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

WORKING PAPER I | EN

OPEN ACCESS

Published 8 Apr 2015

DOI: 10.20376/IDIOM-23665556.15.wp001.en

The "Open Science" Strategy of the Project "Text Database and Dictionary of Classic Mayan"

Katja Diederichs[1] (translated by Mallory Matsumoto)

[1)] Rheinische Friedrich-Wilhelms-Universität, Bonn

The following article presents and explains the publication and knowledge transfer strategy of the research project Text Database and Dictionary of Classic Mayan. The project's goal is to make accessible in a database the epigraphic contents and object biographies of all extant hieroglyphic texts with the aid of digital technology. On the basis of resultant object and text database, a comprehensive dictionary of the Classic Mayan language will be compiled near the end of the project run-time. This research project, which is supported by public funds, sees as its duty to make project knowledge and research results freely available to the public. This openness is regarded as all-encompassing, i.e. as applying to the entire research process as it relates to methodology, contents, and results. The publication policy that this approach represents is understood as digital humanities that reflect the spirit of the age of Open Science.

The Project: "Text Database and Dictionary of Classic Mayan"

The project "Text Database and Dictionary of Classic Mayan" (abbr. TWKM) was established in the Faculty of Arts at the University of Bonn by the North Rhine-Westphalian Academy of Sciences, Humanities and Arts, with the goal of researching the writing and language of the Maya culture. The project aims to use digital technology to compile the epigraphic contents and object history of all known hieroglyphic texts over the next 15 years. Using these data, a comprehensive dictionary of the Classic Mayan language will be compiled and published near the end of the project's runtime (cf. Uni Bonn n.d.).

The dictionary will be compiled on the basis of the thousands of extant text carriers from the pre-Hispanic Maya culture, which existed between the third century B.C. and A.D. 1500 in a region including parts of the contemporary countries of Mexico, Guatemala, Belize, and Honduras. It will reproduce the vocabulary recorded by the hieroglyphic writing system and thus document, in both transliterated and transcribed form, the language, its developmental stages, and its varieties. For these purposes, it is necessary to construct a text database. The database will be available in two forms:

firstly, in an unpublished expert version requiring access authorization, and secondly, in a public form whose contents will be freely accessible worldwide via the Internet.

In addition to its goal of compiling a lexicon for Classic Mayan and analyzing the meaning of its lexemes, the project also aims to create an object description for each text carrier. Database users will thus also be able to retrieve information about the context of a given inscription. The text database will contain descriptions of the text carrier and the text's contents. An additional literature database will provide an overview of which authors have studied or published a monument, discussed a text passage, or been the first to present a linguistic reading of a hieroglyphic sign to the public. Each text carrier will thus receive its own biography that is intimately related to the text contents and will be taken into account when analyzing the meaning of the lexemes. The work environment, which is currently under construction within a Virtual Research Environment (abbr. VRE), is strongly oriented towards the workflow of epigraphic research, which typically begins with documentation of the text carriers and compilation of descriptive data; proceeds to epigraphic analysis (sign classification, transliteration, and transcription); continues with morphological segmentation and linguistic interpretation; and ideally concludes with the inscription's translation and publication.

In addition to conducting corpus analysis, which is fundamental to lexicography, the project will investigate Classic Mayan phonology, morphology, syntax, semantics, and pragmatics for the dictionary that is to be compiled. Questions from historical linguistics and investigation of the script's history and use will also be pursued over the course of this research. The digital dictionary will represent Classic Mayan not only in its transcription, but also in its original spellings. In this manner, the project will simultaneously address the history of both the language and the script (cf. Uni Bonn n.d.).

Given the large quantity of data, these aims can only be realized using the resources and methods of the digital humanities. This long-term project is thus methodologically situated in this field of work and will be conducted in cooperation with the research group for eHumanities TextGrid (from the Department of Research and Development at the Lower Saxon National and University Library of Göttingen) and with the University and Regional Library of Bonn. Due to its orientation towards the digital humanities, the dictionary project constitutes an important point of intersection between the humanities and computer science at the University of Bonn.

"Open Access" as an Avenue for Publication for Science in the Digital Age

The signs of the digital age are recognizable in societal change, as well as in a change in the working methods used in modern science. Today, both are strongly influenced by computer-assisted work and digital communication. For some time now, the sciences have increasingly dedicated themselves to efforts to meaningfully integrate the new digital tools into academic activities. Towards this end, many associations committed to the sciences have developed recommendations for guidelines, which for the time being remain self-imposed. These guidelines are intended to guarantee contemporary avenues for publication and knowledge transfer in the sciences. One such example is provided by the recommendations produced in 2002 by the "Budapest Open Access Initiative", in which the term "open access" was coined to refer to the free publication of scientific work (cf. BOAI 2002). Further recommendations followed, including ones produced in German-speaking countries, such as the 2003 Berlin Declaration, which strives for make scientific knowledge easily accessible (cf. Berliner Erklärung 2003). The Berlin Declaration also initiated a paradigm shift towards "open access" publication and considered open access a worthwhile practice that ideally presupposes the active participation of each individual creator of scientific knowledge and of each individual steward of cultural heritage. In this context, the Declaration extended the description of "open access" to include not only contents in the

form of publications of the results of scientific research, but also raw data, metadata, source materials, digital reproductions of images and graphic material, and scientific material in multimedia form. Open access to and the free availability of knowledge is intended to thus be guaranteed by the Internet, which the Declaration describes as "a comprehensive source of human knowledge and cultural heritage" (Berliner Erklärung 2003). The Declaration additionally addressed the ongoing problem of the structural challenges facing the scientific community of the future, in that it conceded that the vision of a comprehensive and freely accessible representation of knowledge may only be realized if the Internet of the future is sustainable, interactive, and transparent. With these actions, a process was set into motion that led to the creation of a more open scientific environment using information technology and the new tools of the internet, which above and beyond "Open Access" is intended to facilitate access to all scientific knowledge, i.e. "Open Content" (cf. Hilf & Severiens 2013). This process continues into the present, because the contemporary digital sciences also are developing within the advancing digital world. This development in the digital world will lead "to a new quality of scientific information management and thereby of scientific research" (Hilf & Severiens 2013).

Digital Humanities and eHumanities

While these standards do currently indicate a direction for further development, they have not yet been concretely formulated in detail, nor are they obligatory. Thus, in order to formulate these standards for their application to modern and open scholarship, certain technical and methodological measures must be taken, which vary in each scientific discipline and thus require input from various experts.

Within the humanities, new challenges have thereby been formulated for exploring ways in which the methodology and technical manifestation of digital science should be designed and put into practice in each respective discipline. As a result, the boundary between the traditional humanities and computer science and information technology is being breached more and more frequently. This process produced the so-called digital humanities, which, broadly defined, address methods and research questions from both information science and the humanities. On the one hand, with the aid of the digital humanities, research structures and methods from individual disciplines within the humanities are being expanded through digital means. On the other hand, the necessity of solving methodological problems with respect to humanities research questions has stimulated the development of computer science and information science, which in turn drives the implementation of technological solutions (cf. Thaller 2012). This interplay between the humanities and information technology continues to develop. Another field, the eHumanities—a term often used synonymously with the digital humanities—also occupies a very similar space within this interactive sphere. The eHumanities may be understood "as the sum of all approaches that aim to facilitate or improve work in the humanities through the investigation, development, and application of modern information technology" (BMBF 2013). The digital humanities, as well as the eHumanities, contribute to the methodological and structural migration of the humanities into the digital sciences.

Digital Epigraphy of Classic Mayan

The focus of the humanities field to which the TWKM Project belongs is directed towards intellectual property and cultural heritage. Discussions concerning the elements of the Classic Mayan script that have yet to be fully deciphered occupy the forefront of epigraphic analysis. The scientific discourse concerning the epigraphic contents of a monument with an un-deciphered text is reflected in various publications. It is thus inconceivable to create description of a hieroglyph for a dictionary entry, for

example, without linking the relevant lemma to the various hypotheses about its readings, which may have been confirmed or contested over the years. In digital epigraphy, all of this information will be compiled with the aid of computers and made machine-readable as data and metadata in digital form, in order that they may be appropriately linked with each other in a database.

A Database as the Basis for a Dictionary

For the TWKM project, a comprehensive bibliography of research literature from the last several decades will be compiled. Furthermore, the content of the bibliographic entries will be linked to the relevant Maya monument and text carriers in a database. In addition to its function in representing the general development of research on questions regarding the Classic Mayan language, an individual text carrier's object biography serves an important role in the epigraphic analysis of an inscription directed towards acquiring an understanding of its contents. As such, graphical contents and object-specific data, such as the context of archaeological excavation and historical-cultural information about political, social, dynastic, etc. relationships at the time, will be annotated as metadata, in addition to the contents of the text itself. As a result, a database for objects as well as for the linguistic corpus can be constructed. Together with the epigraphic analysis, the database also represents a type of ontology for the cultural history subject "The Classic Period of the Maya Culture", including scientific literature. Moving beyond purely epigraphic research questions, this strategy also allows semantic searches in the database, which, depending on the research question, can provide the appropriate scientific literature, in addition to a content-related response.

Over the course of the project's work, there will surely prove to be an additional benefit of using the text and object database, given that it, strictly speaking, is intended to make the source material available for accomplishing the project's goal, namely for the Dictionary of Classic Mayan. The development of this dictionary, which is to be compiled from the comprehensive corpus of inscriptions in the text database by the end of the project runtime, thus represents to a certain extent the epigraphic workflow in digital epigraphy.

Intended Audience of the Project

As a research project in the digital humanities that is financed by public funds, the project sees as one of its duties the task of making the information gathered during and the results of its research available not only to a scientific audience, but also to the general public.

In 2004, the Ministries of Science of various OECD nations compiled a list of requirements, the "Declaration on Access to Research Data from Public Funding" (cf. OECD 2004). On the basis of this publication, the OECD developed recommendations and principles governing access to research projects financed by public funds, known as the "Principles and Guidelines for Access to Research Data from Public Funding" (cf. OECD 2007). The project's goal of ensuring the greatest possible opportunity for accessing the data compiled and produced by the project is based on these recommendations. However, various national legal and copyright restrictions on data use that protect the rights of the original creator must be taken into account.

In the spirit of these and similar appeals and self-imposed agreements on the implementation of open scholarship, the project hopes that optimal access to the information resulting that it produces will support cooperation within specialized fields of research. Following the basic principle that the value of data lies in their future reuse (cf. OECD 2007: 11), it is intended that such cooperation will allow the project to collaboratively verify its data. In this manner, scientific discourse and progress, which among other things contributes to the creation of a comprehensive dictionary of Classic Mayan, can be promoted to the best possible extent. The resultant dictionary should be made available not only to

scientists, but also to interested members of the public, completely and financially, legally, and technically barrier-free. Thus, the project advocates the belief that its results can be useful even beyond fields of scientific research, because they bring the cultural heritage of the Maya culture to a receptive public, which additionally can increase in the public's eyes the perceived value of data collection projects that are financed with public funds (cf. OECD 2004).

"Open Science" Strategy

In the context of the view of modern science as "Open Science" (cf. OpenScienceASAP n.d.), the opening of the entire scientific process is critical, in addition to access to research results.

This approach means that the project aims to give the interested public and scholars all-encompassing, open access to its scientific publications ("Open Access"), to the free documentation of its methodology and work processes ("Open Methodology"), to its research data ("Open Data"), and to the software used by the project ("Open Source"). In this context, it is anticipated that the transparent and prompt documentation of interim results, as well as making available an infrastructure that may be useful to the general public, will foster collaborative participation in the corresponding scientific discourse and also stimulate intra- and inter-disciplinary exchange with respect to methodology, technology, and contents, which could, in turn, shape the development of the project (cf. OpenScienceASAP n.d.). Uninhibited access to the research contents of the project and the guarantee of their productive reuse must be accomplished through free licenses (cf. DFG 2014). In this sense, the strategy of "Open Science" thus concurrently supports a range of established free access authorizations that cover the entire scientific process. These access authorizations are indicated by the keyword "Open" and may be described as follows:

- "Open Access" describes open publication, which makes the relevant contents accessible and useable for all.
- "Open Methodology" concerns transparent documentation and publication of the application of methods and tools for data compilation and analysis, as well as of the entire work process that it entails.
- "Open Source" applies to the use, presentation, and documentation of one's own, open-source technologies and to the creation of the opportunity for their barrier-free reuse in the future.
- "Open Data" includes the publication of compiled raw and metadata in standardized, open format, as well as the creation of cross-data structures, which allow access to and the free and unrestricted reuse of all of the data (cf. Kraker et al. 2011: 645).

The Various Platforms of the Project

Within the project's work process, the individual "open" strategies converge upon or condition one another. Nonetheless, the individual principles and the project strategies will be discussed in more detail with respect to these strategies. The following section introduces the project's various platforms on which its research results, data, and metadata will be published and made available over the course of the project.

TextGrid Laboratory and Repository

The project's partner, the Department of Research and Development at the Lower Saxon National and University Library of Göttingen, in cooperation with the Society for Scientific Data Processing Göttingen (abbr. GWDG), will ensure the storage of the textual and graphic corpus data, with particular attention to their sustainable and long-term use. The Virtual Research Environment for the humanities that has been developed there since 2006, known as TextGrid, allows materials to be archived. These materials are citable by means of so-called Persistent Identifiers (abbr. PIDs) within the searchable online repository, the TextGrid Repository, in which they are stored. In this long-term archive, the data will be made accessible worldwide at no cost under the "Open Access" license terms (cf. DARIAH.de n.d.).

The TextGrid Laboratory (abbr. TextGrid Lab) will be initially used for creating data and digital representations for the project database. Here, all project data that is to be compiled, such as digital representations and metadata, namely object data and linguistic analyses, will be entered and compiled by the project's own front end. The front end is located within the VRE, in the TextGrid Lab, in which authorized project collaborators can compile and process the data and metadata using an input mask developed specifically for the project. The Laboratory houses a dynamic store, meaning that, for the time being, the data may be altered and edited, and that they may be logically linked to other data and aggregated. Access to TextGrid is granted after registration, for which there is no charge. However, in order to gain full access to the project's data released for work and to edit them as the case may be, the project manager's authorization for project participation is required, in addition to the registration.

After the data in the TextGrid Lab have been checked for consistency and formal validity, they may be eventually published in an additional stage. For these purposes, they will be migrated to the public TextGrid Repository, which contains a static store. As a result, the contents stored in them can no longer be readily changed or deleted. The Repository is freely accessible via the World Wide Web, and the contents transferred from the Laboratory can be searched and the data and metadata downloaded there.

Virtual Inscription Archive

A further forum for presenting select project data will be offered in cooperation with the University and Regional Library of Bonn (abbr. ULB). In the Digital Collections of the ULB- Web presence, a Virtual Inscription Archive will be established. This archive will present exclusively "open access" contents that will be freely accessible via the Web, for which purposes selected data and metadata concerning the text carriers will be transferred. The contents to be transferred must thus be "mapped" onto another metadata format; in other words, the contents of the one format must retain their semantic value when matched to the other format, so that they may later be appropriately represented in the Digital Collections. The format used consists of the open XML-based standards METS and MODS, which were established for digital libraries and whose schema definitions additionally make the contents compatible with the DFG Viewer (cf. DFG-Viewer 2015). The data that is to be matched will thus be transferred from the TextGrid memory using an interface of the ULB software, which represents the contents in their structuration in the Virtual Inscription Archive of the Digital Collections that was previously defined by the project. The interface based on the OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting) will facilitate additional services and functions through free access to project metadata in the Repository and the reuse of these metadata in another context, which moreover gives the metadata additional value.

The Project's Website

Further project information, and in particular publications that are released over the course of the project, will be made available in digital format at no cost on the project's website, which has its own ISSN (International Standard Serial Number). Here, essays, working papers, and additional content from grey literature, such as technical reports and documentation of project work, will be published. "Open Access" initial publications will be published according to the standards of gold open access, and secondary publications according to the standards of green open access (cf. Internet Portal OA n.d.). Thus, the website fulfills the following criteria for releasing scientific publications consecutively under the given ISSN. These strategies comply with the collection guidelines of the German National Library for allocating an ISSN (cf. DNB 2011); furthermore, the contents that are to published will be treated as regularly appearing blog entries with a scientific orientation.

Due to its reception of an ISSN, the project's website will be established as a blog in the catalog of the German National Library and in the international ISSN Portal. This situation guarantees that blog contents will be citable as a network publication with an open target audience (cf. DNB 2011).

"Open Access" Strategy

Currently, the term "open access" describes an internationally-known paradigm of free access and publication of contents in various forms and from various origins. The project considers it important not only to make research results and similar literature available through "open access" publications, but also to release the data and contents that are the subject of the research. These materials include digital data concerning purely textual content and museum and archaeological artifacts. The technical media that are made possible by digitalizing these contents are subject to various regulations, as are the rights that are associated with them.

The Project's Digital Representations

The standards of the 2013 DFG Practical Guidelines on Digitalization (cf. DFG 2013) are used for digitalizing the objects. This publication includes guidelines on digitalizing image-based, museum objects, among others recommendations. The DFG, whose representatives are part of the TextGrid project's consulting committee (cf. TextGrid 2012), aims to define the state of technology with respect to digitalization in the sciences in German-speaking countries, and also anticipates composing addenda to the Practical Guidelines in the near future. In doing so, the DFG, as a signer of the Berlin Declaration, represents an "open access" and "open source" policy that conforms to modern digital scientific work.

In accordance with the DFG Practical Guidelines, it is important to the project that the digitalization of scientifically relevant material be made accessible to research and the interested public worldwide according to the requirements of "open access": "Rights relating to the material must be cleared when project planning begins or at the latest when a proposal is submitted. In particular, any copyright, privacy rights or ancillary rights must be taken into account" (DFG 2013: 6).

Objects digitalized by the project that are subject to licensing restrictions will be limited to the part of the database that is not publically accessible according to the embargo periods. Only contents that have been confirmed as being under free license can be made publically available without further inquiry or registration.

The Project's Publications

All of the image and text materials that are digitalized within the context of and belong to the project, as well as scientific publications, will be published online under the internationally valid Creative

Commons copyright CC BY 4.0 "Open Access". This license level is a so-called "free culture" license because of its openness, whereby the license conforms to the demands of the Berlin Declaration and the "Budapest Open Access Initiative" (cf. Creative Commons n.d.a). CC BY 4.0 indicates that the contents may be readily read, downloaded, expanded, shortened, changed, and further distributed. They may be used for commercial or for non-commercial purposes, without financial, legal, or technical barriers. Nonetheless, care must be taken to name the creator and to reference the original work with a link (cf. Creative Commons n.d.b.). Creative Commons Licenses thus build upon the valid copyright. Contents protected by copyright, such as texts, images, etc., may be reused in the public domain by means of a CC license. To the extent that it is legally possible, the so-called Golden Way of "open access" publication will be used as a license model for publications.

This means that initial publications will simultaneously be made freely available, independent from their publication elsewhere, on the project's own publication platforms via the Internet. The publications will be subject to a quality insurance process in the form of peer reviews or editorial reviews.

For research materials currently published elsewhere that are relevant in the context of the project, as long as the appropriate conditions have been fulfilled, the indispensable/inalienable secondary publication rights will be utilized, "for which the DFG in conjunction with its partners from the Alliance of Science Organizations in Germany (Allianz der Wissenschaftsorganisationen) has lobbied for years, and which came into effect on January 1, 2014 upon the amendment of the German Copyright Act" (Fournier 2015: 6). Thus, contents can be made freely available online as author manuscripts 12 months following their initial publication (cf. Fournier 2015: 6).

Furthermore, for those cases that are more difficult to settle, the project has the option of using additional, different CC license levels, which, for example, increase control over the integrity of the contents, citation requirements, non-commercial use, etc. As a result, these licenses can be more narrowly defined within the CC's scope than the CC BY licenses are, so that dissemination of the contents can nonetheless be made as free as possible under various conditions. In order to generally enhance and technically facilitate content distribution, the CC license model has a structure consisting of three license layers. In addition to a judicially clearly defined and usable definition of the various licenses, each license also has a so-called human readable license layer that generally describes the relevant rights so that they may be easily understood by all interested parties. Furthermore, the third layer is technically structured so that it is machine readable, i.e. can be found and processed by applications in the web (cf. Creative Commons n.d.b.).

Digital Representations and Publications Prepared Outside of the Project

The project attempts to make available to the public in the TextGrid Repository all digitalized images and drawings of Maya monuments, together with their annotated metadata, from various publications whose rights for worldwide publication the project has received or purchased, with appropriate references to the original sources and creatorship. Digital representations and publications with restricted rights will be made available to a select group of researchers for academic purposes within the VRE, in the TextGrid Laboratory.

"Open Methodology" Strategy

The methodology and work process that leads to development of the data structure will be made transparent through open documentation of the project's digital epigraphic workflow. Interested parties may then use this information as a stimulus for developing their own solutions to thematically similar issues.

In this context, TextGrid presents the appropriate platform that permits the entire scientific work process to occur within a single environment, from primary data collection and generation up through archiving and publication of materials (cf. TextGrid n.d.). The VRE TextGrid provides the project with all the digital tools necessary for this process in a single framework. As such, compilation is methodologically linked to object preservation and knowledge transfer. Consequentially, digitalizing images of museum objects, for instance, will not be viewed here as separate from their later usage, which thus should already be taken into account when modeling the data.

The TextGrid lab is modularly constructed and can be expanded using various external "Open Source" "applications, so that all work processes can be conducted in a single environment, if one so desires (cf. TextGrid n.d.). In this context, the various steps of data acquisition within the same workflow converge. The consequences of this fact for the software application is that entering the digital representations, developing the object metadata, linking texts with images, tagging with epigraphic or linguistic metadata as appropriate, and the creation of further metadata is accomplished within the same virtual work environment. Furthermore, presentation of creation of access to the contents will be similarly accomplished within the VRE. Documentation of work processes will be offered online in digital form by means of Usecases and similar, written proceedings on the project's own platforms.

In order to add the data to the clearly defined metadata contents, metadata schemas based on XML-based languages will be designed and constructed, one for the objects and one for linguistic analysis of the data. Documentation of the conceptualization and the version controlled source code of the completed metadata- and analysis metadata schemas will be presented, in that they will be made freely available through openly accessible, web-based hosting services for software development projects, such as "GitHub", for instance.

"Open Source" Strategy

The technical resources for using the project data and metadata in the form of scientific software consist of freely available "open source" applications. As previously described, the software tools utilized in the project for producing data originate in the framework of the "open source" software TextGrid (cf. Gietz et al. 2011), which is thus not subject to any restrictions on use. The source code and versioning of TextGrid are thus publically accessible, and the tools and services have been documented and may be changed or expanded (cf. TextGrid n.d.). Various "open source" licenses, such as GNU, will be established for individual software derivatives under the complete package of the Virtual Research Environment, TextGrid (cf. Gietz et al. 2011). When modelling the metadata schema and saving the metadata according to its guidelines, the project will use only open source data and metadata standards, as well as programming languages whose source codes and documentation have been made public.

Thanks to its service-oriented and modular architecture and to its given interfaces based on "open source" languages and open standards, the TextGrid infrastructure facilitates interoperability between TextGrid and external applications and databases, as well as their contents.

"Open Data" Strategy

In order to facilitate linkage and exchange between the project's assembled data and other data, these data from various sources must be openly available and also comparable to each other. The data can only be sustainably used and linked, independently of application, and incorporated an open data

structure such as the so-called “Linked Open Data” (abbr. “LOD”) or put to use in the larger context of a semantic web, if they can be compared to other data.

Interoperability as a requirement for Linked Open Data

From a technical standpoint, comparability means interoperability. In this context, interoperability between the metadata must be technically applied at various levels: structural, syntactic, and semantic.

- Structural interoperability refers to the fact that a common model, such as the RDF model or the OAI-PMH protocol‘s interface definition, underlies the various metadata formats.
- Syntactic interoperability is guaranteed by means of a common syntax such as that of the mark-up standard XML, which will be represented in the data and are already available.
- Semantic interoperability, on the other hand, requires using the same metadata definitions in such a manner that goes beyond the syntax that provides the form. These definitions will be established in recognized metadata standards, such as TEI, which is based on XML, or using controlled vocabularies for organizing knowledge, such as SKOS (Simple Knowledge Organization System).

Thus, interoperability should define under a common denominator, both technically and with respect to content, how the data can be made interchangeable, while still retaining their semantics and without significant loss of information (cf. Rühle 2012: 2 ff.).

Interoperability as implemented in the project

In order to render the project’s contents interchangeable, the three facets of interoperability described above will be realized by using the TextGrid infrastructure and by appropriately modeling the metadata. The aforementioned open TextGrid infrastructure permits a high degree of structural interoperability. In other words, other applications will be able to receive and reuse data that is stored there, for example. This process occurs by means of interfaces, which establish exchanges between various applications or databases using open standards like REST (Representational State Transfer) or OAI-PMH.

Furthermore, the TextGrid data architecture will make the project’s data technically searchable for the long-term, in that each receive a type of permanent digital address, or, technically speaking, a PID in the form of a URI (Uniform Resource Identifier).

The findability of explicitly free and open contents is just as necessary for their interoperability as it is for their processing and dissemination. The appropriate rights of use will be attached to the project contents in a machine-readable format using the aforementioned third CC license layer. These copyrights are expressed in the standardized CC Rights Expression Language (abbr. CC REL). Thus, the contents and their associated terms of use can be found, recognized, and processed by search engines and software.

The project metadata will be correspondingly modeled in the form of machine-readable RDF Triples. The RDF (Resource Description Framework) format is described as a tripartite data model for semantic web searches. It constitutes a generally recognized standard supported by the web consortium W3C for coding and presenting information in such a manner that they are suitable for the web and for establishing structural interoperability with other contents. The model can be established on a syntactic level using various formal languages, such as the XML-based metadata standard used and supported by TextGrid. The syntactic and semantic interoperability of the data and metadata will be produced accordingly via their technical and content-oriented modeling, for instance when building the metadata schema and the definition of the metadata that it contains. The contents that are defined

in an ontology-like structure should thereby be based not only on established metadata standards. Furthermore, they should be semantically defined in such a way that they are generally comprehensible by means of controlled vocabularies, for example, in order that they may be meaningfully compared and linked to data of other origins. Thus, semantic interoperability permits reference to the same concepts and "things", which is an important point of "Linked Open Data" modeling (cf. Beer et al. 2014).

The XML markup language-based metadata format TEI that the project will use for marking analysis of epigraphic and other contents is supported by TextGrid. In this manner, TEI has established itself as the international standard for text-based digital compilation of humanities data contents (cf. TextGrid n.d.). In addition to TEI, other metadata standards such as EpiDoc that also constitute recognized metadata standards for digital epigraphy will be similarly helpful when designing and modeling the analysis metadata schema. The analysis metadata schema thus guarantees that the metadata to be entered will be stored consistently with respect to their contents, and that they will be intelligible and thus can be meaningfully used.

The object metadata that describe the text carriers in their broader context are also linked with each other in the ontology-like database structure. This structuring is based on common design standards from the realm of cultural heritage, which are created based on the CIDOC Conceptional Reference Model (abbr. CIDOC-CRM) and utilizing widely-used, controlled vocabularies such as SKOS.

For example, the project is preparing and defining Getty Vocabularies in order to expand them. The vocabularies contributed by the project to TGN (Getty Thesaurus of Geographic Names), AAT (Art & Architecture Thesaurus), CONA (Cultural Objects Name Authority), and ULAN (Union List of Artist Names) thus guarantee that the controlled vocabularies will be secure and established in their use beyond the project framework, in fields like Maya art and architecture and in relation to Maya personalities and archaeological sites. Concerted application of recognized standards and vocabularies increases the possiblity of a secure and meaningful description with respect to content, as well as of exchange with other, thematically similar contents.

Finally, the use of such established models and terms for the project guarantees a meaningful and universally valid description of the various metadata on a semantic level. In addition to the XML-based metadata definition, their contents will be defined in RDF triples based on the language Turtle and stored in so-called triplestores. This approach guarantees construction of a further syntactic level of interoperability. From a technical perspective, the database is constructed in the form of a directed graph whose edges link the nodes and leafs, each of which consists of one RDF triple, i.e. of metadata, per resource. The permanent and explicit digital identifications assigned to these resources in the form of PIDs are used to link the resources to each other. The project's use of the triples in Turtle format hereby represents a step undertaken according to the recommendations proposed by the web consortium W3C.

The contents of the directed RDF graph are thereby coded in abbreviated text form. Turtle is compatible with the query language SPARQL, which is also recommended by W3C. The queries formulated in SPARQL facilitate semantic searches within the project's data network. Each node and each leaf of the graph, which is composed of the project's own data, can be triggered and searched using a SPARQL query. The data, which are linked in graph form, openly accessible via the web and searchable using SPARQL queries, thus ultimately constitute the contents described as "open data" or, more explicitly put, "Linked Open Data".

At present, "LOD" is the method established in Open Science for making data freely researchable, available, and linkable to each other. A "Linked Open Data" structure effectively extends the "Open Access" philosophy from publishing research results to publishing the data on which the results are

based. As a result, all people or groups interested in the data have the potential, for example, to test their own hypothesis based on the data or use the data in another manner other than that of the original project (cf. Kraker et al. 2011: 646). For instance, a variety of data visualizations or inference-based corpus linguistic methods, among other efforts, could be undertaken based on the data. The project's "Open Access" strategy thus gives the data a potential value beyond the project itself that can be exploited by anyone at any time.

Concluding Remarks

The TWKM project is currently in its initial phrase, meaning that it is beginning to develop the methods and tools mentioned in this working paper that will shape the project's work process and the publication of research results. These methods and tools are intended to facilitate use of a Virtual Research Environment and compilation of an open database and data infrastructure with the goal of researching the Classic Mayan language. Over the course of the project, there will certainly be some changes to the project's road map. However, the goal of realizing a project that is in every respect open and transparent, and thereby makes a contribution to "Open Science", will remain a recurrent theme throughout the project's work. Thus, the project may confidently work towards the goals that it has undertaken to achieve by the end of the project runtime.

Quellen und Literatur

* Note: All internet resources cited have been accessed on April, 8. 2015. All links are available in the online version of this paper.

Literature

Berliner Erklärung
2003 *Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen*.
Beer, Nikolaos, Kristin Herold, Wibke Kolbmann, Thomas Kollatz, Matteo Romanello, Sebastian Rose, Niels-Oliver Walkowski
2014 Interdisciplinary Interoperability. *DARIAH-DE Working Papers*, 3.
Budapest Open Access Initiative (BOAI)
2002 *Read the Budapest Open Access Initiative*.
Bundesministerium für Bildung und Forschung (BMBF)
2013 *Bekanntmachung des Bundesministeriums für Bildung und Forschung von Richtlinien zur Förderung von Forschungs- und Entwicklungsvorhaben aus dem Bereich der eHumanities*.
Creative Commons
n.d.a *Understanding Free Cultural Works*.
n.d.b *Mehr über die Lizenzen*.
DARIAH.de Digital Research Infrastucture for the Arts and Humanities
n.d. *TextGrid*.
Deutsche Forschungsgemeinschaft (DFG)
2013 *Praxisregeln "Digitalisierung" (DFG-Vordruck)*.
2014 Appell zur Nutzung offener Lizenzen in der Wissenschaft. *Information für die Wissenschaft*, 68
Deutsche Forschungsgemeinschaft (DFG) – DFG-Viewer
2015 *Profil der Metadaten*.
Fournier, Johannes
2015 Open Access und Open Data. Positionen und Perspektiven der Deutschen Forschungsgemeinschaft (DFG). *Archäologische Informationen*, 38.

Gietz, Peter, Markus Widmer, Stefan Funk, Andreas Witt, Oliver Schonefeld, Norman Fiedler
2011 Musterverträge (AP 3.2) und technische Umsetzung. *TextGrid Report* 3.2.1.
Hilf, Eberhard R. & Thomas Severiens
2013 *Vom Open Access für Dokumente und Daten zu Open Content in der Wissenschaft*.
Internet Portal Open Access
n.d. Der freie Zugang zu wissenschaftlicher Information – Open Access Strategien
Kraker, Peter, Derick Leony, Wolfgang Reinhardt & Günter Beham
2011 The Case for an Open Science in Technology Enhanced Learning. *International Journal of Technology Enhanced Learning*, 6 (3): 643–654.
Matthias, Angela & Brigitte Wiechmann
2014 *Sammelrichtlinien*. Deutsche Nationalbibliothek (DNB).
Niedersächsiche Staats- und Universitätsbibliothek Göttingen (SUB)
n.d. *IDIOM*.
OpenScienceASAP
n.d. *Was ist Open Science*?
Organisation for Economic Co-operation and Development (OECD)
2004 *OECD – Declaration on Access to Research Data from Public Funding*.
2007 *OECD Principles and Guidelines for Access to Research Data from Public Funding*.
Rheinische Friedrich-Wilhelms-Universität Bonn, Abteilung für Altamerikanistik (Uni Bonn)
n.d. *Textdatenbank und Wörterbuch des Klassischen Maya*.
Rühle, Stefanie
2012 *Richtlinie für die interoperable Gestaltung von Metadatenprofilen*. Kompetenzzentrum Interoperable Metadaten (KIM), SUB Göttingen.
Schütz, Christian
2015 *Nationales ISSN-Zentrum für Deutschland – häufig gestellte Fragen (FAQ)*. Deutsche Nationalbibliothek (DNB).
TextGrid Konsortium
2012 *Abschlussbericht: Vernetzte Forschungsumgebung in den eHumanities*.
2014 *Das Projekt. TextGrid – Virtuelle Forschungsumgebung für die Geisteswissenschaften*.
Thaller, Manfred
2012 Controversies around the Digital Humanities: An Agenda. *Historical Social Research*, 37 (3): 7-23.

Internet resources and sources

Comité international pour la documentation Conceptual Reference Model (CIDOC CRM)
2014 *What is the Cidoc CRM*. Cidoc CRM Home Page.
Creative Commons
n.d. *Attribution 4.0 International (CC BY 4.0)*.
Creative Commons Wiki
2013 *CC Rel*.
Epidoc: Epigraphic Documents in TEI XML (EpiDoc)
n.d. *EpiDoc: Guidelines*.
Getty Research Institute
n.d. *Getty Vocabularies*.
GitHub, Incorporated
n.d. *GitHub Website*.
GNU General Public Licence
2014 *GNU General Public Licence*.
Hakala, Juha
2010 Persistent identifiers – an overview. *Technology Watch Report (TWR): Standards in Metadata and Interoperability*.
International Standard Serial Number – ISSN International Centre
n.d.a *International Identifier for Serials*.
n.d.b *The ISSN for Electronic Media*.

Library of Congress
2015a *Metadata Encoding and Transmission Standard – METS*.
2015b *Metadata Object Description Schema – MODS*.
Open Archives
n.d. *Open Archives Initiative Protocol for Metadata Harvesting*.
TextGrid Konsortium
n.d. *TextGrid Website*.
Universitäts- und Landesbibliothek Bonn (ULB)
n.d. *Digitale Sammlungen*.
World Wide Web Consortium (W3C)
2006 *W3C. Naming and Addressing: URIs, URLs, …*.
2007 *W3C. SPARQL Query Language for RDF*.
2011a *W3C. REST*.
2011b *W3C. TURTLE – Terse RDF-Triple Language*.
2012 *W3C. SKOS Simple Knowledge Organization System – Home Page*.
2014 *W3C. Resource Description Framework (RDF)*.
2015a *W3C. Website*.
2015b *W3C. Linked Data*.