Interlinked: Schriftzeugnisse der klassischen Mayakultur im Spannungsfeld zwischen Stand-off- und Inlinemarkup in TEI-XML

Sikora, Uwe
sikora@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek, Göttingen, Deutschland

Gronemeyer, Sven
sgronemeyer@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland; La Trobe University, Department of Archaeology and History, Australien

Diehr, Franziska
f.diehr@smb.spk-berlin.de
Stiftung Preußischer Kulturbesitz

Wagner, Elisabeth
ewagner@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Prager, Christian
cprager@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Brodhun, Maximilian
brodhun@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek, Göttingen, Deutschland

Diederichs, Katja
katja.diederichs@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Grube, Nikolai
ngrube@uni-bonn.de
Rheinische Friedrich-Wilhelms-Universität, Abteilung für Altamerikanistik, Deutschland

Die computergestützte Erforschung einer nur teilweise erschlossenen Schrift und Sprache wie im Falle der Hieroglyphenschrift der klassischen Mayakultur steht vor zahlreichen Herausforderungen, insbesondere bei der Erfassung der Komplexität von Schrift- und Bildzeugnissen. Gerade historisierende Wissenschaftsdisziplinen sind auf diverse Informationsquellen angewiesen, um ihre Untersuchungsgegenstände nicht bloß in der historischen Vergangenheit sondern auch in der modernen Gegenwartskultur zu vergesellschaften: Informationen zu ursprünglichen Aufstellungsorten von mit Hieroglyphen reliefierten Stelen, Angaben zum aktuellen Aufbewahrungsort dekorierter und beschriebener Keramiken oder jahrzehntealte Zeichnungen monumentaler Tempelinschriften - das Wissen nicht bloß über die historischen Kontexte sondern auch über die wissenschaftliche Arbeit mit und an den antiken Schriftzeugnissen durch Forscher und Sammler bildet den essentiellen Rahmen, um wissenschaftliche Aussagen und Hypothesen zu formulieren, zu überprüfen und zu plausibilisieren.

Die Grundlage dieses Rahmens bilden Modelle, die zum einen eine formalisierte Beschreibung der benötigten Informationen erlauben und zum anderen eben jene Informationen miteinander in Beziehung setzen. Hier stellen Ontologien und domänenspezifische (Daten-)Modelle unerlässliche Hilfsmittel und notwendige Werkzeuge dar, um Wissen über Objekte, die im Fokus des wissenschaftlichen Interesses stehen, einheitlich und vor allem aussagekräftig zu dokumentieren. Vor diesem Hintergrund verfolgt das Projekt ‘Textdatenbank und Wörterbuch des Klassischen Maya’ (TWKM)[1] das Konzept einer ontologisch-vernetzten Datenbeschreibung: Der antike Text als kulturgeschichtliches Artefakt und somit umfassendes Wissens- und Informationsobjekt wird in einzelne unterschiedliche Informationsbereiche unterteilt, die jeder für sich nach besonderen Anforderungen und eigenen Datenmodellen beschrieben aber aufeinander bezogen werden.

Um den Informationsgehalt der antiken Textressourcen differenziert in maschinenlesbarer Form zu beschreiben, werden verschiedene Informationsbereiche auf unterschiedlichen Ebenen voneinander abgegrenzt: Zunächst werden die Schriftträger anhand eines ontologisch-basierten Metadatenschemas in RDF erfasst: Hier werden Kerninformationen zum Schriftträger (Bezeichnung, Zustand, Material und Technik, Maße etc.), seines archäologischen Kontexts sowie auch darüber hinausgehend historische Ereignisse und Persönlichkeiten der Maya-Kultur dokumentiert (Textdatenbank und Wörterbuch des Klassischen Maya 2017). Die Auszeichnung der textlichen Informationen wird separat in TEI-P5 Dokumenten vorgenommen, die mit den entsprechenden in RDF dokumentierten Ressourcen über persistente URIs (Uniform Resource Identifier) verknüpft werden.

Für die Auszeichnung der etwa 10.000 Maya-Texte dient ein projektspezifisches TEI-P5 Anwendungsprofil. Das TWKM unterscheidet hier zwischen den drei

Informationsbereichen (1) Form, (2) Inhalt und (3) linguistische Analyse, die nach jeweils spezifischen Anforderungen separat erschlossen werden. So wird die Erschließung der Form bzw. der Texttopographie, d.h. der strukturellen Anordnung von Schrift- und Bildbereichen auf dem Schriftträger, unabhängig von der linguistischen Analyse und den in diesem Rahmen verwendeten Beschreibungskategorien durchgeführt.

Das TWKM bedient sich hier am Konzept des sog. Stand-off-Markups[2]: die individuierte Auszeichnung von Informationen, die durch Verweise auf andere ausgezeichnete Informationen virtuell[3] in einen gemeinsamen Zusammenhang gebracht werden. Durch diesen Ansatz kann nicht nur die Komplexität der Auszeichnung einzelner Informationsbereiche individuell angehoben bzw. abgesenkt werden. Auch der direkte Einfluss struktureller Anforderungen des XML-Formats auf die Auszeichnung, die in der Praxis häufig zu Problemen und Herausforderungen führen (z.B. die Wohlgeformtheitsregel und die hiermit einhergehende Maßgabe, dass XML-Elemente sich nicht überschneiden dürfen), wird hierdurch minimiert.[4]

Der Inhalt eines Textes lässt sich somit kohärent d.h. in einem logisch-thematischen Zusammenhang beschreiben, obwohl seine topographische formal-strukturelle Anordnung einer gänzlich anderen Logik folgt: z.B. kann ein Text topographisch in geflochtener Form arrangiert und dementsprechend maschinenlesbar beschrieben werden. Die inhaltlich-logische Struktur des Texts wird separat gemäß ihren eigenen Ordnungsprinzipien beschrieben aber mit den topographischen Strukturen in Beziehung gesetzt:

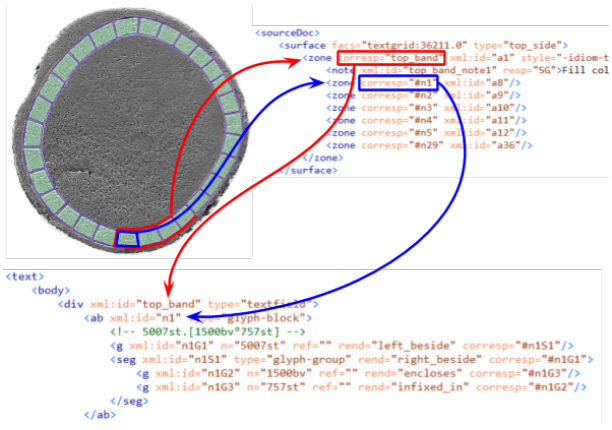

Abbildung 1. Virtueller Zusammenhang zwischen Bildbereich (oben links), Texttopographie (oben rechts) und Inhalt (unten)

Zunächst werden die texttopopgraphischen Eigenschaften des Schriftträgers (tei:sourceDoc) beschreiben. In diesem Kontext werden unterschiedliche Oberflächen (tei:surface) erfasst (z.B. die Flächen einer rechteckigen Stele, die Innen- und Außenseite einer Vase oder Vorder- und Rückseite eines Kodexblattes) und mit digitalen Faksimiles verbunden, die zuvor aus analogen Repräsentationen oder direkt vom Objekt, etwa über 3D-Scanning, erstellt wurden.[5] Diese Oberflächen enthalten einen bis mehrere Textbereiche (tei:zone), die mit Hilfe des Text-Bild-Link-Editors (TBLE) des TextGrid-Laboratorys (Neuroth / Rapp / Söring 2015) mit einzelnen Bereichen des Faksimiles assoziiert wurden. In diesem Rahmen können weitere Informationen, wie bspw. Orientierung oder Ausmaße erhoben werden. Die inhaltliche Erschließung der einzelnen Hieroglyphenblöcke und Schriftzeichen werden wiederum separat in tei:body erfasst.

Die Beschreibung des logo-sylabischen Schriftsystems des klassischen Maya ist aufgrund ihrer Eigenarten und Komplexität eine herausfordernde Aufgabe. Nicht nur aufgrund ihrer vergleichsweise jungen Entzifferungsgeschichte seit den 1950er Jahren gibt es noch eine Reihe von Desiderata bei Lesbarkeit und Verständnis des Schriftsystems. Trotz verschiedener Zeichenkataloge ist die genaue Anzahl von Zeichen noch immer nicht gesichert, damit auch nicht, wie viele Zeichen nicht oder nur unzureichend entziffert sind. Die Inventarisierung war auch bisher eine Herausforderung wegen multipler Klassifizierungsansätze der Schriftzeichen, etwa über die Form (Thompson 1962) oder die Ikonizität (Macri / Looper 2003). Das Projekt hat erstmals eine Systematik anhand einer taxonomischen Beschreibung der Bildung von Graphemvarianten entwickelt und trennt auf diesem Wege auch das bedeutungstragende Zeichen von seiner graphischen Repräsentation. Die wissenschaftliche Deutung des Zeichens, seine linguistische Information, wird dabei über ein System von Kriterien, die durch Aussagenlogik miteinander verbunden sind, qualitativ bewertet. Darauf basierend wird die Plausibilität der jeweiligen Entzifferungshypothese automatisiert eingestuft. Aufgrund der Dynamik der Entzifferungsarbeit ist es demnach nicht möglich, Transliterationen als Basis für die Textauszeichnung zu benutzen (Diehr et al. 2017: 1191-1192).

Deshalb und wegen des Umstands, dass es innerhalb der Mayaforschung keinen Konsens zum Umgang mit den derzeitigen Unicode-Vorschlägen bzw. -Implementierungen zur Mayaschrift gibt (Pallan Gayol / Anderson 2018), verfolgt das TWKM einen eigenen Ansatz zur Schriftbeschreibung gemäß der TEI-P5 Richtlinien.[6] Die Graphe (tei:g) sind in nahezu rechteckigen Blöcken angeordnet (tei:ab[@type='glyph-block']), die ungefähr einem Wort entsprechen. Jedes Zeichen wird mit einer URI-Referenz (@ref) versehen, die auf die konkrete Graphrepräsentation im Zeichenkatalog verweist. Aufgrund der komplexen Graphematik können einzelne Graphe wiederum zu Gruppen im Block zusammengefasst werden (tei:seg[@type='glyph-group']), z.B. bei einer Infigierung. Über weitere Attribute (@corresp und @rend) wird so die Position jedes einzelnen Graphs im Block und in Relation zu den anderen Graphen eindeutig beschrieben.

So kann der hieroglyphische Inhalt eines Texts maschinenlesbar dokumentiert werden, wobei die Entzifferungsgeschichte und hiermit einhergehende Veränderungen und Reinterpretationen einzelner Schriftzeichen auf der Ebene des Zeichenkatalogs abgebildet werden: Sollte sich die Interpretation eines Zeichens beispielsweise hinsichtlich seiner Semantik im Verlauf der fortschreitenden Forschung ändern, so bleibt die hieroglyphische Auszeichnung aller im TWKM erschlossenen Mayatexte unangetastet. Die Veränderung muss lediglich im Zeichenkatalog dokumentiert werden und steht danach als Information für alle Texte zur Verfügung. Gleiches gilt für die Entzifferungsgeschichte eines spezifischen Texts: Sollte die Deutung eines Graphems in einem konkreten Maya-Text revidiert werden, so erfolgt diese Änderung im Zeichenkatalog. Neu klassifizierte Zeichen werden im Zeichenkatalog durch die Relation owl:sameAs[7] mit den betreffenden falsch klassifizierten Zeichen verbunden. Damit ist garantiert, dass vormals falsch klassifizierte Zeichen weiterhin über deren URI auffindbar sind.[8] Die im Korpus referenzierten URIs vormals falsch klassifizierter Zeichen müssen dadurch nicht geändert werden und das kodierte Korpus bedarf keiner nachträglichen Überarbeitung.

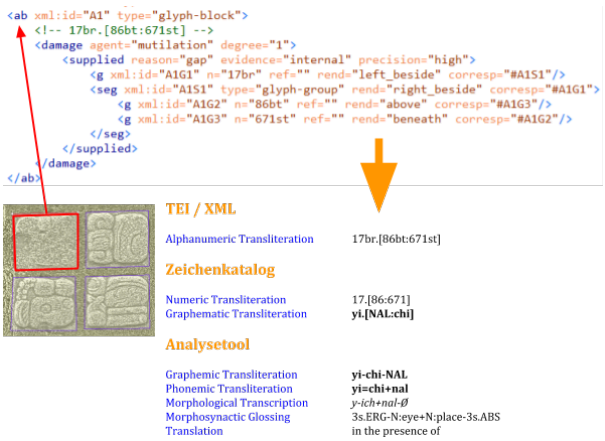

Abbildung 2. Übergang der maschinenlesbaren Textauszeichnung zur linguistischen Analyse

Während die Kodierung des Korpus und die Klassifikation der Zeichen in TextGrid vorgenommen werden, findet die linguistische Analyse der Mayatexte in einem separaten Analysetool statt.[9] Über eine eigens entwickelte Schnittstelle liest das Programm sowohl die TEI/XML-Dokumente als auch die in RDF gespeicherten Transliterationswerte aus TextGrid aus und bereitet sie für den folgenden mehrstufigen Annotationsprozess von Transliteration, Transkription und morphosyntaktischer Glossierung auf. Wo bisher nur vereinzelt Studien vorliegen (vgl. Wichmann 2006), besteht nun das Ziel, eine umfassende Grammatik für das Klassische Maya zu entwickeln. Durch ein Verfahren der mehrstufigen Annotation bei gleichzeitigem Anlegen paralleler Analysepfade, die sich jederzeit auf die entsprechende Belegstelle zurückverfolgen lassen, sind ideale Voraussetzungen zur Durchführung grammatikalischer Bestimmungen und Untersuchungen geschaffen.

Im Fokus des TWKM steht die multiperspektivische Erforschung der Sprache und Schrift des Klassischen Maya. Die über die ganze Welt verstreuten und im Original größtenteils nicht zugänglichen Schriftzeugnisse der antiken Mayakultur werden mittels miteinander verknüpfter Ansätze digital erschlossen. In diesem Rahmen finden zahlreiche Technologien Anwendung, um das derzeitige Wissen über die Maya-Texte nach wissenschaftlichen Standards zu dokumentieren und sukzessive auszuweiten. Durch die kombinierte Verwendung von Ontologien zur Beschreibung und Verknüpfung einzelner Ressourcen auf Metadatenebene einerseits und TEI-P5 XML zur maschinenlesbaren Beschreibung der der Textressourcen andererseits ergibt sich ein engmaschiges Netz aus Informationen zu einem längst vergangenen Kapitel der Menschheitsgeschichte. Ein Netz, das unterschiedliche Zugänge für die wissenschaftliche Forschung sowie für die interessierte Öffentlichkeit bereithält, um tiefe Einblicke in einen vor dem Vergessen bewahrten Teil des kulturellen Erbes zu gewähren - frei, transparent und nachnutzbar.[10]

Fußnoten

1. http://mayawoerterbuch.de
2. Siehe hierzu die Definition in http://uahost.uantwerpen.be/lse/index.php/lexicon/markup-standoff/ . Die Methode wurde erstmalig beschrieben von Thompson / McKelvie (1997).
3. “Virtuell” meint, dass der Zusammenhang nicht durch die hierarchische Struktur der Daten vorgegeben, sondern durch den Verweis strukturell entkoppelter Daten aufeinander hergestellt wird. Es ist dementsprechend ein Zusammenhang, der erst durch die Verarbeitung der verknüpften Daten in einem Informationssystem kultiviert wird, d.h. erst durch die Anwendung von Informationsprozessen wird aus den verknüpften Daten eine zusammenhängende Information erzeugt.
4. Die TEI-Community führt eine anhaltende Diskussion über Vor- und Nachteile des Stand-Off Markups und dessen fortlaufender Entwicklung (vgl. Ba#ski 2010 und Spadini / Turska / Broughton 2015). Dabei sind Flexibilität, Interoperabilität und Nachhaltigkeit der erzeugten Dokumente jene zentralen Faktoren, die in den Diskussionen immer wieder miteinander abzuwägen sind: Durch die Anwendung von Prozessierungsmechanismen, die benötigt werden, um die Daten der unterschiedlichen Dokumente zusammenzuführen, ergeben sich Probleme für die Nachhaltigkeit und Nachnutzung (vgl. Rehm et al. 2010). Dem gegenüber stehen die vielseitigen Möglichkeiten und Potenziale der Datenanreicherung und -verarbeitung: Separate Ressourcen können

unabhängig voneinander bearbeitet und gleichzeitig flexibel ineinander verschränkt werden. Dadurch entstehen semantisch-reichhaltige Dokumente mit hoher Informationsdichte. Diese Vorteile zeigen sich insbesondere im Umgang mit (überlappenden) Hierarchien und Annotationen (z.B. Ide / Romary 2007 und Dipper 2005).

5. Das Projekt bemüht sich um die Nachnutzung von digitalen Faksimiles, die aber für viele Bereiche noch nicht oder in nicht ausreichender Qualität vorliegen. Über Kooperationen werden daher Archive digitalisiert, so etwa die etwa 40.000 Objekte umfassende Fotothek von Prof. Karl Herbert Mayer, Graz, von denen bereits über 5.700 Digitalisate publiziert werden konnten (https://classicmayan.kor.de.dariah.eu/). Für die Arbeiten zum 3D-Scanning siehe https://blog.sketchfab.com/from-the-rainforest-to-virtual-light-scanning-maya-hieroglyphs/ .

6. Diese Herausforderungen sind auch bei anderen antiken, nicht-alphabetischen Schriftsystemen gegeben (vgl. Rossi / De Santis 2019). Zu diesem Zweck wurde zur Vereinheitlichung von Auszeichnungen 2015 die interdisziplinäre Arbeitsgruppe EnCoWS (Encoding Complex Writing Systems) ins Leben gerufen.

7. Zur Definition von owl:sameAs siehe: https://www.w3.org/TR/owl-ref/#sameAs-def .

8. Durch diese Methode werden unter anderem auch Untersuchungen zur Klassifikationsgeschichte der Schriftzeichen ermöglicht.

9. Eine erste Beschreibung des sich in der Entwicklung befindenden Tools ‘ALMAH’ findet sich im Jahresbericht 2017 des Projekts (Grube et al. 2018: 5-7).

10. Die erzeugten Daten werden sukzessive auf dem zukünftigen Projektportal https://www.classicmayan.org/ zugänglich gemacht werden. Des Weiteren werden die Korpusdaten auch frei zugänglich im TextGrid Repository veröffentlicht werden. Die im Projekt entstandenen Schemata sind im öffentlichen Bereich unseres Git-Repositoriums einsehbar und können unter einer CC BY-4.0 Lizenz genutzt werden: https://projects.gwdg.de/projects/documentations/repository .

Bibliographie

Ba#ksi, Piotr (2010): *Why TEI stand-off annotation doesn't quite work and why you might want to use it nevertheless*, in: Proceedings of Balisage 2010. Series on Markup Technologies 5 https://doi.org/10.4242/BalisageVol5.Banski01 [letzter Zugriff: 05.01.2019].

Diehr, Franziska / Gronemeyer, Sven / Prager, Christian / Brodhun, Maximilian / Wagner, Elisabeth / Diederichs, Katja / Grube, Nikolai (2017): *Modellierung eines digitalen Zeichenkatalogs für die Hieroglyphen des Klassischen Maya*, in: **Eibl, Maximilian / Gaedke, Martin (eds.):** *Proceedings der INFORMATIK 2017*, Bonn: Gesellschaft für Informatik, 1185–1196 https://doi.org/10.18420/in2017_120 [letzer Zugriff 1.10.2018].

Dipper, Stefanie (2005): *XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation*, in: Proceedings of Berliner XML Tage 2005 39–50.

Grube, Nikolai / Prager, Christian / Diederichs, Katja / Gronemeyer, Sven / Grothe, Antje / Tamignaux, Céline / Wagner, Elisabeth / Brodhun, Maximilian / Diehr, Franziska (2018): *Arbeitsbericht 2017*, in: Textdatenbank und Wörterbuch des Klassischen Maya, Arbeitsstelle der Nordrhein-Westfälischen Akademie der Wissenschaften und der Künste an der Rheinischen Friedrich-Wilhelms-Universität Bonn http://dx.doi.org/10.20376/IDIOM-23665556.18.pr005.de [letzter Zugriff: 05.01.2019].

Ide, Nancy / Romary, Laurent (2007): *Towards International Standards for Language Resources*, in: **Dybkjaer, Laila / Hemsen, Holmer / Minker, Wolfgang (eds.):** *Evaluation of Text and Speech Systems*, Springer 263–284 https://doi.org/10.1007/978-1-4020-5817-2_9 [letzter Zugriff: 05.01.2019].

Macri, Martha J. / Looper, Matthew G. (2003): *The New Catalog of Maya Hieroglyphs: The Classic Period Inscriptions*, in: The Civilization of the American Indian Series 247. Norman, OK: University of Oklahoma Press.

Neuroth Heike / Rapp, Andrea / Söring, Sibylle (2015): *TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften*, Göttingen: Universitätsverlag Göttingen https://doi.org/10.3249/webdoc-3947 [letzer Zugriff 1.10.2018].

Pallan Gayol, Carlos / Anderson, Deborah (2018): *Achieving Machine-Readable Mayan Text via Unicode: Blending “Old World” Script-encoding with Novel Digital Approaches*, in: **Ortega, Élika / Worthey, Glen / Galina, Isabel / Priani, Ernesto (eds.):** *Book of Abstracts Digital Humanities 2018*, Puentes-Bridges 256–261.

Rehm, Georg / Schonefeld, Oliver / Trippel, Thorsten / Witt, Andreas (2010): *Sustainability of linguistic resources revisited*, in: Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML. Balisage Series on Markup Technologies 6 https://doi.org/10.4242/BalisageVol6.Witt01 [letzter Zugriff: 05.01.2019].

Rossi, Irene / De Santis, Annamaria (2019): *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, Berlin: De Gruyter.

Spadini, Elena / Turska, Magdalena / Broughton, Misha (2015): *TEI Standoff Markup - A work in progress*, in: TEI Members Meeting 2015 urn:nbn:nl:ui:17-f4d0afe1-5c62-4999-8271-7e8cadcd4805 [letzter Zugriff: 05.01.2019].

Textdatenbank und Wörterbuch des Klassischen Maya (2017): Ontology of the Sign Catalogue for Classic Mayan https://classicmayan.org/documentations/idiomschema.html [letzter Zugriff 1.10.2018].

Thompson, J. Eric S. (1962): *A Catalog of Maya Hieroglyphs*, in: The Civilization of the American Indian Series 62. Norman, OK.: University of Oklahoma Press.

Thompson, Henry S. / McKelvie, David (1997): *Hyperlink semantics for standoff markup of read-only documents*, in: Proceedings of SGML Europe.

Wichmann, Søren (2006): *Mayan Historical Linguistics and Epigraphy: A New Synthesis*, in: Annual Review of Anthropology 35: 279-294.