The Person Data Repository

Roeder, Torsten
roeder@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

The *Berlin-Brandenburg Academy of Sciences and Humanities* (Berlin-Brandenburgische Akademie der Wissenschaften, BBAW) is the largest non-university research institution in the region. TELOTA (*The Electronic Life of the Academy*), an initiative for academically applied information technology, was launched here in 2002. The initiative supports the academy's projects by developing IT solutions for research work and digital publications.

The project "Construction of a repository for biographical data on historical persons of the 19th century" – short form: *Person Data Repository* – enhances the existing approaches to data integration and electronically supported research in biographies. It investigates connecting and presenting heterogeneous information on persons of the "long nineteenth century" (1789–1914). The project's aim is to provide a de-central software system for research institutions, universities, archives, and libraries that allows combined access on biographic information from different data pools.

The project is subdivided into three major fields: 1) conceptual design of an adequate data model, which embraces different methods and perspectives; 2) data exchange with national and international cooperation partners; and 3) development of a software solution based on an evaluated framework. The project is funded by the DFG (*Deutsche Forschungsgemeinschaft*, German Research Foundation). The work began in July 2009 with three academic staff members and three student assistants, and will continue for two years.

1. Data Modelling

To structure heterogeneous biographical data, the project pursues a novel approach, which was already presented in a talk at the workshop "Personendateien – Elektronisches Publizieren" in September 2009 in Leipzig (see link below). The approach does not define a person as single data record, but rather as compilation of all statements concerning that person. Thus, it is possible to display complementing as well as contradicting statements in parallel, which meets one of the basic challenges of biographic research.

In the above lecture by Niels-Oliver Walkowski, he notes: "Biographic research, understood as the creation of identifying narrations, performs semantic constructions, which were caused by a human, but which are not identical to it. The consequences are concurring narrations, polysemy and contingency, which are not an expression of lacking knowledge, but are due to the conditions of biographic research."

In order to satisfy different research approaches and perspectives, the smallest entity of the Person Data Repository is not a person, but a single statement on a person, which is named "aspect" in the data model. An aspect bundles references to persons, places, dates and sources. By proper queries it will be possible to create further narrations, whose first dimension is not necessarily a person, but possibly also a time span or a certain location. Those attained insights can enhance the knowledge on persons in turn.

Additionally, all aspects are connected to the corresponding source and to current identification systems respectively, like the LCCN or the German PND. Thus, scientific transparency and compatibility with existing and future systems is guaranteed.

2. Cooperation

As the Person Data Repository acquires data from partners, focusing on data organisation rather than conducting its own research, cooperation is an essential part of the project. The mission statement is guided by the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, which has also been signed by the *Berlin Brandenburg Academy of Sciences and Humanities*, and which promotes the exchange of scientific knowledge through digital media as well as transparency of sources and authors.

Initially, several projects of the BBAW were invited to share their data on the Person Data Repository. Amongst them are the *Marx-Engels-Gesamtausgabe* (MEGA, Complete works by Marx and Engels), the *Alexander-von-Humboldt-Forschungsstelle*, the *Protokolle des Preußischen Staatsministeriums* (Protocols of the Prussian Ministry of State), the *Berliner Klassik* (Classical Berlin) and the *Altmitgliederverzeichnis* (Index of Former Academy Members). Thus, an inventory of several thousand persons is already available; partnerships with further BBAW projects are planned in order to reach about 200,000 entries.

Parallel to this, partnerships with external institutions are being formed. These cooperations will range from sharing unstructured data to data exchange with existing repositories. Contacts with edition projects, image databases and repository databases have already been formed and will be developed during the project's course. A basis for exchange and publication of the gathered person information should be delivered by Open Access oriented agreements, whereas individual arrangements can be made as well.

As it is intended to provide not only the Person Data Repository's contents, but also its infrastructure, it is also of interest for projects whose historical scope is focused outside the 19th century. In this case, it is possible set-up an infrastructure of the same type with freely configurable contents, as it were a sister repository.

3. Development

During the preliminaries for the technical realisation of the Person Data Repository, a list of established software packets has been created. An evaluation shows which of these packets will constitute the core component of the repository. As the software is provided also to other institutions, key elements of the evaluation are documentation, configurability, scalability, expandability and availability of interfaces.

A software which has already been utilized by BBAW projects for gathering person data is the “Archiv-Editor” (Archive Editor), which had been developed by the TELOTA initiative (see link below). This editor will play a central part as an editing tool in the Person Data Repository, and will be developed further according to the project's demands.

Along with the work on the repository and archive software, also the conversion of data is a part of the development field. Manual and automated methods will be utilized in the process. The atomization in single aspects is conducted in three steps: syntax analysis, index based structuring, and manual correction. The first data pool to convert was the person index of the Protocols of the Prussian Ministry of State, a completed project which provides an excellent starting basis, containing over 22,000 person entries. As the second major source for data on persons of the 19th century, the exhaustive indexes of the Complete Works of Marx and Engels have been chosen. Further projects at the academy, like the Alexander-von-Humboldt-Forschungsstelle and Classical Berlin, have placed their data pools at the disposal of Person Data Repository.

4. Perspectives

As the project aims at cooperation to a great extent, it is our wish to communicate with interested parties from all disciplines, in order to build up partnerships between our institutions. Also, we look forward to questions, remarks, proposals, and to exchanging theoretical and technical approaches. A cooperative workshop for partners and interested parties is planned for autumn 2010.

Die *Berlin-Brandenburgische Akademie der Wissenschaften* (BBAW) ist die größte außeruniversitäre Forschungseinrichtung der Region. Im Jahr 2002 wurde hier TELOTA (*The Electronic Life Of The Academy*), eine Initiative für akademisch angewandte Informationstechnologie, ins Leben gerufen. Diese unterstützt die Akademievorhaben mit der Entwicklung informationstechnischer Lösungen für Forschungsarbeit und digitale Publikation.

Mit dem DFG-Projekt „Aufbau eines Repositoriums für biografische Daten historischer Personen des 19. Jahrhunderts“ – kurz: *Personendaten-Repositorium* – werden

bisherige Ansätze der Datenvernetzung und elektronischen Biografik weiterentwickelt. Es erforscht anhand von Personeninformationen des „langen 19. Jahrhunderts“ (1789–1914), wie sich heterogene Datenbestände miteinander verbinden und präsentieren lassen. Ziel des Projektes ist die Bereitstellung eines dezentralen Softwaresystems, welches Lehr- und Forschungseinrichtungen, Archiven und Bibliotheken ermöglicht, biographische Informationen aus verschiedenen Beständen über einen gemeinsamen Zugang zu nutzen.

Das Projekt untergliedert sich in drei Teile: 1) Der Entwurf eines geeigneten Datenmodells, welches unterschiedlichen Perspektiven und Forschungsmethoden gerecht wird, 2) der Datenaustausch mit Kooperationspartnern im In- und Ausland, und 3) die Entwicklung einer Software-Lösung auf der Basis eines zu evaluierenden Framework. Das Projekt wurde von TELOTA bei der DFG beantragt und ist für die Laufzeit von zwei Jahren bewilligt worden. Im Juli 2009 wurde die Arbeit mit drei wissenschaftlichen Mitarbeitern und drei studentischen Hilfskräften aufgenommen.

1. Datenmodellierung

Zur Strukturierung heterogener biographischer Daten verfolgt das Projekt einen neuartigen Ansatz, der bereits in einem Vortrag auf dem Workshop „Personendateien – Elektronisches Publizieren“ im September 2009 in Leipzig vorgestellt wurde (siehe Link unten). Eine Person wird darin nicht als einzelner Datensatz definiert, sondern vielmehr als die Menge aller Aussagen, die zu ihr getroffen werden. Damit ist es möglich, sowohl sich ergänzende als auch sich widersprechende Aussagen nebeneinander abzubilden, was grundlegenden Problemen biografischen Arbeitens Rechnung trägt.

In dem erwähnten Vortrag von Niels-Oliver Walkowski hieß es: „Biografisches Arbeiten verstanden als Erzeugung von identifizierenden Narrationen vollzieht semantische Konstruktionsleistungen, zu denen ein Mensch den Anlass gab, der aber nicht mit ihm zusammenfällt. Eine Folge sind konkurrierende Narrationen, Polysemie und Kontingenz, die nicht Ausdruck mangelnder Kenntnis, sondern den Voraussetzungen biografischen Arbeitens an sich geschuldet sind.“

Gerade also, um verschiedenen Forschungsansätzen und Perspektiven gerecht zu werden, ist die kleinste Dateneinheit des Personendaten-Repositoriums nicht eine Person, sondern eine einzelne Aussage zu einer Person, die in dem Datenmodell „Aspekt“ genannt wird. Ein Aspekt bündelt Bezüge zu Personen, Orten, Daten und einer Quelle. Dadurch wird es möglich sein, durch eine entsprechende Abfrage weitere Narrationen zu erzeugen, bei denen nicht unbedingt eine einzelne Person, sondern auch ein Zeitraum oder ein Ort die erste Dimension bilden könnte. Daraus gewonnene Erkenntnisse erweitern wiederum das Personenwissen.

Zudem werden die Aspekte einerseits mit den jeweiligen Quellen, andererseits mit geläufigen Identifikationssystemen, etwa mit PND und LCCN, verknüpft. Dadurch bleibt die wissenschaftliche Transparenz und die Kompatibilität mit bestehenden und zukünftigen Systemen gewährleistet.

2. Kooperationen

Da das Personendaten-Repositorium die Daten über seine Partner bezieht und sich selbst auf die Organisation der Daten konzentriert, anstatt eigene Datenbestände zu erarbeiten, ist Zusammenarbeit ein essenzieller Bestandteil des Projektes. Von richtungweisender Bedeutung sind dabei die Ziele der *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*, die auch von der Berlin-Brandenburgischen Akademie der Wissenschaften unterzeichnet wurde und welche den wissenschaftlichen Austausch durch digitale Technik unter Gewährleistung von Quellen- und Autorentransparenz befördert.

Zunächst wurden einige Vorhaben der Berlin-Brandenburgischen Akademie der Wissenschaften dazu eingeladen, ihre Daten auf der zukünftigen Repositorien-Plattform verfügbar und verknüpfbar zu machen. Darunter fallen die *Marx-Engels-Gesamtausgabe*, die *Alexander-von-Humboldt-Forschungsstelle*, die *Protokolle des Preußischen Staatsministeriums* sowie das Altmitgliederverzeichnis aus dem Akademiearchiv. Damit liegen bereits Daten zu

mehreren zehntausend Personen vor; weitere Kooperationen mit Akademievorhaben sind in Vorbereitung, um ca. 200.000 Einträge zu erreichen.

Parallel dazu werden externe Institutionen als Partner herangezogen. Die Kooperationsmöglichkeiten reichen von der Übernahme unstrukturierter Datenbestände bis hin zum Austausch mit anderen Repositorien. So sind bereits Kontakte mit mehreren Editionsprojekten, Bilddatenbanken und Verbund-Datenbanken geknüpft worden, die im weiteren Verlauf des Projektes zu vertiefen sind. Die Grundlage für Austausch und Veröffentlichung der Personendaten schafft im besten Falle eine Vereinbarung im Sinne des *Open Access*, wobei auch individuell davon abweichende Verabredungen getroffen werden können.

Da beabsichtigt ist, nicht nur die Daten, sondern auch die Infrastruktur des Personendaten-Repositoriums zur Verfügung zu stellen, ist das Projekt auch für solche Vorhaben interessant, deren historischer Rahmen das 19. Jahrhundert nicht berührt. In diesem Fall kann eine Abmachung über die Einrichtung einer gleichartigen Infrastruktur mit selbst bestimmbaren Inhalten, also eine Art „Schwester-Repositorium“, geschlossen werden.

3. Entwicklung

Im Rahmen der Vorbereitungen für die praktische Umsetzung des Personendaten-Repositoriums wurde eine Liste etablierter Software-Pakete erstellt. Eine Evaluation zeigt, welches die Kernkomponente der Datenhaltung des PDR bildet. Da die Software auch anderen Projekten zur Verfügung steht, gehören Dokumentation, Konfigurierbarkeit, Skalierbarkeit, Erweiterbarkeit und Schnittstellen zu den wesentlichen Anforderungen der Evaluation.

Als bereits längerfristig genutzte Software zur Erfassung von Personendaten existiert innerhalb der BBAW der von der TELOTA-Initiative entwickelte Archiv-Editor (Link s. u.). Dieser wird im Rahmen des PDR eine zentrale Rolle als Werkzeug für die Eingabe von Personendaten spielen und wird entsprechend der veränderten Anforderungen weiterentwickelt.

Neben der Arbeit an der Repositorien- und Archivsoftware fällt auch die Konvertierung von Datenbeständen in den Entwicklungsbereich. Dabei werden sowohl manuelle als auch automatische Verfahren eingesetzt. Die Zerlegung der Biogramme in Einzelaspekte erfolgt zunächst anhand einer einfachen Syntaxanalyse, wird dann über Abkürzungs-, Orts- und Personenverzeichnisse tiefstrukturiert und zum Abschluss manuell korrigiert. Begonnen wurde mit dem Personenregister der *Protokolle des Preußischen Staatsministeriums*, welche sich als abgeschlossenes Projekt und mit über 22.000 Kurzbiogrammen als hervorragende Ausgangsbasis anbot. Als zweite Quelle für Personendaten des 19. Jahrhunderts wurden die umfangreichen Register der *Marx-Engels-Gesamtausgabe* ausgewählt. Weitere Akademienvorhaben, etwa die *Alexander-von-Humboldt-Forschungsstelle*, haben ihre Daten ebenfalls bereits zur Verfügung gestellt.

4. Ausblick

Da unser Projekt in großem Maße auf Kooperationen setzt, ist es unser Wunsch, mit interessierten Projekten aus allen denkbaren Fachgebieten ins Gespräch zu kommen und Partnerschaften zu schließen. Ebenso freuen wir uns auf Fragen, Hinweise und Anregungen sowie auf den Austausch von theoretischen und technischen Lösungsansätzen. Ein Workshop, zu dem sowohl ähnliche Projekte, Kooperationspartner und interessiertes Fachpublikum eingeladen werden, ist für Herbst 2010 geplant.

Further information

- Website of the Person Data Repository `pdr.bbaw.de`
- The TELOTA Initiative of the BBAW `www.bbaw.de/telota`
- The "Archiv-Editor" (Introduction and Download) `http://www.bbaw.de/telota/projekte/personendatenbank-1/archiv-editor`
- "Personendateien" Workshop in Leipzig `www.saw-leipzig.de/aktuelles/personendateien`

Events

Workshop: Personen - Daten - Repositorien (Persons - Data - Repositories), 27th – 29th September 2010 Berlin-Brandenburg Academy of Sciences and Humanities: pdr.bbaw.de/workshop.

References

Brase, J., Klump, J. (2007). 'Zitierfähige Datensätze: Primärdaten-Management durch DOIs'. *Rafael Ball, Wissenschaftskommunikation der Zukunft*. Jülich: Forschungszentrum Jülich. http://books.google.com/books?id=kouJ09GQtbcC&lpg=PA159&ots=YuCzaRjSDM&dq=Zitierf %20hige%20Datens%20tze%3A%20Prim%20rdaten-Management%20durch %20DOI&lr=&pg=PA159#v=onepage&q=&f=false.

Costa, Stefano (2010). *Open Data in Archaeology*. http://blog.okfn.org/2010/02/25/open-data-in-archaeology/.

Dallmeier-Tiessen, S., Dobratz, S., Gradmann, S., Horstmann, W., Kleiner, E., Pampel, H. (et al.). *Positionspapier Forschungsdaten*. http://edoc.gfz-potsdam.de/gfz/13230.

Dallmeier-Tiessen, Sunje, Pfeiffenberger, Hans (2009). 'Umgang mit Forschungsdaten in den Geowissenschaften - Ein Blick in die Praxis'. *Bibliothekartag 2009*. Erfurt: Berufsverband Information Bibliothek e.V.. http://www.opus-bayern.de/bib-info/volltexte/2009/699/pdf/dallmeier-tiessen_geowissenschaften.pdf.

DINI, Arbeitsgruppe "Elektronisches Publizieren" (2009). *Positionspapier Forschungsdaten*. Humboldt- Universität zu Berlin. http://edoc.hu-berlin.de/series/dini-schriften/2009-10/PDF/10.pdf.

Griese, B., Griesehop, H. R. (2007). *Biographische Fallarbeit*. Wiesbaden: VS Verlag für Sozialwissenschaften. http://www.ulb.tu-darmstadt.de/tocs/177279664.pdf.

Hackländer-von der Way, Bettina (2001). *Biographie und Identität*. Berlin. http://dissertation.de.

Henning, Tim (2009). *Person sein und Geschichten erzählen, Quellen und Studien zur Philosophie*. Berlin u.a.: de Gruyter.

Hermann, Elfriede (2003). *Lebenswege im Spannungsfeld lokaler und globaler Prozesse: Person, Selbst und Emotion in der ethnologischen Biografieforschung*. Münster: LIT.

Hoerning, Erika (2000). *Pierre Bourdieu: Die biographische Illusion*. Stuttgart: Lucius & Lucius.

Hoerning, Erika (2000). *Biographische Sozialisation*. Stuttgart: Lucius & Lucius.

Hutto, Daniel D. (2007). *Narrative and understanding persons*. Cambridge University Press. http://books.google.de/books?id=peHYAAAAMAAJ.

Hutto, Daniel (2007). *Framing Narratives*. Cambridge/New York: Cambridge University Press.

Key Perspectives Ltd.. *Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability*. http://www.dcc.ac.uk/docs/publications/SCARP SYNTHESIS.pdf.

Kramer, Christine (2001). *Lebensgeschichte, Authentizität und Zeit*. Frankfurt am Main u.a: Lang.

Kripke, Saul (1981). *Name und Notwendigkeit*. Frankfurt a.M.: Suhrkamp.

Mackenzie, Catriona (2008). *Practical identity and narrative agency*. New York: Routledge.

Marotzi, Winfried (2001). *Methodologie und Methoden der Biographieforschung*. Hohengehren: Schneider.

Moore-Gilbert, B. (2009). *Postcolonial life-writing: culture, politics and self-representation*. London/New York: Routledge.

NESTOR Arbeitsgruppe Grid/e-science und Langzeitarchivierung (2009). *nestor-bericht - Digitale Forschungsdaten bewahren und nutzen - für die Wissenschaft und die Zukunft*. Frankfurt am Main. http://nbn-resolving.de/nbn:de:0008-2009071031.

Neuroth, Heike, Jannidis, Fotis, Rapp, Andrea, Lohmeier, Felix (2009). 'Virtuelle Forschungsumgebungen für e-Humanities. Maßnahmen zur optimalen Unterstützung von Forschungsprozessen in den Geisteswissenschaften'. *Bibliothek, Forschung und Praxis*. **33 (2)**: 161-169. http://www.reference-global.com/doi/abs/10.1515/bfup.2009.017.

Schwiegelsohn, U. (2009). 'Grids als neue Komponenten des Integrierten Informationsmanagements'. *Praxis der Informationsverarbeitung und Kommunikation*. **32 (1)**: 29-32. http://www.reference-global.com/doi/pdfplus/10.1515/piko.2009.006.

Wettlaufer, Jörg. 'Personendateien. Workshop der Arbeitsgruppe Elektronisches Publizieren der Union der deutschen Akademien der Wissenschaft - H-Soz-u-Kult / Tagungsberichte'. *H-Soz-u- Kult*. http://hsozkult.geschichte.hu- berlin.de/tagungsberichte/id=2806&count=126&recno=9&sort=datum&order=down&search=presse& epoche=22.