Textdatenbank und Wörterbuch des Klassischen Maya

Arbeitsstelle der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste an der Rheinischen Friedrich-Wilhelms-Universität Bonn

ISSN 2366-5556

PROJECT REPORT 2 | EN

OPEN ACCESS

Published 19 Dec 2014

DOI: 10.20376/IDIOM-23665556.14.pr002.en

Annual Report for 2014

Nikolai Grube[1], Christian Prager[1], Katja Diederichs[1], Sven Gronemeyer[1,2], Elisabeth Wagner[1], Maximilian Brodhun[3], Franziska Diehr[3] & Petra Maier[4] (translated by Mallory Matsumoto)

[1)] Rheinische Friedrich-Wilhelms-Universität, Bonn
[2)] La Trobe University, Melbourne
[3)] Niedersächsische Staats- und Universitätsbibliothek, Göttingen
[4)] ULB Heinrich-Heine-Universität, Düsseldorf

Project Description, Goals, and Methods

The incompletely deciphered hieroglyphic script and language of the Maya constitutes the primary focus of the research project under discussion here, which is being carried out cooperatively by the Universities of Bonn and Göttingen. The project's goal is to compile a text database and, on this basis, a dictionary of Classic Mayan. Approximately 10,000 text and image carriers have survived from the the time of the Maya culture's fluorescence between A.D. 250 and 950, and their textual and iconographic information provide unique perspectives into the language, culture, and history of pre-Hispanic Maya society. To this day, however, the text and image carriers have yet to be systematically documented and comprehensively analyzed. Such efforts would permit a detailed and precise investigation of the Classic Mayan literary language, for instance by comparing text passages using co-text and co-occurrence analysis, correlating imagery with text passages, or registering the composition or function of a text carrier in the inscription and thus potentially elucidating ambiguous text passages. Until now, such systematic and cross-linked work with text, image, and information carriers was impossible, because the necessary technology did not yet exist in this field of research. Within the framework of this project, the text and image carriers will be systematically described according to uniform standards and the source material will be made machine-readable based on XML, thereby creating the foundation for compiling the dictionary. This undertaking can only be initiated using methods and technologies from the digital humanities, whereby the project is drawing upon tools and technologies that are already available in the virtual research environment (VRE) TextGrid or that are being developed and implemented in the context of the project, as the case may be. An essential prerequisite for this is that not only the linguistic content of the inscriptions and the iconic information from the imagery, but also data concerning the text and image carriers (descriptive or metadata) be taken into account and integrated into the database. Towards this goal, tools and workflows are being developed in TextGrid that facilitate 1) documentation of the text and image carriers with an assessment of the current state of research, 2) epigraphic-linguistic evaluation of the hieroglyphic text,

and 3) editing of the text with the transliteration, transcription, and translation within a single system. The VRE does not simply contain descriptions of the text carriers or information concerning the texts. In addition, the database user also acquires an overview of the authors who have studied or published a monument, discussed a text passage, or were the first to propose an as yet still valid reading for a hieroglyph. The text carrier thus receives its own "biography" that is intimately tied to the text contents and is taken into account when analyzing the meaning of words. The VRE, which is currently under construction, heavily orients itself to the epigraphic workflow, which begins with documentation of the text carrier and registration of descriptive data; proceeds with sign classification, transliteration, and transcription of the texts; continues with morphological segmentation and linguistic interpretation; and ideally concludes with the text's translation and publication.

Start of the Projekt and Beginning Work

The project began on January 1, 2014. The project was contractually and administratively established at the University of Bonn (Project Number: 98050020) and cooperation was initiated between the University of Bonn and the Göttingen State and University Library (SUB) in the period preceding the appointment of the two researchers Dr. Christian Prager and Elisabeth Wagner, M.A. on March 1, 2014 and March 3, 2014, respectively, as well as two research assistants. An on-lending agreement was concluded between the University of Bonn and the SUB that governs the outflow of funds from Bonn to Göttingen to finance the two project workplaces (150%) at the SUB. The position for a computer scientist in Bonn and the metadata and developer positions in Göttingen were posted in the first quarter of 2014 and filled in the beginning of May and June 2014 by Katja Diedrichs in Bonn and Franziska Diehr and Maximilian Brodhun in Göttingen. Sven Gronemeyer, M.A. was hired as an additional researcher on May 1, 2014. The design of the metadata schema as established in the first work package (see project proposal), and the development and programming of the virtual epigraphic research environment in TextGrid were thus undertaken beginning in the first half of 2014. The project's project and the present work report should be evaluated in this context.

Workplace Members, Duties, and Intra-Project Communication

To date, fourteen people are under contract with the project. The three researchers Dr. Christian Prager (project coordination), Elisabeth Wagner, M.A. und Sven Gronemeyer, M.A. are experts on Mayan writing, language, and iconography, and Katja Diedrichs is active in the computer scientific sector of the Bonn workplace. The latter coordinates the cooperation between the project and the ULB with respect to the presentation platform Visual Library, where a virtual inscription archive will be established online beginning in 2016. At the SUB, Franziska Diehr takes care of the design of the metadata schema according to the specialized guidelines from Bonn. On the basis of Franziska Diehr's metadata schema, Maximilian Brodhun develops, programs, and implements the epigraphic workflow, preliminarily entitled "Text-Dokumentation · Text-Analyse · Text-Edition", in the virtual research environment TextGrid. The researchers in Bonn are also supported by research assistants Christiane Bahr (digitalization), Laura Burzywoda (digitalization), Antje Grothe (bibliography), Lena Heise (digitalization), Leonie Heine (digitalization), Jana Karsch (epigraphy, digitalization) Nikolai Kiel (epigraphy, digitalization), and Mallory Matsumoto (linguistics, digitalization).

Also working in association with the project are the eHumanities staff member at the ULB, Jan Kenter, who together with Katja Diederichs organizes the cooperation between the ULB and the project; the librarian Petra Maier from the University and Regional Library of Düsseldorf, a volunteer contributor who has prepared a study on the use of XML to mark-up Maya texts under the direction of Dr. Heike

Neuroth (SUB) and has thus made a guest contribution in this area. Dr. Thomas Kollatz from the Salomon Ludwig Steinheim Institute in Essen, an expert on the XML-based mark-up standards TEI and Epidoc, has also supported the project as a contact person since November 2014.

Written communication between Bonn and Göttingen is conducted on an private wiki from the collaborative software Confluence, which is integrated into the research infrastructure DARIAH-DE and by means of which all participating scientists exchange data and discuss controlled vocabularies, authority controls, and epigraphic norms and standards, among other issues. The wiki contains the complete project documentation, including all aggregated technical and specialized contents to date, which must remain accessible as reference material over the long term. This includes documents concerning project organization and the various fields of work, as well as materials that can be stored there and accessed and collectively worked on by collaborators. Documents such as lectures, presentations, or work lists are stored in a cloud server and can thus be collaboratively edited.

Verbal communication is central to the project's work: weekly telephone conferences allow project participants to discuss urgent and short-term duties and tasks. At monthly project meetings in Bonn and Göttingen, intermediate-term milestones and goals are established and tasks for the coming months are planned. The first project meeting involving all participants took place on January 14, 2014 in Bonn and facilitated coordination between the project partners. On May 19 and 20, 2014 the Bonn researchers led a workshop on the Maya script and calendar in Göttingen. In this way, the project employees at the SUB were able to familiarize themselves with the subject matter. The first meeting about metadata and vocabulary occurred on June 19, 2014 in Bonn. Development of the virtual research environment was the focus of the meeting on August 28, 2014, and the metadata schema was addressed at additional project meetings on October 15 and November 27, 2014. The meeting between the ULB, SUB, and the Bonn workplace that is planned for December 10, 2014 has on its agenda the import and presentation of the epigraphic research data on the ULB's presentation platform Visual Library.

In addition to their work in the project, the project greatly values the professional qualification of its staff. In order to orient themselves in the field of Digital Humanities, the staff from the Bonn workplace furthered their education by participating in various national and international conferences and courses (Passau, Lausanne, London). The focus of these events was particularly on the field of Digital Humanities, and by participating they became familiar with and applied methods from this field, with particular emphasis on acquisition of the mark-up language XML. Additionally, the staff members were briefed on the application of TextGrid in the humanities at the regular meetings for TextGrid users.

Workshops, Conferences, und the Kick-Off Meeting

In 2014, the project was presented to the scientific community and the public at five national and international conferences and meetings. Between January 31 and February 2, 2014, Nikolai Grube, Christian Prager, and Elisabeth Wagner introduced the project to a scientific audience at the 17th Conference of Mesoamericanists in Basel, where they presented the goals and tasks of the dictionary project and its technical realization. On June 4 and 5, 2014, the members of the research center presented the project in the context of a public exhibition of the Academy of Sciences, Humanities and the Arts in the state parliament in Düsseldorf. In preparation for this event, an exhibition column with textual and graphic material was designed. As part of the German Open Monument Day (Tag des offenen Denkmals) on September 14, the staff was again able to present the project to a broad audience. Colleagues and experts were able to inform themselves about the project at the kick-off meeting on October 14 in Düsseldorf. In the context of the international workshop "Words in Context:

Perspectives and Strategies for the Lexicography of Classic Mayan" in the Karl Arnold House of Sciences in Düsseldorf, all staff members from the Bonn and Göttingen workplaces presented their specialized work areas to an international audience of experts. Colleagues from the United States (David Stuart and Marc Zender), Australia (Peter Mathews), and Germany (Gordon Whittaker) were invited to participate in this workshop and to present results of their own, current research. At a subsequent workshop on standards and conventions in transliterating, transcribing, and analyzing Maya texts, the specialists agreed upon common norms and standards that will be utilized in the project. A follow-up workshop is planned for 2015 to discuss the project's progress and results. Finally, the project was introduced to specialists in the field of digital humanities in Essen, when the workplace staff presented the project to colleagues and experts from the eHumanities at the 5th TextGrid User Meeting on November 25 and 26, 2014. Another international presentation is planned at a digital humanities meeting in Graz, in order that additional contacts to other DH projects may be made. In December 2015, the academy project will be represented at the 20th European Maya Conference as a co-organizer of this meeting about digital methods in Maya research.

Scientific Advisory Board und Collaboration

In the framework of our workshop on Maya writing on October 14, we were able to recruit renowned international researchers as scientific advisors to our project: Prof. David Stuart (Austin), Prof. Marc Zender (New Orleans), Prof. Peter Mathews (Melbourne), and Prof. Gordon Whittaker (Göttingen). A meeting of the scientific advisory board is planned in the coming year in the context of the 20th European Maya Conference in Bonn, for which the project will organize its own panel about the dictionary project.

Tasks and Work in Progress in Bonn

In this section of the work report, members of the Bonn workplace report on their activities and the current state of their work.

Text Archive

At the Bonn workplace, the **Text Archive** and various data collections and work lists are currently being developed and compiled, respectively. These are being promptly published on the project's website www.mayawoerterbuch.de, and are also being added to the virtual research environment TextGrid. The inscription archive, which is currently housed in file cabinets and remains to be digitalized, is under construction and will be established at the beginning of 2015. As soon as the University of Bonn makes available the necessary space in early 2015, the analog archive will be established in file format. Berthold Riese's documentation of Maya documentation, which consists of 135 file folders with photographs, drawings, and notes and was handed over to the workplace in the summer of 2014, will be incorporated into the Bonn Maya inscription archive. The indexation and digitalization of this portion of the archive is currently underway and will be complete by the end of 2015. The 40,000-photo archive of Karl Herbert Mayer (Graz) will also be incorporated into the workplace's inscription archive in the coming years. For this purpose, the archive was examined on its original site and an inventory was taken, the results of which are summarized in a project report and will be published on the project website as a Working Note. The archive of Karl Herbert Mayer, which comprises slides, film negatives, and prints, will be gradually transported to Bonn, where it will be indexed, digitalized, and archived by research assistants. As soon as the space are ready for occupation, construction of the

analogue inscription archive will also be initiated. Due to the large quantity of data, digitalization and indexation of the photos will proceed at least until the end of 2016. The workplace's analogue and digital inscription archive will be enriched not only by the project director's and the staff members' own materials, but also by gifts from colleagues in Germany and abroad, who will make their image inventories available for digitalization and open access publication. A presentation platform for these data will be provided by the Digital Collections of the ULB, with which the project has established a cooperation for this reason. Preparations for use of the Visual Library Software were discussed this year will all those involved, and detailed examination will begin in the course of 2015.

Site List

In direct conjunction with the construction of the inscription archive, Christian Prager and a research assistant began to compile an **Inventory** of all text and image carriers. The foundation for this work is **site list** that currently contains 515 entries with information about the site of discovery (coordinates, bibliography, etc.). For this purpose, all relevant publications, databases, and websites of collections, museums, and other research institutions have been and will continue to be systematically examined and searched for Maya text carriers in order to thereby obtain an overview of all published or documented inscriptions with their corresponding metadata (such as site of discovery, dimensions, or bibliographical references). This work is still underway and will be sufficiently far along by mid-2015 that it will include all published and documented text carriers. The list of sites and inscriptions constitutes the basis of a data pool that will be added to the virtual research environment, as well as be published on the website with the aid of an SQL database. Both lists will be continuously administered and made available to colleagues and other interested parties via the project website. The foundations for the database were prepared by Sven Gronemeyer and put into place with the help of an external programmer (see below). Work on the inventory and the database has not yet been concluded. The foundation of this documentary work was provided by development of a detailed metadata or description schema for the text carriers on the basis of previous documentation in our field of research, which was already outlined in the research proposal and constitutes the basis of the work carried out by metadata specialist Franziska Diehr (SUB). Further details and the current state of work are addressed in the section about the metadata schema.

Bibliography

The results of previous research must be represented in the virtual research environment and will be integrated with the aid of a bibliographic database. The goal is to create an object description and biography for every text carrier. Database users can thereby retrieve information regarding the relevant context of an inscription. The virtual research environment thus contains descriptions of the text carriers or information about the texts' contents. Furthermore, the user may access the bibliographic database to obtain an overview of which authors have studied or published a monument, discussed a text passage, or been the first to publically propose a linguistic reading for a hieroglyph or sign that is still valid today. The text carrier thus receives a biography that is intimately linked to the text content and must be taken into account when analyzing the meaning of words. Examination of the relevant reference material is thus a central point that was addressed from the beginning of the project. The goal is in principle to search on site through the entire bibliographic inventory that has been used for the construction of the inscription archive and the metadata mark-up and to refer to this inventory in the virtual research environment, as well as on the project website. The free and open-source application Zotero will be used to collect, manage, and cite various online and offline sources in compiling the bibliography. This application supports processing and editing of bibliographic

references and bibliographic lists and permits collaborative work from different locations. Using an API, bibliographic data can be directly inputted into the website, with the result that the bibliography can be searched, displayed, and downloaded via the project website. The goal is to construct a comprehensive bibliographic database concerning the Maya culture that can be inputted into the virtual research environment TextGrid and then linked with the datasets. At present, references from before 1960 are being incorporated on the basis of the bibliography for Mesoamerican archaeology, linguistics, and anthropology by Ignacio Bernal, which includes over 10,000 citations. Each entry is checked for completeness and accuracy with the aid of database and the original literature and – when available – linked with a URL through which an online version of the relevant monograph or article may be accessed. To date, 4,836 datasets have been compiled, and as soon as Bernal's bibliography has been completely incorporated, additional scientific bibliographies from websites and other sources will be added. In cooperation with the ULB, the project obtained a two-year subscription to the bibliographic database "Anthropology Plus", which aids in checking the datasets and in compiling and adding additional literature. The costs for this service are being divided between the project workplace and the ULB. We anticipate with an inventory of a total of 30,000-40,000 thematically tagged datasets that may be referenced later. The dataset can be searched in accordance with library standards and is available for open access download via the project website www.mayawoerterbuch.de. For this purpose, two external programmers are currently developing a Zotero API, which will be complete by year's end, in order that the literature database will be put online and will be searchable over the course of next year (see section about the website).

Grapheme Inventory and Concordance

For the project's epigraphic work, various work lists, discussion papers, and data are currently being compiled that will be added to the virtual research environment and be made directly recallable on the project website over the course of the project. One project involves compiling a concordance of all sign classifications that have been published since 1931. The goal is to use this work as the foundation for a complete sign catalogue of the Maya script that integrates pre-existing catalogues and, as an online version, can be revised and modified at any time. For this purpose, all individual signs were isolated from previously known catalogues as individual files and compiled in a table with their corresponding catalogue number. Creating this overview facilitates better comparison of the catalogues and, as a result, the development of a current sign catalogue that remedies the shortcomings of previous catalogues. The table also serves as the primary reference for linguistic readings, in addition to sign classification. The table records each sign's phonetic and/or logographic reading, as appropriate; the degree to which the relevant reading is considered secure; its derivation; and a bibliographic reference. The list is addressed and critically discussed in team meetings. The contents of the table are normed data and are being saved in a gazetteer. This method ensures that all project staff use the same sign reading when processing texts. The table, which records over 10,000 entries on almost 200 pages, is currently being processed and will be posted online as soon as the necessary database structure for WordPress has been programmed and installed. This should take place in the first quarter of 2015.

Linguistics

Standards and norms desperately need to be established in the field of Mayan linguistics, and they are also essential to being able to conduct joint research. An inventory of bound, grammatical morphemes with their function and meaning is being compiled at present. It summarizes the current state of research of all presently known grammatical elements of Classic Mayan, including citations, and

provides a foundation for further research in this field. The list will also be available online in the first quarter of 2015, whereby the morphemes may be publically discussed. Bound and lexical morphemes are classified and named when they are glossed, although unified glossing standards have yet to be established for Mayan linguistics (as mentioned in the research proposal). For this reason, the first workshop on glossing rules and standards in Mayan languages was held in the Department for Anthropology of the Americas at the University of Bonn from September 4-9, 2014. The annual workshop is a joint initiative of the research projects "Text Database and Dictionary of Classic Mayan" and "XML-Based Compilation Standards for Colonial Lexicographies of Amerindian Languages as Exemplified by K'iche'" (Prof. Frauke Sachse and Dr. Michael Dürr). The results of this workshop are currently being discussed internally and will be published on the project's website as a working paper "The Bonn Glossing Rules". Another work in progress is a concordance of all word lists that have hitherto been compiled for Classic Mayan (e.g. Boot, Mathews). The goal is not only to summarize the current state of research, but also to compare the various transcriptions of Classic Mayan and to integrate them into the virtual research environment.

Presentation of Research Data in the Digital Library

The ULB's software Visual Library, which was created by the software company Semantics, typically contains metadata from the union catalogue of the University Library Center of the State of North Rhine-Westphalia (hbz) by means of an OAI interface. The metadata will be linked to the digital representation within the Visual Library Manager and made publically accessible online in the ULB's Digital Collections. Both metadata and digital representations presented in the Digital Inscription Archive originate from the TextGrid database. Hence, in preparation, many possible routes were technically designed through which the data are to be represented and flow into the Visual Library, like the use of an OAI-PMH interface for transmitting the repository's public data, for instance. Use of the publication architecture SADE, which is able to transfer filtered data that flow into the VL as well as into the project's website, was also considered. Similarly, the project contemplated applying the SPARQL endpoint in TextGrid. In this case, only those data would be transmitted whose access level had been defined in the TextGrid-CRUD service as public. The design of the aforementioned approaches was prepared especially by Katja Diederichs, Jan Kenter (ULB Bonn), Max Brodhun (SUB) and Semantics representatives in various meetings with the responsible colleagues. In collaboration with the ULB and Semantics, presentation of the inscriptions will be implemented in the ULB's Digital Collections on the basis of the pre-existing TextGrid data. This implementation affects the technical installation of the content presentation, as well as the thematic arrangement and browsing structure. The first data will be represented in the ULB's Digital Collections at the end of 2015.

Online Presentation and Social Media

Online Presentation

A project website (www.mayawoerterbuch.de) is being set up for the purposes of presenting the project, its research questions, and its goals, and disseminating its research results. Although the website is primarily directed at an academic audience, the project also addresses the general public by also giving basic information about the Maya script and decipherment history, as well as particular research questions that are intimately related to the project's goals.

The website's construction and development are being conducted in three phrases. Phase I consists of conceptual designing and functional drafting on the basis of directly coded PHP pages with SQL queries. The latter interact with a MySQL database design that provides currently available research data and work lists. Phase I thus serves above all as a feasibility study. Work on this phrase commenced in late May 2014 under Sven Gronemeyer and were largely completed in October 2014. A basic version of this static draft went live in September 2014. Thereafter, Phase II was initiated with the porting of the draft to a CMS (Content Management System), which was scheduled as an intermediate-term milestone, but was preferred for many reasons. Porting the styles to CMS templates and content migration were easier, given that the website draft was still quite rudimentary at that point in time. Due to its simple front-end, its PHP-based template creation, and the potential for creating versions in multiple languages, WordPress was chosen as the CMS, a decision also inspired by the modern structure of its pages that encourages interactivity. A second component, namely the incorporation of an API for the project's bibliographic database on Zotero that will comprise some 30,000-40,000 entries, was just as important. A Zotero plugin is already available for WordPress, making it easier for the developers to develop a solution specifically adapted to the project's needs on the basis of this foundation. The composition and adaptation of the CMS and the programming of the Zotero API are being undertaken by the firm Beuse Project Management in Cologne. The Phase II website's go-live is scheduled for the end of 2014. Phase III of the website represents a future development that will only become relevant when a sufficient number of completed objects is available in the TextGrid Lab. Publication of RDF/XML data from the TextGrid Rep on any website is possible by means of TextGrid's own module SADE (Scalable Architecture for Digital Editions). In addition to making adjustments in SADE, an interface in the WordPress CMS must be created in order to be able to publish the RDF/XML on the website. Conceptual design of Phase III can likely be initiated towards the end of 2015.

For Phases I and II, the project's previously compiled work lists (which continue to be updated) are important as research data to be published. They comprise a site list and a list of museums with Maya artifacts (both of which will be enhanced in Phase II with a maps plugin), a concordance list of all ten sign catalogues for the Maya script, as well as a morpheme list for Classic Mayan. All work lists may be sorted, filtered, and searched according to various criteria. Furthermore, the project will continuously publish technical reports (concerning archive digitalization, for instance) and working papers (for example about glossing rules).

Social Media

One advantage of using the website is quick and extensive dissemination of research data and questions, when possible in interaction with the user, i.e. the scientific community. In this way, feedback can be given on all work lists in order to initiate discussions, although these exchanges will be conducted outside of observable channels of communication. In order to create a less former and faster format with a wide reach, the project will use channels in social networks, particularly Facebook and Twitter, in addition to its own website. It is anticipated that these will be developed in the coming year. In these channels, news can be posted, decipherments discussed, and events announced, in addition to many other possibilities for allowing experts and the general public to participate in the project's work essentially in real-time.

Tasks and Ongoing Work in Göttingen

Design and development of the project's computational infrastructure constitutes part of the duties of our project partner SUB in Göttingen. Development of the metadata schema and subsequent

programming and adjustment of the virtual research environment TextGrid to the project's tasks, needs, and goals occupy the highest priority. The following sections summarize the work completed up to this point in time and provide a glimpse of upcoming work in 2015.

Metadata: Demands and Design

The considerable demands that the project places on the virtual research environment also apply to the metadata schema. A domain with an extremely heterogeneous data volume must be described: the inscriptions, text and image carriers, actors, events, places, sources, among many others, must be represented in a logically structured form in a metadata schema, in order that the data are machine-readable and –processable. Two broad levels may be distinguished in designing the metadata schema: 1) textual information, including transliteration, transcription, translation, and text-markup, as well as components for compiling the dictionary; and 2) the non-textual objects that comprise an inscription's context: text and image carriers, actors (both those mentioned in the text and those relevant for research history), events, places, and sources relevant to research. The metadata schema will be constructed based on internationally recognized standards. One always strives for subsequent usage of standards in order to facilitate interoperability with other systems and to simultaneously improve the quality of one's own schema by re-using existing designs.

Two standards essentially serve as the basis for the metadata schema: CIDOC CRM, for representing non-textual objects, and XML-TEI / EpiDoc, for text mark-up. CIDOC CRM (CIDOC Conceptual Reference Model) is an ISO standard that was developed in the field of museums studies in order to be able to formally describe museum objects and processes. This reference model is thus optimally suited for describing epigraphic processes and objects, as well as for documenting research history. The comprehensive standard TEI (Text Encoding Initiative) has been successfully used for years for marking up texts, particularly in the humanities. EpiDoc represents a version of TEI specialized for describing inscriptions. Unlike TEI, it makes available elements that are particularly necessary for marking up texts, such as for indicating text passages that are missing due to erosion.

Since each of the two levels alone is very complex in and of itself, development of the schema will take place in two steps: in the first year (2014), the section of the schema for non-text objects was and continues to be developed, and in 2015, the section for text mark-up and analysis, as well as the dictionary components will be created. Before a schema can be designed and constructed, the demands being made of it must be posed and clarified. For these tasks, knowledge of schema modeling and domain-related expertise are needed. It is important for the modeling process to understand the domains that need to be represented. Exchange of knowledge between the Bonn and Göttingen is thereby essential. For this, the wiki is being employed in addition to telephone conferences and especially in-person meetings (see above). A catalogue of demands, which lists and clarifies the use of all elements that need to be represented in the domain comprises the base for this work. The elements are examined and revised step for step. This process is very time-consuming and labor-intensive in some respects, but critical for later use in the virtual research environment: problems with technical implementation that, at this point in time, could only be solved with difficulty, if at all, can only be addressed if all questions regarding the demands being made of the metadata schema have been finally settled.

At present, two-thirds of the catalogue of demands have been examined and implemented in the metadata schema. After the demands have been successfully clarified, a phrase of intensively researching additional standards (e.g. for describing visual representations) will be initiated. Thereafter, a machine-readable form of the schema and extensive documentation will be compiled.

The TEI Metadata Schema for Registering Textual Data

(by Petra Maier, project volunteer) The goal of the project conducted in the framework of the extra-occupational Master's degree in Library and Information Science (MALIS) was to create a metadata schema using the data format TEI for registering Maya texts. The TEI metadata schema provides a foundation that can be adjusted and expanded to address any demands that remain open as the project continues to develop. Firstly, the demands that were being made of the metadata schema concerning the texts were formulated, such as the exact order of the text fields (single/double columns, etc.) or colored regions. As a second step, relevant modules were selected from the among the very extensive TEI metadata sets that were under consideration for recording the texts. In addition, the project consulted the EpiDoc guidelines, which represent a selection of TEI guidelines specifically for describing inscriptions. In order to adequately implement the demands, the metadata schema was divided into sections that build upon each other: "Inscription" and three sub-sections, "TextDivision", "Block", and "Sign". This strategy for dividing the TEI metadata schema facilitates representation of the exact order of the glyph blocks. Colored text areas and lacunae can be described by the metadata. As a whole, this makes possible a (rough) reproduction of the text structure using the metadata. The problem of clearly identifying the signs as individual components of the glyph blocks has yet to be solved. According to the present mark-up method, the signs are written one after another, similar to a running text. Presently, there are various proposals and approaches under consideration for solving this problem, and these will be examined over the course of the project. Similarly, the TEI schema as it currently stands does not yet provide a possibility for marking up images belonging to the text, their relative dimensions, and information concerning their exact positioning on the text carrier.

Vocabularies and Normed Data

Controlled vocabularies are being used to support the mark-up of non-text as well as text objects. Controlled vocabularies are essential for orderly documentation: they permit representational control and continuity, and they prevent ambiguity and errors. Standard vocabularies are used whenever possible, for which there are two possible scenarios: 1. an existing vocabulary can be used in full, or 2. concepts may be integrated into vocabulary developed in-house via matching. For example, some concepts from the Getty Art and Architecture Thesaurus are being used to develop a thesaurus for object types and forms.

It will not be clear how many vocabularies will actually be needed until both components of the metadata schema have been developed. The vocabularies are being developed in parallel with the design of the metadata schema. SKOS (Simple Knowledge Organization System) is being used for machine-readable representation of the vocabularies. A tool is being used to develop the vocabularies and to simplify matching to standard vocabularies. Evaluation of the most appropriate tool has yet to be conducted and will take place at the beginning of next year. The project's normed data will be included in the domain of georeferencing. We are using the normed data of Getty TGN (Thesaurus of Geographic Names) and Geonames in order to be able to indicate places in a standardized manner and visualize them on a map in the application.

Informationstechnologie

TextGrid-Lab

The TextGrid Lab[1] constitutes the front-end of the technical infrastructure. All objects are created using this application in order to be able to work with them later. Various aspects that are relevant in this context will be described below.

User and Access Management

TextGrid offers an authorization and authentication service (TG-auth[2]) in order to facilitate collaborate work between all project members. Using this service, each project member can be assigned the most specific role possible. In this context, distinctions are made between the project manager, the editor, and the observer. It is also possible to give each user the ability to delete objects.

Object Administration

Server-side preparation and storage of the data set are necessary in order to be able to make the objects available to all project users at the same time. For this, the fields of compiling, reading, updating, and deleting are necessary mechanisms. To facilitate this, TextGrid offers the Services TG-crud[3]. For the project "Text Database and Dictionary of Classic Mayan", the following objects are relevant:

- descriptive Metadata Objects
- Text Analysis Files
- Image Objects, in order to mark up and reference them

The applications being used for these objects are described below.

Metadata Entry Mask

The metadata objects have a heterogeneous character and the metadata objects as a whole are proving to be complex. The relationships between the individual metadata objects are thereby of primary interest. The data are being generated in the file type Resource Description Framework[4] (RDF) in XML[5] Representation in order to be able to effectively store, and thereby also have the possibility efficiently retrieve, these relationships. With these file forms, logical statements concerning resources are stored in the linguistically familiar format Subject-Predicate-Object (for example, Konrad Zuse – engineered – the Z3). A special entry mask that makes an HTML form available was created to make entry of the metadata as comfortable and uncomplicated as possible for the user. From this form, the relevant data are subsequently converted into RDF/XML and stored in TextGrid. Storing the data in RDF format allows them to be linked to a data network (Linked Data)[6]. This aspect allows connections to be recognized that were previously not apparent. In addition, the data are largely interchangeable with data from other projects. This point is further strengthened by an open

[1] https://dev2.dariah.eu/wiki/display/TextGrid/Main+Page#MainPage-Frontend:TextGridLaboratory
[2] https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=8131296
[3] https://dev2.dariah.eu/wiki/display/TextGrid/TG-crud
[4] http://www.w3.org/RDF/
[5] http://www.w3.org/XML/
[6] http://www.w3.org/wiki/LinkedData

arrangement/configuration/composition of the data and the link to an open data network (Linked Open Data).

The entry mask is intended to be based as closely as possible on the epigraphic workflow "Documentation – Analysis – Editing". For this purpose, additional adjustments were enabled, such as the creation of a specific object at a specific point in time. Correct entry of the metadata is critical. In order to support this process, aids are made available to the person entering the data. The selection of particular terms from a menu, which are generated from a controlled vocabulary, prevents spelling errors, for instance. When referencing another object, the database is queried in order to transfer the relevant unique identifier from the database and thus avoid spelling errors at this stage as well. In addition, functions are used to validate certain contents. Thus, for instance, storage of an incomplete URL is prevented.

Database and Storage

In the first instance, all objects are stored on the TextGrid server. The data are redundantly archived in a special database for RDF data for efficient storage of the objects with their relationships and retrieval of the data. These databases are described as graph databases. The term triplestores[7]. is used for the RDF triples in particular. The project chose BigData[8] for the triplestore, given that it can store a large quantity of triples and offers the possibility of concurrency to protect against possible database breakdowns. The standard SPARQL Protocol and RDF Query Language[9] (SPARQL) is used when retrieving data. The combination of RDF – Tripelstore and SPARQL has proved to be efficient and effective in current practice when working with complex and heterogeneous data landscapes.

Storage

For the first project phase, 3 TB of memory were requested from the project DARIAH-DE[10]. This storage is located in the Göttingen computing center GWDG[11] and offers a cost-neutral solution.

Publication of the Objects

The compiled RDF metadata objects, as well as the complete texts that have been analyzed and marked up in TEI, are to be published on various platforms following their completing. Mechanisms for publishing in the TextGrid Repository are integrated into the TextGrid Lab and are optimized for the entire workflow. In the future, additional mechanisms will be made available for publishing. The application Scalable Architecture for Digital Editions[12] (SADE) offers the potential to publish all objects in external locations as well. This opportunity allows forgoing publication in the TextGrid Repository. Nonetheless, it also forgoes the mechanisms for long-term archiving. At this point, the objects are being published using SADE in order to be able to go ahead and present the data in an edited form that has also been adapted to the design of the project website (using a CSS[13]) , which has been adapted as needed). This publication is being conducted in parallel with publication in the TextGrid Repository.

Extracting the metadata objects using an OAI-PMH[14] interface represents a third possibility.

[7] http://www.w3.org/wiki/LargeTripleStores
[8] http://bigdata.com/
[9] http://www.w3.org/TR/rdf-sparql-query/
[10] https://de.dariah.eu/
[11] http://www.gwdg.de/
[12] http://www.bbaw.de/telota/software/sade/sade-1
[13] http://www.w3.org/Style/CSS/
[14] http://www.openarchives.org/pmh/

Indexing the Full Text

Subsequently, the marked-up full text should be searchable by means of a full text search. The TG search[15], which is already integrated into the TextGrid infrastructure, is being used for this function, and it, in turn, employs ElasticSearch for search queries. For this purpose, the data are initially converted into the **J**ava**S**cript **O**bject **N**otation Format[16] (JSON) during indexing.

Bonn and Göttingen, December 1, 2014

Dr. Christian Prager, on behalf of the authors

[15] http://www.openarchives.org/pmh/
[16] http://www.w3schools.com/json/