Einbindung von Primärdaten in Digitale Bibliotheken

Stempfhuber, Maximilian; Zapilko, Benjamin

Postprint / Postprint
Sammelwerksbeitrag / collection article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Stempfhuber, M., & Zapilko, B. (2009). Einbindung von Primärdaten in Digitale Bibliotheken. In R. Kuhlen (Hrsg.), *Information: Droge, Ware oder Commons? Wertschöpfungs- und Transformationsprozesse auf den Informationsmärkten* (S. 1-6). Boizenburg: Hülsbusch. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-464422

Einbindung von Primärdaten in Digitale Bibliotheken

Maximilian Stempfhuber & Benjamin Zapilko

GESIS – Leibniz-Institut für Sozialwissenschaften
D-53111 Bonn
E-Mail: {max.stempfhuber | benjamin.zapilko}@gesis.org

Zusammenfassung

Obwohl auf Nutzerseite ein hoher Bedarf an einer integrierten Recherche in verschiedenen Informationstypen besteht, ist eine konsequente Einbindung von Primärdaten in Digitale Bibliotheken bislang nur selten anzutreffen. Am Beispiel sozialwissenschaftlicher Studien wird gezeigt, wie Faktendaten auf konzeptioneller Ebene in einer Digitalen Bibliothek verfügbar gemacht und mit den dort vorhandenen textuellen Informationen integriert werden können. Dabei werden auch Konzepte des Semantic Web berücksichtigt.

1 Motivation

Neuere Studien (vgl. [Poll 2004]) belegen, dass bei Wissenschaftlern ein dringender Bedarf am integrierten Zugriff auf fachlich gebündelte, wissenschaftliche Informationen besteht. Bislang sind diese Informationen (z. B. sozialwissenschaftliche Studien, statistische Daten, Publikationen, usw.) häufig auf spezialisierte Angebote verteilt, wobei die Erschließungsdaten unterschiedlich strukturiert und für die Inhaltserschließung verschiedene Verfahren verwendet werden. Aus Nutzersicht ergibt sich dadurch das Problem, dass zur Befriedigung eines Informationsbedürfnisses mehrere Informationssysteme genutzt werden müssen, die sich jeweils in Bedienung, Inhalt und anzuwendender Recherchestrategie unterscheiden. Nutzerstudien im Projekt ELVIRA[1], bei dem ein Konzept zur integrierten Recherche in Texten und statistischen Zeitreihen erarbeitet wurde (vgl. [Krause et al. 1997a]) ga-

[1] ELVIRA; gefördert durch das BMWi

ben deutliche Hinweise darauf, dass sich Texte und Fakten komplementär verhalten, Texte z. B. Lücken in Faktendaten schließen oder für deren Interpretation dienen können und damit häufig zusammen recherchiert werden.

Daneben sehen sich Digitale Bibliotheken in jüngster Zeit immer mehr in der Pflicht, den kompletten wissenschaftlichen Forschungszyklus zu dokumentieren und zugänglich zu machen (vgl. [DELOS 2005] und [Gold 2007]).

2 Einbindung des Datenbestandskataloges in das Fachportal SOWIPORT

Das sozialwissenschaftliche Fachportal SOWIPORT[2] bietet Nutzern einen integrierten Zugriff auf 15 Datenbanken verschiedener Anbieter, die wiederum unterschiedliche Informationsarten wie Literaturnachweise, Nachweise von Forschungsprojekten, Veranstaltungen, Institutionsprofile etc. bereitstellen. Ergänzt werden soll dieses Angebot durch den Datenbestandskatalog von GESIS, der Studienbeschreibungen von internationalen, sozialwissenschaftlichen Studien enthält. Dabei sollen die Studienbeschreibungen mit den übrigen Informationstypen für die integrierte Suche verbunden werden. Hierfür muss der dabei auftretenden Heterogenität begegnet werden:

- Strukturelle Heterogenität: Die Metadatenschemata für Forschungsdaten (z. B. DDI[3] oder SDMX[4]) unterscheiden sich erheblich von den für Publikationen gebräuchlichen Formaten (z. B. DublinCore[5]).
- Semantische Heterogenität: Bei Literaturinformation und bei Primärdaten finden unterschiedliche Sacherschließungssysteme (z. B. Nomenklaturen, Klassifikationen und Thesauri) Anwendung. Dies erschwert es, Nutzeranfragen zwischen den Informationstypen, aber auch zwischen unterschiedlichen Sammlungen eines einzigen Informationstyps abzubilden.

In SOWIPORT wird der strukturellen Heterogenität dadurch begegnet, dass für jeden einzelnen Informationstyp ein Standardformat entwickelt wurde,

[2] http://www.sowiport.de
[3] Standard der Data Documentation Initiative; http://www.ddialliance.org
[4] Statistical Data and Metadata Exchange Standard; http://www.sdmx.org
[5] http://www.dublincore.org

das alle relevanten Felder sämtlicher Einzelstandards sowie Verknüpfungsinformationen zu externen Daten und Systemen auf ein gemeinsames Datenschema normiert. Im Falle sozialwissenschaftlicher Studien orientiert sich dieses interne Standardformat am DDI-Standard. Daraus ergeben sich mehrere Vorteile: Verringerung der Komplexität der Datenhaltung (pro Informationstyp existiert genau ein Standardformat), Minimierung der vom System parallel zu generierenden Teilanfragen sowie die Reduzierung des Aufwands zur Zusammenführung der Teilergebnisse und zur Generierung einer einheitlichen Ergebnisliste.

Um die Beziehungen zwischen den einzelnen Informationstypen zu modellieren, bietet sich die Verwendung gängiger Modelle wie das CERIF-Modell[6] (Common European Research Information Format) oder die Policy Grid Ontologie[7] (vgl. [Chorley et al. 2008]) an. Dadurch können formale Relationen zwischen allen Akteuren, Aktionen und Ergebnissen abgebildet werden, die am wissenschaftlichen Forschungsprozess beteiligt sind oder daraus entstehen. Dies erlaubt es z. B. ein Projekt und die darin durchgeführten Studien zu relationieren und diese Relationen dann im Information Retrieval zu nutzen. Diese explizit ausgedrückten Verbindungen stellen den intellektuell kontrollierten Kern eines Forschungsinformationssystems dar, auf dem die nachfolgend beschriebenen weiteren Verfahren der semantischen Heterogenitätsbehandlung basieren.

Um der semantischen Heterogenität entgegen zu wirken, werden in SOWIPORT die Terme einer Nutzeranfrage mithilfe von Crosskonkordanzen automatisch zwischen den kontrollierten Vokabularen der Informationssammlungen abgebildet. Hierdurch erhöht sich der Recall einer Anfrage insbesondere bei den Informationssammlungen, in denen der vom Nutzer verwendete Suchbegriff nicht für die Indexierung verwendet wurde (vgl. [Mayr&Petras 2008]). Bei der Integration des Datenbestandskataloges in SOWIPORT kommt in einem ersten Schritt eine direkte Abbildung von dessen eigenem Kategoriensystem auf den Thesaurus Sozialwissenschaften zum Einsatz, von dem aus wiederum eine Reihe von Crosskonkordanzen zu anderen Thesauri bestehen.

---

[6] http://www.eurocris.org/cerif/introduction/

[7] http://www.policygrid.org/

Eine inhaltliche Erschließung auf der Ebene von Thesauri und Klassifikationen reicht im Kontext von Studienbeschreibungen jedoch nicht aus, da diese komplexer aufgebaut sind und neben der eigentlichen Studienbeschreibung in der Regel noch Fragebögen und Codebücher enthalten, die die einzelnen Fragen und die damit erhobenen Variablen beschreiben. Die gängige Praxis, die Studien in ihrer Gesamtheit mit Schlagwörtern zu indexieren, reflektiert allerdings nicht die sozialwissenschaftliche Intention einzelner Fragestellungen und führt daher zu einem sehr hohen Recall bei gleichzeitig sinkender Precision, was für den Nutzer einen deutlich höheren Aufwand zur Identifikation relevanter Ergebnisse bedeutet.

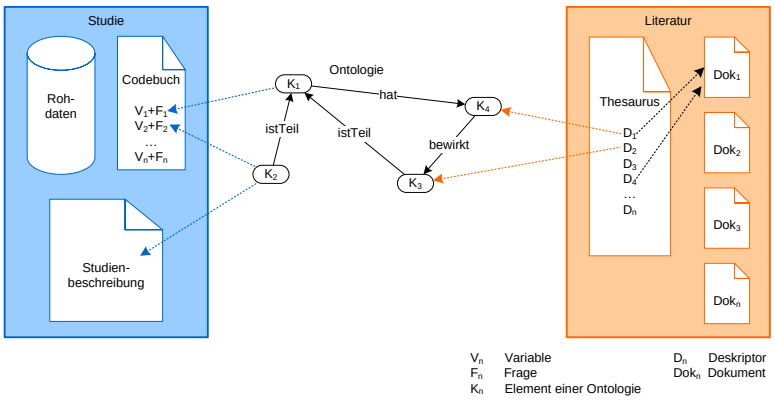

Abb. 1: Inhaltserschließung mittels Ontologie

Hier ist der Einsatz von Ontologien zu untersuchen, mit denen die Studien insgesamt sowie deren einzelnen Fragen und Variablen semantisch präziser erschlossen werden könnten als dies mit Thesauri der Fall ist. Dieser Ansatz verknüpft die semantisch reichhaltige Inhaltsrepräsentation für Primärdaten mittels Ontologien mit der Inhaltserschließung für Texte mittels Thesauri ohne Präkombination (vgl. [Krause&Stempfhuber 2005]) und gleicht Vor- und Nachteile hinsichtlich Indexierungsaufwand und notwendiger Präzision miteinander aus.

3 Ausblick

Der Datenbestandskatalog von GESIS wurde in einer ersten Version in die integrierte Recherche des sozialwissenschaftlichen Fachportals SOWIPORT

eingebunden. Die mit den geschilderten Verfahren zur Termtransformation erreichbare Qualität der Recherche wird in einem nächsten Schritt evaluiert und in Abhängigkeit vom Ergebnis optimiert. Es ist zu erwarten, dass vor allem alternative Verfahren der Anfrageerweiterung mittels Crosskonkordanzen entwickelt werden müssen, um dem großen mengenmäßigen Unterschied im Umfang der kontrollierten Vokabulare von Forschungsdaten und Thesauri zu begegnen und den Recall kombinierter Anfragen zu erhöhen. Darüber hinaus soll auf Basis der gewonnenen Erfahrungen ein allgemeines Modell für einen integrierten Text-Fakten-Zugriff in Digitalen Bibliotheken entwickelt werden.

Literatur

[Chorley et al. 2008] Chorley, A.; Edwards, P.; Hielkema, F.; Philip, L.; Farrington, J.: Developing Ontologies to Support eSocial Science: The PolicyGrid Experience. In: Proceedings of the 4th International Conference on e-Social Science 2008 Manchester.

[DELOS 2005] The DELOS Network of Excellence on Digital Libraries: Recommendations and Observations for a European Digital Library (EDL). 4th DELOS Brainstorming Workshop on Digital Libraries. Dezember 2005.

[Gold 2007] Gold, Anna: Cyberinfrastructure, Data and Libraries. Part 1 & 2. In: D-Lib Magazine, Volume 13 Number 9/10.

[Krause et al. 1997] Krause, Jürgen; Mandl, Thomas; Stempfhuber, Maximilian: Text-Fakten-Integration in ELVIRA. Bonn: GESIS, IZ-Arbeitsbericht Nr. 12.

[Krause 2003] Krause, Jürgen: Standardisierung von der Heterogenität her denken – Zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken. Bonn: GESIS, IZ-Arbeitsbericht Nr. 28.

[Krause&Stempfhuber 2005] Krause, Jürgen; Stempfhuber, Maximilian: Nutzerseitige Integration sozialwissenschaftlicher Text- und Dateninformationen aus verteilten Quellen. In: König, C.; Stahl, M.; Wiegand, E. (Hrsg.): Datenfusion und Datenintegration: 6. Wissenschaftliche Tagung. Bonn: GESIS, Tagungsberichte, Bd. 10. S. 141-158.

[Mayr&Petras 2008] Mayr, P.; Petras, V.: Cross-concordances: terminology mapping and its effectiveness for information retrieval. IFLA World Library and Information Congress. (erscheint)

[Poll 2004] Poll, Roswitha: Nutzungsanalyse des Systems der überregionalen Literatur- und Informationsversorgung. Teil 1: Informationsverhalten und Informationsbedarf der Wissenschaft. In: ZfBB 51 (2004), S. 59-75.