Von IIIF zu IPIF? Ein Vorschlag für den Datenaustausch über Personen

Vogeler, Georg
georg.vogeler@uni-graz.at
Universität Graz, Österreich

Vasold, Gunter
gunter.vasold@uni-graz.at
Universität Graz, Österreich

Schlögl, Matthias
matthias.schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Wien, Österreich

Gesellschaften fügen sich aus Individuen zusammen. Das gilt auch für die Vergangenheit, aus der die Mehrzahl der Individuen nur schlecht bis gar nicht dokumentiert ist. Es hat sich deshalb ein eigenständiger historischer Forschungsbereich entwickelt, die "Prosopographie", die sich der Aggregation von Einzelinformationen zu Individuen aus historischen Quellen und ihrer Auswertung widmet (Keats-Rohan 2007). Dieses Forschungsgebiet hat früh digitale Methoden eingesetzt. Der Beitrag widmet sich der Frage, ob die Methoden vergleichbar zu IIIF (International Image Interoperability Framework) in ein „International Proposography Interoperability Framework" (IPIF) integriert werden können.

Ein IPIF muss von den Personendatenbanken abweichen, die sich als kontrollierte Vokabularien und Referenzen für Linked Open Data in den Digital Humanities etabliert haben (GND/VIAF, deutsche-biographie), bzw. im Begriff sind, sich zu etablieren (wikidata). Diese berücksichtigen nämlich nicht den Vorgang, mit dem Informationen über eine Person aus historischen Quellen aggregiert werden. Der Ansatz weicht damit auch von der „personography" der TEI ab, die, wie die Linked-Data-Ressourcen, eine Person mit einer Liste an Eigenschaften beschreiben. Ein IPIF muss dagegen ein Modell realisieren, für das Bradley/Short (2005) die Bezeichnung „Factoid"-Model eingeführt haben. Es geht von drei Informationseinheiten aus: Quelle, Individuum und Aussagen der Quelle über das Individuum. John Bradley hat das Modell mehreren Projekten des King's College London zu Grunde gelegt (PASE, DPRR, CCEd). Auch das Persondendatenrepositorium (PDR) der Berlin-Brandenburgischen Akademie der Wissenschaften (Neumann et al. 2011) und Projekte, die die Software der BBAW weitergenutzt haben, verwenden das gleiche Modell, auch wenn das PDR nicht explizit auf Bradley referenziert. Ebenso verwendet das Repertorium Academicum Germanicum ein solches dreiteiliges Modell (Andresen 2008).

Das dreiteilige Modell impliziert auch, dass (auch widersprechende) Aussagen über dasselbe Individuum aus verschiedenen Quellen an verschiedenen Orten publiziert werden können. Es erscheint also als ein Paradebeispiel für das *Web of Data* des W3C. Das *Web of Data* ist die Fortführung der Semantic-Web-Aktivitäten des W3C. Es konzentriert sich auf die Öffnung von Datenbanken und erhebt insbesondere den Anspruch, individuelle kleine Datenmengen als RDF über das Semantic Web abfragbar zu machen. Technisch ist RDF, die Grundlage des *Web of Data*, eine weit verbreitete und gut unterstützte Technologie. Es ist deshalb auch eine Technologie, mit deren Hilfe immer häufiger Maschinen auf prosopographische Datenbanken zugreifen können. Deshalb haben Bradley/Pasin (2015) eine CIDOC-CRM basierte Version des Factoid-Modells vorgeschlagen und entsprechende Ontologien veröffentlicht (Bradley 2017). Das Basismodell ist aber auch mit anderen Vokabularien realisiert worden: SNAP verwendet z.B. Vokabularien aus dem Linking-Ancient-Wisdom-Projek[1]). Das King's Digital Lab hat jüngst mit Hilfe von *Ontop*[2] die prosopographische Datenbank zur römischen Republik als LOD-Ressource incl. eines SPARQL-Endpoints veröffentlicht.[3]

Diese Strategie teilt jedoch das Problem vieler RDF-Ressourcen: Die technische Pflege eines SPARQL-Endpoints ist sehr anspruchsvoll. SPARQL-Endpoints sind häufig nur unzuverlässig verfügbar. Nicht zuletzt deshalb stellen große Lieferanten von RDF-Ressourcen wie die Deutsche Nationalbibliothek (GND) bis dato keine SPARQL-Endpoints für ihre Daten zur Verfügung. Als alternative Technologie etabliert sich in den Digitalen Geisteswissenschaften zunehmend die Publikation von eigenen RESTful APIs, die zwar weit weniger flexible Abfragen erlauben, dafür aber deutlich stabiler funktionieren und einfacher implementiert werden können. RESTful APIs sind ein Quasi-Industriestandard und werden von jedem Webentwicklungsframework und jeder Programmiersprache unterstützt. Mit OpenAPI (ehemals swagger)[4] und Core API[5] liegen auch Vorschläge vor, derartige API-Definition standardisiert zu beschreiben, so dass die Implementation von einschlägigen API-Anbietern und API-Konsumenten teilweise sogar automatisiert werden kann.[6] Aus Sicht des Software-Engineering erscheint es also angemessen, auf eine eigene API-Definition statt auf einen SPARQL-Endpoint zurückzugreifen. Gleichzeitig wird es damit erschwert, Daten aus verschiedenen Datenquellen zu aggregieren, da für jeden Datenanbieter ein eigener API-Konsument programmiert werden müsste. Im Bereich der Bibliotheken hat sich deshalb für die Bereitstellung von Bildern von Büchern mit IIIF eine Kombination aus einem Datenstandard und einer Adressierungs-API durchgesetzt. Es ist an der Zeit, auch für personenbezogene Daten über

einen solchen technischen Standard nachzudenken, der die Implementation von Anwendungen erleichtert und die Daten auch praktisch interoperabel macht.

Ein solcher Standard muss von konkreten Anwendungsszenarien ausgehen. Sie können unter den Überschriften „Biographical Lexicon“, „Careers”, „Source Editing“, „Fact Checking“, „New Interpretation“, „Publish a Database”, „Integrate Other Databases“, „Analysis“, „Tool User“, „Tool Builder” zusammengefasst werden. Die Szenarien bilden sowohl Forschung mit prosopographischen Daten wie die Erzeugung solcher Daten ab. Zusätzlich achten die Szenarien darauf, nicht nur explizit prosopographische Workflows zu berücksichtigen, sondern schließen auch wissenschaftliches Edieren als Szenario mit ein, in dem der edierte Text als Beleg für eine Person betrachtet werden kann. In einem Workshop in Wien im Februar 2017 haben Forscher aus dem Themengebiet der Prosopographie religiöser Orden solche Anwendungsszenarien diskutiert und einen Entwurf für eine API entwickelt.

Ein Ergebnis dieser Arbeit ist eine nach den Standards von OpenAPI beschriebenen Definition einer prosopographischen API.[7] Die API baut auf dem dreiteiligen Factoid-Modell auf und erlaubt den Zugriff auf Personen, Aussagen, Quellen und ihr Aggregat, einem „Factoid“. Für alle diese Objekte gibt es eigene Pfade zur Suche und Ausgabe der Daten über die zu ihnen abgelegten IDs. Im Kern der API steht deshalb der Zugriff auf Factoide (/factoid). Sie können individuell über bekannte IDs adressiert werden (/factoid/id). Wichtiger sind aber inhaltliche Filtermöglichkeiten. Sie ergeben sich einfach aus den Eigenschaften des Factoids, als Aussage über eine Person. Die Parameter *s, st* und *f* lassen also die Suche in den Inhalten der mit dem Factoid verknüpften Quellen (source), Aussagen (statement) und den Metadaten des Factoids selbst (factoid) zu. Dabei ist der Standard eine Volltextsuche. Ebenso lassen sich die Quellen und Personen abfragen. Als Parameter können aber auch Identifikatoren für die einzelnen Informationsgruppen übergeben werden, also z.B. mit / statement/?p_id=Placidus_Seiz alle Aussagen über die Person mit einem Identifikator „Placidus_Seiz“ in einem beliebigen Kontext. Die Anwendung liefert dann ein JSON-Objekt zurück, in dem diese Aussagen formalisiert sind. Zu jeder Aussage gehört eine ID, mit der Entwickler z.B. über die API überprüfen können, woher die jeweilige Aussage stammt.

Als Rückgabewert der API-Definition sind JSON-Serialisierungen vorgesehen. Die Statements können Daten als Text (z.B. der Quelle) ebenso wie strukturiert als Graph enthalten. Die Graphen sollen den Spezifikationen von JSON-LD folgen. Damit können zwei Ziele erreicht werden: Erstens ist damit die Ausgabe der API direkt in Linked-Open-Data-Umgebungen nutzbar, kann prinzipiell auch in einer FROM-Klausel einer SPARQL-Abfrage integriert werden oder in Caching-Mechanismen wie im 2011 als Linked Data Middleware von Virtuoso vorgeschlagenen URI-Burner verwendet werden. Zweitens wird damit ein Standard verwendet, der die Referenzierung der verwendeten Vokabularien und ihre formale Beschreibung mit RDFS und OWL ermöglicht.

Der Workshop in Wien hat als Kernproblem eines echten Datenaustausches die divergierenden Datenmodelle für die Einzelaussagen über die Individuen identifiziert. Während die Individuen selbst im Factoid-Modell keine beschreibenden Metadaten tragen und damit kaum Probleme beim Datenaustausch erzeugen, sind für die Aussagen über die Individuen je nach Projekt, Verwendungszweck und Forschungsdomäne eine Vielfalt von Vokabularien im Einsatz. Einen Ausweg aus dieser Situation bietet die 2017 gegründete dataforhistory-Initiative.[8] Die Initiative arbeitet daran projekt- und domänenspezifische Modellierungen zu erleichtern, die zum CIDOC CRM kompatibel sind. Die derzeitige API-Definition sieht deshalb vor, dass die zurückgegebenen Daten eine Referenz auf ein Schema (in JSON-LD als @context) enthalten müssen, das die verwendeten Klassen und Eigenschaften auf Definitionen im CIDOC CRM abbildet, der es der die API konsumierenden Anwendung erlaubt, die Daten als CIDOC CRM zu interpretieren und darauf aufbauende Operationen durchzuführen. Ergänzend dazu ist ein Parameter format=json/cidoc-crm vorgesehen, bei dem die Transformation serverseitig stattfindet. Die Abbildung auf CIDOC CRM soll insbesondere die grundlegenden Suchoperationen ermöglichen, die Katerina Tzompanaki und Martin Doer 2012 formuliert haben und die im Projekt researchspace[9] realisiert werden. Die API definiert die Objekteigenschaft graph für die strukturierte Repräsentation der Daten über Personen.

John Bradley und Michele Pasin haben 2015 eine OWL basierte Ontologie vorgestellt, in der eine „temporal entity documented“ (TED) als Ereignis (E4 und E5 im CIDOC-CRM) oder als eine zeitliche Einheit oder klare zeitliche Grenzen (E3: condition, state) modelliert sind. Das entspricht dem Stand der Diskussion über prosopographische Datenmodelle (Lind 1994, Andresen 2008, Tuominen / Hyvönen / Leskinen 2018).

Nicht zuletzt der Erfolg von IIIF belegt, dass eine solche API aber auch Referenzimplementationen benötigt. Dabei ist entsprechend der oben beschriebenen Benutzungsszenarien sowohl an Ressourcen zu denken, die Daten bereitstellen, als auch an Anwendungen, die diese Daten konsumieren. Die Nachnutzung des „Archiveditors“, eines zunächst projektinternen Werkzeugs der BBAW, in anderen Projekten zeigt, dass dabei nicht nur an Datenextraktion und –anzeige sondern auch an Datengenerierung zu denken ist. Im Rahmen der Arbeit an der Personendatenbank der Österreichischen Akademie der Wissenschaften ist deutlich geworden, dass gerade automatische Informationsextraktion von „Personenrelationen“ (Schlögl et al. forthcoming, Schlögl et al. 2018) von einer solchen API profitieren kann. Die automatisch generierten Aussagen können als eigenständige Factoide in die Personendatenbanken eingehen. Die Metadaten des

Factoids und die Referenz auf die verwendete Quelle stellen sicher, dass sie als automatisch generierte Daten identifizierbar bleiben. Der Vortrag wird Beispiele für Datenangebote aus dem Umfeld mittelalterlicher Urkunden (Register der Urkundenempfänger von Papsturkunden nach den Regesten von August Potthast, Daten aus monasterium.net) und Steuererhebungen (England) vorstellen, und Prototypen für Anwendungen benennen, welche die mit der API bereitgestellten Daten konsumieren können.

Fußnoten

1. http://lawd.info/ontology/
2. https://github.com/ontop/ontop
3. http://romanrepublic.ac.uk/rdf , Dokumentation von John Bradley: http://romanrepublic.ac.uk/rdf/doc
4. https://www.openapis.org/
5. http://www.coreapi.org
6. z.B. das Python-Framework Flask in Verbindung mit https://github.com/zalando/connexion , vgl. weitere Tools: https://swagger.io/tools/open-source/open-source-integrations/
7. https://github.com/GVogeler/prosopogrAPhI
8. http://dataforhistory.org
9. https://www.researchspace.org/

Bibliographie

Andresen, Suse (2008): *Das 'Repertorium Academicum Germanicum'. Überlegungen zu einer modellorientierten Datenbankstruktur und zur Aufbereitung prosopographischer Informationen der graduierten Gelehrten des Spätmittelalters*, in: **Sigrid Schmitt u. Michael Matheus (eds.):** *Städtische Gesellschaft und Kirche im Spätmittelalter* (Geschichtliche Landeskunde 62). Stuttgart: Steiner 17-26.

Bradley, John (2017): *Factoids. A site that introduces Factoid Prosopograph*, http://factoid-dighum.kcl.ac.uk/ und https://github.com/johnBradley501/FPO

Bradley, John / Pasin, Michele (2015): *Factoid-based Prosopography and Computer Ontologies. Towards an integrated approach*, in: DSH 30,1: 86-97.

Bradley, John / Short, Harold (2005): *Texts into databases. The Evolving field of New-style Prosopography*, in: LLC 20, suppl. 1: 3-24.

CCEd: *Clergy of the Church of England Database*, King's College London http://theclergydatabase.org.uk/

DPRR: *Digital Prosopography of the Roman Republic*, King's College London

Keats-Rohan, Katherine S.B. (ed.) (2007): *Prosopography. Approaches and Applications. A Handbook* (Prosopographica et genealogica 13). Oxford: P&G.

Lind, Gunner (1994): *Data Structures for Computer Prosopography*, in: Yesterday: Proceedings from the 6th International Conference of the Association of History and Computing, Odense 1991. Odense: University Press of Southern Denmark. 77-82.

Neumann, Gerald / Körner, Fabian / Roeder, Torsten / Walkowski, Niels-Oliver (2011): *Personendaten-Repositorium*, in: Berlin-Brandenburgische Akademie der Wissenschaften. Jahrbuch 2010: 320-326.

PASE: *Prosopography of Anglo-Saxon England*, King's College London, URL: http://www.pase.ac.uk/jsp/index.jsp

Schlögl, Matthias / Katalin Lejtovicz (2018): *A Prosopographical Information System (APIS)*, in: **Antske Fokkens / ter Braake Serge / Sluijter, Ronald / Arthur, Paul / Wandl-Vogt, Eveline (eds.):** *BD-2017. Biographical Data in a Digital World 2017. Proceedings of the Second Conference on Biographical Data in a Digital World 2017*. Linz, Austria, November 6-7, 2017. Budapest: CEUR (CEUR Workshop Proceedings 2119): 53-58.

Schlögl, Matthias / Lejtovicz, Katalin / Bernád, Ágoston Zénó / Kaiser, Maximilian / Rumpolt, Peter (2018): *Using deep learning to explore movement of people in a large corpus of biographies*. Zenodo. http://doi.org/10.5281/zenodo.1149023

Tuominen, Jouni / Hyvönen, Eero / Leskinen, Petri (2018): *Bio CRM. A Data Model for Representing Biographical Data for Prosopographical Research*, in: BD-2017. Biographical Data in a Digital World 2017, hg. v. Antske Fokkens, Serge ter Braake, Ronald Sluijter, Paul Arthur, Eveline Wandl-Vogt, Budapest: CEUR (CEUR Workshop Proceedings 2119): 59-66.

Tzompanaki, Katerina / Doerr Martin (2012): *Fundamental Categories and Relationships for intuitive querying CIDOC-CRM based repositories*, Technical Report ICS-FORTH/TR-429, April 2012, <http://cidoc-crm.org/docs/TechnicalReport429_April2012.pdf>