

GAMS: Geisteswissenschaftliches Asset Management System

Funktionalitäten und Prozesse eines FEDORA-basierten Digitalen Archivs

Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities
(ZIM-ACDH)

Einleitung

Moderne Infrastrukturen für die Verwaltung und Bereitstellung von geisteswissenschaftlichen Daten stehen in einem Spannungsfeld: Einerseits soll die Nachhaltigkeit und Verfügbarkeit im Sinne der Langzeitarchivierung gewährleistet werden, andererseits stellt sich die Anforderung einer flexiblen und individualisierbaren Nutzung vorgehaltener Inhalte. Mit dem auf FEDORA basierenden Geisteswissenschaftlichen Asset Management System (GAMS) und dem am ZIM-ACDH dafür entwickelten Client *Cirilo* wird der Archivierungsaspekt mit einer Präsentations- und Managementschicht verbunden, sodass idealerweise alle genannten Aufgaben von einer gemeinsamen Infrastruktur erfüllt werden können. Hier handelt es sich um ein institutionelles Repository, das eine Vielzahl von Digitalen Editionen, Bild- und Quellensammlungen enthält.

Die Trennung von Inhalt und Präsentation als grundlegendes Merkmal XML-basierter Textformate impliziert einerseits einen hohen Grad an Flexibilität bei der Analyse und der Transformation so vorliegender (Text-)Daten in unterschiedliche Darstellungs- und Präsentationsformen, erfordert gleichzeitig aber auch darauf abgestimmte Verarbeitungsworkflows im täglichen Umgang mit diesen Daten.

In der Definition des Open Archival Information System Reference Model (OAIS-Referenzmodell) bemüht man sich um eine Standardisierung von Workflows in Digitalen Archiven. Es wurde 2002 als ISO-Norm akzeptiert und hat sich seither breit durchgesetzt. Dabei gibt das Referenzmodell lediglich einen formal-strukturellen Rahmen für das Design von Digitalen Archiven vor und definiert einen offenen Standard für ein dynamisches, erweiterbares Archivinformationssystem mit dem Anspruch auf Allgemeingültigkeit. Es fokussiert auf die nachhaltige Nutzbarmachung digital vorliegender Bestände und dies unter der Berücksichtigung sich ständig verändernder Technologien. In GAMS wurde versucht, zentrale Aspekte dieses Modells zu realisieren.

GAMS

GAMS ist ein Asset Management System zur Verwaltung, Publikation und Langzeitarchivierung digitaler Ressourcen. Mit seinem objektorientierten Ansatz bietet das Open Source Projekt FEDORA einen geeigneten Rahmen für die Implementierung des OAIS-Referenzmodells im Rahmen von GAMS: Ein flexibles, erweiterbares Framework zur

Speicherung, zum Management und zur Dissemination von komplexen digitalen Objekten, das die Versionierung der Inhalte und eine METS-basierte Serialisierung der Objekte für die Langzeitarchivierung ermöglicht. Das der FEDORA-Architektur zugrundeliegende Datenmodell unterstützt die enge Verknüpfung von Inhalt und Metadaten und ermöglicht es so, zusammengehörige Informationen auf einfache und nachvollziehbare Weise gemeinsam zu verwalten. FEDORA bringt dabei eine Vielzahl von nützlichen Merkmalen mit, als Beispiel ist hierbei die standardisierte OAI-PMH-Schnittstelle zu nennen.

In der einfachsten Form stellt eine FEDORA-Objektinstanz lediglich die Aggregation einzelner in einem Klassenmodell beschriebener Inhalte dar und bietet somit eine Speicherinfrastruktur in der zusammengehörige Daten gemeinsam verwaltet werden können: Ein TEI-Dokument und XSL-Stylesheets zur Erzeugung von Präsentationsformen (Analyseansichten, Fassungssynopsen u.Ä.) des Objekthinhalts, ein RDF-Metadatensatz zur Beschreibung semantischer Relationen des Objektes, Bild- und Bildmetadaten etwa der Faksimiles einer im TEI-Dokument transkribierten Handschrift sowie PREMIS-basierte Langzeitarchivierungsmetadaten, die Eigenschaften des Lebenszyklus eines Objektes beschreiben u.v.m. In einer funktionalen Sichtweise beschreibt das FEDORA-Objektmodell aber auch (webservicebasierte) programmgesteuerte Abläufe, die auf den Daten der Objekte operieren: Etwa eine XSLT-Transformation zur dynamischen Erzeugung einer bestimmten Objektsicht beim Zugriff auf das Objekt oder die automatisierte Extraktion semantischer Relationen aus einem TEI-Dokument beim Upload einer Datei.

Cirilo Client

Cirilo ist eine Java-Applikation, die über das Management API (API-M) auf das FEDORA-Repository zugreift und die vor allem für Massenoperationen wie Ingestprozesse oder Ersetzungsvorgänge entsprechende Funktionalitäten mitbringt. Der Client unterstützt den Ingest sowohl aus dem Dateisystem, einer eXist Datenbank oder einer Excel-Tabelle. Während des Ingestprozesses werden Metadaten aber auch semantische Informationen regelbasiert und automatisch aus dem Inhalt extrahiert und in die neu erzeugten Objekte übernommen (zum Beispiel im Dublin Core Format).

Über das Design der oben vorgestellten Inhaltsmodelle (Objektklassen) können in einem FEDORA-Repository komplexe (Objekt-)Klassenhierarchien konstruiert werden. Inhaltsmodelle beschreiben dabei nicht nur die inhaltliche Struktur von digitalen Objekten, sondern können über WSDL (Web Service Description Language) auch so genannte Disseminatoren an die Daten eines Archivobjektes binden. Damit ist es möglich Services mit Datenströmen im Objekt zu verbinden und dynamische Abläufe (z.B. Migrationsvorgänge, XSLT-Transformationen) sowie eventgesteuerte Workflows (z.B. eine Textnormalisierung im Uploadvorgang eines TEI-Dokumentes) zu beschreiben und in die Objektlogik zu verpacken. Der Vorteil dieses Konzepts liegt darin, dass komplexe Datenquellen und Workflows einfach abgewickelt werden können.

Derzeit bietet der Client mehrere Inhaltsmodelle für spezifische Zwecke an, besonders umfangreich ist das TEI-Modell. Der Ingest von TEI-Objekten kann flexibel konfiguriert werden: semantische Informationen können extrahiert werden, referenzierte Bilder werden zum Objekt hinzugefügt oder mit den Textdaten in Verbindung stehende Ontologiekonzepte werden aufgelöst. Ein Abfrageobjekt ermöglicht Suchen mit

spezifischen Parametern in dem in FEDORA integrierten Mulgara Triplestore. Mit Hilfe von Ontologie- und Abfrageobjekten können dynamische Register erstellt werden. Zur Strukturierung des Bestandes können Kollektionen und hierarchische Sammlungen angelegt werden. Einige Inhaltsmodelle sind für spezifische Primärquellen optimiert, beispielsweise METS/MODS, HTML, PDF, BibTeX oder externe Ressourcen, die über URL erreichbar sind. Ein Modell für linguistische Primärquellen wird derzeit in Kooperation mit dem ICLTT der ÖAW entwickelt. Zusätzlich wird getestet, wie kontrollierte Vokabularien und Thesauri wie geonames.org sinnvoll in die Infrastruktur integriert werden können.

Den Einsatzmöglichkeiten von Disseminatoren im Design von Objektklassen sind eigentlich keine Grenzen gesetzt, somit können über Webservices einfach vorhandene Werkzeuge an Objekthinhalte gebunden werden. Systemadministratoren erhalten damit ein höchst flexibles Werkzeug zum Design und zur Parametrierung eines Digitalen Archives. Aus einer NutzerInnenperspektive ist lediglich die Instanziierung einer Objektklasse im Repository nötig, um die in einer Objektklasse gekapselte Funktionalität zur Verfügung zu haben. Das METS/MODS-Objekt ist für die Anzeige im DFG-Viewer optimiert, TEI-Objekte können als Grundlage für die Anwendung der Voyant Tools oder der Versioning Machine dienen; Inhalte, die placeName-Elemente enthalten, können auf eine Google Map projiziert oder im Geo-Browser dargestellt werden. Häufig kommen projektspezifische XSL-Stylesheets zum Einsatz, die die Objekte in eine gewünschte Anzeigeform bringen.

Der Client Cirilo wird noch 2014 als österreichischer Beitrag zu DARIAH-EU unter einer Open-Source Lizenz zusammen mit einer umfangreichen Dokumentation zur Verfügung stehen.

Ablauf

Der Workshop ermöglicht es den TeilnehmerInnen, den Client Cirilo auf einer Instanz des GAMS Repositoriums zu testen. Dabei wird zwischen vortragendenzentrierten Impulseinheiten und praktischen Übungen der TeilnehmerInnen abgewechselt: Zunächst werden bestimmte Arbeitsabläufe und Funktionalitäten anhand von konkreten, in Projekten mit unterschiedlichen Quellenmaterialien und Forschungsinteressen bereits umgesetzten Lösungen demonstriert und erläutert. Danach können die TeilnehmerInnen, entweder anhand eigener Projektdaten oder mittels vorbereiteter Beispieldateien diese in der Testumgebung erproben.

Die Vortragenden stehen während des gesamten Workshops für Fragen zur Verfügung und unterstützen die TeilnehmerInnen während der praktischen Übungen.

Literatur

DARIAH-EU: www.dariah.eu [2013-10-28]

DFG-Viewer: <http://dfg-viewer.de/ueber-das-projekt/> [2013-10-28]

FEDORA Commons: <http://www.fedora-commons.org/> [2013-10-28]

Geisteswissenschaftliches Asset Management System: <http://gams.uni-graz.at/> [2013-10-28]

Geo-Browser: <http://geobrowser.de.dariah.eu/> [2013-10-28]

Google Maps: <https://maps.google.at/> [2013-10-28]

Carl Lagoze, Sandy Payette, Edwin Shin, Chris Wilper, Fedora. An Architecture for Complex Objects and their Relationships. 2005. <http://arxiv.org/ftp/cs/papers/0501/0501012.pdf> [2013-10-28]

Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012 <http://public.ccsds.org/publications/archive/650x0m2.pdf> [2013-12-02].

Johannes Stigler. Think global, act local. Reale Probleme und virtuelle Lösungen. Eine Bestandsaufnahme anlässlich 50 Jahre Österreichische Mediathek und des UNESCO-World-Day for Audiovisual Heritage 2010. Hg. von. G. Fröschl & R. Hubert & E. Murlasits & S. Steinlechner, LIT Verlag Wien 2012, S. 113-126.

Johannes Stigler (gemeinsam mit W. Hofmeister). Edition als Interface. Möglichkeiten der Semantisierung und Kontextualisierung von domänenspezifischen Fachwissen in einem Digitalen Archiv am Beispiel der XML-basierten Augenfassung zur Hugo von Montfort-Edition, in Editio 24/2010, S. 39-56.

Versioning Machine: <http://v-machine.org/> [2013-10-28]

Voyant Tools: <http://voyant-tools.org/> [2013-10-28]