Erweiterung eines Forschungsdaten-repositoriums um ein Modul für die Nachnutzbarkeit und Analyse von Textressourcen

Schneider, Gerlinde
gerlinde.schneider@uni-graz.at
Universität Graz, Österreich

Vasold, Gunter
gunter.vasold@uni-graz.at
Universität Graz, Österreich

Digitale Editionen stellen als digitalisierte und tiefenerschlossene Textressourcen eine wertvolle Quelle zur Nachnutzung innerhalb großflächiger linguistischer und literaturwissenschaftlicher Analysen dar (Rybicki, 2019). Zusätzlich werden innerhalb von digitalen Editionsprojekten selbst immer öfter textanalytische Verfahren eingesetzt.

Das am Zentrum für Informationsmodellierung der Universität Graz entwickelte und betriebene Repositorium GAMS (Geisteswissenschaftliches Asset Management System)[1] umfasst als Forschungsdateninfrastruktur Daten von mehr als hundert Forschungsprojekten aus verschiedenen Wissenschaftsbereichen. Digitale Editionen und Textsammlungen machen dabei, neben neben digitalen Sammlungen aus dem Kulturerbebereich, den Großteil der im Repositorium vorhandenen Bestände aus.

Um die bereits im Repositorium vorhandenen Textressourcen in geeigneten Formaten nachnutzbar bereitzustellen, beziehungsweise diesen Aspekt im Zuge laufender und zukünftiger Projekte berücksichtigen zu können, wurden während der letzten Monate Adaptierungen an der GAMS-Infrastruktur vorgenommen, die mit diesem Poster erläutert und dargestellt werden sollen.

Technischer Hintergrund

GAMS ist eine registrierte[2], trusted[3] Repositoriumsinfrastruktur, die auf der Free and Open Source Software Fedora Commons[4] basiert. Sie setzt auf eine OAIS-konforme Architektur und verfolgt eine weitgehend XML-basierte Content-Strategie. GAMS ermöglicht seinen Benutzer*innen die Verwaltung und Veröffentlichung von Ressourcen aus Projekten mit permanenter Identifizierung und Anreicherung mit Metadaten. Ein speziell entwickelter Client (*Cirilo*) stellt Funktionalitäten für Massenoperationen an den gespeicherten Objekten zur Verfügung. (Stigler/Steiner 2018)

Objekt Modell

Content Models definieren komplexe digitale Objekte, die dem Fedora-Objektmodell entsprechen. Sie sind speziell auf die Anforderungen, die Forschungsdaten aus unterschiedlichen geisteswissenschaftlichen Bereichen an Langzeitarchivierung und Datendissemination stellen, ausgelegt. Für wissenschaftliche Editionen wird beispielsweise ein speziell entworfenes *TEI Content Model* eingesetzt.

Jedes Modell enthält einen primären Datenstrom, der die Inhaltsdaten des Objekts enthält, zum Beispiel ein TEI-Dokument. Zusätzliche Datenströme können Metadaten (z.B. Dublin Core), weitere Inhaltsdaten oder aus dem primären Datenstrom derivierte Daten enthalten (z. B. aus dem TEI-Dokument extrahierte RDF Daten).

Für die jeweiligen Modelle definierte Services kombinieren und transformieren Datenströme zu Präsentationsinhalten, auf die in verschiedenen Ausgabeformaten über im Content Model definierte Schnittstellen zugegriffen werden kann. Ein häufig verwendetes Format zur Dissemination ist HTML, was die Präsentation der Daten über eine dynamisch erzeugte Webseite ermöglicht.

Contexte, als spezielle Containerobjekte, ermöglichen es, einzelne Inhaltsobjekte in größere Einheiten zusammenzufassen und zu organisieren. Sie enthalten wiederum eigene Datenströme und Disseminationsmethoden.

Anpassungen für Textressourcen

Zur Verwaltung und Bereitstellung von im GAMS vorliegenden Textressourcen wie auch dezidiert linguistischen Forschungsdaten wurde das bestehende TEI Content Model angepasst und erweitert. Über den Cirilo Client können Objekte als Text- bzw. Sprachressourcen gekennzeichnet werden. So gekennzeichnete Objekte werden dann automatisch mit für das CMDI Framework (Goosen et al., 2015) aufbereiteten, komponentenbasierten Metadaten und einem eigenen Handle Identifier versehen. Solche Daten können dann geharvestet werden und über das Virtual Language Observatory (Van Uytvanck et al., 2012) der CLARIN Infrastruktur[5] als Sprachressource gefunden werden.

Der OAI-Endpoint des Repositoriums wurde dementsprechend angepasst. Auf inhaltlicher Ebene wurde ein XML-basiertes Konfigurationsformat eingeführt, das es erlaubt, auf den Ausgangsdaten operierende Pipelines bzw. Toolchains zu definieren und als Massenoperation zu triggern. Ein Anwendungsfall hierfür ist beispielsweise Preprocessing zur Aufbereitung der Daten für darauf

aufbauende Analyseschritte. Per Default wird eine, auf dem an der Österreichischen Akademie der Wissenschaften entwickelten XSL-Tokenizer (Schopper, 2019) basierende Pipeline ausgeführt, was einerseits ein tokenisiertes TEI-Dokument als separaten Datenstrom im Objekt erzeugt, und andererseits die Daten als Plain Text, im von den im Rahmen von CLARIN entwickelten Weblicht Tools verwendeten *Text Corpus Format* (TCF)[6], sowie im von gängigen Corpus Tools verwendeten *Vertical*-Format bereit stellt. Diese Daten können daraufhin direkt mit den genannten Tools verarbeitet und analysiert werden. Wie der Tokenizer ist auch die Pipeline selbst projektbezogen anpassbar und kann aus mehreren Transformationsschritten bestehen, darunter beispielsweise auch die Möglichkeit, die jeweiligen Texte via TreeTagger (Schmid, 1995) zu annotieren.

Die über diese Pipelines erzeugten Datenformate können benutzerdefiniert gekapselt und mit dem primären TEI-Datenstrom als Objekt im Repositorium langzeitarchiviert werden. Durch die Speicherung der Verarbeitungspipeline gemeinsam mit den zu verarbeitenden Daten wird jeder Prozessierungsschritt dokumentiert und nachvollziehbar gemacht, was wesentlich für die Nachnutzung ist.

Für die Aggregation mehrerer TEI Objekte zu einem verarbeitbaren Corpus wurde ein sogenanntes *Corpus Context Model* geschaffen. Diesem Modell entsprechende Objekte können vom Benutzer selbst über den *Cirilo* Client angelegt und mit entsprechenden Textobjekten befüllt werden.

Dieser spezielle Context stellt über die entsprechenden Datenströme Dublin Core wie auch CMDI Metadaten bereit. Die VERTICAL Datenströme der zugeordneten TEI-Objekte werden zu einem Datenstrom aggregiert, welcher bei Bedarf in einem Corpus Management System (Vorzugsweise NoSketch Engine) indiziert und über dieses abgefragt werden kann. Das aggregierte Corpus kann außerdem als ZIP-Datei heruntergeladen werden.

Die beschriebenen Features stehen für sämtliche im Repositorium vorhandenen Textressourcen, also nicht nur für genuin linguistische Daten zur Verfügung. Das bedeutet, dass etwa bestehende, im Repository vorhandene Digitale Editionen mit geringem Aufwand auch für linguistische Analysen verfügbar gemacht werden können.

Fußnoten

1. http://gams.uni-graz.at/
2. https://www.re3data.org/
3. https://www.coretrustseal.org
4. Flexible Extensible Digital Object Repository Architecture, https://duraspace.org/fedora
5. European Research Infrastructure for Language Resources and Technology, https://www.clarin.eu/
6. http://weblicht.sfs.uni-tuebingen.de

Bibliographie

Goosen T., et al. 2015. CMDI 1. 2: Improvements in the CLARIN Component Metadata Infrastructure. Selected papers from the CLARIN 2014 Conference, pp. 36-53. https://hdl.handle.net/20.500.11755/91536b93-31cb-4f4a-8125-56f4fe0a1881 .

Rybicki, J. (2019). Keynote at the 2019 TEI Conference and members meeting “What is text, really? TEI and beyond”.

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German . Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schopper, D. (2019). XSLT-Tokenizer (Software), https://github.com/acdh-oeaw/xsl-tokenizer .

Van Uytvanck, D., et al. (2012). Semantic metadata mapping in practice: the Virtual Language Observatory. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 1029-1034. http://www.lrec-conf.org/proceedings/lrec2012/pdf/437_Paper.pdf .