Wege zur Verknüpfung von eSciDoc und TextGrid (erstes Konzept) (R 1.3.2)

Version 1.0 vom 5.3.2012
Arbeitspaket 1.3
verantwortlicher Partner SUB Göttingen

TextGrid

Vernetzte Forschungsumgebung in den eHumanities

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Projekt: TextGrid - Vernetzte Forschungsumgebung in den eHumanities
BMBF Förderkennzeichen: 01UG0901A
Laufzeit: Juni 2009 bis Mai 2012

Dokumentstatus: final
Verfügbarkeit: öffentlich
Autoren:

Stefan E. Funk (DAASI)
Peter Gietz (DAASI)
Patrick Harms (SUB Göttingen)
Felix Lohmeier (SUB Göttingen)
Wolfgang Pempe (SUB Göttingen)

Revisionsverlauf:

Datum	**Autoren**	**Kommentare**
19.2.2012	Lohmeier	Entwurf
29.2.2012	Funk, Harms, Pempe	Überarbeitungen
1.3.2012	Gietz	Ergänzungen, Überarbeitung
5.3.2012	Lohmeier	Endredaktion

Inhaltsverzeichnis:

1. Problemstellung

Die Virtuelle Forschungsumgebung TextGrid und die Forschungsinfrastruktur eSciDoc bieten Wissenschaftlerinnen und Wissenschaftlern jeweils eigene Werkzeuge und Dienste für die gemeinsame Arbeit an elektronischen Daten. Trotz unterschiedlicher technischer Lösungen, Zielgruppen und Anwendungsszenarien, die im Laufe der langjährigen Projekte gewachsen sind, gibt es einige Überschneidungspunkte. Dieser Bericht widmet sich der Fragestellung, auf welcher technischen und semantischen Ebene eine Kooperation zwischen eSciDoc und TextGrid realisiert werden kann.

eSciDoc hat in den letzten Jahren eine produktive, digitale Forschungsinfrastruktur für die Max-Planck-Gesellschaft aufgebaut. Auf Basis dieser Infrastruktur werden die Daten der Max-Planck-Institute archiviert und in Virtuellen Forschungsumgebungen zugänglich gemacht. Diese auf eSciDoc aufbauenden Systeme werden von eSciDoc „Applikationen“ genannt. Diese richten sich an unterschiedliche Zielgruppen, bieten unterschiedliche Leistungen, basieren aber alle auf der Infrastruktur von eSciDoc mit mehr oder weniger großen Anteilen von Eigenentwicklungen. Zu den umfangreicheren Lösungen gehören insbesondere PubMan, ViRR, Imeji (Nachfolger von Faces) und das relativ neue Projekt „Digitization Lifecycle“[1], das bis Januar 2013 läuft, als Nachfolger von ViRR beschrieben wurde und im Juli 2012 in einer produktiv nutzbaren Form veröffentlicht werden soll. Weiterhin bietet das FIZ Karlsruhe mit KnowEsis professionellen Support für die eSciDoc Infrastruktur, Hosting und auch spezifische Anpassung und Neuentwicklung von auf eSciDoc basierenden Lösungen an.[2]

Eine Zusammenarbeit mit eSciDoc wird auf zwei verschiedenen Ebenen betrachtet:

a) Gemeinsame Werkzeuge und Dienste: Einige Werkzeuge, die in TextGrid oder eSciDoc entwickelt wurden, sind auch für die jeweils andere Zielgruppe passend. Einige Basisdienste können interoperabel und über einen gemeinsamen Dienstekatalog zugänglich gemacht werden.

b) Austausch der Daten: Es könnte eine übergreifende Recherche über Daten in beiden Systemen aufgebaut werden. Weitergehend können gemeinsame Empfehlungen für Datenstandards und gemeinsame Objektmodelle die Wiederverwendbarkeit der Forschungsdaten über Systemgrenzen hinweg fördern.

Die grundlegenden TextGrid-Konzepte sind in Report 1.2.1 beschrieben.[3] Die eSciDoc-Konzepte sind auf der Webseite nachzulesen.[4] Im Laufe der Kooperationsgespräche, die im Februar 2010 intensiviert wurden, hat sich herausgestellt, dass es lohnenswerter ist, in den Bereichen der Toolentwicklung und des Datenaustauschs zu kooperieren. Die Erfahrungen im Bereich Publikationsmanagement, Sammlungskonzept und Suchperformanz sind darüber hinaus in die Fortentwicklung der TextGrid-Dienste eingeflossen.

---

[1] Projekt Digitization Lifecycle: http://colab.mpdl.mpg.de/mediawiki/Digitization_Lifecycle
[2] Bekanntmachungen auf den eSciDoc Days 2011, 26./27. Oktober, Berlin: https://www.escidoc.org/JSPWiki/en/ESciDocDays2011Program
[3] TextGrid Report 1.2.1: Roadmap Integration Grid / Repository http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R121_v1.0.pdf
[4] eSciDoc – General Concepts: https://www.escidoc.org/JSPWiki/en/GeneralConcepts

2. Gemeinsame Werkzeuge und Dienste

Beide Systeme, TextGrid und eSciDoc, setzen auf eine service-orientierte Architektur. Daher können eigenständige, generische Dienste prinzipiell relativ leicht auch im jeweils anderen System zum Einsatz kommen. Schwieriger wird es bei Werkzeugen mit einer grafischen Benutzerschnittstelle, da diese nicht nur technisch auf das verwendete Framework portiert, sondern auch in die bestehende Ergonomie eingepasst werden müssen. Dieser Aufwand kann sich jedoch lohnen, wenn durch die Bündelung gemeinsamer Entwicklungsressourcen nach Abzug des Portierungsaufwands noch zusätzliche Ressourcen übrig bleiben oder durch eine zusätzliche Zielgruppe die Community wesentlich erweitert wird.

2.1. DigiLib

Das Bildbetrachtungs- und Referenzierungstool DigiLib umfasst die Galerieansicht mehrerer Bilder, Zoom, Skalierungs-, Markierungs- und Referenzfunktionen. Es wurde am Max-Planck-Institut für Wissenschaftsgeschichte entwickelt und kommt in vielen Kontexten von eSciDoc zum Einsatz.[5] In der aktuellen Beta-Version des TextGridLabs wurden die Komponente zur Anzeige von Bildern sowie die Zoom-Funktion bereits integriert. Hierzu wurde ein eigenes Eclipse-Plugin für die Benutzerschnittstelle programmiert. Der DigiLib-Scaler, der die Daten für die Anzeige prozessiert, greift über das TextGrid-Modul TG-crud auf die Daten in TextGrid zu. Eine weitergehende Integration mit dem Text-Bild-Link-Editor ist geplant. Weiterhin sollen über die Workflow-Komponente Funktionalitäten des Scalers zum automatischen Prozessieren größerer Bildmengen angeboten werden. Weiterhin ist für die TextGrid-Community die Bildverwaltungs-Applikation imeji[6] interessant.

2.2. CoNE (Control of Named Entities)

CoNE[7] ist ein im Rahmen von eSciDoc entwickelter Web-Service zur Verwaltung kontrollierter Vokabulare, wobei der Schreibzugriff auf einzelne Vokabulare via Rechtemanagement gesteuert werden kann. Das Nutzerinterface ist über eine Web-Browser-Applikation realisiert. In TextGrid könnte dieser Service für die Verwaltung von URI-Präfixen (u.a. für Identifikatoren aus anderen, ggf. externen Vokabularen) dienen sowie zur Verwaltung von Beziehungen (projektspezifisch und allgemein) und Objekttypen (/object/provided/format) verwendet werden. Über ein Eclipse-Plugin für das TextGridLab könnte eine nahtlose Einbindung in XML-Editor, Text-Text-Link-Editor und Metadaten-Editor angeboten werden. Das Rechtemanagement müsste durch TG-auth* ersetzt werden. CoNE ist unabhängig von der restlichen eSciDoc-Architektur nutzbar und benötigt nur eine PostgreSQL-Datenbank. . Vorarbeiten in Bezug auf Normdaten sind in DARIAH eingeflossen und werden dort im AP 1.4 weiter verfolgt. Beispielsweise könnte DARIAH CoNE einsetzen, um die Normdatendienste zu optimieren.

---

[5] Eintrag zu DigiLib im Wiki der Max Planck Digital Library: http://colab.mpdl.mpg.de/mediawiki/Digilib
[6] Eintrag zu imeji im Wiki der Max Planck Digital Library: http://colab.mpdl.mpg.de/mediawiki/Imeji
[7] Eintrag zu CoNE im Wiki der Max Planck Digital Library:
http://colab.mpdl.mpg.de/mediawiki/Control_of_Named_Entities

3. Austausch der Daten

Die Daten vieler Max-Planck-Institute sind von großem Interesse für die TextGrid-Community und auch umgekehrt sind in TextGrid hinterlegte und mit TextGrid erstellte Forschungsdaten relevant für die Max-Planck-Institute. Eine gegenseitige Referenzierung der Daten auf Präsentationsebene anhand von Persistent Identifiern ist natürlich problemlos möglich. Evaluiert wurde im Frühjahr 2010 ein Szenario, in dem TextGrid-Objekte in eSciDoc nachgewiesen werden, während die Daten selbst zur Langzeitarchivierung im TextGrid Repository liegen.. Eine direkte Veröffentlichung von TextGrid-Objekten in eSciDoc wäre über SWORD[8] zu realisieren. Hierzu müsste im TextGrid Backend, d.h. den Komponenten TG-crud und/oder TG-publish ein Atom-Feed implementiert werden.

Das Objektmodell beider Systeme ist ähnlich: Der „Context“ in eScidoc korrespondiert mit dem TextGrid „Project“ und der „Container“ ist das Äquivalent zu „Aggregation“ (vgl. Abb. 1).

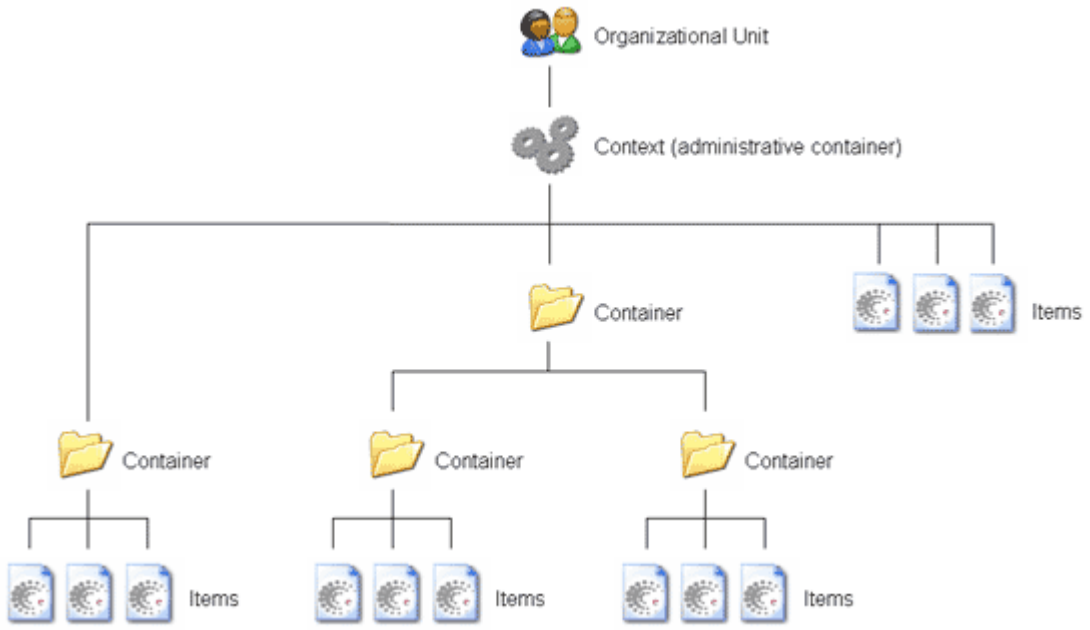

Abbildung 1 Objektmodell von eScidoc

Auch das Versionierungskonzept von eScidoc ist fast identisch mit dem von TextGrid.

Schnittstellen und Adaptoren zwischen beiden Systemen könnten in einem ersten Schritt eine übergreifende Recherche in den Daten beider Repositorien ermöglichen. Weitergehend würde die Entwicklung gemeinsamer Daten- (Sammlungen/Kollektionen) und Objektmodelle (Dateien, Metadaten) die Wiederverwendbarkeit der Forschungsdaten über Systemgrenzen hinweg fördern.

Allgemeine Überlegungen zum Austausch zwischen Datenrepositorien sind ausführlich in TextGrid Report 1.3.1 “Development of a Federated Repository Infrastructure for the Arts

[8] SWORD: http://www.swordapp.org/docs/sword-profile-1.3.html

and Humanities in Germany" beschrieben. TextGrid strebt eine auf Atom-Feeds basierende Repository Föderation an, die in DARIAH entwickelt werden könnte.[9]

4. Empfehlungen

Das britische Schwesterprojekt TEXTvre[10] hat das TextGrid Repository so weiterentwickelt, dass es als unterliegende Storage-Komponente das auf Fedora basierende institutionelle Repository am King's College London nutzt, und generell auch mit anderen Fedora-Repositories arbeiten kann. Da eSciDoc auch auf Fedora basiert, böte sich hier ein Ansatzpunkt für die Integration beider Systeme auf Repository-Ebene. Das komplexe Retrievalmodell und die Rechteverwaltung von eSciDoc sind jedoch schwierig auf TextGrid-Konzepte abzubilden. Weiterhin ist nicht klar, ob eSciDoc bei Fedora bleiben wird. Auf den eSciDoc Days 2011 wurde berichtet, dass überlegt wird, Fedora durch Hbase oder Hdsl zu ersetzen.[11]

Es scheint daher zielführender, die Zielgruppen und Anwendungsszenarien von TextGrid und eSciDoc weiter auszudifferenzieren, gemeinsam benötigte Werkzeuge und Dienste, wie bereits bei DigiLib geschehen, gemeinsam zu entwickeln. Die Integration von CONE in TextGrid wäre eine sinnvolle Ergänzung.

Für die gemeinsame Nutzung von Daten existieren bereits wichtige Voraussetzungen. Im Rahmen von DARIAH, an dem viele Projektpartner von TextGrid und auch die MPDL als einer der beiden Träger von eSciDoc beteiligt sind, kann diesbezüglich ein übergreifendes Zukunftsszenario für eine Aufgabenverteilung in einer nationalen Forschungsinfrastruktur diskutiert werden. Hier können sich beide Forschungsinfrastrukturen einbringen und an einer Föderation von Daten-Repositorien arbeiten. Übergreifende Authentifizierungs- und Autorisierungsinfrastrukturen, wie die SAML-basierte DFN-AAI, sowie deren verschiedene Pendants in Europa, bieten eine wichtige Voraussetzung für eine solche Datenföderation.

[9] Vgl. TextGrid Report 1.3.1, S. 15ff: http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_R131_Development_of_a_Federated_Repository_Infrastructure_for_the_Arts_and_Humanities_in_Germany.pdf
[10] TEXTvre: http://textvre.cerch.kcl.ac.uk/
[11] Vgl. Projekt SCAPE, an dem das FIZ Karlsruhe beteiligt ist: http://www.scape-project.eu/