Einführung in das PANDORA Linked Open Data Framework.

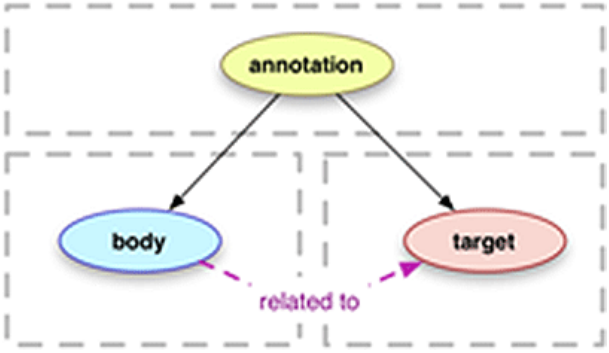
Johnson, Christopher
christopher.johnson@uni-goettingen.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Wettlaufer, Jörg
jwettla@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Beschreibung des Workshops [Zeitrahmen 4h]

Der Workshop stellt eine Softwarearchitektur vor, die zurzeit im Rahmen des Projekts „Johann Friedrich Blumenbach – online“ der Göttinger Akademie der Wissenschaften im Zusammenhang mit der geplanten digitalen Edition der gedruckten Werke und naturhistorischen Sammlungen J.F. Blumenbachs (1752-1840) entwickelt wird. Bei der Konzeption stehen Interoperabilität, Erweiterbarkeit und Nachnutzung als zentrale Entwicklungsziele im Vordergrund. Ausgangspunkt des **PANDORA** [**P**resentation (of) **AN**notations (in a) **D**igital **O**bject **R**epository **A**rchitecture] Linked Open Data (LOD) Frameworks sind digitale Abbildungen von Texten und Objekten, die in einem Fedora Commons Repository(1) gespeichert und über das International Image Interoperability Framework (IIIF) visualisiert werden. Das Framework ist insbesondere für den Einsatz im Museumskontext und im Bereich der digitalen Präsentation von Kulturgutüberlieferung geeignet. Dabei können sowohl text- also auch objektbasierte Fragestellungen untersucht bzw. Kulturgüter präsentiert und digital verfügbar gemacht werden. Ein besonderer Vorteil ist dabei die Bereitstellung der Daten als LOD und die Möglichkeit der Einbindung der Ressourcen in andere Kontexte. In dem Workshop sollen die Einsatz- und Nachnutzungsmöglichkeiten sowie die Nachhaltigkeit dieser Architektur vorgestellt, diskutiert und anhand von Beispielanwendungen zusammen mit den Teilnehmerinnen und Teilnehmern erprobt werden.

PANDORA ist zunächst einmal eine Sammlung von Open Source Anwendungen, die über ein gemeinsames „Manifest“ Dokument die Präsentation der Daten für den Anwender organisieren. Das „Manifest“ besteht aus einem JSON-LD(2) Dokument und wird aus einem digitalen Objektrepositorium über die dynamische Verwendung von SPARQL-Abfragen(3) erzeugt. Es orientiert sich dabei an der Semantik und dem Konzept der „IIIF Presentation API“(4). Diese Schnittstelle definiert, wie die Struktur und das Layout eines komplexen und bild-basierten Objekts in einem Standardformat dargestellt werden kann und zielt darauf ab, die Interoperabilität und Erweiterbarkeit von Präsentationen basierend auf dem Open Annotation Datenmodell(5) zu erleichtern. In diesem Modell ist oa:Annotation jede Ressource, die aus zwei Komponenten besteht, einen „body“ und einen „target“:



[Abb. 1: Annotation Datenmodell]

In der IIIF Presentation API ist das Ziel ein "canvas" (eine Leinwand), der eine Abstraktion des Client-Arbeitsplatz oder Sichtbereichs darstellt. Die Annotation (body) kann mit jedem verknüpften oder eingebetteten Objekt wie einem Bild, einer Beschreibung oder einem semantischen Tag verlinkt sein. Die assoziative Beziehungen zwischen verschiedenen Annotation-„bodies“ auf einem „canvas“ sind mit der Linked-Data Semantik im Manifest instanziiert. Die Segmentierung ermöglicht die Auswahl eines Bereichs eines Bildes oder eines „canvas“ unter Verwendung rechteckiger Begrenzungsrahmen oder mit der „IIIF Image API“(6), einem „stream“ von Bildausschnitten. Hotspot Verknüpfungen ermöglichen es die Auswahl auf ein Anmerkungsobjekt zu lenken, um eine Zustandsänderung in einem anderen Annotationsobjekt auszulösen.

Die Annotationen existieren im Fedora-Repository als LDP Container(7), der in einer Hierarchie von Ressourcen eine HTTP-adressierbare Ressource ist. Wenn der LDP Container in einem Triple-Store überführt wird, existiert er dort als RDF Ressource und als sog. „Named Graph“(8). Der IIIF Manifest Service unterstützt die Serialisierung bzw. „Kanonikalisierung“(9) des JSON-LD Dokuments in Form einer geordneten Liste, die im Ressource Description Framework als „collection" bezeichnet wird. Die Darstellung einer Manifest-Sequenz eines „canvas“ als RDF Sammlung erfordert die Verwendung von leeren Knoten, sog. „blank nodes“, die wie folgt miteinander verwoben sind:

<LDP_Manifest_Sequence_Container> sc:hasCanvases _:c11 .

_:c11 rdf:first <http://localhost:8080/fcrepo/rest/edition/base/canvas/c000> .

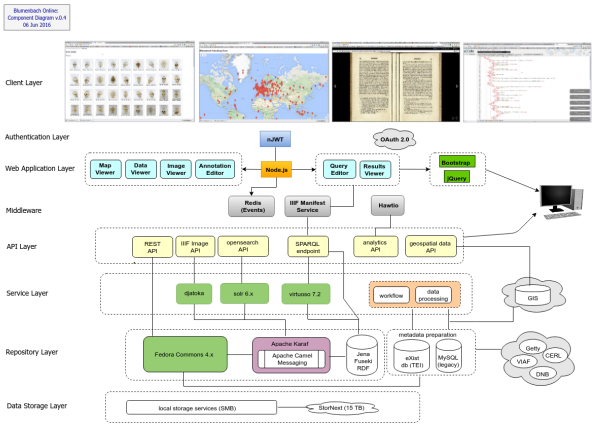
_:c11 rdf:rest _:c001.

_:c001 rdf:first <http://localhost:8080/fcrepo/rest/edition/base/canvas/c001> .

_:c001 rdf:rest …

Im Fedora-Repository wird ein „blank node“ mit einer bekannten Skolem IRI [nach RFC5785(10)] repräsentiert.

Durch die Verwendung des PANDORA IIIF Manifest Services(11) wird die Konstruktion von Präsentationen aus SPARQL Abfragen erlaubt, die eine sehr differenzierte Darstellung der Annotationen über JSON-LD ermöglichen. Der Entwurf einer LDP Container-Hierarchie und von Sammlungs-Definitionen im Einklang mit der Semantik der IIIF Presentation API "Annotation-Liste"(12) und "Layer"(13) für die Darstellung von Textsequenzen (Zeilen, Wortgruppen, Absätze, Seiten, Kapitel, etc.) ist ein integraler Bestandteil von PANDORA. Das folgende Schaubild verdeutlicht die Architektur des Frameworks und die Verknüpfung und das Zusammenspiel der einzelnen Komponenten:



[Abb. 2: PANDORA Architektur]

Mit einer klaren Trennung der Domain- und Client-Rollen bietet das PANDORA Framework Flexibilität und Erweiterbarkeit für alle möglichen Web-Client Präsentationsmethoden. Darüber hinaus unterstützt PANDORA Node.js Instanzen, die durch socket.io und Redis Pub/Sub(14) Ereignisse verbunden sind und dadurch Redundanz und Durchsatz für dezentrale asynchrone Operationen bieten. Das Framework besteht aus aktueller Open Source Software nach Industriestandards für Linked Data. Dazu gehören das Fedora-Repository, Apache Jena, Apache Camel, Apache Karaf, Open Virtuoso und Solr. Es ist gekennzeichnet durch Interoperabilität, Flexibilität und Erweiterbarkeit und erlaubt, durch die Verwendung von Standard-Software, ebenfalls eine Nachnutzung der Forschungsdaten über Linked Open Data Schnittstellen. Diese Daten können über den SPARQL-Endpoint entweder lokal integriert oder extern zur Nachnutzung angeboten werden. Weitere Informationen finden sich im GitHub Repository.(15) Eine ausführliche Dokumentation sowie eine Webseite mit Links zum Download der Komponenten befinden sich in Vorbereitung.

Eine zentrale Herausforderung für langfristig angelegte Forschungsprojekte, wie sie im Akademienprogramm der Bund-Länder-Kommission in Deutschland mit Laufzeiten zwischen 15 und 25 Jahren üblich sind, ist die Nachhaltigkeit von Systemarchitekturen in einer ständig fortschreitenden Entwicklung von Standardisierung und Versionierung. PANDORA begegnet dieser Herausforderung mit einem entkoppelten Aufbau auf der Grundlage von relativ unabhängigen voneinander agierenden Systemkomponenten, die bei Bedarf einfach ausgetauscht werden können, ohne die Grundfunktionalität zu gefährden. Auf der Ebene der Viewer können verschiedene Entwicklungen wie z.B. mirador(16) eingesetzt werden, ohne dass eine spezielle Anpassung notwendig ist. PANDORA setzt in Hinblick auf die langfristige Verfügbarkeit auf Standards aus dem Bereich des Semantik Web, die sich inzwischen weltweit durchgesetzt haben und damit sehr wahrscheinlich auch in Zukunft eine aktive Weiterentwicklung des Frameworks erlauben. Darüber hinaus ermöglichen diese Standards eine effiziente Vernetzung mit anderen Ressourcen im Web.

In dem Workshop sollen die einzelnen Komponenten des PANDORA Frameworks vorgestellt und deren Installation und Konfiguration erklärt werden. In einer Testumgebung, die für die Teilnehmer auf einem Server im Internet zur Verfügung stehen wird, können Beispieldatensätze gespeichert und die Funktionalität des Frameworks erprobt werden. Ebenfalls ist vorgesehen, die vorgestellte Architektur der Software intensiv zu diskutieren und mit anderen Lösungen für digitale Repositorien/Präsentationsumgebungen zu vergleichen.

Für die gewinnbringende Teilnahme sind Grundkenntnisse in Semantik Web Technologien sowie Kenntnisse der verwendeten Standards und/oder Open Source Software von Vorteil. Der Workshop eignet sich für eine Gruppe bis etwa 15 Personen. Die Teilnehmer sollten einen eignen Rechner/Laptop mit Verbindung zum Internet zur Verfügung haben, um im interaktiven Teil des Workshops die Funktionalitäten von PANDORA selber ausprobieren zu können. Die lokale Installation von zusätzlicher Software wird voraussichtlich nicht notwendig sein. Wichtige Informationen über die PANDORA Architektur können auch schon vorab in einem Video angesehen werden. (17)

Workshop Programm

Time	Title	Notes
9:00-9:30	Einführung	Gegenseitige Vorstellung. Einführung in das PANDORA Framework und Überblick zum Workshopverlauf
9:30-10:05	Ziele und Anwendungsbeispiele	Zielbeschreibung für den Workshop und Vorstellung von Anwendungsbeispielen
10:05-10:30	Technologische Herausforderungen	Vertiefende Einführung in die technologische Architektur von PANDORA - IIIF Presentation API, IIIF Image API, Manifest Service
10:30-10:45	Pause	
10:45-11:15	Einführung in den Übungsteil	Es wird eine Einführung für die Übung in kleinen Gruppen mit den IIIF Image und Presentation API gegeben.
11:15-11:40	Übungsteil / hands on	Hands-on Übung in kleinen Gruppen / Individuell Konkret: - Ein Manifest mit dem Javascript Client generieren - Ein Manifest mit dem IIIF Client visualisieren
11:40-12:00	Erfahrungsaustausch und Diskussion	Berichte aus den Kleingruppen
12:00-12:15	Pause	
12:15-1:00	Abschlussdiskussion, Fragen	Offene Fragen und Diskussion.

Linkliste

http://fedorarepository.org/
https://www.w3.org/TR/json-ld/
https://www.w3.org/TR/sparql11-query/
http://iiif.io/api/presentation/2.1/
http://www.openannotation.org/spec/core/core.html
http://iiif.io/api/image/2.1/
https://json-ld.github.io/normalization/spec/
http://www.rfc-editor.org/rfc/rfc5785.txt
https://github.com/blumenbach/iiif-manifest-service
http://redis.io/topics/pubsub
https://github.com/blumenbach/
http://github.com/IIIF/mirador

Für ein einführendes Video zur PANDORA Architektur siehe: https://youtu.be/TEqUkiO6tcA

Organisatoren des Workshops:

Christopher Hanna Johnson, MA.
Projekt “Johann Friedrich Blumenbach-online” der ADW Göttingen
Geiststraße 10
37073 Göttingen
oder
http://github.com/blumenbach
Forschungsinteressen: Semantik Web Technologien, Digitale Editionen, Softwareentwicklung, Cultural Heritage Studies

---

Dr. Jörg Wettlaufer
Digitisation Coordinator / Researcher
Akademie der Wissenschaften zu Göttingen (ADWG)
Göttingen Centre for Digital Humanities (GCDH)
Papendiek 16
37073 Göttingen
Germany
Tel. +49 551 39 20477 | 39 5366
/ skype: joewett
Forschungsinteressen: Digitale Geschichtswissenschaft, Semantik Web Technologien, Digitale Editionen