120 km Papier – trotz Linked Data

Die Schweiz hat umfangreiche Datenbestände. Bisher wurden sie separat von Gemeinden, Kantonen und Bund gespeichert. Mit dem Projekt Lindas werden nun viele Informationen von Behörden verknüpft und frei zugänglich. Dabei übernimmt das Schweizerische Bundesarchiv die führende Rolle. Warum er auf Open und Linked Data setzt, und wie er zu Pilotprojekten inspiriert, erläutert Jean-Luc Cochard, der Leiter der Abteilung IT.

Interview:
Anne-Careen Stoltze-Siebmann
Kommunikation
anne-careen.stoltze-siebmann@bfh.ch

Warum steigt eine analoge Institution, die eher ein verstaubtes Image hat, in ein ultramodernes Thema wie Linked Data ein?

Jean-Luc Cochard Wir sind gar nicht so altmodisch. Im Gegenteil, wir sind quasi an der Front. Wir müssen uns mit den modernen Entwicklungen beschäftigen, schon deshalb, weil immer mehr Daten digital zu uns kommen. Schon seit ein paar Jahren archivieren wir immer mehr Datenbanken. Linked Data ist für Datenbanken wichtig, aber auch für Open-Government-Data (OGD), wo wir ebenfalls aktiv sind.

Wie archivieren Sie Datenbanken?

JC Immer häufiger ist heute der gesamte Weg der Dokumente von der Entstehung über die lokale Speicherung bis zur Archivierung elektronisch. Dementsprechend bekommen wir immer mehr elektronische Daten von Behörden, Ämtern und Institutionen. Wir übernehmen eine Auswahl der Dokumente. Das sind meistens Word- und Excel-Dateien. Im Weiteren sind Datenbanken wichtige Informationsquellen, deren Archivierung sich lohnt. Um die Daten dauerhaft zu sichern, wandeln wir sie in das Format SIARD (Software Independent Archiving of Relational Databases) um, welches wir 2008 entwickelt haben.

Wie sieht es mit Ihren Papierbeständen aus?

JC Unser nationales Archiv ist im Vergleich zu jenen in anderen Ländern ein relativ kleines. Unsere Bestände reichen zurück bis zur Helvetik Ende des 18. Jahrhunderts. Das ist nichts im Vergleich etwa zu Frankreich, dessen Archive im Mittelalter beginnen. Wir haben etwa 60 Kilometer Dokumente auf Papier und werden weitere 60 Kilometer erhalten, weil zahlreiche Unterlagen noch in den Ämtern lagern. Wir werden diese Dossiers als Papier behalten. Die Akten lagern in vier Stockwerken unter unserem Gebäude. Wir scannen bisher nicht systematisch, sondern auf Wunsch unserer Kunden. Anschliessend haben wir von einer Datei eine Papier- und eine elektronische Version. In der Regel wird ein Kunde zukünftig die digitale Version erhalten.

Wie sehen Sie die Zukunft? Ist das Bundesarchiv in zehn Jahren papierlos?

JC Nicht völlig. Wie erwähnt werden wir nach wie vor 120 Laufkilometer Papierakten haben. Aber wir werden fast nur noch digitale Dokumente erhalten. Gewisse Dokumente werden noch länger auf Papier konserviert werden, zum Beispiel Staatsverträge. Der Zugang zum Archiv soll zukünftig aber primär elektronisch erfolgen. Das heisst, wir digitalisieren analoge Dokumente auf Bestellung und stellen diese online zur Verfügung.

Von welchen Wechselwirkungen zwischen Technologie und Verwaltung profitieren Sie bei Ihrer Arbeit?

JC Wir begleiten technische Entwicklungen aktiv mit, zum Beispiel SIARD, das eine Antwort auf die Verbreitung der relationalen Datenbanken ist. Bei Linked Data könnten wir warten, bis wir mit diesem Datenformat konfrontiert werden. Wir bevorzugen jedoch, eine aktive Rolle bei der Einführung dieser Technologie einzunehmen, um auf ihren Einsatz in der Verwaltung Einfluss nehmen zu können. Und wir sehen uns auch als Anwender. Die Linked Data-Technologie ist ausgereift. Wenn wir etwas Neues implementieren, brauchen wir neben der technischen Umsetzung auch Beratung und Support sowie weiterführende Lösungen. Dabei spielen Hochschulen, Forschung und Unternehmen eine sehr wichtige Rolle. Dieses Zusammenspiel braucht es.

Welche Rolle sollte von Ihrem Standpunkt aus der Staat einnehmen, die Forschung vermehrt fördern oder als Gesetzgeber den Rahmen vorgeben?

JC Das ist eine schwierige Frage. Der Staat muss verschiedene Rollen spielen. Eigentlich sollte ein Archiv wichtiger oder gar zentraler Teil einer (nationalen) Dateninfrastruktur sein. Neben diesem Leistungsangebot können wir auch eine inspirierende Rolle spielen. Aber

Jean-Luc Cochard leitet die IT-Abteilung des Schweizerischen Bundesarchives und ist unter anderem verantwortlich für die Digitalisierung der Dokumente.

es ist sicher auch von grosser Bedeutung, dass der Staat in die Ausstattung der Hochschulen und in die Forschung investiert, damit sie Pilotprojekte, zum Beispiel mit uns, starten können.

Haben Sie dafür Beispiele?

JC Das belgische ICT-Unternehmen Cetic hat alte, typografisch aufwändig gestaltete Tabellen der digitalisierten Staatsrechnung für uns in Daten konvertiert. Das ist eine sehr anspruchsvolle Aufgabe, denn diese Tabellen lassen sich nicht einfach in Excel-Dateien umwandeln. Die Firma hat dafür eine Nischenlösung gefunden, die sich ursprünglich aus der Arbeit des belgischen Nationalarchivs ergeben hat. Anschliessend haben sie diese Dienstleistung an Kunden wie uns verkauft. Wir haben das auf unsere Bundesrechnungen angewendet. Auf diese Weise kann sich aus einem praktischen Problem ein Spin-off eines Forschungsinstituts entwickeln. Dafür braucht es keine grosse Softwarefirma.

Ausserdem sind wir im Besitz des vollständigen Inventars der Rechnungen der Bundesverwaltung seit 1848. Diese wurden digitalisiert, sind aber unbrauchbar, weil der Inhalt der Tabellen nur sehr schlecht von den OCR-Systemen erkannt werden kann. Ein Institut der katholischen Universität von Louvain hat eine Expertise entwickelt, um dieses Problem zu lösen. Solche Dienstleistungen, die wir genutzt haben, können die Entwicklung von Start-ups begünstigen. Es besteht bei den Archiven eine Nachfrage nach Applikationen zur Aufwertung von solchen analogen Inhalten.

Wie nutzt das Bundesarchiv schon heute die entsprechenden Technologien?

JC Im Fall von Linked Data wandeln wir parallel mit einigen Kantonsarchiven unseren Katalog schrittweise in dieses Format um. Unser Ziel ist es, dass wir mithilfe von Linked Data die Inhalte der Dossiers miteinander verknüpfen können. Ein Beispiel: Eine historische Person kann in verschiedenen Archivdokumenten genannt werden. Eine erste Spur findet sich vielleicht im Kanton Genf, dann gibt es einen Hinweis in Neuenburg und einen Bundesratsbeschluss. Diese Dokumente werden bisher separat in den Archiven aufbewahrt und Historikerinnen und Historiker müssen diese mühsam einzeln suchen. Erkennen, dass es sich um die gleiche Person handelt, diese Information speichern und veröffentlichen, damit künftige Forschende davon Nutzen ziehen können, ist eine Verwendungsmöglichkeit des Potentials dieser Technologie. Wenn einmal alle Archivkataloge in Linked Data umgewandelt sind, werden die Verknüpfungen die Recherche wesentlich erleichtern.

Wie weit sind Sie damit?

JC Wir sind in der Pilotphase. Wir wollen nun herausfinden, welche Vorteile diese Umstellung uns bringt und auch welche Schwierigkeiten. Ein weiteres Ziel dieser Pilotphase ist es, den Archivinstitutionen aufzuzeigen, wie sie diese Technologie einsetzen können. Im vergangenen September haben wir die aktuellen Resultate unserer Arbeiten an einer Fachtagung des Vereins Schweizerischer Archivarinnen und Archivare vorgestellt. Das Echo war sehr positiv. Besonders wichtig ist uns eine einfache Bedienbarkeit, ein klares Interface, damit wir eben beispielsweise Historikerinnen und Historikern einen guten Service liefern können.

Wo sehen Sie Herausforderungen?

JC Wir stehen vor vielfachen Herausforderungen. Welchen Stellenwert wollen wir Linked Data gegenüber dem aktuellen Angebot von öffentlichen Websites wie www.swiss-archives.ch geben? Es gibt verschiedene Optionen, die untersucht werden können. Sie hier alle zu erklären, würde zu weit führen. Linked Data erlaubt auch Mechanismen einzuführen, die es einer Community von Nutzerinnen und Nutzern ermöglicht, Daten zu ergänzen, ähnlich wie die Foren, die man auf vielen Websites findet. Weiter: Wollen wir solche Möglichkeiten umsetzen? Welche Vorteile können wir daraus ziehen und zu welchen Kosten? Und natürlich, wie werden wir solche Daten in Zukunft archivieren?

Und wie schützen Sie die Daten vor Missbrauch?

JC Die Publikation von Daten im Web geht zwangsläufig mit einem Verlust an Kontrolle einher, was mit diesen Daten gemacht wird. Dabei spielt es keine Rolle, ob es sich um Linked Data, ein Dokument oder statistische Daten in Tabellenform handelt. Die Quelle ist die Referenz und diese muss deshalb geschützt werden.

Könnte jemand von aussen die Daten verändern?

JC Jede Website hat einen Administratoren-Zugang, um Inhalte der Site ändern zu können. Eine Linked Data Infrastruktur funktioniert nach dem gleichen Prinzip. Es müssen deshalb Mechanismen eingeführt werden, die das Eindringen in das System über diesen Zugang verhindern. Bei Linked Data ist aber auch zu bedenken, dass es sich um Kopien von Daten handelt, die in einer nicht öffentlichen und sehr gut geschützten Datenbank verwaltet werden. Die Ursprungsquelle der Daten kann somit nicht geändert werden. Zudem können wir auf unser Papierarchiv zurückgreifen. Wir werden sicher noch für eine lange Zeit die Ursprungsdaten auf unserem internen System behalten genau wie die Daten in Papierform. So können wir leicht Fälschungen enttarnen, wenn wir die Originale mit einer verdächtigen Datei abgleichen können.

Wie viele Daten haben Sie in digitaler Form?

JC Wir haben einige Terabyte auf mehreren Servern an verschiedenen Standorten gespeichert. Zudem sind die Daten dreifach kopiert. Diese Server funktionieren getrennt voneinander, falls einer ausfällt. Es werden ständig mehr.

Wo findet die Suche momentan statt – auf Ihrem Server oder bei einem Cloud-Anbieter?

JC Im Moment kann der Archivkatalog – das sind nur die Metadaten – auf www.swiss-archives.ch durchsucht werden. Diese Website ist auf einem Bundesserver gehostet. Wie es zukünftig bei Linked Data aussehen könnte, ist noch offen.

Seit einigen Wochen betreibt das Bundesarchiv den Linked Data-Service Lindas. Wie kam es dazu?

JC Vereinfacht gesagt, ermöglicht Lindas, verschiedene, dezentral zur Verfügung gestellte Datenbestände zentral zu sammeln und online zu stellen. Das Datenformat RDF für Linked Data ist gleichzeitig das Zielformat für OGD, weil es viele Freiheiten für die maschinelle Weiterverarbeitung lässt. Seit Herbst 2015 stellt Lindas auf www.lindas-data.ch Daten zum freien Gebrauch zur Verfügung. Lindas war zunächst als Pilotlösung vorgesehen und wurde vom Staatssekretariat für Wirtschaft (Seco) entwickelt. Dann meldeten auch das Bundesamt für Statistik und andere Behörden Interesse an. Bald stellte sich die Frage, wer diese Lindas nach der Pilotphase weiter betreibt. Da kamen wir ins Spiel, weil wir auch aktiv in OGD sind. Frei verfügbare Behördendaten können so vermehrt auch in einem möglichst einfach weiter verwendbarem Dateiformat angeboten werden. Im Dezember hat das Bundesarchiv deshalb eine Vereinbarung mit dem Seco unterschrieben. Nun betreiben wir Lindas und entwickeln es auch weiter. ■

Projekt Lindas

Das Staatssekretariat für Wirtschaft (Seco) hat den Linked Data Service Lindas in Zusammenarbeit mit IT-Unternehmen entwickelt. Ziel von Lindas ist es, Daten als Linked Data im RDF-Format zu speichern und zur Verfügung zu stellen. Dieses Format ist besonders für die automatische Verarbeitung grosser Datenmengen aus unterschiedlichen Quellen geeignet. Beispiel: Bisher wurden die Informationen über Verwaltungsstellen, Behördenleistungen bei Bund, Kantonen und Gemeinden dezentral in vielen verschiedenen Datenbanken erfasst. Entsprechend aufwändig und komplex ist es, bei einem konkreten Anliegen auf Anhieb zu den richtigen Informationen bzw. Daten zu gelangen. Dank Lindas könnten diese Daten verknüpft und zentral zur Verfügung gestellt werden. So würde eine Bürgerin aus dem Kanton Thurgau rasch erfahren, an welche Behörde sie sich wenden muss, wenn sie einen neuen Ausweis braucht.

Glossar

- **SIARD:** Software Independent Archival of Relational Databases ist eine offene Auszeichnungssprache zur Langzeit-Archivierung von relationalen Datenbanken in Form von Textdaten basierend auf XML.
- **OCR:** ist die Abkürzung der englischen Bezeichnung Optical Character Recognition, die die automatisierte Texterkennung innerhalb von Bildern bezeichnet.
- **OGD:** Open Government Data, englisches Synonym für Daten, die von Staat und Verwaltung im Interesse der Allgemeinheit zur freien Nutzung und Weiterverbreitung zugänglich gemacht werden.
- **RDF:** «System zur Beschreibung von Ressourcen» bezeichnet eine technische Herangehensweise im Internet zur Formulierung logischer Aussagen über beliebige Dinge (Ressourcen). Ursprünglich wurde RDF als Standard zur Beschreibung von Metadaten konzipiert. Mittlerweile gilt RDF als grundlegender Baustein des semantischen Webs.