

The Creation of a TEI Metadata Schema for Cataloging Classic Mayan Texts

Petra Maier¹ (translated by Mallory Matsumoto)

¹ ULB Heinrich-Heine-Universität, Düsseldorf

The present paper was first published as DARIAH-DE Working Paper 8 under CC BY 4.0*. The present version was translated from German, with some of the original figures replaced.

Preliminary Remark: The present report is based on a project that is being conducted as part of the extra-occupational Master's degree program in Library and Information Science (MALIS) at the University of Applied Science in Cologne.

Background

Early 2014 saw the initiation of the project "Textdatenbank und Wörterbuch des klassischen Maya" (TWKM, *Interdisciplinary Dictionary of Classic Mayan*) under the direction of Prof. Dr. Nikolai Grube (Department of Anthropology of the Americas, Faculty of Humanities, University of Bonn), with funding from the North Rhine-Westphalian Academy of Sciences, Humanities and Arts. The project, which is being conducted in cooperation with the TextGrid research group (under the direction of the Göttingen State and University Library) and the Bonn University Library, has a projected runtime of 15 years. The overarching project structure is divided into five stages of three years each. The ultimate goal of the project is to catalog all known Mayan hieroglyphic texts in a digital corpus that will serve as the foundation for future epigraphic and linguistic analysis. Over the course of the TWKM project, a dictionary – in both digital and printed format – will be compiled that will contain all known vocabulary words and also reflect their use in the written language (see Grube 2011: 13).

A partial goal of the first stage of the TWKM project is the creation of a working version of the dictionary in electronic format. One necessary component of this sub-project was the conception of a data model in an electronic research environment. The research project requires such a complex metadata design due to its comprehensiveness, as it aims to catalog all known inscribed objects and their texts, as well as to continue researching signs that have not yet been undeciphered or are

* Petra Maier: *Die Erstellung eines TEI-Metadatenschemas für die Auszeichnung von Texten des Klassischen Maya*. DARIAH-DE Working Papers Nr. 8. Göttingen: DARIAH-DE, 2015. URN: urn:nbn:de:gbv:7-dariah-2015-1-6.

polyvalent. The project had already expressed its intention to catalog the hieroglyphic texts using the standards of the TEI (Text Encoding Initiative) Consortium in its initial proposal (see Grube 2011: 13).

Brief Overview of Classic Mayan

In order that the reader may understand the project's documentation and become acquainted with the topic of research, the following section briefly outlines the Classic Mayan language and its spatiotemporal context.

From a geographic perspective, the region of the Maya extends across an area that includes parts of what are now the Mexican states of Chiapas, Tabasco, Campeche, Quintana Roo and Yucatan, as well as the nations of Belize, Guatemala, and western portions of Honduras and El Salvador (Figure 1) (see Grube & Gaida 2006: 23).



Figure 1. Geographic Location of the Maya Region, drafted by Sven Gronemeyer after Grube & Gaida (2006: 23) with height relief by Shuttle Radar Topography Mission (SRTM), PIA03364, courtesy NASA/JPL-Caltech.

The pre-Columbian Maya used the writing system to represent rulers and their families: events such as birth and accession to the throne are frequently described in inscriptions. These events are usually associated with calendrical dates, which permit the inscriptions and the events that they record to be dated to the very day. These dates can be converted to the Gregorian calendar using a correlation that has been widely established within Maya studies (see Grube and Gaida 2006: 22-24).

The Maya writing system is a hieroglyphic script that is first attested in the third century B.C. The writing system spread throughout the Maya region beginning in the Classic Period (A.D. 250-900) (Grube 1993: 222-225). Over the course of the script's history, it continued changing and adapting to the needs of its writers and commissioners. New signs were invented, old signs fell out of use, and the readings of other signs changed over time (Grube 1993: 225ff.).

Following the conquest of the Maya region by the Spaniards beginning in the early sixteenth century, the hieroglyphic script fell into disuse and knowledge of the writing system was lost (see Grube 1993: 215ff.).

The Maya script is a so-called logosyllabic writing system, meaning that it consists of two types of signs: logograms and syllabograms (see Gronemeyer 1999: Chapter 2.1). In most cases, a hieroglyphic block corresponds to a word and consists, on average, of three to four signs, usually a combination of logograms and syllabograms. In contemporary Maya studies, 650 distinct signs have been identified. Syllables that are frequently used have multiple variant signs, which allowed the scribe to avoid sign repetition. Most hieroglyphic texts can now be read and interpreted, although not all hieroglyphs in the writing system have been deciphered. Some sign collocations can be read phonetically, but their meaning has not (yet) been identified (see Grube 2011: 6, 11). The Classic Mayan language is thought to be related to the contemporary Ch'ol languages of the Mayan language family, which are primarily spoken in the Maya area of what is now Mexico, and to the Yucatekan languages, spoken on the Yucatan peninsula (see Grube 1993: 222). Correspondences between Classic Mayan and contemporary Mayan languages thus contribute to decipherment efforts.

Maya hieroglyphic texts and iconography have been preserved on various classes of objects. Due to the warm, humid environment of the Maya region, many of the objects which have survived are those constructed of imperishable materials, such as stone and ceramics. Such text carriers include free-standing monuments, architectural elements (e.g. lintels, hieroglyphic stairways), jewelry, ceramics, and small sculptures. Additional texts have been found in caves, either as painted murals or rock carvings (e.g. the caves of Naj Tunich). Bark-paper codices are much more rarely preserved, with only three being known today.

Current State of Research

Research into the Classic Mayan language and script has traditionally lacked comprehensive documentation. The individual vocabularies that have been published are restricted in scope to the investigation of specific research questions, or incorporate only select hieroglyphs. Since the end of the 1990's, several lexicographic catalogs have been produced that contain commentary above and beyond a simple, alphabetic list, but documentation of the spatial distribution and of changes to the script over time is still lacking. Thus, existing hieroglyphic vocabularies do not permit investigation of current research questions concerning topics such as the development of the hieroglyphic writing system.

In Maya studies, these research deficits arise from incomplete documentation and a lack of digital editions of source materials to date. In other areas of language studies, there are existing projects that

provide researchers with access to comprehensive inscription corpora in digital format; for instance, the digital corpus *Thesaurus Linguae Aegyptiae* (TLA)¹ permits searching through ancient Egyptian textual materials, and thus facilitates investigation of relevant research questions by using specific analytical queries (e.g. regarding word frequencies). The corpus also contains a translation of each text. The project *Pennsylvania Sumerian Dictionary* (PSD)² of the University of Pennsylvania represents another such undertaking, which has produced a comprehensive Sumerian dictionary. A unique aspect of the latter project is that the tools developed for compiling the corpus and working with the Sumerian language have been made freely available for use. As such, they may be utilized by subsequent projects.

The TEI Format

As per the specifications of the project's original proposal, the Maya hieroglyphic texts are being cataloged using a TEI metadata schema. In this context, metadata can be generally defined as structured information concerning the Maya texts as a whole, as well as the mark-up of special features in the texts. The metadata schema thus also includes local annotations of the texts themselves.

Text Encoding Initiative (TEI) is an international organization that was founded in 1987 in order to develop guidelines for coding machine-readable texts, particularly for the social sciences and humanities³. The abbreviation TEI is also used to indicate the metadata set itself, as in the following documentation of the TWKM project⁴.

TEI employs the mark-up language "Extensible Markup Language" (XML), which has established itself as the standard for digitally describing source materials in contemporary humanities research and thus permits targeted queries and further processing. Due to its standardized element set, TEI offers the advantage of long-term and clear interpretability of datasets. Furthermore, the utilization of TEI in projects such as TWKM promotes recognition of the format as the standard and thus facilitates data exchange (Rouché and Flanders 2007-2014; see Werning 2013:3).

The TEI metadata schema of the current version P 5 represents a defined quantity of XML elements. The schema is divided into various modules, each of which marks up specific elements and attributes. For example, elements are defined for coding digital dictionaries in the module "dictionaries". An element can contain other elements or pure text. Each TEI-compliant text is introduced by the element `<teiHeader>`. This strategy effectively creates the title page of the electronic text file and contains the file description (required) or specifications regarding amendment of the text (optional), among other things. Within a TEI file, the header can be used repeatedly. The body text follows the header and can differ greatly according to the text that is being described.

TEI pursues two goals: firstly, to allow researchers to digitally represent their source materials using a description language; and secondly, to represent this digital information by using a shared, widely understood code. By using a comprehensive code, TEI can be very detailed and specialized for use with various source materials. Similarly, it is possible to restrict the code to essential information without

¹ *Thesaurus Linguae Aegyptiae*. Arbeitsstelle Altägyptisches Wörterbuch. Berlin-Brandenburg Academy of Sciences and Humanities. <http://aaew.bbaw.de/tla/index.html> (04.08.2014).

² *Pennsylvania Sumerian Dictionary*. University of Pennsylvania. <http://psd.museum.upenn.edu/epsd1/index.html> (04.08.2014).

³ See "TEI: Frequently Asked Questions". TEI Consortium. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TitlePageVerso.html> (04.08.2014).

⁴ In order to more easily distinguish between the two projects, the overarching project will be denoted as the TWKM project.

specializing in particular phenomena. An advantage of the detailed code is that the described text offers more possibilities for application, such as targeted queries; however, one must keep in mind that this code also makes inputting more demanding and requires greater technical expertise. Use of TEI in various fields is also encouraged by the potential for defining the mark-up language using adaptations specific to the purposes of individual projects. This characteristic inspires the subsequent use and spread of the TEI standard, and it has the potential to facilitate mutual stimulation between different research areas, while at the same time differentiating them from one another (see Rouché & Flanders 2007-2014). The metadata schema for Classic Mayan texts thus represents a metadata set that was compiled for this purpose and that is capable of describing specific information.

Numerous projects that set out to catalog digital texts of various genres draw upon the TEI metadata schema. On the homepage of the TEI initiative, a list of selected projects is available. These projects also include projects that aim to catalog digital text versions of epigraphic source materials, such as the *Inscriptions of Aphrodisias* project of King's College London⁵.

Project Definition and Planning

Goals

The goal of this sub-project was to develop the foundation for the TEI metadata schema for cataloging all known Classic Mayan texts. The TEI metadata schema thus constitutes a component of the metadata concept as a whole. Because the TWKM project was still in its initial phase and many questions related to the data contents remained unanswered, this TEI metadata schema was intended as a foundation that could be further adapted over the course of the TWKM project. The sub-project therefore did not aim to create a final, complete metadata schema.

General Procedures

Within the TWKM project, responsibilities are divided into two areas: specialized tasks related to Classic Mayan, and technical and computer science support.

In order to catalog the Classic Mayan texts, it is necessary to know the basic structure of the language. This prerequisite has a two-fold justification: firstly, this knowledge is a foundational requirement for cataloging the relevant data; and secondly, it is essential for communicating with scholars in order to better understand their needs. As such, it was necessary to become acquainted with the Classic Mayan language, in order to learn about its structure and become familiar with the relevant technical terms.

In order to catalog information that is important to scholars, and to address various aspects of research, several levels were taken into account for the metadata schema:

- Material object: including Maya artifacts, as well as modern documents such as rubbings, reports of discoveries, etc.
- Inscription: cataloging the hieroglyphic texts themselves and all of the information pertaining to them

⁵ See "Projects Using the TEI." TEI Consortium. <http://www.tei-c.org/Activities/Projects/> (04.08.2014) and Reynolds, Roueché & Godard 2007, <http://insaph.kcl.ac.uk/iaph2007/>.

- Place: relevant to this level are the location of discovery, as well as the current place of storage (e.g. museums)
- “Actor”: including actors named in the text (e.g. rulers, gods) and depicted in the iconography, as well as modern actors, such as researchers participating in excavations or the museum housing the objects
- Time: this category includes dating the objects (with the requisite conversion of Maya calendrical dates into Gregorian dates), date of discovery, etc.

Different metadata standards are drawn upon to catalog all the necessary data and information, in order to do justice to their diverse facets. As such, the text carriers are primarily described using CIDOC CRM⁶. The TEI metadata schema was drawn upon in order to catalog the inscriptions themselves; later, the schema will also form for the foundation for the analysis of the Maya script and for the compilation of the dictionary. This component will be described below, given that this sub-project is related to the development of relevant metadata concerning the texts. The field descriptions of the elements, as well as the terms and definitions relating to the text structure, are in English, the preferred language of the TWKM project and also the language of the later TWKM database.

Procedures for Cataloging the Texts

The requirements of the metadata schema with respect to the texts were formulated based on the goals and conceptions of scientific experts, which had arisen from the project proposal submitted to the Academy of Sciences, Humanities and Arts and from related discussions. An assortment of modules relevant to cataloging the texts was selected, in order to limit the very extensive TEI metadata set. Given that TEI currently serves as the foundation for other epigraphic cataloging projects, inquiries were made into comparable projects with the goal of acquiring more information about their metadata structure.

Scientific Challenges

The demands of the scientific experts can be divided into two categories: 1. those relating to the metadata schema as a whole, and 2. those that need to be taken into account particularly when describing the texts.

1. General Requirements

- All metadata elements for cataloging all texts that have been found and will be found in the future, i.e. different representations have to be taken into account.
- Incorporation of temporal and spatial parameters, i.e. discovery location and dating must always be retrievable.
- Script variants in correlation with the relevant time (dating) must be readable; in other words, the exact notation of hieroglyphs and signs, respectively, must be associated with the corresponding (dated) text.

⁶ The CIDOC Conceptual Reference Model (CRM) constitutes a documentation format for the field of cultural heritage and has been the official ISO Standard (ISO 21127:2006) since 2006. This format was selected in order to be able to appropriately represent the numerous aspects of the object itself, such as history of discovery, provenance, and relevant figures, such as excavators, curators, etc.

- Facilitation of a language- and script-based search function in the database, i.e. original spelling, transcription, and translation must be cataloged.
- Accounting for undeciphered text passages with an image of the original spelling.
- References to secondary literature (abbreviated citation with a URN linked to a bibliography).
- The metadata schema should be able to be used by subsequent projects; in other words, the metadata schema should be as flexible as possible.

2. Text-specific Requirements

- Representing the relationship between text and image
- The number of text fields, and of hieroglyphic blocks and signs per text field on an individual text carrier, must be calculable.
- Form/representation of the texts (single-/double-column, rectangular, etc.) must be apparent.
- Ability to define colored text areas.
- Description of difference in block size, i.e. “capital letters” and blocks that are depicted on a smaller scale must be differentiable.
- Cataloging of the texts must be separated from their interpretations.
- Reading order and orientation of individual signs must be indicated.

Metadata Schema for Cataloging the Texts

The TEI description language should be suitable for as many humanities fields as possible, according to the ideas originally underlying its development, and presents a very extensive element set. As such, the initial search for appropriate elements is time-consuming.

EpiDoc (Epigraphic Documents) offers a more restricted scope specific to epigraphy. EpiDoc is an international community of scholars whose research concentrates on ancient inscriptions. This community has developed recommendations for coding inscriptions with XML that constitute a subset of the *TEI P5 Guidelines* and are specially oriented towards working with ancient and medieval texts. By now, the recommendations have been extended from ancient Greek and Latin inscriptions to describing papyri and manuscripts (see Elliott, Bodard & Cayless et al. 2006-2013). These recommendations are advantageous because TEI elements that are inappropriate for describing inscriptions can be eliminated from the outset, and because the project provides optimal support for the description of epigraphic materials with its own amendments to definitions (see Rouché & Flanders 2007-2014).

In order to initially select elements that could be used for professionally describing the texts, the modules of the *TEI P5 Guidelines* that appeared relevant were probed (see TEI Consortium 2014:2). The following areas were identified:

- *header*: each TEI-compliant text must specify certain descriptions of the file itself, so that the module is relevant to each TEI file.
- *core*: the module contains elements that can appear in all text genres to be described. Many of these core elements can be flexibly employed and can appear in every text passage.

- *textstructure*: the elements of this module are used to describe the external text structure. Given that the texts are structured in the arrangement of the hieroglyphs, elements of this module can be relevant to the description.
- *gaiji*: this module contains elements for describing unusual script types, symbols, and hieroglyphs. This module is taken into consideration because the Maya script is hieroglyphic and consists of individual signs.
- *figure*: The elements for reproducing images, tables, etc. that appear in a text are defined in this module. Images are often present on objects inscribed with Classic Mayan texts. Because these images are related to the text, they must be sufficiently represented, along with the text itself.
- *transcr*: This module defines elements for representing primary sources, i.e. the texts themselves. This module was taken into consideration because images of sources (e.g. digital photographs of text carriers) have to be included in the TWKM project.

The modules that appear to relate to analytical aspects or that are highly focused on individual text genres were not taken into consideration during this initial orientation.

In EpiDoc, described elements are divided into various areas that could be relevant to epigraphic publications. As in the case of the TEI modules, areas appropriate to the TWKM project were probed as well (see Rouché & Flanders 2007-2014):

- *the edition of the epigraphic text itself*: instructions for describing text structure, the display format, and the transcription are provided.
- *history of the discovery, documentation, and interpretation*: the code of bibliographic references is explained here. The TWKM project aims to associate particular readings with references to scientific literature in which they are mentioned.

The remaining areas specified in EpiDoc are related either to information concerning the text carrier itself (history of discovery, etc.), or to elements affecting text analysis. These areas would be redundant here, since data regarding text carriers will be contained in separate data containers within the master plan of the metadata schema.

This selection of elements was then ultimately evaluated according to scholarly requirements: what elements are available for describing text structure? Which elements are appropriate for describing the hieroglyphs?

While developing the TEI schema for representing the structure of hieroglyphic texts, it became clear that scientific terms and the scientific relevance of particular specifications needed to be clarified. What is the most useful description for the side of a text carrier, for instance; is there a front and a back side? How can the relationship between text and image be established? Which specifications belong to the factual representation of the text, and which are already on the level of interpretation? And: how can individual hieroglyphs be clearly addressed without anticipating a particular interpretation?

Representing the Structure of the Text

One of the challenges of reproducing the text structure is the large number of forms the design of a text may assume; all of these have to be represented by the metadata. The arrangement of hieroglyphic blocks varies, as does the form of the text field (Table 1).

Arrangement of Hieroglyphic Blocks	Single-column Double-column Combination of single- and double-column Horizontal lines Combination of columns and horizontal lines Initials
Form of Text Fields	Square Hieroglyphic band Angled Cartouche (i.e. with outer frame) “Captions” (hieroglyphs as internal components of an image) Speech bubble

Table 1. Overview of Possible Structural Configurations of Classic Mayan Texts.

In order to account for all of these facets using the described data, the metadata schema was arranged in sections that build upon each other (Figure 2). This division is intended to facilitate selection of relevant metadata elements and to make the procedure more transparent for further use. Elements for describing the “Inscription” as well as the three sub-sections “TextDivision”, “Block”, and “Sign”, will be discussed and described below.

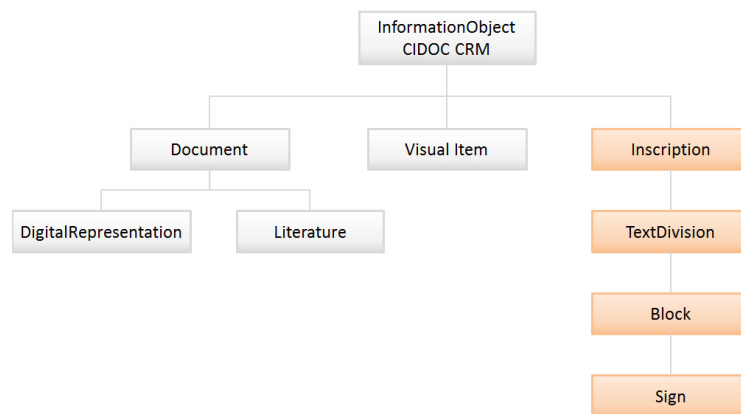


Figure 2. Excerpt from the Master Plan for the Metadata Schema (Colored Mark-Up: TEI as the Basis).

TEI Elements

The TEI header and a text element form the basic pair of a TEI element. The header contains metadata that describe the document as a whole and can either be very comprehensive or kept rather “narrow”. The text element contains the metadata of the document itself. The element `<teiHeader>`, together with its descriptive and explanatory information, constitutes the electronic title page, as it were, whereas the element `<text>` contains the textual content of the object with annotations that clarify its structure and additional characteristics.

`<teiHeader>`

According to the *TEI P5 Guidelines*, the element `<teiHeader>` must minimally contain the element `<fileDesc>` (file description), which describes the electronic file. This element, in turn, is assigned three obligatory components: `<titleStmt>`, `<publicationStmt>` and `<sourceDesc>`.

The `<title>` subelement `@type`, which can indicate alternative forms of names, is redundant here; alternative designations for the text carriers that appear in the scientific literature will be stored in a so-called vocabulary⁷, for which reason the conventional designation alone is considered to be sufficient.

Similarly, the representation of people who are associated with the object will be foregone here. These specifications will be stored in the CIDO CRM category “Actor” and “Appellation”, respectively, and the explicit URI of the TWKM-ID will ensure that they are connected to the object in the metadata schema. This approach offers the advantage of not having to re-develop data that are already represented elsewhere. The approach to the object data is similar: mass, context of discovery, dating, etc. can be marked appropriately and in detail using the CIDOC metadata set. As a result, only a few elements are used for the `teiHeader`; for instance, the specifications `<extent>`, `<notesStmt>`, `<author>`, and `<geoDecl>` for the find coordinates can be eliminated – data entry is therefore marginal and minimally taxing.

Consequently, for the TWKM project, the element `<teiHeader>` could be reduced to the following specifications:

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>[TWKM-ID]</title>
    </titleStmt>
    <publicationStmt>
      <authority>[name]</authority>
      <idno type="URI">[link to object-ID]</idno>
    </publicationStmt>
    <sourceDesc>
      <p>[e.g. Copan, Stela D]</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

The identification number within the element `<publicationStmt>` uses a hyperlink to connect to the corresponding object itself, and thereby to all of the metadata relating to the text carrier.

Additionally, in the TextGrid recommendations, `<encodingDesc>` (code description) and `<editorialDecl>` (description of the editorial principles) are specified with the element `<normalization>`, which represents the degree of standardization and normalization (see Blümm and Wegstein 2008: 22ff.). It remains to be determined whether or not these elements would be viable options for the TWKM project at this stage.

Inscription

When describing the texts, the possibility that one object may contain multiple texts and that individual texts can refer to images must be representable. The text description must reflect the overall picture, i.e. the arrangement of the texts and related images.

⁷ The vocabularies that are being compiled for the TWKM project are being coded according to the “Simple Knowledge Organisation System” (SKOS).

Prior to the text description, a reference will be made to the digital facsimile (digitalization of a rubbing, drawing, or digital photograph) using the element `<facsimile>` and the corresponding URI of the digitalization, following the example of EpiDoc (see Bodard 2007-2014).

The text will be identified with the tag `<text>`. This element does not contain an individual, stand-alone text, nor a text consisting of multiple sections. In the case of multiple texts that belong together, the element `<text>` will be enclosed by `<group>` in order to represent the larger unit (see TEI Consortium 2014: 150, 1445). This strategy could prove useful for describing two corresponding fragments of a Maya inscription. The text itself is represented in the element `<body>`, although this element in each case only contains the stand-alone texts. In other words, from this descriptive level onward, only individual texts are addressed.

Two additional elements of the corpus are `<front>` and `<back>`: `<front>` serves to describe all contents that precede the actual text (e.g. title page, foreword, dedication), whereas `<back>` refers to all components that follow. However, it is certainly possible that introductory or even concluding formulae (e.g. the naming of the artist who created the text) could also be differentiated from the text description itself using these elements. Because this process already entails interpreting the text contents, the use of the tags `<front>` and `<back>` should be avoided. For the Maya texts, use of the element `<body>` is sufficient.

Due to the fact that Maya texts may appear on different areas of an object, the side will be defined next. For scholars, it is customary to speak of the front and back sides of a text carrier. The front side is identified by the image of a ruler, if present, or otherwise by the indication of the date. This distinction resulted in the descriptions of the sides: front, right, left, back. However, these descriptions should not be confused with TEI elements, which are already excluded from use. This specification is a component of the element `<body>`, not `<text>`. In the case of a cohesive text that continues across multiple sides, the specification is a component of the text division (see below).

Abbreviations for describing images that may be used analogously for describing text fields were established for the TWKM project, which permits a unified designation (Table 2):

Abbreviation for	Clarification
f or b front or back	The side with the image of the ruler or specification of the date is generally regarded as the front side. It remains to be clarified how objects should be handled for which these details are not known or visible.
l or r left or right	The sides to the left and right of the front side.
t or u top or underside	Description for the upper and lower sides of the text carrier. Texts on the bottom side include lintels and the base of ceramic vessels, for instance.
g girth	Used for a running text, e.g. in the case of circular altars.

Table 2: Designations of the Specification @type of `<body>`.

It remains to be determined whether the designation “girth” should also be used for ceramic roll-outs and running texts, respectively. The implementation of these designations in the case of irregular objects, such as inscriptions on zoomorphs (sculptures in animal form) or in caves, remains similarly under debate.

Converting the many possible arrangements of the hieroglyphic blocks, as well as of forms of the text field, presents a particular challenge. For example, in the case of a column, it must be clearly documented where a new line begins, where the column begins, and where the reading sequence begins in the next column. This process is comparable to reading a newspaper. How can single- and double-columns be converted? A general method for representing text structure was sought based on these “simple” examples. This foundation could then be tested on further forms, such as that of a rectangular text, and expanded.

TextDivision

“TextDivision” constitutes the sub-section of “Inscription” and describes one text passage in particular or a text field on an object. The element `<div>` from the TEI Standard lends itself to description. This element can either be used in numbered or un-numbered style. The un-numbered variant reflects a hierarchy of individual text passages, in which `<div1>` describes the uppermost level, `<div2>` the following level, etc. The variant without numeration is used here, given that there is no hierarchy of individual text passages in the hieroglyphic texts and that all passages are seen as equal to each other. The text may be classified using the attributes `@type` and `@subtype`, respectively. As such, individual text components can be described separately, for instance; similar to the element `<body>`, “passages” can be more exactly defined using `@n`. Differentiation according to arrangement type is useful for classification (see Table 2). However, an explicit vocabulary would have to be compiled, indicating for instance the possible arrangements of hieroglyphic blocks as a value of the attribute `@type` and the form description as a value of `@subtype`:

```
<div type="combination-column-line" subtype="right-angled">
```

By expanding a numeration, the corresponding text field within the side of the inscription can be more exactly described:

```
<div n="B1-D3" type="combination-column-line" subtype="right-angled">
```

Frequently, only fragments of inscriptions are available in archaeological research. For such cases, the EpiDoc recommendations provide the `@type` “fragment”, which is positioned before the corresponding description of the text passage:

```
<div type="fragment">
```

The end of a column is tagged with `<cb>` (column break). In addition, description of change in sides is necessary to account for the three extant codices. The beginning of a new page is indicated using `<pb>` (page break).

In scholarly research, individual hieroglyphic blocks are referred to using a grating similar to that used to partition a chess board. This denotation must be reflected in the TEI elements. Under “Inscription”, the entire grating of the inscription is represented, permitting each individual hieroglyph to be specifically referenced, e.g. the identity of block D3 of Fig. 3 is clearly established. Nonetheless, in some instances, the position of a block relative to the coordinates of the grating is not clear, or two blocks are located at the same coordinates. In this case, a sub-classification is employed, so that the “sub-blocks” are designated “A2a” and “A2b”, for example. The basic structure of the inscription can be described using the “coordinates”.

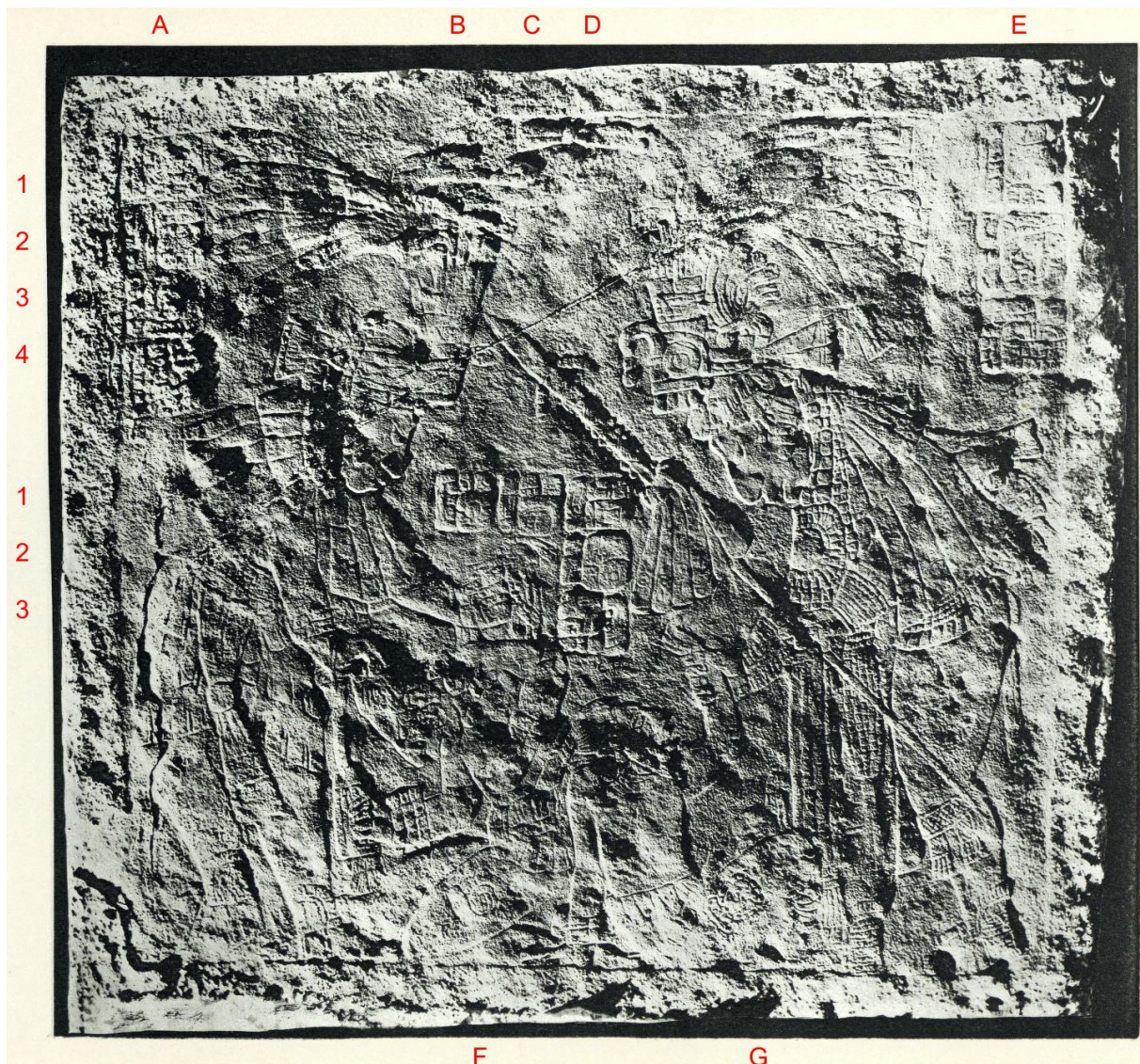


Figure 3: Maya Inscription with Pictorial Representation and Labeling of the Hieroglyphs (Matrix). Yaxchilan Lintel 8.⁸

The relationship between text and image is relevant not only at the level of the text passage, but also at the level of individual block. Different combinations exist: the text passage as a whole can refer to a pictorial representation, the text passage serves as a “speech bubble” of an actor, or one or more blocks are positioned on an actor or an object. A controlled vocabulary will be compiled for unambiguous designation of these variants.

A comparison with the “comic” genre came to mind when considering how to describe relationship between text and image. A search indicated the existence of the TEI-based *Comic Book Markup Language* (CBML; Walsh 2012). The tag `<balloon>`⁹ is introduced into a distinct CBML module to mark

⁸ After Maler 1903: pl. 52, the block designations are added after the CMHI.

<https://www.peabody.harvard.edu/CMHI/flash/detailview.swf?num=8&site=Yaxchilan&type=Lintel>

⁹ „<balloon>“. In: Walsh 2012, <http://dcl.slis.indiana.edu/cbml/schema/cbml.html#TEI.balloon> (10.08.2014).

“speech bubbles”, and `<caption>`¹⁰ to indicate text belonging to an image. Whether or not the description of inscriptions and pictorial representations can be conducted analogously is still under debate. For this purpose, the CBML module would have to be integrated or a distinct typification would have to be defined. However, `<caption>` is also defined in TEI, meaning that the TEI elements alone may be sufficient.

According to the *TEI P5 Guidelines*, the representation of text-image relationships is realized using `<figure>`. The pictorial representation is defined using `<graphic>` and a URL. A description of the image using the element `<figDesc>` is not required, because it is already included in the CIDOC CRM.

For Example:

```
<figure>
  <graphic url="..." />
  <ab type="caption">[signs with relation to an image]</ab>
</figure>
```

Signs are frequently depicted on images of individuals (people, gods, animals) or objects. The results of a discussion indicated that the location of the segment of text is significant within the context of the representation: a sign indicating the ruler is found on the headdress, and signs represented on the thighs of individuals are exclusively associated with social subordinates (see Fig. 4). The script thus expresses sociocultural structure, and thereby provides important information to researchers. In order to describe the distinction using metadata, the scholars in Bonn compiled an additional vocabulary to facilitate specification by the `type`-attribute.

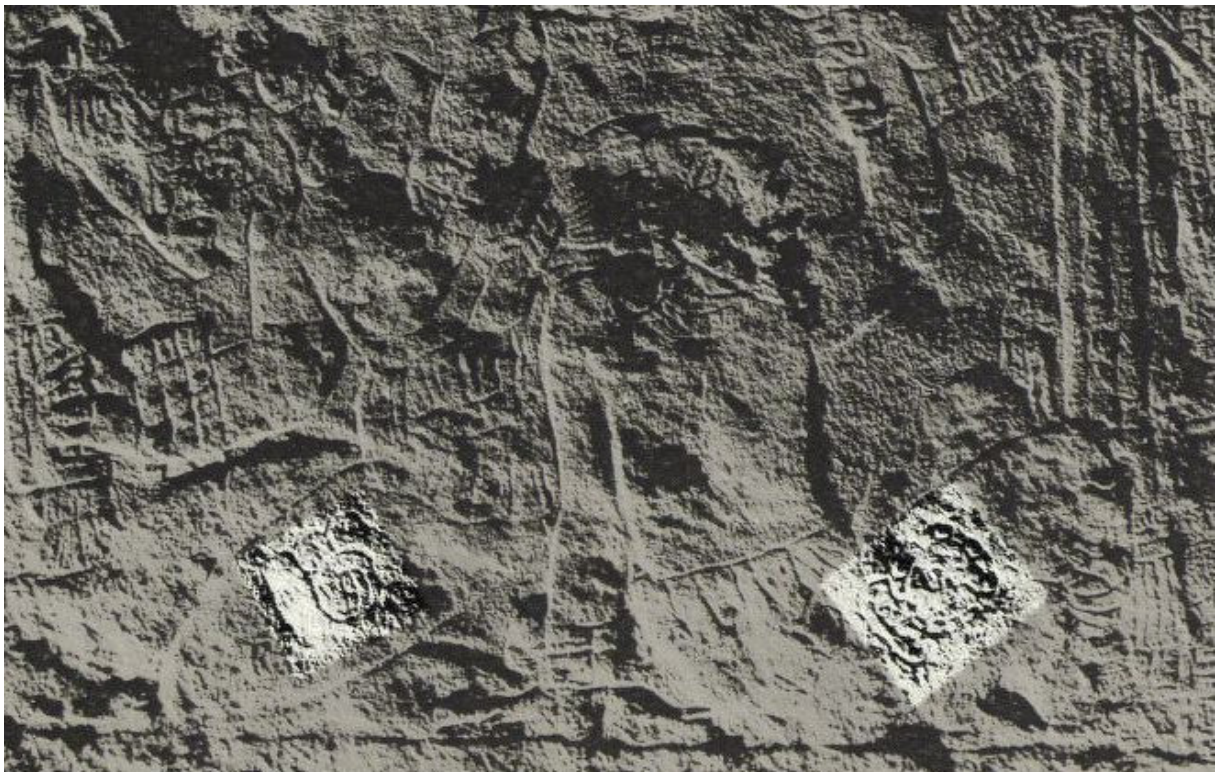


Figure 4: Hieroglyphs on Individuals as an Expression of Social Status (excerpt from Yaxchilan Lintel 8).

¹⁰ „<caption>“. In: Walsh 2012, <http://dcl.slis.indiana.edu/cbml/schema/cbml.html#TEI.caption> (10.08.2014).

Block

An attempt to address the problem of describing the blocks resulted in a subdivision of the `<div>` element using a defined attribute that specifies the exact block coordinates (e.g. A1), whereby a block would be described as follows: `<div type="block" n="coordinates">`. According to the same schema, individual logograms or syllabograms would be defined as `@subtype=sign`. This approach already proved to be unusable during the development of additional, relevant descriptive criteria, such as highlighting individual signs. According to the *TEI P5 Guidelines*, very few core elements such as `<gap>` are permitted within the element `<div>`. The tag `<hi>` (highlighted) required for identifying colored blocks, however, is not allowed. Thus, another solution had to be found.

After examining the elements and searching for comparable cases in the EpiDoc guidelines, the solution appeared to be to insert an element in between. `<l>` (line) or `<ab>` (anonymous block) would come into consideration for this purpose, although `<l>` serves to describe verses according to the *TEI P5 Guidelines*. In contrast to `<l>`, `<ab>` can be more freely used, for which reason this element was selected (see TEI Consortium 2014:508):

```
<div n=A type="column">
  <ab type="Block" n=A1>
    T1:257.1:624:178
  </ab>
  ...
</div>
```

An alternative representation of the blocks enables the element `<milestone>`:

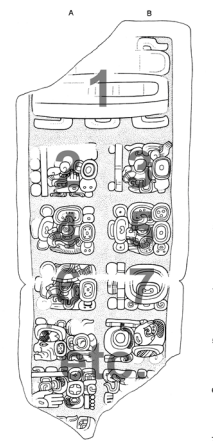
```
<milestone unit="block" n=A1>T1:257.1:624:178
<milestone unit="block" n=A2>...
```

However, use of the `<milestone>` tag should be discussed before it is used. “Since it is not structural, validation of a reference system based on milestones cannot readily be checked by an XML parser, so it will be the responsibility of the encoder or the application software to ensure that they are given in the correct order” (TEI Consortium 2014: 114 ff.).

In order to achieve a clearer description of the structure, it would be wise to mark line breaks. For this purpose, the element `<lb>` (line break) is used in place of “end-of-line”, i.e. after the second hieroglyphic block in a double-column structure.

A variant of the TEI metadata schema for a double column whose first block is represented as larger than the others could thus appear as follows:

```
<text>
  <body type="front">
    <div type="column" n=A>
      <ab type="block" n=A1.B1>
        <hi rend="tall">[grapheme] </hi>
      </ab>
      <ab type="block" n=A2>
        </lb> [grapheme]
      </ab>
      <ab type="block" n=B2>
        </lb> [grapheme]
      </ab> ...
    </div>
  </body>
</text>
```



Sign

A hieroglyphic block usually consists of three to four (maximally five) signs in different combinations. Thus, it is important to be able to reproduce the reading order. For this procedure, scholars have established a standard according to which adjacent signs are separated by a period, for example, and signs that are stacked atop each other are separated by a colon. This convention also indicates whether an individual sign is vertically or horizontally oriented within the block. Thus, this standard can be used for indicating sign order¹¹.

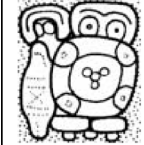





					
T1	T257	T1	T624	T178	
T1:257.1:624:178					

Figure 5: Representation of the Reading Order of Individual Signs within a Hieroglyphic Block (see Grube 2011: 7).

The representation of Classic Mayan hieroglyphic signs is diverse, and also varies and develops across time and space. For this reason, it was necessary to link each hieroglyph with its original spelling. Only thus could the development and variants of each sign be made tangible. In order to reproduce the inscription, the signs were represented using a classification, according common scientific methods: for example, T178 would represent the syllable *la* according to Thompson's classification. This procedure already represents a step towards interpretation of the signs and therefore must be regarded critically.

There are other classification systems in addition to Thompson's, which will be combined and supplemented to create a unique sign concordance for the TWKM project. Each sign will receive a unique identification number that will later be used as its primary reference. The concordance will be compiled over the course of the TWKM project and expanded as needed. Uninterpretable signs will not be indicated with a question mark, but instead will receive their own unique number within the concordance; the reading, transcription, etc. can then be updated to reflect the current state of knowledge. Because concordance numbers will be allotted to the standardized form of each sign, each variant form must also be given a unique ID, in order to be able to trace the geographic and temporal distribution of the variants' use. This method allows a unique number to always be used in the metadata description; thus, undeciphered texts can be taken into account by referring to the original spelling, as per the project requirements. No solution was yet available for representing numbers, which constitute a separate sign category within the Maya hieroglyphic script; as such, not all dates in the inscriptions could be represented according to the current state of research.

It would be possible to describe the concordance according to TEI, for example in accordance with a taxonomy (see TEI Consortium 2014: 46ff.). The individual signs could thus be referenced using an ID. Then, for instance, the attribute `xml:id="I156"` would be synonymous with a TWKM number in the description.

The TEI elements `<g>` and `<glyph>` (reference to `<g>`), respectively, could potentially be used to describe signs according to their original manifestation, particularly in the case of signs for which no Unicode exists (see TEI Consortium 2014:181). The EpiDoc recommendations restrict themselves to

¹¹ The reading order typification that had been included in the proposal was not pursued. See Grube 2011: Attachment 11.

using `<g>` only “where a symbol is non-meaning-bearing”; the symbol, such as a crucifix¹², is described in a subsequent `@type` attribute. It would be conceivable to create a TWKM project-specific module for the concordance that would be structured similarly to the XML schema for describing the tag `<glyph>`.

Because representing a sign entails an interpretation, it is important to document each reading with references to secondary literature. A bibliography will be generated using the open-source reference management software program Zotero, which additionally allows data to be exported in TEI format. References to a particular entry are realized using the `<ref>` tag, which links to the corresponding entry in the bibliography:

```
<ref target="#Stuart 2008">158-159</ref>
```

Missing and Illegible Text Passages, Hieroglyphs, and Signs

Lacunae can be represented in all three subsections in the text, depending on the extent of the missing text passage, i.e. as part of the descriptions of `<div>`, `<block>`, and `<sign>`. In each case, they are introduced by the element `<gap>` and more exactly defined by an attribute. According to the *TEI P5 Guidelines*, the attributes are optional; nonetheless, it is wise in this case to follow the EpiDoc recommendations, according to which the attribute `@reason` is mandatory. ‘Lost’, ‘illegible’, ‘omitted’, and ‘ellipsis’ are intended as values¹³. EpiDoc offers very comprehensive specifications for describing text passages that cannot be represented. Among other options, it is possible to also indicate the size of a gap, at least to the extent that this information is known:

```
<gap reason="illegible" quantity="1" unit="block"/>
```

The so-called Leiden Conventions are also used in Maya studies to convert the original inscriptions, whereby lacunae and their respective sizes can be represented. Thus, the implementation of EpiDoc lends itself to the process of utilizing the Leiden Conventions.

Critical Evaluation of the Metadata Schema

The danger of dividing the inscription into so many small components is that the TEI structure becomes confusing—as such, one should consider whether some elements can be omitted while still producing the same result. Another consideration is whether an individually adapted selection of elements should be defined for each of the various possible arrangements (single-, double-column, etc.), comparable to the *TEI P5 Guidelines* in their subdivision according to genre. It is wise to define multiple optional elements, in order that they may be selected from the available set as needed.

Comparison of the element set with the demands that researchers have formulated over the course of the project indicates that the demands are largely accounted for in the element set. The problem of clearly identifying the signs as individual components of the hieroglyphic blocks remained unsolved. According to the current mark-up, the signs are written one after another, as in a running text. A remedy to this problem may be provided by the sign concordance, which uses an `xml:id` to mark individual signs and their variants. Using this, it would also be possible to calculate the number of signs preserved in an inscription. There are various possibilities for describing the signs: at the conclusion of

¹² “Symbol (Non meaning-bearing)”. In: EpiDoc-Guidelines. <http://www.stoa.org/epidoc/gl/latest/trans-symbol.html> (22.07.2014).

¹³ “<gap>”. In: EpiDoc-Guidelines. <http://www.stoa.org/epidoc/gl/latest/ref-gap.html> (15.08.2014).

this sub-project, it could not yet be determined whether the element `<g>` / `<glyph>` or `<milestone>`, respectively, was appropriate for marking individual signs. It is possible that a viable solution may be identified after the concordance has been compiled. It was not possible to divide pure description from interpretation of the texts as was demanded of the project, because no clearly coded language is available for the individual signs. To some degree, explicit assignments were not possible, because the Classic Mayan script and language have not yet been completely investigated.

Another question was whether or not representing the arrangement of the hieroglyphs within a block using the previous standard (period for two adjacent signs, etc.) was sufficient for research purposes, or whether a precise mark-up that permits targeted queries would be necessary. It is also possible that the sign arrangement typification mentioned in the TWKM project proposal could offer a satisfactory solution (see Grube 2011: Attachment 11).

Given that many texts include iconography that is highly relevant to analyzing, and thereby interpreting, the content of the text, too few elements were used for marking pictorial representations. Representing the relative proportions of images and indicating their precise position were not possible at that stage. An additional raster would have to be defined in order to describe the exact position and also to indicate empty spaces. The raster would not only locate the hieroglyphs using coordinates, but also employ the same aspect ratio for all inscriptions. As such, it would be possible to clearly describe the pictorial representations and potentially to thereby convey the images' proportions.

The selection of the elements can serve as the foundation for further elaboration. Over the course of the project, even more issues will have to be taken into account – new demands and challenges are constantly being identified in project discussions. Furthermore, one should reconsider whether this metadata schema will also be able to describe unusual inscription forms that have not been accounted for in the extant examples. As soon as the optimal representation has been determined this working base, the transcription and transliteration of the signs may be carried out. These later processes are also supposed to be described in TEI.

The metadata schema indicates that the attributes of the *TEI P5 Guidelines* require additional adjustments for cataloging the texts, for example with regard to attribute values. The specifications of the EpiDoc Guidelines were consulted frequently. However, TWKM-specific adjustments proved to be necessary, particularly as regards descriptions of the relationship between text and image. The selection of elements also indicates that the schema consists of a mix of various TEI modules, which was necessary in order to take into account the different aspects of the inscriptions.

Conclusions

Familiarizing oneself with this project entails two very complex goals. A basic understanding of Classic Maya is a prerequisite to be able to follow scientific discussions in this field and to understand the demands of the project. In addition, it is essential to have a (basic) knowledge of the TEI format. In this respect, preselection according to module was helpful for attaining an initial overview. The TEI modules, as well as sections of the EpiDoc recommendations, facilitate examination of these hitherto unfamiliar materials. The *TEI P5 Guidelines* provide a quick introduction and, in the online version, allow rapid searches for individual elements whose possible applications are always indicated in examples. However, it was difficult to identify obligatory elements of a module: it is not apparent from the survey of the individual elements which element in the hierarchy is obligatory and which is optional. These distinctions are indicated only in the explanation of the *Guidelines*. Thus, it is always

necessary to check the corresponding chapter of the selected elements¹⁴. Inquiries into other projects that are digitally cataloging inscriptions also failed to produce further information in the case of problems for which the EpiDoc recommendations do not offer a solution, such as when converting individual signs. Nonetheless, the example of “Comic book markup language” indicates that possible approaches may not only be found in epigraphic projects.

Regular exchange between all project contributors was a basic prerequisite for the success of a project such as this – requirements that are not accounted for in the metadata schema or the technical infrastructure (layout, search functions, etc.) would otherwise require great effort to correct. Thus, precise and diligent collaboration was important from the beginning. In meetings between the TWKM project teams, it became apparent which data and information were significant for executing the TWKM project.

In summary, the process of developing the metadata concept indicates that this field greatly resembles to that of library cataloging: the processes of preparing of standardized data for names, compiling of controlled vocabularies, and recognizing common structures within the data are visible in descriptive and subject cataloguing in scientific libraries, as well as in the preparation of authority control – even if the mark-up language for the TWKM project presumably hardly plays a role in scientific universal libraries¹⁵. Thinking outside of the box is also worthwhile for librarians, as their expertise allows them to provide meaningful support to research projects in the field of so-called Digital Humanities.

References

Arbeitsstelle Altägyptisches Wörterbuch

n.d. Thesaurus Linguae Aegyptiae. Berlin Brandenburgische Akademie der Wissenschaften.

<http://aaew.bbaw.de/tla/index.html>

Blümm, Mirjam, and Werner Wegstein

2008 The TEI header for Texts in Baseline Encoding. In *TextGrid's Baseline Encoding for Text Data in TEI P5 (2007-2009)*, edited by Mirjam Blümm et al., pp. 19–27.

<http://www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf>

Bodard, Gabriel

2007 Structure of an EpiDoc Edition. In *EpiDoc Guidelines: Ancient documents in TEI XML (Version 8)*, edited by Tom Elliott, Gabriel Bodard, and Hugh Cayless. <http://www.stoa.org/epidoc/gl/latest/supp-structure.html>

Elliott, Tom, Gabriel Bodard, and Hugh Cayless

2006 EpiDoc: Epigraphic Documents in TEI XML. <http://epidoc.sf.net>

Gronemeyer, Sven

1999 Das Schriftsystem der Maya. Hausarbeit im Rahmen des Proseminars „Schriftsysteme Amerikas“.

<http://www.sven-gronemeyer.de/research/schrift.html>

Grube, Nikolai

1993 Schrift und Sprache der Maya. In *Die Welt der Maya*, edited by Reiss-Museum der Stadt Mannheim. 3rd ed. Zabern, Mainz.

2011 Textdatenbank und Wörterbuch des Klassischen Maya (TWKM). Antrag für ein Forschungsprojekt im Rahmen des Forschungsprogramms der Deutschen Akademien der Wissenschaften (Akademieprogramm) für 2013. Bonn.

Grube, Nikolai, and Maria Gaida

2006 *Die Maya: Schrift und Kunst*. SMB-DuMont, Berlin & Köln.

¹⁴ When using an XML editor such as that of *oXygen*, the data can be easily checked for validity and well-formedness.

¹⁵ According to the German Research Foundation's 2009 Practical Guidelines for Digitalization, the TEI format should be used for cataloging medieval manuscripts (q.v. pg. 18). The Herzog August Library in Wolfenbüttel and the University Library of Heidelberg, among others, are following these recommendations.

Maler, Teobert

1903 *Researches in the Central Portion of the Usumatsintla Valley: Reports of Explorations for the Museum*. Vol. 2. Memoirs of the Peabody Museum of Archaeology and Ethnology, Harvard University 2. Peabody Museum, Cambridge, MA.

Pennsylvania Sumerian Dictionary

n.d. The Pennsylvania Sumerian Dictionary. University of Pennsylvania.
<http://psd.museum.upenn.edu/epsd1/index.html>

Reynolds, Joyce, Charlotte Roueché, and Gabriel Bodard

2007 *Inscriptions of Aphrodisias (2007)*. <http://insaph.kcl.ac.uk/iaph2007>

Roueché, Charlotte, and Julia Flanders

2007 Gentle Introduction to Mark-up for Epigraphers. In *EpiDoc Guidelines: Ancient documents in TEI XML (Version 8)*, edited by Tom Elliott, Gabriel Bodard, and Hugh Cayless.
<http://www.stoa.org/epidoc/gl/latest/intro-eps.html>

TEI Consortium

n.d. *Projects Using the TEI*. <http://www.tei-c.org/Activities/Projects/>

n.d. *TEI: Frequently Asked Questions*.

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TitlePageVerso.html>

2014 *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.6.0., 20.01.2014.

<http://www.tei-c.org/Guidelines/P5/>

Walsh, John A.

2012 *Comic Book Markup Language*. School of Library and Information Science, Indiana University.

<http://dcl.slis.indiana.edu/cbml/>

Werning, Daniel A.

2013 *Datenkodierung in TEI XML im Rubensohn-Projekt (Arbeitsbericht)*.

http://www.gwdg.de/~dwernin/drafts/Werning-TEI_im_Rubensohn_Projekt.pdf



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>