eg. in order to have fine-grained bundling of software packages. While this would typically include "normative requirements", tests can potentially capture and reveal optional features of specifications. Prospective consumers of the compliance reports can be application developers finding appropriate software for use, as well as automatic software package managers.

The following figure *Linked Specifications Reports* depicts an overview of linking the LDN specification, its test suite, the generated implementation report for the dokieli project, reports summary, and an article citing the specification.

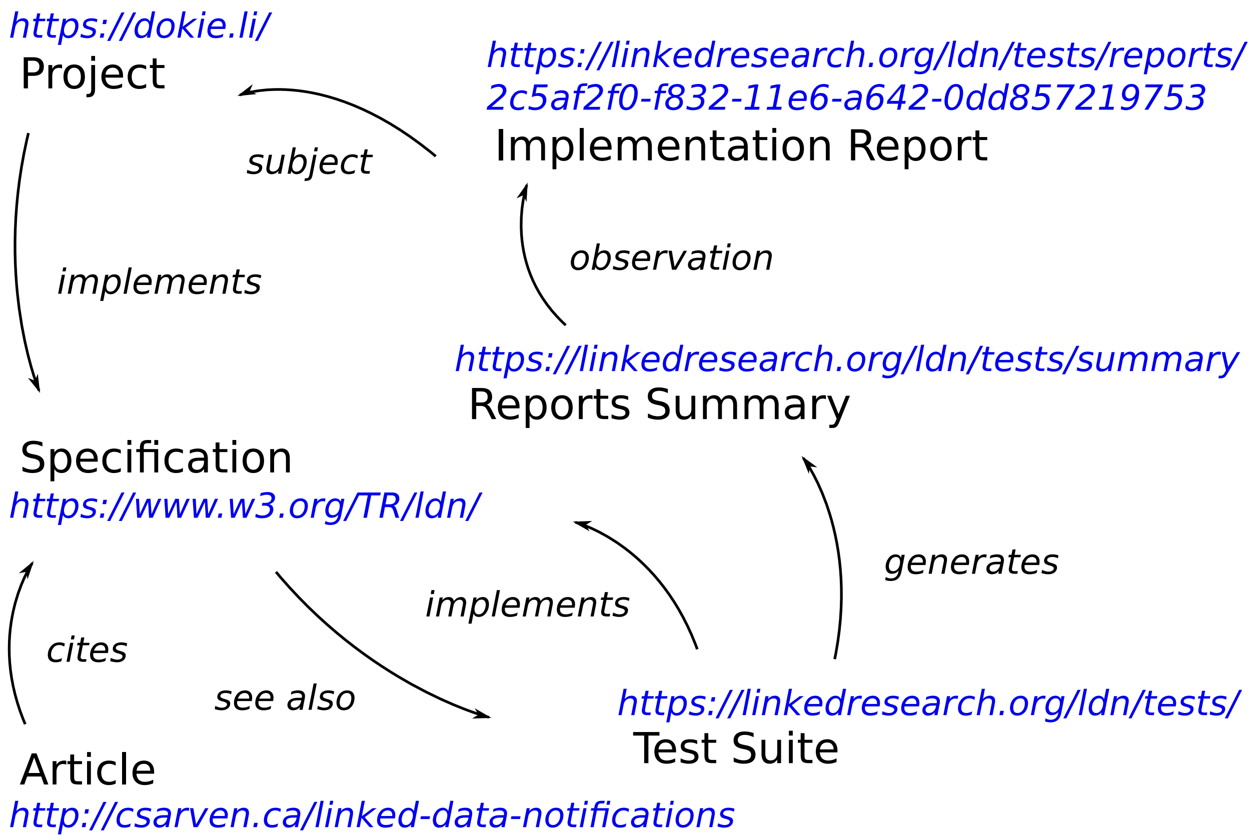

https://dokie.li/
Project
https://linkedresearch.org/ldn/tests/reports/2c5af2f0-f832-11e6-a642-0dd857219753
Implementation Report
subject
implements
observation
https://linkedresearch.org/ldn/tests/summary
Reports Summary
Specification
https://www.w3.org/TR/ldn/
generates
implements
cites
see also
https://linkedresearch.org/ldn/tests/
Test Suite
Article
http://csarven.ca/linked-data-notifications

Figure 16. An overview of linking a specification, test suite, generated implementation report for the project, reports summary, and an article citing the specification

The information patterns in the LDN specification and the implementation reports sections are reusable across other specifications and related components. Sources are available from:

- The specification: https://www.w3.org/TR/ldn/
- The test suite: https://linkedresearch.org/ldn/tests/
- Implementation reports summary and individual reports: https://linkedresearch.org/ldn/tests/summary

The test suite uses the *mayktso* [449] server as the LDN receiver, but any conformant receiver implementation will work here.

The notifications carrying the reports have an HTML+RDFa representation (alternatively in other RDF serialisations upon content negotiation). The scholarly article on *Linked Data Notifications* [558] uses the *CiTO* vocabulary to cite the specification with `cito:citesAsAuthority`. Another peer reviewed article, *Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli* [559], contextually cites the specification with `cito:citesAsPotentialSolution` from its architectural overview section, as well as the LDN Test Suite with `cito:citesAsAuthority` from its adoption section. This is useful in that we can have articles linked to what is already available with minimal effort. Including this article that you are currently reading and interacting with.

The realisation here is that we have everything operating in a way that is interoperable: the specification, test suite, discovery of the reports, and academic articles, all reusing existing vocabularies.

7.3 Forces and Functions of Linked Research

The forces and functions of scholarly communication in context of the Linked Research *Design Principles* are described as follows. The following is also a *conceptual* response to Van de Sompel's call to investigate "how to fulfill the core functions of scholarly communication using the technical paradigm" around self-publishing (*Scholarly Communication: Deconstruct & Decentralize?* [560]).

The *forces*:

Actor
Any “Web agent” — person or software — may independently participate using identities they control, manage the accessibility of Web resources, shape content, as well as perform all of the functions in scholarly communication.

Accessibility
The retrievability and reusability of Web resources can be managed by actors with access controls.

Content
Content can be shaped as Linked Data and offer suitable presentations and interactions on the Open Web Platform.

Applicability
The need and relevance for research and their application can be mined from available content.

The *functions*:

Registration
The ecosystem permits any actor to register units of communication by associating a URI to new or nonexistent resources.

Awareness
Any actor can distribute, discover, and reuse registered units of communication.

Certification
Any actor can perform quality-control and participate in a certification process with other actors.

Archiving
Any actor may trigger an archiving service to create snapshots of the content of registered Web resources.

7.4 Contextualising Linked Research

There is an abundance of scientific information on the Web. However, humans and machines have restricted access, are required to pay, or hindered to efficiently discover relations. The knowledge industry can better address societal challenges by embracing the virtues of the Web via decentralisation and interconnectivity of our collective knowledge.

Mirroring the Web: LR is an initiative to strive to experience the full effects of the Web by observing and making use of its affordances. LR serves as a conceptual framework that describes how actors can control their interactions and data, make use of human- and machine-readable information through Web-interfaces, as well as how decentralised and socially-aware applications can exchange information on route to fulfilling the core functions of scholarly communication.

Traps: Are there ideological and technology traps in LR? The traps can be seen as analogous to McLuhan’s *Reverses Into* effect. When the LR design principles are "pushed to the extreme", highly specialised and fine-grained centralised systems may emerge while still operating within the larger decentralised system on the Web. The degree of interoperability among the autonomous systems will vary. The artifacts may be controlled or managed by a new generation of entities in information spaces, eg. agents delegated to perform certain operations on behalf of an actor. Just as centralised publishing and distribution services currently exist for the dominating print-centric industry, the medium can still accommodate similar control over semantic and granular scientific resources. Individual autonomy without cooperation among the actors can make it difficult to establish and maintain trust.

Reconfiguring Forces and Functions: The existing or prior models for information exchange that has been described so far primarily focus on institutional or subject repository driven scholarly communication. They inherently require the involvement of third-party service providers – for- or non-profit – which impose a particular arrangement of the coupling of the scholarly information space per *forces and functions*. For instance, traditional journals most commonly couple the registration and certification functions, whereas the configuration based on LR allows different actors to activate different functions on demand.

Basic Costs: From the perspective of an actor, the costs for a personal domain name along with data storage can be in the range of 100 USD per year with sufficient space for common Web publishing – "unlimited" amount of registration of units of communication and publishing any hypermedia. As for archiving services, it would be typically handled by "external" stakeholders with respect to the actor sharing their registered content.

Social Barriers: At this time, promotion and tenure are integral to the academic profession. The incentives in the academic profession predominantly relies on an ecosystem that is largely led by third-party publishers and services. LR can be seen as market driven from the perspective of self-driven actors in the system – similar to proponents of the preprint movement.

7.5 Linked Research as a Paradigm

With the arrival and the adoption of the Web by the masses, a “paradigm shift” – from print to electric – had occurred. Sociotechnical advancements in the past three decades in context of the Web serves as an evidence of the shift towards using the Web as a primary environment (medium) for "normal" scholarly communication. Put into historical perspective, we are indeed operating in the early phase of this new paradigm. Thus, I content that all advancements – no matter how radical they appear or claim to be – still operate within the framework of the Web, regardless of content shaped as if for print, being accessible to the privileged, controlled by third-party entities, and so forth.

I will use Kuhn’s five characteristics – “accuracy, consistency, scope, simplicity, and fruitfulness” from *Objectivity, Value Judgment, and Theory Choice* – to measure the quality of *Linked Research* as a shared *sub*-paradigm for scholarly communication on the Web. The criteria is not intended to be exhaustive but, rather, used to cover varied aspects to situate LR. I provide these descriptions for the purpose of assisting *stakeholders* gain a perspective on LR’s values, facilitate determining its adequacy for use, as well as to evaluate LR against other models.

LR is based on the notion that the Web is inherently decentralised, social, and for everyone. Its design principles shapes data and systems to be built on open Web standards to carry out scholarly activities.

Accuracy: LR is based on the core mechanisms for identification, interaction, and representation of units of information on the Web. For instance, *semantic citations* enables actors to build precise and contextual relations between atomic units, preserve author’s intentions at the time of citing, create a snapshot of the target while citing, as well as announce the citation activity. Thus, the creation, dissemination, and discovery of typed citations (as opposed to generic citations), brings forth an accurate representation of a network of shared knowledge. In this respect, LR is empirically adequate within the frame of what Linked Data vocabularies can potentially express irrespective of where the cited entities are located on the Web.

Consistency: LR is consistent within itself as per the alignment with the *Architecture of the Web* and using open Web standards (in contrast to print-centric approaches on the Web). LR adopts the notion of actors having free and unrestricted access to knowledge, and thus consistent with several community principles and initiatives with the same goal, eg. the Open Access movement, self-archiving of literature so that it is readable by anyone for free, Semantic Web publishing. In this respect, LR is a continuation or aligned with the theme to democratise knowledge through the use of the Web. On the other hand, with respect to the common social practices around "publishing", LR is explicit about actors (content creators) ultimately deciding how to make their contributions. Hence, LR is inconsistent with the existing practice or market-driven notion of requiring actors to use particular third-party services in order to participate in scientific and knowledge exchange, as well as the way in which contributions are acknowledged by relevant stakeholders. From this perspective, LR’s notion of an "actor" can be interpreted to be *incommensurable* with the "actor" in the contemporary scholarly system. Furthermore, LR is currently a technology driven change as opposed to market driven change in that the added value comes from a way to reconfigure the interplay of forces and functions. However, it is not a paradigmatic change in terms of strategic repositioning and development of the forces and functions eg. who can ultimately be an actor and their potential competences.

Scope: LR has broad scope in that knowledge can be incrementally connected and extended (at any time) by using shared identifiers for units of communication and compatible data models. Through the use of languages like RDF, information discovery mechanism remains uniform.

Simplicity: LR is a "lossless" approach to representing and sharing information on the Web in that the content can be registered independently from its presentation and behaviour, thus making it possible to refer and connect units of communication at any level of abstraction. As machine-readable representation is encoded through the serialisation of the RDF language, the underlying information can be transformed into alternative serialisations without compromising its integrity. The Open Web Platform makes it possible to treat and handle content "as is" without requiring supplemental protocols, proprietary plug-ins, or external tools to create and share content.

Fruitfulness: Machine-readable information facilitates discovery and reuse of units of communication. As in the case of semantic citations, LR opens up the possibility to create as well as detect precise relations which were previously unnoted or disclosed through appropriate interoperable standards. Units of communication in the form of a knowledge graph being machine-identifiable and queryable has the affordance to detect or inspect patterns. For instance, open research questions can be found by analysing semantically structured literature, as well as detecting topical trends in research communities. LR’s design principles enables the use of existing and future open Web standards. Its continuous interoperability would permit LR to be evolvable.

Whether to use the LR model will depend on the "value judgment" of stakeholders and the research community. LR is merely a reflection of using the Web to its full potential for scholarly communication. LR in and itself is not a new paradigm, but rather a conceptual specification that can be used as an *exemplar* reflecting the technical advancements.

8. Conclusions

The *Research Goals* of this thesis have been to investigate how decentralised, interoperable, and socially-aware Web applications can be designed to fulfill the core functions in scholarly communication while enabling actors with high degree of control in shaping and sharing their content. In this concluding section, I will present the *Research Questions Review*, provide *Interpretations* of the findings, and offer *Perspectives* for future research.

8.1 Research Questions Review

The sections *Scholarly Communication on the Web*, *Structure of Scholarly Information*, and *Decentralising Scholarly Communication* served to explain the state of sociotechnical affairs in scholarly communication, as well as to identify knowledge gaps by answering the *Knowledge Questions* in order to address the *Technical Research Problems*. In this section I will review how the technical research problems have been answered.

What technical mechanisms, standards or protocols can be employed for decentralised information exchange on the Web?

Linked Data Notifications is a communication protocol that describes how applications can perform the roles of *senders*, *receivers*, and *consumers* when exchanging notifications. Such decoupling of application roles in context of a client–server model permits interoperability between two classes of interactions, senders and receivers, or between consumers and receivers. These are atomic interactions in that they facilitate the design of loosely coupled systems that can exchange structured notifications as Linked Data. As such, the LDN protocol affords a high *degree of control* to actors in that they can use their preferred applications, store notifications where they prefer, switch between applications without relocating the notifications, and relocate the notifications without having to change applications. LDN's underlying mechanism triggers *registration*, fulfills *awareness*, and can support fulfilling the *certification* and *archiving functions in scientific communication* in a decentralised ecosystem.

How can Web technologies be employed to fulfill the core functions of scholarly communication in an open and interoperable way?

A decentralised client application based on the *Open Web Platform* can be devised to allow independent actors to exchange units of communication with decentralised storage services. *dokieli* demonstrates how an assembly of Web standards can come together in a cohesive way in order to enable actors to reach a high *degree of control* over their creations and social activities. HTTP is used to perform CRUD operations, (HTTP) URIs to identify Web resources, and Linked Data design principles to uniformly represent, link, and discover information. Actors can identify themselves through WebID (HTTP URI) whether tied to a real or fictitious entity, extend their profile descriptions, and authenticate against servers with mechanisms like WebID-TLS and WebID-OIDC. Machine-interpretable profile descriptions also enables the client application to adapt to accommodate actor's capabilities. It is feasible to create and consume human- *and* machine-readable units of (scholarly) communication using standards such as RDFa in HTML, and be combined with other RDF serialisations based on technical specifications. A variety of Linked Data vocabularies can be used for semantic content authoring. The Linked Data Notifications protocol is used to exchange structured notifications. The Web Annotation Data Model and Protocol is used to associate resources and exchange annotations. The ActivityPub protocol is used to exchange social activities. The resources that dokieli generates are findable, accessible, interoperable, and reusable by other standards-compliant Linked Data applications. The resulting technical design operates under the notion of a scholarly ecosystem where applications are decoupled to fulfill the registration, awareness, certification, and archiving *functions in scholarly communication*.

In summary, Linked Data Notifications is a building block for diverse decentralised and loosely coupled Web applications to exchange notifications, and fulfills the *awareness* function in scholarly communication. dokieli demonstrates how RACA-aware Linked Data applications can be designed while fulfilling essential *decentralised Web publishing* requirements.

8.2 Interpretations

In this section I reflect on what the findings of this thesis can mean to us and why we should care.

PAYGO Linked Data: Publishing units of communication by following the Linked Data principles is similar to the *pay as you go* approach in dataspace systems in that knowledge on the Web can be extended incrementally without having everything up front or even a commitment to do so. This is particularly evident when disparate collection of units (or graphs) published by different entities can be connected when they share globally identifiable entities, as well as being universally useful when the content they identify is accessible.

FAIR Linked Data: For openly accessible resources, I note that the *Linked Data* approach and technology stack can fulfill the *FAIR guiding principles*. One generalisation that could be made from this is that Linked Data design principles can be adopted as a base requirement when building interoperable Web applications, which would also fulfill FAIR. After all, the FAIR principles does not prescribe the mechanisms for Web data, and as such, interoperability across applications for independently produced FAIR-conformance is not systematically guaranteed. In contrast, the Linked Data approach specifies the mechanisms within the framework of the *Architecture of the Web* and guarantees interop.

Decentralised Identity and Authentication: Applications primarily operate in a read-only information space in the absence of decentralised identities and authentication mechanisms. Client applications can be loosely coupled with servers when the source of authentication is distributed. Actors can use independently built client applications to perform CRUD operations provided that they can authenticate using their preferred WebID, as well as have the possibility to decide on the storage location of their data. Thus, it is possible to distribute scholarly contributions on the Web based on an actor (or contributor) centric model.

Interoperable Applications: Decoupling of the forces and functions in scholarly communication in context of the Web necessitates interoperable applications through the use of standard data formats and communication protocols. These class of applications need to only reflect user’s intent within the frame of a loosely coupled system design. That is, applications having semantic understanding of the affordances of the system makes it possible to automate information exchange. Consequently enabling users to switch between standards-compliant applications without interfering with existing data. For instance, HTTP, URI, and HTML are widely used standards on the Web enabling us to use any user agent (like a Web browser) on the Open Web Platform. Similarly, one layer up on the interoperability ladder would be to employ communication protocols and data formats that are capable of addressing specific use cases.

Sociotechnical Centralisation: We can observe different forms of centralisation within decentralisation. As we adopt mechanisms with the intention of decentralisation and interoperability, are we free of all forms of centralisation and vendor lock-ins? For instance, *URI Ownership* discussed the inherent social centralisation of DNS in that individuals merely rent domain names from registrars and maintain the mapping to physical servers. *Semantic Organisation of Notifications* highlights one form of organic centralisation around the choice of Linked Data vocabularies that the applications make use of. Technical Web specifications for the most part reflect the group’s goals, cultural norms or biases. For instance, LDN was initially incubated in the *Solid project* [561] within the understanding of a socially-aware read-write Linked Data paradigm, and then developed to be a technical specification through the W3C SWWG as part of a broader context of decentralisation efforts.

AAA in RACA: *Function of Peer Review* noted the intentions and limits of peer review in context of the Web. It is neither the case that the right, the privilege, or the practice of the notion “anyone being (technically) allowed to say anything about anything” is exempt from ethical considerations. As we have seen in modern Web social media, management and distribution of misinformation and disinformation is profitable and poses a danger to civilisation. For instance, as electronic publication has also been blurring the lines between academic work and personal opinions, making the distinction between the two as part of the scientific endeavour has only become more vital. In that respect, the notion of *self-publishing* does not entail quality-controlled and certified units of communication in and itself, but rather focuses on increasing actors' immediate role and control. Thus, trust and accountability remain as social processes.

School of Thought: Some of dokieli's design decisions were driven by the need to enable actors to autonomously participate in a decentralised scholarly ecosystem. With respect to the four functions in scientific communication, the expectations are to operate under a *deconstructed* or *decoupled* model while providing high degree of control to content creators and consumers, as well as generating and reusing interoperable data. This approach shares some conceptual similarities to the "self-archiving" approach in that scholars have the right to freely share their Web resources from open repositories. However, a distinction between the "self-archiving" initiative and *self-publishing* in context of the *Linked Research* goal can be made in that while an actor may prefer to share their contributions through third-party controlled components, the baseline for "self-publishing" is that they are not required to. Thus, the network location and the applications that are used to make the content available are orthogonal to the notion of what constitutes a scholarly contribution. Lastly, parallel characteristics between dokieli's approach and the *Republic of Letters* can be made in that both operate under the idea of scholarly activities to be a researcher-centric model – direct information exchange between authors and readers – without requiring intermediaries.

8.3 Perspectives

> Radical changes of identity, happening suddenly and in very brief intervals of time, have proved more deadly and destructive of human values than wars fought with hardware weapons.
>
> *Laws of Media* [211], p. 97, Marshall McLuhan, 1989

Sociotechnical Advancement: Given our half a millennia of experience in mass print publishing, and half a century on electronic publishing, we are in a position to situate, examine and question our assumptions as to what constitutes scholarly knowledge exchange and participation in context of the Web. If research communication is predominately and increasingly driven by the Web, how can we approach the integration of scholarly processes with the inherent properties of the Web? One way to look at this is by *examining history*: the printing press played a major role in moving the non-literate public to learn to read and write for themselves. Given the Web, we find ourselves in a similar situation with different modes of communication. For instance, the Web – an environment for simultaneous read and write operations on hypermedia – inherently enables anyone to be a publisher, where each actor can exercise their full creativity and ability to express the shape of their content. In order to use the Web effectively – as far as taking advantage of its built-in characteristics, the research community needs to better position itself to self-publish and disseminate their knowledge, as well as be more attributable and accountable in their social engagement. This kind of ongoing evolution is necessary for the advancement of research communication and consequently the society. Moving towards Web-centric information flow requires the scholarly ecosystem to unchain itself from antiquated practices and inadequate use of the available medium. In context of major communication shifts, we have not scratched the surface – 30 years into the advent of the Web is still a brief period in history.

Addressing Social Implications: *Social Implications of LDN* looked at a particular area of design considerations for notifications. *dokieli's Activities* describes how a "friend of a friend" social network can work. An AAA-based ecosystem while can potentially enable everyone to voice themselves, the Web is not immune to the distribution of propaganda, cyberbullying or hate speech. How can we cultivate positive behaviour, and design systems to discover and filter content based on ethical norms or personal preferences? Thus, in context of *decentralised Web publishing*, future research can investigate:

- How can harassment and abuse be prevented or handled?
- How would moderation or civil discourse work?
- How can a resource prohibit or consent to be annotated at fine granularity?
- How can the conditions to annotate and notify be indicated and monitored?
- How can a resource's annotation policy and rules be respected?
- What may be the conditions to associate different identities?
- How can domain expertise be factored in?

The W3C TAG Finding, *Ethical Web Principles* [562], 2019, sets out ethical principles which will guide the Technical Architecture Group. These principles can be applied when designing systems in order to provide a net positive social benefit.

Extended Profiles: WebID Profile descriptions can be extended to enable richer decentralised clientside applications in that both the user interface and the content can be a reflection of actor's preferences and abilities. What kinds of "preference negotiation" can be made?

Privacy Aware Information Spaces: Actors can be better supported in their decision-making if their applications are privacy-aware and respecting, for example through the use of standards like ODRL. This leads to the possibility of observing the interplay of forces and functions in scholarly communication, and in particular, how privacy-aware applications can generate and consume new forms of content and interactions. How can a decentralised ecosystem balance privacy, integrity and availability?

Lossless Citations: Applications supporting semantic content authoring enables fine-grained connections to be created between units of communication. Approaches like in *contextual citations* (with semantically typed relations) make it possible to preserve authors' intended connection between units of communication. Further study on creating, annotating, as well as extracting citation Linked Data can provide further insights into (scholarly) content. How can open research questions be systematically discovered by machines?

RACA-Aware Applications: *Forces and Functions in Specifications* characterised open Web protocols, standards, and models in context of scholarly communication. While this contribution is not intended to be a complete collection of approaches, future work can extend the current *RDF Data Cube dataset* and the EARL results including additional subjects with new observations. The availability of such structured dataset can help answering the knowledge questions of future research when designing "*RACA-Aware*" decentralised and interoperable applications. Thus, potentially answering a research question along the lines of *How can we systematically remix open Web specifications in order to design RACA-Aware applications?* Another area of research would be for applications to systematically and adaptively determine, as well as apply the right abstraction level for content based on actor's needs.

On-Demand Archiving: While there are publicly usable on-demand archiving services eg. Internet Archive, there is a need for client applications to be able to both choose from existing archiving services as well as allow users to specify their preferred services when making *contextual citations* or archiving any unit of communication. To that end, future research can investigate how personal or community-controlled archiving services can be designed in order to enable decentralised client applications.

Linked Data-Aware User Agents: At the moment, common user agents like the Web browser are not natively equipped to parse and re-serialise data expressed with the RDF language, or have the ability to construct SPARQL queries and process query results, yet alone adapt the user interface based on content. Consequently, client applications and Web Extensions are forced to include (JavaScript) code libraries to handle these features until there is built-in support. Such method of inclusions increases the complexity of using as well as maintaining the libraries. Future research can investigate effective ways to natively integrate RDF and SPARQL libraries and support corresponding APIs in the Web browser so that client applications can focus on higher-level user intentions as opposed to lower-level data management.

Third-Party Control Considered Harmful: The current state of scholarly communication on the Web is ravelled with sociotechnical limits and complications. The overarching technical research problem aimed to investigate and design Web-centric artifacts that would enable content creators and consumers to have more control over their activities, as well as to effectively exchange information while respecting actors' privacy. The Web continues to enable anyone to be a publisher. The findings of thesis reveals that third-party publishers are not required and that *decentralised Web publishing* is, and ultimately continues to be, feasible *for everyone*.

⁂

To boldly go where we have not gone before: I contend that one way to advance research communication on the Web in its totality is that, social policies, funding, research exchange, as well as professional incentives, all need to accurately reflect the affordances of the media. The Linked Research initiative ironically indicates the need for a *counter-environment* as a means of practising the dominant and unnoticed environment, that is the Web as a socially-aware communication medium. The social conventions that surrounds the notion of a "scholarly contribution" can be anticipated to be revisited in order to improve the accessibility of knowledge, and be more inclusive and equitable for everyone. The evolution of sociotechnical systems require people's adaptability at the same time. In order for the scholarly ecosystem to be genuinely “open” and inclusive, systems need to be ethically grounded and interoperate with the notion of “universal design” – actor diversity with variations in needs, capabilities, and aspirations.

I propose the notion of *Linked Research* as a potentially useful shared paradigm for scholarly activities on the Web. You are cordially invited to *self-publish* and share your activities related to any unit of significance in this thesis.

So there we have it. It is all sort of simple.

-Sarven

References

1. ^ *The Medium is the Massage: An Inventory of Effects*, *http://worldcat.org/isbn/9781584230700*
2. ^ a b *The Matrix Reloaded*, *https://en.wikiquote.org/wiki/The_Matrix_Reloaded*
3. ^ *Connections*, *https://en.wikipedia.org/wiki/Connections_%28TV_series%29*
4. ^ *Connections*, *https://en.wikipedia.org/wiki/Connections_%28TV_series%29*
5. ^ *information society*, *https://en.wikipedia.org/wiki/Information_society*
6. ^ *Design Science Methodology for Information Systems and Software Engineering*, *https://www.springer.com/cda/content/document/cda_downloaddocument/9783662438381-c1.pdf*
7. ^ *Dynamic and context-sensitive linking of scholarly information*, *https://biblio.ugent.be/publication/522209/file/1873597.pdf*
8. ^ *Practical Assessment, Research & Evaluation*, *https://pareonline.net/pdf/v14n13.pdf*
9. ^ *Understanding Me*, *http://www.worldcat.org/oclc/249804097*
10. ^ *The Gutenberg Galaxy*, *http://www.worldcat.org/oclc/993539009*
11. ^ *Connections*, *http://www.worldcat.org/oclc/174040346*
12. ^ *The Printing Press as an Agent of Change*, *http://www.worldcat.org/oclc/611804138*
13. ^ *Objectivity, Value Judgment, and Theory Choice*, *https://iweb.langara.bc.ca/rjohns/files/2018/03/Kuhn_theory_choice.pdf*
14. ^ *The Structure of Scientific Revolutions*, *http://www.worldcat.org/oclc/857115808*
15. ^ *Objectivity, Value Judgment, and Theory Choice*, *https://iweb.langara.bc.ca/rjohns/files/2018/03/Kuhn_theory_choice.pdf*
16. ^ *Commentariolus*, *https://en.wikipedia.org/wiki/Commentariolus*
17. ^ *De revolutionibus orbium coelestium*, *https://en.wikipedia.org/wiki/De_revolutionibus_orbium_coelestium*
18. ^ *geocentric model*, *https://en.wikipedia.org/wiki/Geocentric_model*
19. ^ *The Gutenberg Galaxy*, *https://localhost:8443/lr-thesis/the-gutenberg-galaxy*
20. ^ *The Structure of Scientific Revolutions*, *https://localhost:8443/lr-thesis/the-structure-of-scientific-revolutions*
21. ^ *This is Marshall McLuhan: The Medium is the Massage*, *https://archive.org/details/thisismarshallmcluhanthemediumisthemessage*
22. ^ *WorldWideWeb: Summary*, *https://groups.google.com/forum/message/raw?msg=alt.hypertext/eCTkkOoWTAY/bJGhZyooXzkJ*
23. ^ *social machines*, *https://en.wikipedia.org/wiki/Social_machine*
24. ^ *Weaving the Web*, *http://www.worldcat.org/oclc/729965164*
25. ^ *Evolution of the Web*, *https://www.w3.org/DesignIssues/Evolution*
26. ^ *Universal Document Identifier*, *https://www.w3.org/People/Berners-Lee/ShortHistory.html*
27. ^ *Uniform Resource Identifier*, *https://tools.ietf.org/html/rfc3986*
28. ^ a b *Uniform Resource Locators*, *https://tools.ietf.org/html/rfc1738*
29. ^ *Hypertext Transfer Protocol*, *https://tools.ietf.org/html/rfc2616*
30. ^ *Hypertext Markup Language*, *https://www.w3.org/TR/html/*
31. ^ *Technical Architecture Group*, *https://tag.w3.org/*
32. ^ *Architecture of the World Wide Web, Volume One*, *https://www.w3.org/TR/webarch/*
33. ^ *Orthogonality*, *https://www.w3.org/TR/webarch/#orthogonal-specs*
34. ^ *Uniform Resource Identifier (URI): Generic Syntax*, *https://tools.ietf.org/html/rfc3986*
35. ^ *Universal Resource Identifiers -- Axioms of Web Architecture*, *https://www.w3.org/DesignIssues/Axioms#uri*
36. ^ *URIs, URLs, and URNs: Clarifications and Recommendations 1.0*, *https://www.w3.org/TR/uri-clarification/*

37. ^ *Generic Resources*, *https://www.w3.org/DesignIssues/Generic*
38. ^ *URI ownership*, *https://www.w3.org/TR/webarch/#uri-ownership*
39. ^ *Domain Name System*, *https://en.wikipedia.org/wiki/Domain_Name_System*
40. ^ *Internet Corporation for Assigned Names and Numbers*, *https://www.icann.org/*
41. ^ *Domain Names*, *https://tools.ietf.org/html/rfc1034*
42. ^ *countries, categories, multiorganization*, *https://tools.ietf.org/html/rfc920*
43. ^ *URNs, Namespaces and Registries*, *https://www.w3.org/2001/tag/doc/URNsAndRegistries-50*
44. ^ *Internet Protocol Suite*, *https://tools.ietf.org/html/rfc1123*
45. ^ *Architectural Styles and the Design of Network-based Software Architectures*, *http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm*
46. ^ *HTTP/1.1 Message Syntax and Routing*, *https://tools.ietf.org/html/rfc7230*
47. ^ *Open Web Platform*, *https://www.w3.org/2001/tag/doc/IAB_Prague_2011_slides.html#owp*
48. ^ *The Rule of Least Power*, *http://www.w3.org/2001/tag/doc/leastPower.html*
49. ^ *Principle of Least Power*, *https://www.w3.org/DesignIssues/Principles#PLP*
50. ^ *Cascading Style Sheets*, *https://www.w3.org/TR/CSS/*
51. ^ *Mathematical Markup Language*, *https://www.w3.org/TR/MathML/*
52. ^ [a] [b] *RDFa*, *https://www.w3.org/TR/rdfa-core/*
53. ^ *Scalable Vector Graphics*, *https://www.w3.org/TR/SVG/*
54. ^ *Information Management: A Proposal*, *https://www.w3.org/History/1989/proposal.html*
55. ^ *Resource Description Framework*, *http://www.w3.org/TR/rdf11-concepts*
56. ^ *Sense and Reference on the Web*, *http://www.ibiblio.org/hhalpin/homepage/thesis/*
57. ^ *open-world assumption*, *https://en.wikipedia.org/wiki/Open-world_assumption*
58. ^ *closed-world assumption*, *https://en.wikipedia.org/wiki/Closed-world_assumption*
59. ^ *Triples*, *https://www.w3.org/TR/rdf11-concepts/#section-triples*
60. ^ *International Resource Identifier*, *https://tools.ietf.org/html/rfc3987*
61. ^ *subject–verb–object*, *https://en.wikipedia.org/wiki/Subject%E2%80%93verb%E2%80%93object*
62. ^ *linguistic typology*, *https://en.wikipedia.org/wiki/Linguistic_typology*
63. ^ *entity–attribute–value*, *https://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model*
64. ^ *The Self-Describing Web*, *http://www.w3.org/2001/tag/doc/selfDescribingDocuments*
65. ^ *follow your nose*, *https://www.w3.org/2001/sw/wiki/Linking_patterns#.E2.80.9CFollow_your_nose.E2.80.9D*
66. ^ *RDF/XML*, *https://www.w3.org/TR/rdf-syntax-grammar/*
67. ^ *XML*, *https://www.w3.org/TR/xml/*
68. ^ *Turtle*, *https://www.w3.org/TR/turtle/*
69. ^ *JSON-LD*, *http://www.w3.org/TR/json-ld/*
70. ^ *JSON*, *https://tools.ietf.org/html/rfc7159*
71. ^ *SPARQL Protocol and RDF Query Language*, *https://www.w3.org/TR/sparql11-overview/*
72. ^ *SPARQL 1.1 Query Language*, *https://www.w3.org/TR/sparql11-query/*
73. ^ *SPARQL 1.1 Federated Querying*, *https://www.w3.org/TR/sparql11-federated-query/*
74. ^ *SPARQL 1.1 Update*, *https://www.w3.org/TR/sparql11-update/*
75. ^ *SPARQL 1.1 Protocol*, *https://www.w3.org/TR/sparql11-protocol/*
76. ^ *SPARQL 1.1 Service Description*, *https://www.w3.org/TR/sparql11-service-description/*
77. ^ *SPARQL 1.1 Entailment Regimes*, *https://www.w3.org/TR/sparql11-entailment/*
78. ^ *SPARQL 1.1 Graph Store HTTP Protocol*, *https://www.w3.org/TR/sparql11-http-rdf-update/*
79. ^ *Linked Data – Design Issues*, *https://www.w3.org/DesignIssues/LinkedData*
80. ^ *httpRange-14: What is the range of the HTTP dereference function?*, *https://www.w3.org/2001/tag/issues.html#httpRange-14*

81. ^ *Still hates computers*, *http://www.theinquirer.net/inquirer/feature/1735712/professor-wendy-hall-speaks*
82. ^ *Creating a Science of the Web*, *https://eprints.soton.ac.uk/262615/1/Web%2520Science.htm*
83. ^ *A Framework for Web Science*, *https://eprints.soton.ac.uk/id/eprint/263347*
84. ^ *A Framework for Web Science*, *https://eprints.soton.ac.uk/id/eprint/263347*
85. ^ *Credibility of the Web: Why We Need Dialectical Reading*, *https://www.ideals.illinois.edu/bitstream/handle/2142/13425/credibility.pdf*
86. ^ *The Future of the Humanities*, *http://www.worldcat.org/oclc/469418474*
87. ^ *The Printing Press as an Agent of Change*, *http://www.worldcat.org/oclc/611804138*
88. ^ *Republic of Letters*, *https://en.wikipedia.org/wiki/Republic_of_Letters*
89. ^ *The Royal Society*, *https://en.wikipedia.org/wiki/Royal_Society*
90. ^ *Journal des sçavans*, *https://en.wikipedia.org/wiki/Journal_des_s%C3%A7avans*
91. ^ *Philosophical Transactions of the Royal Society*, *https://en.wikipedia.org/wiki/Philosophical_Transactions_of_the_Royal_Society*
92. ^ *Mémoires de l'Académie des Sciences*, *https://en.wikipedia.org/wiki/Comptes_rendus_de_l%27Acad%C3%A9mie_des_Sciences*
93. ^ *Journal des sçavans*, *https://en.wikipedia.org/wiki/Journal_des_s%C3%A7avans*
94. ^ *Philosophical Transactions of the Royal Society*, *https://en.wikipedia.org/wiki/Philosophical_Transactions_of_the_Royal_Society*
95. ^ *Encyclopedia of International Media and Communications*, *http://www.worldcat.org/oclc/773482913*
96. ^ *The Sociology of Science: Theoretical and Empirical Investigations*, *http://www.worldcat.org/oclc/695759425*
97. ^ *The Many Meanings of Open*, *https://www.w3.org/DesignIssues/Open*
98. ^ *The Open Definition*, *https://opendefinition.org/*
99. ^ *The Open Definition 2.1*, *https://opendefinition.org/od/2.1/en/*
100. ^ *What is free software?*, *https://www.gnu.org/philosophy/free-sw.html*
101. ^ *Free Cultural Works*, *http://freedomdefined.org/Definition*
102. ^ *Open Content*, *http://opencontent.org/definition/*
103. ^ *Creative Commons*, *https://creativecommons.org/*
104. ^ *CC licenses*, *https://creativecommons.org/licenses/*
105. ^ *Creating a global knowledge network*, *http://www.cs.cornell.edu/~ginsparg/physics/blurb/pg01unesco.html*
106. ^ *arXiv*, *https://arxiv.org/*
107. ^ *linear increase in submission rate*, *https://arxiv.org/stats/monthly_submissions*
108. ^ *The Subversive Proposal*, *http://www.arl.org/sc/subversive/i-overture-the-subversive-proposal.shtml*
109. ^ *Budapest Open Access Initiative*, *http://www.budapestopenaccessinitiative.org/read*
110. ^ *Budapest Open Access Initiative*, *http://www.budapestopenaccessinitiative.org/read*
111. ^ *The effect of open access and downloads ('hits') on citation impact: a bibliography of studies*, *http://opcit.eprints.org/oacitation-biblio.html*
112. ^ *The deliverance of open access books*, *https://www.universiteitleiden.nl/en/news/2019/01/open-access-books-attract-many-more-readers-and-slightly-more-citations*
113. ^ *Modularity: the next form of scientific information presentation?*, *http://www.science.uva.nl/projects/commphys/papers/jkmodul.htm*
114. ^ *The deconstructed journal – a new model for academic publishing*, *http://eprints.rclis.org/5841/1/DJpaper.pdf*
115. ^ *Decoupling the scholarly journal*, *https://www.frontiersin.org/articles/10.3389/fncom.2012.00019/full*

116. ^ *First meeting of the Open Archives initiative*, *http://www.openarchives.org/ups1-press.htm*
117. ^ *The UPS Prototype*, *http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html*
118. ^ *Open Archives Initiative*, *http://www.openarchives.org/*
119. ^ *OAI-PMH*, *http://www.openarchives.org/OAI/openarchivesprotocol.html*
120. ^ *ResourceSync*, *http://www.openarchives.org/rs/resourcesync*
121. ^ *OAI-ORE*, *http://www.openarchives.org/ore/toc*
122. ^ *Memento*, *https://tools.ietf.org/html/rfc7089*
123. ^ *Reminiscing About 15 Years of Interoperability Efforts*, *http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html*
124. ^ *Open Science Definition*, *https://www.fosteropenscience.eu/taxonomy/term/100*
125. ^ *Data on the Web Best Practices*, *https://www.w3.org/TR/dwbp/*
126. ^ *FAIR Guiding Principles*, *https://www.nature.com/articles/sdata201618*
127. ^ *Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud*, *https://content.iospress.com/articles/information-services-and-use/isu824*
128. ^ *Cost of not having FAIR research data*, *https://publications.europa.eu/portal2012-portlet/html/downloadHandler.jsp?identifier=d375368c-1a0a-11e9-8d04-01aa75ed71a1&format=pdf&language=en*
129. ^ *General Data Protection Regulation*, *https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679*
130. ^ *Revision of the PSI Directive*, *https://ec.europa.eu/digital-agenda/en/news/revision-psi-directive*
131. ^ *Office Journal C 240/2014*, *https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:C:2014:240:FULL&from=EN*
132. ^ *Ted Nelson Demonstrates XanaduSpace (by Arthur Bullard)*, *https://www.youtube.com/watch?v=1yLNGUeHapA&start=50&end=90*
133. ^ *Portable Document Format*, *https://www.iso.org/obp/ui/#iso:std:iso:32000:-2:ed-1:v1:en*
134. ^ *TeX*, *https://en.wikipedia.org/wiki/TeX*
135. ^ *LaTeX*, *https://en.wikipedia.org/wiki/LaTeX*
136. ^ *XMP Specification Part 1: Data Model, Serialization, and Core Properties*, *https://wwwimages2.adobe.com/content/dam/acom/en/devnet/xmp/pdfs/XMP%20SDK%20Release%20cc-2016-08/XMPSpecificationPart1.pdf*
137. ^ *Media Queries*, *http://www.w3.org/TR/css3-mediaqueries/*
138. ^ *Principle of Least Power*, *https://www.w3.org/DesignIssues/Principles#PLP*
139. ^ *Semantic publishing: the coming revolution in scientific journal publishing*, *https://onlinelibrary.wiley.com/doi/pdf/10.1087/2009202*
140. ^ *Genuine Semantic Publishing*, *http://www.tkuhn.org/pub/sempub/sempub.dokieli.html*
141. ^ *Why Linked Data is Not Enough for Scientists*, *https://www.research.manchester.ac.uk/portal/files/29982854/POST-PEER-REVIEW-NON-PUBLISHERS.PDF*
142. ^ *Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications*, *https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-28*
143. ^ *Argument graphs: Literature-Data Integration for Robust and Reproducible Science*, *https://www.isi.edu/ikcap/sciknow2015/papers/Clark.pdf*
144. ^ *Measuring academic influence: Not all citations are equal*, *https://arxiv.org/pdf/1501.06587.pdf*
145. ^ *Why (almost) Everything We Know About Citations is Wrong: Evidence from Authors*, *https://openaccess.leidenuniv.nl/bitstream/handle/1887/65227/STI2018_paper_241.pdf*
146. ^ *An annotation scheme for citation function*, *https://www.cl.cam.ac.uk/~sht25/papers/sigdial06.pdf*
147. ^ *FaBiO and CiTO: ontologies for describing bibliographic resources and citations*, *http://speroni.web.cs.unibo.it/publications/peroni-2012-fabio-cito-ontologies.pdf*

148. ^ Corpora for the conceptualisation and zoning of scientific papers, http://www.lrec-conf.org/proceedings/lrec2010/pdf/644_Paper.pdf
149. ^ Genuine Semantic Publishing, http://www.tkuhn.org/pub/sempub/sempub.dokieli.html
150. ^ Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data, https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2001414
151. ^ Digital Object Identifier, https://doi.org/
152. ^ DOI Registration Agencies, http://www.doi.org/registration_agencies.html
153. ^ CrossRef, https://www.crossref.org/
154. ^ Open Researcher and Contributor ID, https://orcid.org/
155. ^ Persistent Uniform Resource Locator, https://purl.org/
156. ^ Permanent Identifier Community Group, http://www.w3.org/community/perma-id/
157. ^ w3id.org, https://w3id.org/
158. ^ a b Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping, http://www.ijdc.net/article/download/9.1.331/362/
159. ^ archive.is, http://archive.is/
160. ^ WebCite, http://webcitation.org/
161. ^ Perma.cc, https://perma.cc/
162. ^ Webrecorder, https://webrecorder.io/
163. ^ Portico, https://www.portico.org/
164. ^ , https://www.lockss.org/
165. ^ CLOCKSS, https://clockss.org/
166. ^ Archive-It, https://archive-it.org/
167. ^ Swiss Federal Archives, https://www.bar.admin.ch/
168. ^ Library and Archives Canada, https://www.bac-lac.gc.ca/
169. ^ TIB, https://www.tib.eu/
170. ^ Zenodo, https://zenodo.org/
171. ^ Internet Archive, https://archive.org/
172. ^ Wayback Machine, https://web.archive.org/
173. ^ Physics World article for end March 1992, https://www.w3.org/History/1994/WWW/Journals/PhysicsWorld/PW.ascii
174. ^ Academics and their online networks: Exploring the role of academic social networking sites, https://firstmonday.org/ojs/index.php/fm/article/view/4937/4159
175. ^ Open Access Directory, http://oad.simmons.edu/oadwiki/Main_Page
176. ^ citizen science, https://en.wikipedia.org/wiki/Citizen_science
177. ^ PASTEUR4OA Briefing Paper: Infrastructures for Open Scholarly Communication, http://www.pasteur4oa.eu/sites/pasteur4oa/files/resource/Scholarly%20Platforms%20Briefing%20Paper_FINAL.pdf
178. ^ a b Encyclopedia of International Media and Communications, http://www.worldcat.org/oclc/773482913
179. ^ Free at Last: The Future of Peer-Reviewed Journals, http://cogprints.org/1685/1/12harnad.html
180. ^ Implementing Peer Review on the Net: Scientific Quality Control in Scholarly Electronic Journals, http://cogprints.org/1692/1/harnad96.peer.review.html
181. ^ What is open peer review? A systematic review, https://f1000research.com/articles/6-588
182. ^ A multi-disciplinary perspective on emergent and future innovations in peer review, https://f1000research.com/articles/6-1151
183. ^ The preregistration revolution, https://www.pnas.org/content/115/11/2600
184. ^ Changing the culture of scientific publishing from within , http://neurochambers.blogspot.com

/2012/10/changing-culture-of-scientific.html

185. ^ *Goodhart’s law*, *https://en.wikipedia.org/wiki/Goodhart%27s_law*
186. ^ *Citations impacts*, *https://en.wikipedia.org/wiki/Citation_impact*
187. ^ *Journal Impact Factor*, *https://en.wikipedia.org/wiki/Impact_factor*
188. ^ *h-index*, *https://en.wikipedia.org/wiki/H-index*
189. ^ *Prestigious Science Journals Struggle to Reach Even Average Reliability*, *https://www.frontiersin.org/articles/10.3389/fnhum.2018.00037/full*
190. ^ *Deep impact: unintended consequences of journal rank*, *https://www.frontiersin.org/articles/10.3389/fnhum.2013.00291/full*
191. ^ *Altmetrics*, *https://altmetrics.org/*
192. ^ *The Altmetrics Collection*, *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0048753*
193. ^ *2019 Big Deals Survey Report: An Updated Mapping of Major Scholarly Publishing Contracts in Europe*, *https://eua.eu/downloads/publications/2019%20big%20deals%20report.pdf*
194. ^ a b *SPARC Landscape Analysis*, *https://osf.io/preprints/lissa/58yhb/download*
195. ^ *Is the staggeringly profitable business of scientific publishing bad for science?*, *https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science*
196. ^ *Inequality in Knowledge Production: The Integration of Academic Infrastructure by Big Publishers*, *https://hal.archives-ouvertes.fr/hal-01816707/document*
197. ^ *The Oligopoly of Academic Publishers in the Digital Era*, *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0127502*
198. ^ *Opening the Black Box of Scholarly Communication Funding: A Public Data Infrastructure for Financial Flows in Academic Publishing*, *https://olh.openlibhums.org/articles/10.16995/olh.72/*
199. ^ *Open Access, the Global South and the Politics of Knowledge Production and Circulation*, *https://www.openlibhums.org/news/314/*
200. ^ *Group: Open Access Irony Award*, *http://www.citeulike.org/group/13803*
201. ^ *Practicing What You Preach: Evaluating Access of Open Access Research*, *https://mfr.osf.io/render?url=https%3A%2F%2Fosf.io%2Fdownload%2F4k9zd%2F%3Fdirect%26mode%3Drender*
202. ^ *Highlights from the SOAP project survey. What Scientists Think about Open Access Publishing*, *https://arxiv.org/pdf/1101.5260.pdf*
203. ^ *Vertical Integration in Academic Publishing*, *https://books.openedition.org/oep/9068*
204. ^ *Publisher, be damned! From price gouging to the open road*, *https://www.tandfonline.com/doi/full/10.1080/08109028.2014.891710*
205. ^ *Sci-Hub*, *https://en.wikipedia.org/wiki/Sci-Hub*
206. ^ *Open Access, the Global South and the Politics of Knowledge Production and Circulation*, *https://www.openlibhums.org/news/314/*
207. ^ *Forces and functions in scientific communication*, *http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html*
208. ^ *Rethinking Scholarly Communication*, *http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html*
209. ^ *A Perspective on Archiving the Scholarly Web*, *https://hvdsomp.info/papers/Papers/2014/iPres2014_Sompel_Treloar.pdf*
210. ^ *Requirements for Digital Preservation Systems*, *http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html*
211. ^ a b c *Laws of Media*, *http://www.worldcat.org/oclc/981396559*
212. ^ *Laws of Media*, *http://www.worldcat.org/oclc/981396559*
213. ^ *The Medium is the Massage*, *http://www.worldcat.org/oclc/47679653*
214. ^ *Gestalt laws of grouping*, *https://en.wikipedia.org/wiki/Principles_of_grouping*

215. ^ *mechanical movable type printing press*, *https://en.wikipedia.org/wiki/Printing_press*
216. ^ *Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web*, *http://drops.dagstuhl.de/opus/volltexte/2019/10328/pdf/dagrep_v008_i009_p029_18371.pdf*
217. ^ *Practical Knowledge Representation for the Web*, *https://www.cs.vu.nl/~frankh/postscript/IJCAI99-III.html*
218. ^ *The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey*, *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC442179/*
219. ^ *Content Negotiation*, *https://tools.ietf.org/html/rfc7231#section-3.4*
220. ^ *RDFa 1.1*, *https://www.w3.org/TR/rdfa-core/*
221. ^ *RDFa in HTML*, *https://www.w3.org/TR/rdfa-in-html/*
222. ^ *The Self-Describing Web*, *http://www.w3.org/2001/tag/doc/selfDescribingDocuments*
223. ^ a b *RDFa Use Cases: Scenarios for Embedding RDF in HTML*, *https://www.w3.org/TR/xhtml-rdfa-scenarios/*
224. ^ *RDF/A Syntax*, *https://www.w3.org/MarkUp/2004/rdf-a.html*
225. ^ *RDF in HTML: Approaches*, *http://infomesh.net/2002/rdfinhtml/*
226. ^ *Don’t Repeat Yourself*, *https://en.wikipedia.org/wiki/Don't_repeat_yourself*
227. ^ *user interaction*, *https://www.w3.org/TR/html5/editing.html*
228. ^ *loading Web pages*, *https://www.w3.org/TR/html5/browsers.html*
229. ^ *text-based Web browsers*, *https://en.wikipedia.org/wiki/Text-based_web_browser*
230. ^ *progressive enhancement*, *http://en.wikipedia.org/wiki/Progressive_enhancement*
231. ^ *Progressive Enhancement and the Future of Web Design*, *http://hesketh.com/publications/progressive_enhancement_and_the_future_of_web_design.html*
232. ^ *extensibility mechanisms*, *http://w3c.github.io/html/single-page.html#design-notes-extensibility*
233. ^ *common infrastructure*, *http://w3c.github.io/html/infrastructure.html#infrastructure*
234. ^ *DroppedAttributeProfile*, *https://www.w3.org/html/wg/wiki/DroppedAttributeProfile*
235. ^ *Gleaning Resource Descriptions from Dialects of Languages*, *https://www.w3.org/TR/grddl/*
236. ^ *XSL Transformations*, *https://www.w3.org/TR/xslt/*
237. ^ *DCMI Metadata Terms*, *http://www.dublincore.org/documents/dcmi-terms/*
238. ^ *schema.org*, *http://schema.org/*
239. ^ *Bibliographic Ontology Specification*, *http://purl.org/ontology/bibo/*
240. ^ *Semantic Publishing and Referencing Ontologies*, *http://purl.org/spar*
241. ^ *The SPAR Ontologies*, *https://w3id.org/spar/article/spar-iswc2018/*
242. ^ *FaBiO*, *http://purl.org/spar/fabio*
243. ^ *FRBR-DL*, *http://purl.org/spar/frbr*
244. ^ *DoCO*, *http://purl.org/spar/doco*
245. ^ *DEO*, *http://purl.org/spar/deo*
246. ^ *DataCite*, *http://purl.org/spar/datacite*
247. ^ *CiTO*, *http://purl.org/spar/cito*
248. ^ *BiRO*, *http://purl.org/spar/biro*
249. ^ *C4O*, *http://purl.org/spar/c4o*
250. ^ *PRO*, *http://purl.org/spar/pro*
251. ^ *PSO*, *http://purl.org/spar/pso*
252. ^ *PWO*, *http://purl.org/spar/pwo*
253. ^ *SCoRO*, *http://purl.org/spar/scoro*
254. ^ *FRAPO*, *http://purl.org/cerif/frapo*
255. ^ *BiDO*, *http://purl.org/spar/bido*
256. ^ *FiveStars*, *http://purl.org/spar/fivestars*
257. ^ *Simple Knowledge Organization System*, *https://www.w3.org/TR/skos-reference/*

258. ^ *Friend of a Friend*, *http://xmlns.com/foaf/0.1/*
259. ^ *vCard Ontology - for describing People and Organizations*, *https://www.w3.org/TR/vcard-rdf/*
260. ^ *SIOC Core Ontology Specification*, *http://rdfs.org/sioc/spec/*
261. ^ a b *Activity Vocabulary*, *https://www.w3.org/TR/activitystreams-vocabulary/*
262. ^ *Web Annotation Vocabulary*, *https://www.w3.org/TR/annotation-vocab/*
263. ^ a b *Embedding Web Annotations in HTML*, *https://www.w3.org/TR/annotation-html/*
264. ^ *Selectors and States*, *https://www.w3.org/TR/selectors-states/*
265. ^ a b c *Linked Data Notifications*, *https://www.w3.org/TR/ldn/*
266. ^ *The Cognitive Characteristics Ontology*, *http://purl.org/ontology/cco/core*
267. ^ *Creative Commons Rights Expression Language*, *http://creativecommons.org/ns*
268. ^ *Creative Commons licenses*, *https://creativecommons.org/licenses/*
269. ^ *Open Digital Rights Language*, *https://www.w3.org/ns/odrl/2/*
270. ^ *ODRL Vocabulary & Expression*, *https://www.w3.org/TR/odrl-vocab/*
271. ^ *Web Access Control*, *https://www.w3.org/wiki/WebAccessControl*
272. ^ a b *Access Control List*, *http://www.w3.org/ns/auth/acl*
273. ^ a b *The Cert Ontology 1.0*, *https://www.w3.org/ns/auth/cert*
274. ^ *The PROV Ontology*, *https://www.w3.org/TR/prov-o/*
275. ^ *The OPMW-PROV Ontology*, *http://www.opmw.org/model/OPMW/*
276. ^ *P-Plan Ontology*, *http://www.opmw.org/model/p-plan/*
277. ^ *Wf4Ever Research Object Model*, *https://w3id.org/ro/*
278. ^ a b *Verifiable Claims Data Model 1.0*, *https://www.w3.org/TR/verifiable-claims-data-model/*
279. ^ *The Memento terms vocabulary*, *http://mementoweb.org/ns*
280. ^ *Link Relation Types for Simple Version Navigation between Web Resources*, *https://tools.ietf.org/html/rfc5829*
281. ^ *Design Intent Ontology*, *https://w3id.org/dio*
282. ^ *RDF Data Cube vocabulary*, *https://www.w3.org/TR/vocab-data-cube/*
283. ^ *DDI-RDF Discovery Vocabulary*, *http://rdf-vocabulary.ddialliance.org/discovery.html*
284. ^ *Semantic Sensor Network Ontology*, *https://www.w3.org/TR/vocab-ssn/*
285. ^ *Describing Linked Datasets with the VoID Vocabulary*, *https://www.w3.org/TR/void/*
286. ^ *Evaluation and Report Language*, *https://www.w3.org/TR/EARL10-Schema/*
287. ^ *Semanticscience Integrated Ontology*, *http://sio.semanticscience.org/*
288. ^ *STATO: the statistical methods ontology*, *http://purl.obolibrary.org/obo/stato.owl*
289. ^ a b *Nanopublication Guidelines*, *http://nanopub.org/guidelines/working_draft/*
290. ^ *Micropublications*, *http://purl.org/mp*
291. ^ *Description of a Project*, *http://usefulinc.com/*
292. ^ *Web Content Accessibility Guidelines*, *https://www.w3.org/TR/WCAG/*
293. ^ *Web Content Accessibility Guidelines (WCAG) 2.1*, *https://www.w3.org/TR/WCAG21/*
294. ^ *Accessible Rich Internet Applications*, *https://www.w3.org/TR/wai-aria-1.1/*
295. ^ *Digital Publishing WAI-ARIA Module 1.0*, *https://www.w3.org/TR/dpub-aria/*
296. ^ *WAI-ARIA Graphics Module*, *https://www.w3.org/TR/graphics-aria/*
297. ^ *Authoring Tool Accessibility Guidelines*, *https://www.w3.org/TR/ATAG/*
298. ^ *ATAG 2.0 Guidelines*, *http://www.w3.org/TR/ATAG20/#guidelines*
299. ^ *User Agent Accessibility Guidelines*, *https://www.w3.org/TR/UAAG20/*
300. ^ *internationalization*, *https://www.w3.org/International/*
301. ^ *localization*, *https://www.w3.org/International/questions/qa-i18n#l10n*
302. ^ *CLEAR: a credible method to evaluate website archivability*, *http://purl.pt/24107/1/iPres2013_PDF/CLEAR%20a%20credible%20method%20to%20evaluate%20website%20archivability.pdf*

303. ^ The impact of JavaScript on archivability, https://www.cs.odu.edu/~mln/pubs/ijdl-archivability-2015.pdf
304. ^ ArchiveReady, http://archiveready.com/
305. ^ Significance is in the Eye of the Stakeholder, https://www.planets-project.eu/docs/papers/Dappert_Significant_Characteristics_ECDL2009.pdf
306. ^ Journal Article Tag Suite, http://jats.niso.org/
307. ^ From Markup to Linked Data: Mapping NISO JATS v1.0 to RDF using the SPAR (Semantic Publishing and Referencing) Ontologies, https://www.ncbi.nlm.nih.gov/books/NBK100491/
308. ^ Text Encoding Initiative, http://www.tei-c.org/
309. ^ Electronic Publication, http://www.idpf.org/epub
310. ^ EPUB Content Documents, http://www.idpf.org/epub3/latest/contentdocs
311. ^ Research Articles in Simplified HTML, https://w3id.org/people/essepuntato/papers/rash-peerj2016.html
312. ^ W3C Scholarly HTML Community Group, https://www.w3.org/community/scholarlyhtml/
313. ^ Scholarly HTML, https://w3c.github.io/scholarly-html/
314. ^ Statistical Linked Dataspaces, https://csarven.ca/statistical-linked-dataspaces
315. ^ Statistical Data and Metadata eXchange, http://www.iso.org/iso/catalogue_detail.htm?csnumber=52500
316. ^ Linked SDMX Data, https://csarven.ca/linked-sdmx-data
317. ^ Australian Bureau of Statistics, https://www.abs.gov.au/
318. ^ Swiss Federal Statistical Office, http://www.bfs.admin.ch/
319. ^ Bank for International Settlements, http://www.bis.org/
320. ^ European Central Bank, http://www.ecb.int/
321. ^ Food and Agriculture Organization of the United Nations, http://www.fao.org/
322. ^ Federal Reserve Board, https://www.federalreserve.gov/
323. ^ International Monetary Fund, http://www.imf.org/
324. ^ Organisation for Economic Co-operation and Development, http://www.oecd.org/
325. ^ UNESCO Institute for Statistics, http://www.uis.unesco.org/
326. ^ 270a.info, https://270a.info/
327. ^ Linked Open Data Cloud, https://lod-cloud.net/
328. ^ Linked Statistical Data Analysis, https://csarven.ca/linked-statistical-data-analysis
329. ^ Well-formed cubes, http://www.w3.org/TR/vocab-data-cube/#wf
330. ^ Oh yeah?, https://www.w3.org/DesignIssues/UI#OhYeah
331. ^ Enabling Scientific Data on the Web, https://www.era.lib.ed.ac.uk/bitstream/handle/1842/9957/Milowski2014.pdf
332. ^ A Declaration of the Independence of Cyberspace, https://www.eff.org/cyberspace-independence
333. ^ Decentralization: The Future of Online Social Networking, http://dig.csail.mit.edu/2008/Papers/MSNWS/index.html
334. ^ DuckDuckGo, https://duckduckgo.com/
335. ^ Understanding Knowledge as a Commons, http://www.worldcat.org/oclc/731904330
336. ^ Systematizing Decentralization and Privacy: Lessons from 15 Years of Research and Deployments, https://content.sciendo.com/downloadpdf/journals/popets/2017/4/article-p404.pdf
337. ^ From Databases to Dataspaces: A New Abstraction for Information Management, https://people.eecs.berkeley.edu/~franklin/Papers/dataspaceSR.pdf
338. ^ Principles of Dataspace Systems, https://homes.cs.washington.edu/~alon/files/pods06.pdf
339. ^ Web-scale Data Integration: You can only afford to Pay As You Go, http://cidrdb.org/cidr2007/papers/cidr07p40.pdf
340. ^ A decentralized architecture for consolidating personal information ecosystems: The WebBox,

https://eprints.soton.ac.uk/id/eprint/273200

341. ^ An Architecture of a Distributed Semantic Social Network, http://svn.aksw.org/papers/2011/SWJ_DSSN/public.pdf
342. ^ Distributed Semantic Social Networks: Architecture, Protocols and Applications, http://ul.qucosa.de/api/qucosa%3A12983/attachment/ATT-0/
343. ^ SWAT0, https://www.w3.org/2005/Incubator/federatedsocialweb/wiki/SWAT0
344. ^ SWAT1, https://www.w3.org/2005/Incubator/federatedsocialweb/wiki/SWAT1_use_cases
345. ^ Federated Social Web Incubator Group, https://www.w3.org/2005/Incubator/federatedsocialweb/
346. ^ Data ownership and interoperability for a decentralized social semantic web, https://tel.archives-ouvertes.fr/tel-00917965/document
347. ^ Amber: Decoupling User Data from Web Applications, https://pdos.csail.mit.edu/papers/amber:hotos15.pdf
348. ^ A Demonstration of the Solid Platform for Social Web Applications, http://gdac.uqam.ca/WWW2016-Proceedings/companion/p223.pdf
349. ^ 400+ Tools and innovations in scholarly communication, https://docs.google.com/spreadsheets/d/1KUMSeq_Pzp4KveZ7pb5rddcssk1XBTiLHniD0d3nDqo
350. ^ Socially Aware Cloud Storage, https://www.w3.org/DesignIssues/CloudStorage
351. ^ Read-Write Linked Data, https://www.w3.org/DesignIssues/ReadWriteLinkedData
352. ^ Linked Data Platform, https://www.w3.org/TR/ldp/
353. ^ a b c Linked Data Platform Use Cases and Requirements, https://www.w3.org/TR/ldp-ucr/
354. ^ LDP Paging, https://www.w3.org/TR/ldp-paging/
355. ^ Hydra, http://www.hydra-cg.com/spec/latest/core/
356. ^ Linked Data Fragments, http://linkeddatafragments.org/
357. ^ Triple Pattern Fragments, http://www.hydra-cg.com/spec/latest/triple-pattern-fragments/
358. ^ Linked Data Templates, https://atomgraph.github.io/Linked-Data-Templates/
359. ^ RDF/POST Encoding for RDF, http://www.lsrn.org/semweb/rdfpost.html
360. ^ Fedora API Specification, https://fcrepo.github.io/fcrepo-specification/
361. ^ HTTP Framework for Time-Based Access to Resource States -- Memento, https://tools.ietf.org/html/rfc7089
362. ^ Web Annotation, https://www.w3.org/annotation/
363. ^ Web Annotation Protocol, http://www.w3.org/TR/annotation-protocol/
364. ^ Web Annotation Data Model, http://www.w3.org/TR/annotation-model/
365. ^ Activity Streams 2.0, https://www.w3.org/TR/activitystreams-core/
366. ^ ActivityPub, https://www.w3.org/TR/activitypub/
367. ^ Social Personal Data Stores: the Nuclei of Decentralised Social Machines, https://www.research.ed.ac.uk/portal/files/20028374/vanKleek_et_al_2015_Social_Personal_Data_Stores.pdf
368. ^ The Presentation of Self on a Decentralised Web, http://dr.amy.gy/
369. ^ ORCID community, https://orcid.org/about/community
370. ^ such initiatives, https://orcid.org/content/requiring-orcid-publication-workflows-open-letter
371. ^ [GOAL] eLife collects ORCIDs from authors of accepted papers at proofing, http://mailman.ecs.soton.ac.uk/pipermail/goal/2017-March/004432.html
372. ^ User Tracking on Academic Publisher Platforms, https://www.codyh.com/writing/tracking.html
373. ^ Ad Tech Surveillance on the Public Sector Web, https://www.cookiebot.com/media/1121/cookiebot-report-2019-medium-size.pdf
374. ^ W3C Workshop on the Future of Social Networking Report, https://www.w3.org/2008/09/msnws/report.html
375. ^ FOAF+SSL, http://blogs.sun.com/bblfish/entry/foaf_ssl_creating_a_global

376. ^ RDFAuth, http://blogs.sun.com/bblfish/entry/rdfauth_sketch_of_a_buzzword
377. ^ sketch of a simple authentication protocol, https://lists.w3.org/Archives/Public/semantic-web/2008Mar/0207.html
378. ^ FOAF+SSL: RESTful Authentication for the Social Web, http://dig.csail.mit.edu/2009/Papers/SPOT/foaf-ssl-spot2009.pdf
379. ^ WebID, https://www.w3.org/2005/Incubator/webid/spec/
380. ^ Web Identity and Discovery, https://www.w3.org/2005/Incubator/webid/spec/identity
381. ^ WebID Profile, https://www.w3.org/2005/Incubator/webid/spec/identity/#dfn-webid_profile
382. ^ The 'Basic' HTTP Authentication Scheme, https://tools.ietf.org/html/rfc7617
383. ^ Web of Trust, https://en.wikipedia.org/wiki/Web_of_trust
384. ^ Pretty Good Privacy, https://en.wikipedia.org/wiki/Pretty_Good_Privacy
385. ^ key signing parties, https://en.wikipedia.org/wiki/Key_signing_party
386. ^ X.509, http://www.itu.int/rec/T-REC-X.509/en
387. ^ HTTP over TLS, https://tools.ietf.org/html/rfc2818
388. ^ Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, https://tools.ietf.org/html/rfc5280
389. ^ The Transport Layer Security Protocol Version 1.2, https://tools.ietf.org/html/rfc5246
390. ^ WebID-TLS, https://www.w3.org/2005/Incubator/webid/spec/tls
391. ^ Extending the WebID Protocol with Access Delegation, http://ceur-ws.org/Vol-905/TrampEtAl_COLD2012.pdf
392. ^ OpenID Connect, https://openid.net/specs/openid-connect-core-1_0.html
393. ^ The OAuth 2.0 Authorization Framework, https://tools.ietf.org/html/rfc6749
394. ^ WebID-OIDC, https://github.com/solid/webid-oidc-spec
395. ^ Prohibiting Delegation, http://erights.org/elib/capability/delegations.html
396. ^ Persistent Domains, https://www.w3.org/DesignIssues/PersistentDomains
397. ^ Cool URIs don't change, http://www.w3.org/Provider/Style/URI
398. ^ Philosophical Engineering and Ownerhip of URIs, https://www.w3.org/DesignIssues/PhilosophicalEngineering
399. ^ URI Persistence Policy, https://www.w3.org/Consortium/Persistence
400. ^ Decentralized Identifiers, https://w3c-ccg.github.io/did-spec/
401. ^ Analyzing the Persistence of Referenced Web Resources with Memento, https://arxiv.org/pdf/1105.3459.pdf
402. ^ Hiberlink, http://hiberlink.org/
403. ^ Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253
404. ^ Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115253
405. ^ Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0167475
406. ^ Persistent URIs Must Be Used To Be Persistent, https://arxiv.org/pdf/1602.09102.pdf
407. ^ cite-as: A Link Relation to Convey a Preferred URI for Referencing, https://tools.ietf.org/html/draft-vandesompel-citeas-04
408. ^ Signposting the Scholarly Web, http://signposting.org/
409. ^ Robust Links, http://robustlinks.mementoweb.org/spec/
410. ^ Trusty URI, http://trustyuri.net/
411. ^ Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data, https://arxiv.org/pdf/1401.5775.pdf
412. ^ Naming Things with Hashes, https://tools.ietf.org/html/rfc6920

413. ^ *Trusty URI Specification – Version 1, http://trustyuri.net/spec/v1.FADQoZWcYugekAb4jW-Zm3_5Cd9tmkkYEV0bxK2fLSKao.md*
414. ^ *The Anatomy of a Nanopublication, https://content.iospress.com/download/information-services-and-use/isu613?id=information-services-and-use%2Fisu613*
415. ^ *Decentralized provenance-aware publishing with nanopublications, https://peerj.com/articles/cs-78/*
416. ^ *Signing HTTP Messages, https://tools.ietf.org/html/draft-cavage-http-signatures-11*
417. ^ *Linked Data Proofs, https://w3c-dvcg.github.io/ld-proofs/*
418. ^ *The Presentation of Self on a Decentralised Web, http://dr.amy.gy/*
419. ^ *Why is the Web Loosely Coupled? A Multi-Faceted Metric for Service Design, http://www2009.eprints.org/92/1/p911.pdf*
420. ^ *web hosting service, https://en.wikipedia.org/wiki/Shared_web_hosting_service*
421. ^ *self-hosting, https://en.wikipedia.org/wiki/Self-hosting*
422. ^ *client–server model, https://en.wikipedia.org/wiki/Client%E2%80%93server_model*
423. ^ *Virtuoso, http://www.chakoteya.net/Voyager/613.htm*
424. ^ *Linked Data Notifications: a resource-centric communication protocol, https://csarven.ca/linked-data-notifications*
425. ^ *Extended Semantic Web Conference, http://2017.eswc-conferences.org/*
426. ^ *Linked Specifications, Test Suites, and Implementation Reports, https://csarven.ca/linked-specifications-reports*
427. ^ *Linked Data Notifications, https://www.w3.org/TR/ldn/*
428. ^ *Semantic Pingback, https://aksw.github.io/SemanticPingback/*
429. ^ *Webmention, https://www.w3.org/TR/webmention/*
430. ^ *Pingback, http://www.hixie.ch/specs/pingback/pingback*
431. ^ *Provenance Pingback, http://www.w3.org/TR/prov-aq/#provenance-pingback*
432. ^ *WebSub, https://www.w3.org/TR/websub/*
433. ^ *PubSubHubbub, http://pubsubhubbub.github.io/PubSubHubbub/pubsubhubbub-core-0.4.html*
434. ^ *DSNotify, http://www.cibiv.at/~niko/dsnotify/*
435. ^ *sparqlPuSH, https://www.w3.org/2001/sw/wiki/SparqlPuSH*
436. ^ *ResourceSync Change Notification, http://www.openarchives.org/rs/notification/1.0/notification*
437. ^ *W3C Social Web Working Group, https://www.w3.org/Social/WG*
438. ^ *user stories, https://www.w3.org/wiki/Socialwg/Social_API/User_stories*
439. ^ *RDF Calendar, https://www.w3.org/TR/rdfcal/*
440. ^ *PROV Ontology, https://www.w3.org/TR/prov-o/*
441. ^ *W3C Recommendation, https://www.w3.org/TR/ldn/*
442. ^ *Overview of Linked Data Notifications, https://www.w3.org/TR/ldn/linked-data-notifications-overview.svg*
443. ^ *used to JSON-based APIs but not RDF, http://manu.sporny.org/2014/json-ld-origins-2/*
444. ^ *Shapes Constraint Language (SHACL), https://www.w3.org/TR/shacl/*
445. ^ *ShEx, https://shexspec.github.io/spec/*
446. ^ *Security and Privacy Review, https://www.w3.org/TR/ldn/#security-and-privacy-review*
447. ^ *Self-Review Questionnaire: Security and Privacy, https://www.w3.org/TR/2015/NOTE-security-privacy-questionnaire-20151210/*
448. ^ *test suite, https://linkedresearch.org/ldn/tests/*
449. ^ a b *mayktso, https://github.com/csarven/mayktso*
450. ^ a b *dokieli, https://dokie.li/*
451. ^ *LDN Tests Summary, https://linkedresearch.org/ldn/tests/summary*
452. ^ *Exit Criteria, https://www.w3.org/TR/ldn/#exit-criteria*

453. ^ *LDN Test Suite*, *https://linkedresearch.org/ldn/tests/*
454. ^ *LDN Test Reports*, *https://linkedresearch.org/ldn/tests/reports/*
455. ^ *LDN Test Reports and Summary*, *https://linkedresearch.org/ldn/tests/summary*
456. ^ *LDN Test Reports and Summary*, *https://linkedresearch.org/ldn/tests/summary*
457. ^ *Sloph*, *https://dr.amy.gy/chapter5#building-pwo*
458. ^ *Scholastic Commentaries and Texts Archive*, *https://scta.info/*
459. ^ *International Image Interoperability Framework*, *https://iiif.io/*
460. ^ *comparison of notification mechanisms*, *https://csarven.ca/linked-data-notifications#comparison-of-notification-mechanisms*
461. ^ *BasicContainer*, *https://www.w3.org/TR/ldp/#dfn-linked-data-platform-basic-container*
462. ^ *Social Web Protocols*, *https://www.w3.org/TR/social-web-protocols/*
463. ^ *The WebSocket Protocol*, *https://tools.ietf.org/html/rfc6455*
464. ^ *Generic Event Delivery Using HTTP Push*, *https://tools.ietf.org/html/rfc8030*
465. ^ *Follow Activity*, *https://www.w3.org/TR/activitypub/#follow-activity-inbox*
466. ^ *Followers Collection*, *https://www.w3.org/TR/activitypub/#followers*
467. ^ *Information Management: A Proposal*, *https://www.w3.org/History/1989/proposal.html*
468. ^ *Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli*, *https://csarven.ca/dokieli-rww*
469. ^ *Social Linked Data*, *https://solid.mit.edu/*
470. ^ *node-solid-server*, *https://github.com/solid/node-solid-server*
471. ^ *ORCID source*, *https://github.com/ORCID/ORCID-Source/*
472. ^ *WorldWideWeb*, *https://worldwideweb.cern.ch/*
473. ^ *Polyglot Markup*, *https://www.w3.org/TR/html-polyglot/*
474. ^ *TimeTravel*, *http://timetravel.mementoweb.org/*
475. ^ *ISO 8601*, *https://www.iso.org/obp/ui#iso:std:iso:8601:-1:ed-1:v1:en*
476. ^ *XML Patch*, *https://tools.ietf.org/html/rfc5261*
477. ^ *A Media Type for XML Patch Operations*, *https://tools.ietf.org/html/rfc7351*
478. ^ *SPARQL Update*, *https://www.w3.org/TR/sparql11-update/*
479. ^ *Mixed Content*, *https://www.w3.org/TR/mixed-content/*
480. ^ *Social Annotations in Digital Library Collections*, *http://www.dlib.org/dlib/november08/gazan/11gazan.html*
481. ^ *Digital Publishing Annotation Use Cases*, *https://www.w3.org/TR/dpub-annotation-uc/*
482. ^ *Annotating All Knowledge Coalition*, *https://hypothes.is/annotating-all-knowledge/*
483. ^ *A Coalition for Scholarly Annotation*, *https://hypothes.is/blog/a-coalition-of-scholarly-annotators/*
484. ^ *vendor lock-in*, *https://en.wikipedia.org/wiki/Vendor_lock-in*
485. ^ *A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management*, *http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf*
486. ^ *Cross-Origin Resource Sharing*, *https://www.w3.org/TR/cors/*
487. ^ *sparkline*, *https://en.wikipedia.org/wiki/Sparkline*
488. ^ *single-page application*, *https://en.wikipedia.org/wiki/Single-page_application*
489. ^ *Web browser extension*, *https://en.wikipedia.org/wiki/Browser_extension*
490. ^ *User Interfaces for Semantic Authoring of Textual Content: A Systematic Literature Review*, *http://svn.aksw.org/papers/2011/JWS_SemanticContentAuthoring/public.pdf*
491. ^ *Table 6: Overview of results*, *http://svn.aksw.org/papers/2011/JWS_SemanticContentAuthoring/public.pdf*
492. ^ *FAIR Metrics*, *https://github.com/FAIRMetrics/Metrics*
493. ^ *A design framework and exemplar metrics for FAIRness*, *https://www.nature.com/articles*

/sdata2018118

494. ^ *web agent*, *https://www.w3.org/TR/webarch/#def-web-agent*
495. ^ *Google Docs*, *https://docs.google.com/*
496. ^ *Medium*, *https://medium.com/*
497. ^ *Twitter*, *https://twitter.com/*
498. ^ *Facebook*, *https://facebook.com/*
499. ^ *WordPress*, *https://wordpress.org/*
500. ^ *MediaWiki*, *https://www.mediawiki.org/*
501. ^ *Wikipedia*, *https://en.wikipedia.org/*
502. ^ *Wikidata*, *https://www.wikidata.org/*
503. ^ *Hypothes.is*, *https://hypothes.is/*
504. ^ *LibreOffice Online*, *https://github.com/libreoffice/online*
505. ^ *Amaya*, *https://www.w3.org/Amaya/*
506. ^ *Annotea: An Open RDF Infrastructure for Shared Web Annotations*, *https://www.w3.org/2001/Annotea/Papers/www10/annotea-www10.html*
507. ^ *Annotea*, *https://www.w3.org/2001/Annotea/*
508. ^ *protocol*, *https://www.w3.org/2001/Annotea/User/Protocol.html*
509. ^ *XML Pointer Language*, *https://www.w3.org/TR/xptr-element/*
510. ^ *Open Annotation Data Model*, *http://www.openannotation.org/spec/core/*
511. ^ *Embedding Web Annotations in HTML*, *https://www.w3.org/TR/annotation-html/*
512. ^ *Model*, *https://www.w3.org/TR/annotation-model/*
513. ^ *Vocabulary*, *https://www.w3.org/TR/annotation-vocab/*
514. ^ *Linked Data Notifications*, *https://www.w3.org/TR/ldn/*
515. ^ *Editor’s Draft*, *https://linkedresearch.org/ldn/*
516. ^ *Test Suite*, *https://linkedresearch.org/ldn/tests/*
517. ^ a b *Semantic Statistics*, *http://semstats.org/*
518. ^ a b *CEUR-WS.org*, *http://ceur-ws.org/*
519. ^ *ceur-make*, *https://github.com/ceurws/ceur-make*
520. ^ *SemStats 2016 Proceedings*, *http://ceur-ws.org/Vol-1654/*
521. ^ *examples in the wild*, *https://github.com/linkeddata/dokieli/wiki#examples-in-the-wild*
522. ^ *Solid Panes*, *https://github.com/solid/solid-panes*
523. ^ *Linked Research*, *https://linkedresearch.org/*
524. ^ *https://csarven.ca/*, *https://csarven.ca/*
525. ^ *social machine*, *https://en.wikipedia.org/wiki/Social_machine*
526. ^ *Buckminster Fuller*, *https://en.wikiquote.org/wiki/Buckminster_Fuller*
527. ^ *Linked Research*, *https://linkedresearch.org/*
528. ^ a b *Universal Declaration of Human Rights*, *http://www.un.org/en/universal-declaration-human-rights/*
529. ^ *Tower of Babel*, *https://en.wikipedia.org/wiki/Tower_of_Babel*
530. ^ *As We May Think*, *http://www.ps.uni-saarland.de/~duchier/pub/vbush/vbush.txt*
531. ^ *Freedom of Information*, *https://en.wikipedia.org/wiki/Freedom_of_information*
532. ^ *free-culture movement*, *https://en.wikipedia.org/wiki/Free-culture_movement*
533. ^ *Accessibility, Usability, and Inclusion*, *https://www.w3.org/WAI/fundamentals/accessibility-usability-inclusion/*
534. ^ *Call for Linked Research*, *https://csarven.ca/call-for-linked-research*
535. ^ *individual researchers*, *https://linkedresearch.org/calls#reviewers*
536. ^ *Call for (Enabling) Linked Research*, *https://linkedresearch.org/calls*
537. ^ *Examples in the Wild*, *https://github.com/linkeddata/dokieli/wiki#examples-in-the-wild*

538. ^ *Linked Open Research Cloud*, *https://2018.eswc-conferences.org/call-for-papers/*
539. ^ a b *Extended Semantic Web Conference*, *https://eswc-conferences.org/*
540. ^ *International Semantic Web Conference*, *https://en.wikipedia.org/wiki/International_Semantic_Web_Conference*
541. ^ *Decentralizing the Semantic Web*, *https://iswc2019.desemweb.org/*
542. ^ *Enabling Decentralised Scholarly Communication*, *https://linkedresearch.org/events/eswc2017/*
543. ^ *Enabling Open Semantic Science*, *https://semsci.github.io/SemSci2018/*
544. ^ *Linked Data on the Web*, *http://events.linkeddata.org/ldow2017/*
545. ^ *Research Objects*, *https://researchobject.github.io/ro2019/*
546. ^ *Researcher-Centric Scholarly Communication*, *https://linkedresearch.org/events/the-web-conf-2018/*
547. ^ *SAVE-SD*, *http://cs.unibo.it/save-sd/2017/*
548. ^ *Web Observatories, Social Machines and Decentralisation Workshop*, *http://sociam.org/wow2017/*
549. ^ *Authoring, annotations, and notifications in a decentralised Web*, *https://indico.cern.ch/event/405949/contributions/2486377/*
550. ^ *Authoring, Annotations, and Notifications in a decentralised Web with Dokieli*, *http://swib.org/swib17/*
551. ^ *Linked Open Research Cloud*, *https://linkedresearch.org/cloud*
552. ^ *Web Annotations*, *https://www.w3.org/TR/annotation-model/*
553. ^ *Creative Commons CC0 1.0 Universal*, *https://creativecommons.org/publicdomain/zero/1.0/*
554. ^ *Linked Open Data Cloud*, *https://lod-cloud.net/*
555. ^ *Re: Please update your resource in the LOD Cloud Diagram*, *https://lists.w3.org/Archives/Public/public-lod/2017Jun/0020.html*
556. ^ *Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud*, *http://events.linkeddata.org/ldow-lddl/papers/LDOW-DeBattista-et-al.pdf*
557. ^ *Linked Specifications, Test Suites, and Implementation Reports Related Work*, *https://csarven.ca/linked-specifications-reports#related-work*
558. ^ *Linked Data Notifications*, *https://csarven.ca/linked-data-notifications*
559. ^ *Decentralised Authoring, Annotations and Notifications for a Read-Write Web with dokieli*, *https://csarven.ca/dokieli-rww*
560. ^ *Scholarly Communication: Deconstruct & Decentralize?*, *https://www.youtube.com/watch?v=o4nUe-6Ln-8&t=2792*
561. ^ *Solid project*, *https://solid.mit.edu/*
562. ^ *Ethical Web Principles*, *https://www.w3.org/2001/tag/doc/ethical-web-principles/*