TextGrid
Virtuelle Forschungsumgebung
für die Geisteswissenschaften

Aufbereitung der Enzyklopädien der Digitalen Bibliothek (R 5.5.1)

Version 31.05.2015
Arbeitspaket 5.4
Verantwortlicher Partner Universität Würzburg

TextGrid
Virtuelle Forschungsumgebung in den Geisteswissenschaften

Ein Projekt im D-GRID

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Projekt: TextGrid – Institutionalisierung einer Virtuellen Forschungsumgebung in den Geisteswissenschaften
BMBF Förderkennzeichen: 01UA1203<Buchstabe>
Laufzeit: Juni 2012 bis Mai 2015

Dokumentstatus: < final >
Verfügbarkeit: <öffentlich >
Autoren:

<Betz Katrin>

Inhaltsverzeichnis

2 Überblick 4
3 Struktur der Ausgangsdaten 4
3.1 Ordnerstruktur 4
4 XML-Markup der Ausgangsdaten 4
4.1 Zielstruktur 6
5 Aufbereitung der Enzyklopädien 6
5.1 Vorarbeiten 6
5.2 Umsetzung des Workflows 7

1 Überblick

Während der letzten Projektförderphase wurden die Enzyklopädien der Digitalen Bibliothek aufbereitet. Insgesamt wurden 22 Enzyklopädien für den Datenbankimport in das TextGrid-Repository vorbereitet. Im Folgenden soll die Struktur der Originaldaten, die Zielstruktur, die einzelnen Bearbeitungsschritte und die Abbildung der Datenstruktur auf das TextGrid-Metadatenmodell dargestellt werden.

2 Struktur der Ausgangsdaten

2.1 Ordnerstruktur

Für jede Enzyklopädie ist ein eigener Ordner vorhanden. Innerhalb dieser Ordner ist die Aufteilung der einzelnen Teile einer Enzyklopädie uneinheitlich: Die Ordner können eine oder mehrere Dateien beinhalten. Enthält ein Ordner nur eine Datei wie z.B. beim Damenkonversationslexikon, sind in dieser die Enzyklopädie-Einträge, die Metadaten der Enzyklopädie sowie eventuelle Nachtragsbände oder Vorwörter enthalten.
Bei anderen Enzyklopädien sind im Ordner mehrere Dateien enthalten. Hier sind die enzyklopädischen Einträge in einer Datei gespeichert, in einer weiteren Datei finden sich die Metadaten zusammen mit weiteren Informationseinheiten wie Vorwörtern oder Anhängen anderer Art (z.B. Beziehungs- und Nachweisartikel, Verzeichnisse von Bildern etc.). Als einzelne Datei können außerdem Nachtragsbände vorhanden sein.
In weiteren Fällen wie z.B. dem Heiligenlexikon sind zusätzlich die enzyklopädischen Einträge selbst auf zwei Dateien verteilt: Eine erste Datei enthält die Einträge A-L, eine zweite Datei enthält die restlichen Einträge.

3 XML-Markup der Ausgangsdaten

Auf einer ersten groben Strukturierungsebene sind die Enzyklopädien durch die Elemente *catdiv*, *articlegroup* und *article* gegliedert. *catdiv* und *articlegroup* sind hierarchisch übergeordnete Elemente, die zumeist mehrere *article*-Elemente umfassen. Die Verwendung dieser Strukturelemente ist insgesamt jedoch nicht einheitlich: z.B. kann eine *articlegroup* ein *catdiv* enthalten und umgekehrt. Teilweise dienen außerdem rekursiv verwendete *article*-Elemente in den Anhangsdateien zur weiteren Untergliederung.
Das *article*-Element wird zusätzlich verwendet um die einzelnen Lemmata zu kodieren. Eine mögliche Ausprägung der Struktur der Originaldaten sei im Folgenden am Beispiel des Damenkonversationslexikon verdeutlicht.

```
<catdiv name="Lizenz: Gemeinfrei" >
        <articlegroup name="-" sort="yes“>
                […]
        </articlegroup>
```

```
<articlegroup name="M">
        <article>
                <lem>Vorrede</lem>
                […]
        </article>
        <article>
                <lem>Stahlstiche</lem>
                […]
        </article
</articlegroup>
<articlegroup name="A" sort="yes">
        <catdiv name="Lexikalischer Artikel">
                <article>
                        <lem>A</lem>
                        <text><p><lemfloat>A</lemfloat>, der erste Buchstabe im Alphabete, ist
zugleich der erste klare Sprachlaut, welchen die menschliche Zunge aussprechen lernt, der zuerst vom
Kinde gelallt wird. Was unseren Urältern von höchster Bedeutung und Heiligkeit war, bezeichneten sie mit
A: so die Indier das Licht, die Deutschen das Wasser, die Griechen die Luft. à auf Briefen, in Rechnungen,
Preiscouranten etc. bedeutet in, zu, für. <b>A</b> in der Musik bezeichnet die sechste diatonische
Klangstufe der ersten oder tiefsten Octave unseres Tonsystems. <b>A-dur</b> siehe Tonarten.</p>
                        </text>
                </article>
        </catdiv>
</articlegroup>
</catdiv>
```

Erläuterungen:

- articlegroup name="-": enthält Informationen, die von zeno für die Einstiegsseite auf der Internetseite verwendet werden. Dieses Element wird nicht mitverarbeitet.
- articlegroup name="M": enthält Vorwörter oder Anhänge. Diese *articlegroup* kann durch die rekursive Verwendung des Elementes *article* weiter untergliedert sein.
- articlegroup name="A": Enthält die einzelnen Enzyklopädieeinträge, die wiederum durch das Element *article* kodiert werden.
- Die Strukturierung der anderen Enzyklopädien kann von der beschriebenen Struktur abweichen: Im Anhang können unterhalb der *articlegroup*-Ebene weitere *article*-Elemente zur Strukturierung dienen, der Hauptteil kann – z.B. wenn ein Nachtragsband vorhanden ist – durch mehrere *articlegroup* strukturiert sein etc.
- *article*: Einzelne Lemmata sind durch das *article*-Element kodiert. Innerhalb der Lemmata ist das Lemma als *lem* oder *lemfloat* kodiert. Die Bedeutungserklärungen werden von *text* umfasst. Daneben ist das Markup innerhalb der Einträge rudimentär und enthält außer der Markierung von Absätzen nur layoutorientierte Informationen. Semantische Einheiten wie

z.B. eine Untergliederung der Einträge in unterschiedliche Bedeutungserklärungen sind nicht erfasst. In einigen Fällen können einzelne Bedeutungen zwar durch eine als Überschrift markierte Einheit getrennt sein. Auch hierbei ist die Überschrift aber nur als solche markiert, Überschrift und zugehörige Bedeutung sind nicht als zusammengehörige Einheit markiert.

- Desweiteren sind in den Daten verschiedene Verlinkungen kodiert, die auf Fußnoten oder auf andere *article*-Elemente in der Enzyklopädie selbst oder in den Anhangsdateien verweisen.

3.1 Zielstruktur

Die Daten sollen am Ende des Workflows als valide *TEI*-Dateien vorliegen. Die Grobgliederung der Enzyklopädien soll durch *TEI* und *teiCorpus*-Elemente kodiert sein. Es soll eine weitere Gliederungsebene hinzugefügt werden, sodass die Buchstabenstrecken der Enzyklopädien durch *TEI*-Elemente erfasst sind.
Auf der Ebene der Feinauszeichnung sollen das layoutorientierte Markup soweit wie möglich auf ein semantisches Markup abgebildet werden. Hierbei wird jedoch nur das vorhandene Markup transformiert, eine Aufwärtstransformationen findet nicht statt.
Die Metadatendaten sollen in strukturierter Form im *teiHeader* kodiert sein.
Die Datenstruktur soll auf das TextGrid-Metadatenmodell abgebildet werden und auf dieser Basis für den Datenimport in das Repository gesplittet werden.

4 Aufbereitung der Enzyklopädien

4.1 Vorarbeiten

1. Für jede einzelne Enzyklopädie wurde ermittelt, ob sie einen Anhang o.Ä. enthält, ob dieser Anhang innerhalb oder außerhalb der Hauptdatei liegt, und es wurde die interne Strukturierung des Anhangs durch *articlegroup* oder *article*-Elemente festgehalten.
2. Manuelles Kodieren der Metadaten als *teiHeader*: Die Metadaten für die einzelnen Enzyklopädien enthalten zum Teil recht komplexe Informationen über das Erscheinen einzelner Bände zu unterschiedlichen Zeitpunkten oder auch in unterschiedlichen Verlagen. Diese Informationen sind in den Originaldaten nur als String vorhanden. Für jede einzelne Enzyklopädie wurde deshalb manuell ein *teiHeader* kodiert, in dem alle verfügbaren Metadaten in strukturierter Form erfasst sind. Die kodierten *teiHeader* wurden in einer Hilfsdatei gespeichert, die denselben Dateinamen trägt wie die Enzyklopädie. Im Rahmen der Transformationen werden die erstellten *teiHeader*-Hilfsdateien in die Gesamtenzyklopädie kopiert.
3. Um die Enzyklopädien auf sinnvolle Art und Weise in das TextGrid-Repository zu importieren, war es notwendig, die Enzyklopädien nach Buchstabenstrecken zu splitten. Hierzu musste eine Möglichkeit gefunden werden, jeweils den ersten Eintrag einer Buchstabenstrecke als Markierungsgrenze herauszufiltern. Erschwert wurde dieses Ziel, weil in den Originaldaten nicht nur das Lemma an sich, sondern das Lemma zusammen mit seinem Artikel und/oder weiteren Attributgruppen kodiert wurde (z.B. die Quadratur des Kreises). Deshalb wurden die Buchstabenstrecken halbautomatisch extrahiert: In einem

ersten

Schritt wurden mit Hilfe einfacher Heuristiken mögliche Buchstabenstreckengrenzen ermittelt und in einer Zwischendatei gespeichert. Die in der Zwischendatei gespeichert Ergebnisse wurden manuell nachkorrigiert, sodass am Ende eine Datei entsteht, in der jeweils die ersten Einträge einer Buchstabenstrecke korrekt erfasst sind. Diese Datei wird während den folgenden Verarbeitungsprozessen ausgelesen und so in den Gesamtworkflow integriert.

4. In den Originaldaten sind mehrere Arten von Verlinkungen zu finden. Eine Verlinkung kann z.B. auf ein anderes Lemma verweisen, auf Bilder oder aber auf eine Fußnote. Die Verlinkung kann entweder innerhalb eines Lemmas eindeutig sein oder innerhalb der Hauptdatei. Manche Verlinkungen verweisen aber auch auf ein Ziel in der ausgelagerten Appendix-Datei. Das Verweisziel der zeno-Verlinkung wird nicht durch Id-Referenzen festgelegt, sondern durch einen Pfadausdruck, der auf den Zielartikel verweist und der im Falle von Fußnoten um eine Fußnotennummerierung ergänzt werden kann. Jede Art dieser Verlinkungen wurde durch eine ID-Referenzierung ersetzt.
   Für die verschiedenen Verlinkungstypen werden in den Originaldaten zwar teilweise unterschiedliche Verlinkungselemente verwendet, die Anwendung dieser Elemente ist aber nicht konsistent. Auch für die Verlinkung musste deshalb jede Enzyklopädie einzeln analysiert werden, um zu ermitteln, welche Art von Verlinkung wie kodiert wird.

4.2 Umsetzung des Workflows

Die Enzyklopädien wurden in mehreren Transformationsschritten verarbeitet. Diese Schritte sind modularisiert und bauen aufeinander auf. Die folgende Liste gibt einen Überblick über die einzelnen Verarbeitungsschritte und deren Abhängigkeiten

1. ID-Vergabe

In einem ersten Transformationsschritt werden IDs vergeben, die innerhalb aller Enzyklopädien einzigartig sind. Die ID-Vergabe wird in den Folgeschritten für die Auflösung der Verlinkung benötigt.

2. Erstellen der TEI/teiCorpus-Struktur

Ziel des zweiten Transformationsschrittes ist es, sowohl für den Hauptteil der Enzyklopädien als auch für ihre Anhänge eine sinnvolle *TEI/teiCorpus*-Gliederungsstruktur zu generieren. Dabei werden einerseits alle Hauptteile der Enzyklopädien in einem Ordner ausgegeben, die Anhänge in einen anderen, wobei Anhang und Hauptteil jeweils denselben Dateinamen erhalten. Die einheitliche Trennung von Anhang und Hauptteil ist die Voraussetzung für die weitere Verarbeitung der Enzyklopädien. Durch die Trennung von Anhang und Hauptenzyklopädie können für diese in den weiteren Transformationen eigene Stylesheets aufgerufen werden.
Neben der Trennung von Anhang und Hauptteil der Enzyklopädien wird in diesem Transformationsschritt außerdem die grobe Binnenstruktur dieser Teile erstellt. Einerseits werden eventuelle Untereinheiten des Anhangs durch *teiCorpus*- und *TEI*-Elemente gebündelt. Wie oben

beschrieben ist die Verschachtelung der zeno-Gliederungselemente *article*, *articlegroup* und *catdiv* z.T. uneinheitlich. Deshalb wurde für die Erstellung der Binnenstruktur ein Hauptstylesheet mit allgemeingültigen Regeln verfasst. Anschließend wurden für jede einzelne Enzyklopädie Stylesheets integriert, die gegebenenfalls die allgemeingültigen Regeln überschreiben und ergänzen.
Neben der Abbildung der zeno-Gliederungseinheiten auf eine *TEI-teiCorpus*-Struktur wurden die Enzyklopädien in diesem Schritt außerdem in Buchstabenstrecken untergliedert, sodass jede Buchstabenstrecke von einem *TEI*-Element umfasst ist. Hierbei werden die erstellten Hilfsdateien eingelesen und auf der Basis der enthaltenen Informationen die Grenzen der Buchstabenstrecken ermittelt.
Bei der Erstellung der groben *TEI-teiCorpus*-Struktur werden außerdem die manuell erstellten *teiHeader* in die Daten integriert. Das Einbinden der *teiHeader* erfolgt über einen Abgleich der Dateinamen.
Die *TEI*/*teiCorpus*-Struktur und die in den *teiHeader* kodierten Metadaten sind die Grundlage für das Splitten der Dateien gemäß des TextGrid-Metadatenmodells.

3. Mapping der Feinstruktur auf TEI

Aus Gründen der Übersichtlichkeit wurde die Abbildung der zeno-Elemente unterhalb der Gliederungsebene nach TEI in einem eigenen Modul vorgenommen. Das Modul ist untergliedert in Regeln für die Verarbeitung der Anhänge und Regeln zur Verarbeitung der Lexikoneinträge. Wie beim vorherigen Transformationsschritt wurden auch hier auf Grund von Unregelmäßigkeiten in den Daten sowohl für den Anhang als auch für Hauptenzyklopädie ausgehend von einem generellen Hauptstylesheet Ausnahmeregeln für jede Enzyklopädie formuliert.
Zur Bearbeitung der Elemente unterhalb der Gliederungsebene zählt außerdem die Auszeichnung und Auflösung der Verlinkung. Die Referenzierung von Verweiszielen über Pfadausdrücke wird durch die Referenzierung von IDs ersetzt. Auch hierbei wurde wieder ein generelles Stylesheet verfasst, dessen Regeln gegebenenfalls durch spezielle Regeln für die einzelnen Enzyklopädien überschrieben werden.

4. Abbildung der Struktur auf das TextGrid-Metadatenmodell/Splitting

Das TextGrid-Metadatenmodell stellt für den Datenimport die Objekttypen *item*, *edition*, *work* und *collection* zur Verfügung. Jedes der Objekte besteht jeweils aus einer Metadatendatei und einer Inhaltsdatei. Ein *item* kann als Inhaltsdatei eine XML-Datei, eine *jpeg*-Datei oder aber eine Aggregation enthalten. Die Metadatendatei des Items enthält als obligatorische Angabe neben einem Titel und beim Import automatisch generierten Daten nur den Rechteinhaber. Ein *work*-Objekt besteht aus einer leeren Inhaltsdatei, die zugehörige *work*-Metadatendatei enthält Informationen über den Autor/Herausgeber, das Entstehungsdatum etc. Zu einem *work*-Objekt muss notwendigerweise zusätzlich ein *edition*-Objekt erstellt werden. Die Inhaltsdatei dieses Objektes ist im Falle der digitalen Bibliothek eine Aggregation, die Bestandteile (Items oder weiter untergliedernde Aggregationen) auflistet. Besteht das Werk aus mehreren Einzelwerken und/oder mehreren *items*, werden diese aggregiert. In der Metadatendatei der *edition* sind schließlich die

Angaben zu der Digitalisierungsquelle kodiert. Die Inhaltsdatei der *edition* verweist auf das *item* und kann zusätzlich weitere Unterwerke oder *items* aggregieren. Abbildung 5 gibt einen Überblick über die verwendeten Objekte.

Abbildung 1: Übersicht über die verwendeten Objekte des TextGrid-Metadatenmodells

Editionen: - Metadatendatei: Angaben zur Digitalisierungsquelle - Inhaltsdatei: Aggregiert die in der Edition enthaltenen Objekte	Aggregationen: - Metadatendatei: Rightsholder - Inhaltsdatei: Aggregiert verschiedene andere Objekte
Werk: - Metadatendatei: enthält Titel des Werkes und Entstehungsdatum - Inhaltsdatei: leer	Item: - Metadaten: Rightsholder - Inhaltsdatei: TEI-Dokument/jpg

Zur Überführung der *TEI*/*teiCorpus*-Struktur in das TextGrid-Metadatenmodell wurden die einzelnen Enzyklopädien gesplittet. Dabei wurden folgende Abbildungsregeln entwickelt:

- Jedes *TEI*-Element, das eine Buchstabenstrecke erfasst, wird als *item* importiert. Dazu müssen die Dateien für ein *item*-Objekt, erstellt werden.[1] Die *item*-Inhaltsdatei ist die *TEI*-Datei selbst. Deren Metadatendatei wird auf der Basis der Metadaten des *teiHeaders* generiert.
- Eine Enzyklopädie als Ganzes wird als *Edition* erfasst, die entweder die untergeordneten Buchstabenstreckendateien aggregiert oder aber auf einer ersten Ebene die Hauptenzyklopädie und den Anhang aggregiert, die dann wiederum die untergeordneten Buchstabenstrecken bzw. Unterkapitel des Anhangs aggregieren. Die Metadaten der *edition* werden auf der Basis der im *teiHeader* kodierten Daten erzeugt.
- *teiCorpus*-Elemente, die selbst nicht als Werk gelten wie z.B. Strukturelemente die die einzelnen Unterkapitel des Anhangs umfassen, werden auf ein *item* abgebildet, dessen Inhaltsdatei wiederum eine Aggregation ist.
- *teiCorpus*-Elemente werden vollständig durch die Aggregationsmechanismen des TextGrid-Metadatenmodells ersetzt.
- Grafiken, die in der Enzyklopädie verwendet werden, werden der Übersichtlichkeit halber nicht bei den Buchstabenstrecken, sondern in einer dedizierten Aggregation auf Ebene der Edition aggregiert.

5. Link-Rewriting

[1] Zur logischen Struktur und zur Verbindung zwischen den einzelnen Objekttypen vgl. <https://dev2.dariah.eu/wiki/display/TextGrid/Metadata>.

Nach dem Splitten der Dateien gemäß des Metadatenmodells musste die Verlinkungen erneut bearbeitet werden, da sich das Verweisziel nach dem Splitten eventuell in einer anderen Datei befindet. Hierzu wurden alle möglichen Verweisziele in einem Zwischenschritt zusammen mit dem jeweiligen Dateinamen in eine Mappingtabelle geschrieben. In einem weiteren Schritt wird diese Mappingtabelle eingelesen und die Verlinkung um den passenden Dateipfad ergänzt.

6. Validierung

Nach der Verarbeitung der Enzyklopädien werden die Ergebnisse in zwei Validierungsschritten überprüft. In einem ersten Schritt wird überprüft, ob es sich bei allen Ergebnisdateien um valide *TEI*-Daten handelt. In einem weiteren Schritt wird über verschiedene Diffing-Prozesse die Vollständigkeit der Texte überprüft.