UNTER DER SPITZE DES EISBERGES: HANDLUNGSEMPFEHLUNGEN ZUR ANREICHERUNG VON (META)DATEN MIT LINKED OPEN DATA

Felix Sasaki[a]

[a] Bereich Sprachtechnologie, DFKI, Deutschland, felix.sasaki@dfki.de

Dieser Artikel thematisiert den Umgang mit Linked Open Data Quellen bei der Anreicherung von Daten oder Metadaten. Hierbei werden oft Aspekte von Anwendungen diskutiert, die für den Endnutzer sichtbar sind und unmittelbare Mehrwerte erzeugen. Der Artikel hingegen befasst sich Herausforderungen, die oft im Verborgenen liegen. Man muss sich diesen Herausforderungen stellen, um eine wiederholbare Wertschöpfung aus öffentlichen Datenquellen zu erreichen.

1. EINFÜHRUNG

Initiativen wie Coding Da Vinci zeigen den Erfolg und Mehrwert von öffentlich verfügbaren Datenquellen, welche zunehmend im technischen Paradigma „Linked Data“ zur Verfügung gestellt werden. Für den Endnutzer ist es wichtig, welche neuen Anwendungen durch Datenquellen möglich werden. In vielen Projekten spielen deshalb Fragen hinsichtlich Datenvisualisierung, Navigation in Datenräumen oder die Interaktion zwischen Daten und dem Nutzer eine große Rolle. Dieser Vortrag betrachtet derartige Fragen keinesfalls als unwichtig. Sie sind aber nur die Spitze des Eisberges. Um den Erfolg beim Umgang mit öffentlichen Daten insbesondere für kleinere Institutionen aus dem Bereich des kulturellen Erbes wiederholbar zu machen, muss man unter die Wasseroberfläche schauen. Dabei tut sich eine Vielzahl von Herausforderungen auf, die teilweise technischer, teilweise organisatorischer Natur sind. Sie sollen hier diskutiert werden.

2. HINTERGRUND: DAS FREME FRAMEWORK

Die Erfahrungen im Umgang mit öffentlich verfügbaren Daten, von denen hier berichtet werde soll, wurden in den letzten 1 ½ Jahren bei der Entwicklung des FREME Frameworks im Rahmen eines europäischen Projektes [1] gesammelt. FREME ist ein technisches Framework zur semantischen und mehrsprachigen Anreichung digitaler Inhalte. Die Entwicklung wurde vorangetrieben durch verschiedene Use Cases, zu denen auch die Anreichung bibliographischer Metadaten gehört. Aus der oben beschriebenen Nutzerperspektive ergeben sich hierdurch neue, vielfältige Möglichkeiten, z.B. die Disambiguierung von Autorenamen, facettierte Suche hinsichtlich wissenschaftlicher Themengebiete, oder die Verknüpfung mit generellen Wissensquellen wie Wikipedia. Im Verlauf der Entwicklung von FREME haben sich allerdings viele Herausforderungen „unter der Wasseroberfläche“ gezeigt, welche zum einen durch technische Lösungen, zum anderen durch die Dokumentation von Handlungsempfehlungen angegangen wurden. Die gelernten Erfahrungen sind im Folgenden zusammengefasst. Das Ziel ist dabei, eine wiederholbare Wertschöpfung aus öffentlichen Datenquellen, über den einzelnen Projektkontext hinaus zu unterstützen. Die Handlungsempfehlungen fokussieren den beschrieben Use Case von bibliographischen Metadaten, sind aber auf andere Use Cases und entsprechende Daten aus dem Bereich des kulturellen Erbes übertragbar.

3. HANDLUNGSEMPFEHLUNGEN ZUR ANREICHERUNG VON (META)DATEN MIT LINKED OPEN DATA

Nutzertypen definieren. Man kann nicht von jedem Mitarbeiter in kulturellen und anderen öffentlichen Institutionen ein Expertentum für Linked Data erwarten. Bei der Entwicklung technischer Infrastrukturen für den Umgang mit Linked Data muss deshalb eine Reihe potentieller Nutzer auch „unter der

Wasseroberfläche“ definiert werden: Vom Systemadminstrator, der Infrastrukturen aufsetzt bis zum Entwickler von Anwendungen, der sich nicht unbedingt mit Linked Data Details auskennen muss.

Linked Open Data Tooling für die nicht LOD Experten bereitstellen. Der Linked Data Technologiestack ist von großer Komplexität. Nicht jede Einrichtung kann sich aber einen Linked Data Experten leisten – im wahrsten Sinne des Wortes. Linked Data Infrastrukturen sollten derart konfigurierbar sein, dass man Anwendungen ohne detaillierte Linked Data Programmierung entwickeln kann. Für den Bereich kulturelles Erbe relevante Datenquellen müssen unmittelbar verfügbar sein, inklusive prototypischer Abfragen (z.B. Geburtsdaten von Personen, Geokoordinaten von Denkmälern, Links zu Wikipedia etc.).

Ausgewählte Datenquellen per Default bereitstellen. Die wachsende Menge von öffentlichen Datenquellen macht es unmöglich, jede Datenquellen für Anwendungsentwickler bereit zu stellen. Deshalb sollte es möglich sein, ausgewählte Datenquellen per Default zugänglich zu machen, mit dem oben beschriebenen Linked Open Data Tooling. In FREME, für den Use Case der Metadatenanreicherung, werden folgenden Datenquellen unterstützt: DBpedia, ONLD, Geopolitical Ontology, VIAF, ORCID, Library of Congress Author Names, Europeana und GRID.

Metadaten zu Linked Open Data Quellen erstellen. Für die breite, eventuell auch kommerzielle Nutzung sind Metadaten zu Linked Open Data Quellen essentiell. Abzudecken sind dabei z.B. Informationen hinsichtlich Datenprovenienz, Lizenzen, Stabilität der Daten und Update-Zyklen.

Auffrischung von Datensätzen in adäquaten Zyklen ermöglichen. Manche öffentlichen Datenquellen werden vom Bereitsteller in jeder Sekunde aktualisiert, andere Quellen nur jedes Jahr. Zudem ist das Auffrischen oft nicht komplett automatisierbar, da aktuelle Datensätze nicht immer via stabilen Downloadlinks zugänglich sind. Deshalb sollten für jeden Datensatz der jeweilige Auffrischungszyklus und die entsprechenden Schritte zur dokumentiert werden, durch welche die Daten zugänglich werden.

Umgang mit bestehenden Workflows und Formaten erleichtern. In vielen Einrichtungen ist es nicht machbar, die Informationsverarbeitung von heute auf morgen auf Linked Data umzustellen. Eine nachhaltige Nutzung von Linked Data sollte deshalb die Integration in existierende Workflows unterstützen. Dabei sind XML basierte Formate von großer Bedeutung. Eine Integration von Linked Data Informationen wird leichter, wenn sie bestehende XML Verarbeitungsschritte nicht behindert.

Grundlegende Infrastruktur Open Source bereitstellen. Grundlegende Softwarekomponenten, welchen den Umgang mit Linked Data erleichtern, sollten als Open Source bereitgestellt werden. Nur so kann man von auch kleineren Institutionen eine breitere Nutzung erwarten. Das FREME Framework ist dem entsprechend unter einer Lizenz verfügbar, die kommerzielle als auch nicht kommerzielle Nutzung erlaubt.

4. WEITERE SCHRITTE

Die grundlegende Entwicklung von FREME, welche die genannten Handlungsanforderungen umsetzt, ist abgeschlossen. Nun gilt es, Feedback zu den Anforderungen, Bezüge zu ähnlichen Anforderungen (z.B. formuliert im Rahmen von NESTOR [2] oder der W3C „Data on the Web Best Pratices“ [3]), sowie eine breite praktische Nutzung des Frameworks zu erlangen. Diese Nutzung und auch Weiterentwicklung außerhalb des Projektkontext hat schon begonnen, im DFKI geführten Projekt „Digitale Kuratierungstechnologien“ sowie in der Bibliothek des Trinity College Dublin. Wir hoffen, dass dieser Artikel zu einer Weiternutzung und zu einer Umsetzung der beschriebenen Handlungsanforderungen führt – damit wir uns in Zukunft mehr um die Spitze des Eisberges kümmern können.

5. LITERATURHINWEISE

1 Vgl. http://www.freme-project.eu zum FREME Projekt und http://api.freme-project.eu/doc/current/ zur Dokumentation der technischen Infrastruktur.

2 Vgl. http://www.langzeitarchivierung.de/

3 Vgl. https://www.w3.org/TR/dwbp/