Metadaten im Zeitalter von Google Dataset Search

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Data Center for the Humanities, Universität zu Köln

Rau, Felix
frau@uni-koeln.de
Institut für Linguistik, Universität zu Köln

Perspektiven auf Kuratieren, Suchen, Finden und Präsentieren aus der Praxis

Relevanz des Themas

Metadaten sind Teil einer jeden datenbezogenen Forschungspraxis und ein essentieller Bestandteil einer jeden Forschungsdatenmanagementstrategie. Auch wenn Metadatenproduktion und -kuration oft einen von den Forscher*innen ungeliebten Aspekt der Projektarbeit darstellt, ist die durch Metadaten sichergestellte Find- und Zitierbarkeit von Daten für die Forschungsgemeinschaft und insbesondere auch für die Geldgeber von höchstem Interesse. Einzelne Datenzentren und nationale und internationale Forschungsinfrastrukturen haben seit Jahren in die Verbesserung der Findmechanismen investiert. Mit dem Eintritt von großen kommerziellen Anbietern wie Google und Elsevier in die Suche wissenschaftlicher Daten(sätze) kündigen sich in den letzten Monaten massive Änderungen in der Landschaft an. Als einzelnes fach- oder datentypspezifisches Repositorium müssen wir in dieser Landschaft den Erwartungen und Anforderungen von Depositor, Geldgebern, und potentiellen Nachnutzern gerecht werden, während wir mit unseren Geld- und Personalressourcen verantwortungsvoll umgehen.

Problemstellung

Forschungsdatenzentren und Forschungsinfrastrukturen wenden einen bedeutenden Teil ihrer personellen und finanziellen Ressourcen für die Modellierung und Kuration von Daten und Metadaten auf. Bereitstellung und Erhalt von oft komplexen Suchfunktionalitäten und die Schaffung von Anschlussfähigkeit der Metadatenschemata mit externen wissenschaftlichen und fachspezifischen Suchportalen ist arbeitsaufwändig und wartungsintensiv.

Viele Forschungsdatenzentren haben in den letzten Jahren komplexe Metadatenschemata entwickelt, die die Vielschichtigkeit des Gegenstandes detailliert abbilden. Dabei gibt es durchaus eine Tendenz zur fachspezifischen Übermodellierung von Metadaten in hochkomplexen Schemata. Diese repositoriums- und gegenstandsspezifischen Schemata bilden sowohl den grundlegenden Inhalt der Webseiten des einzelnen Repositoriums als auch das Ausgangsmaterial um externe wissenschaftliche Datensuchportale zu bedienen.

Da diese komplexen Datensätze der einzelnen Portale in den gängigen Findmechanismen des Webs – insbesondere in der noch stets dominanten Google Web Search – aus verschiedenen Gründen schlecht erfasst werden, wurden in den letzten Jahrzehnt Metasuchportale entwickelt um eine repositoriumsübergreifende Findbarkeit sicherzustellen. Diese Metaportale sind normalerweise an eine Domäne gebunden. Die Domäne kann der Nationalstaat sein, wie beim niederländischen Portal NARCIS[1], sie kann an eine Forschungsinfrastruktur gebunden sein, wie im Falle des CLARIN VLO[2], sie kann fachspezifisch sein, wie im Falle des OLAC-Portals für Spracharchive[3], oder der Service kann an einen anderen Infrastrukturservice gekoppelt sein, wie die Koppelung der DataCite-Suche[4] an die DOI-Registrierung.

Es ist unumstritten, dass eine einfache Auffindbarkeit und Zugänglichkeit von Datensätzen ein zentraler Baustein der Nachnutzung von Forschungsdaten ist und deshalb die Möglichkeit gefunden zu werden höher zu bewerten ist, als die Genauigkeit der zu grunde liegenden Metadaten, und die verschiedenen Metasuchportale versprechen dies zu leisten. Die Existenz dieser Metaportale hat dazu geführt, dass einzelne Repositorien die Investition in maßgeschneiderte Webinterfaces verzichtet und sich Datenzentren auf Kuration und Bereitstellung der Metadaten über (Harvesting-)Schnittstellen konzentrieren. Dies ist auch eine Reaktion auf die Tatsache, dass die in wissenschaftlichen Projekten erstellten Portale oft nicht die Qualität und Nutzerfreundlichkeit erreichen, die Nutzer*innen aus kommerziellen Webangeboten gewöhnt sind, und anspruchsvolle Webseiten, bedingt durch die Kurzlebigkeit modernen Webtechnologien sehr wartungs- und kostenintensiv sind.

Um die Diversität und Komplexität der einzelnen Metadatenschemata handhabbar zu machen, reduzieren die Metaportale die Angaben auf kompakte und flache Strukturen, die oft aus nicht mehr als Listen von Key-Value-Paaren bestehen. Das DARIAH-DE Repository hat sich für den geradezu radikalen Schritt entschieden für das Anlegen einer Kollektion im Deposit-Interface “Publikator” gerade mal drei Pflichtfelder zu definieren (Mache & Klaffki 2018). Diese sind *Titel*, *Urheber*in* und *Rechteverwaltung* (Linzenzbestimmung).

Die komplexen Metadatenschemata in Kombination mit der Verarbeitung der Metadaten in den Metaportale hat aber auch den Effekt, dass Angaben, die nicht mit Blick auf die vereinfachte Repräsentation in den Metaportalen angelegt wurden, ohne Kontext nicht mehr verständlich sind. So kann ein Feld “Description”, das sich auf ein Objekt wie Sprache oder Sprecher in einem komplexen Metadatenschemas bezieht, in einem

Metaportal so dargestellt werden, als würde es sich auf den gesamten Datensatz beziehen.

Zustand

Das *Kölner Zentrum Analyse und Archivierung von AV # Daten* (KA³) wird seit 2015 als Verbundvorhaben vom BMBF gefördert. Partner sind das *Max-Planck-Institut für Psycholinguistik* in Nimwegen (Niederlande), das *Fraunhofer-Institut IAIS* in Sankt Augustin, das *Archiv "Deutsches Gedächtnis"* der FernUniversität Hagen sowie drei Akteure an der Universität zu Köln - das *Regionalen Rechenzentrum*, das *Institut für Linguistik* und das *Data Center for the Humanities*. Ein zentraler Bestandteil des Vorhabens ist die Etablierung eines Forschungsdatenrepositoriums am Kölner Standort. Hier werden vor allem Daten aus der linguistischen Sprachdokumentation kuratiert und archiviert und eine Ausweitung auf annotierte oder transkribierte audiovisuelle Daten aus methodisch verwandten Fachdisziplinen wie z.B. den ethnologischen Fächern oder der Oral History vorbereitet. Ein festgeschriebenes Projektziel ist die vollständige Integration des Repositoriums und Archivs *Language Archive Cologne* in die CLARIN-Forschungsinfrastruktur. Die Nähe zum Forschungsalltag der linguistischen Sprachdokumentation und die Anbindung an CLARIN bringen Bedingungen und Traditionen mit, die zu berücksichtigen sind.

Die Mehrheit der Metadaten in diesem Fachbereich liegen im IMDI-Format[5] vor. Das Format wird aktiv nur noch in wenigen Spracharchiven weltweit eingesetzt, hat aber nachhaltig die Konzeptualisierung von Metadaten in der Fachgemeinschaft geprägt. Der ausladende Charakter und das akademische Eingabetool, das primär genutzt wurde,[6] wurden in dem Wunsch geschaffen, die Bandbreite der Forschungstätigkeit zu inkludieren. Hier verschwimmen die Grenzen zwischen Daten und Metadaten.

Die CLARIN-Forschungsinfrastruktur macht klare Vorgaben bezüglich der Vorhaltung und Publikation von Metadaten durch die angeschlossenen Zentren. Metadaten müssen im CMDI-Metadatenformat verfügbar und über eine OAI-PMH Schnittstelle harvestbar sein. Die in der CMDI Component Metadata Registry definierten Profile können frei im Rahmen der existierenden Best Practices modelliert werden (CMDI 2018). Durch ein semantisches Mapping über die CLARIN Concept Registry soll eine semantische Interoperabilität von Metadatenkategorien gewährleistet werden.

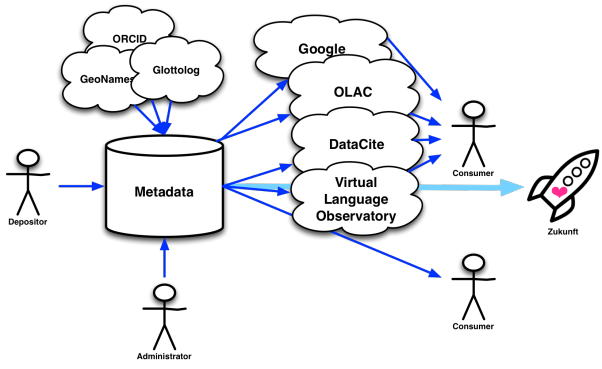
Metadaten in der CLARIN-Infrastrukur werden durch das Metasuchportal *Virtual Language Observatory* (VLO) geharvestet und dargestellt. Durch eine festgelegte Anzahl von Metadatenkategorien, die in jedem spezifischen Metadatenprofil durch Vergabe der entsprechenden Konzepten kenntlich gemacht werden müssen, werden die durchsuchbaren und facettierbaren Kategorien festgelegt. Mit der Zuordnung von bestehenden Metadatenfeldern zu definierten Kategorien (facet mapping) bestimmt der Metadatenprovider also letztendlich das Erscheinungsbild seiner Daten im VLO.

Zu den relevanten internationalen Infrastrukturen und Findmechanismen, die nicht in CLARIN organisiert sind, zählt vor allem der Katalog und die Metasuchmaschine der Open Language Archives Community (OLAC). OLAC hat ein eigenes sprachressourcen-orientiertes Metadatenformat und funktioniert ebenfalls nach dem Harvester-Prinzip über die Bereitstellung einer OAI-PMH-Schnittstelle durch den Metadatenprovider.

Ein wichtiger Gravitationspunkt für eine nicht-fachspezifische übergreifende internationale Adressierbarkeit und Discoverability von Forschungsdatensätzen sind die Dienste des DataCite Konsortiums in Verbindung mit der Registrierung von Digitial Object Identifier (DOIs) für jeden Datensatz. Mit jeder DOI wird ein einfacher, aber erweiterbarer, Metadatensatz registriert, der durch die DataCite Metadatensuche auffindbar wird. DOI und DataCite ist es gelungen, dass die Vergabe mit positiven und progressiven Effekten assoziiert wird, wie Anerkennung akademischer Leistungen und Wissenschaftlichkeit des Inhalts, sodass erfolgreich eine Nachfrage aus der Forschung generiert wird. Das Metadatenschema wird damit als Referenz für eine nicht-fachspezifische Beschreibung von Forschungsdatensätzen weiter an Bedeutung gewinnen. Metadatensätze werden hier aktiv durch den Metadatenprovider über eine Schnittstelle bei DataCite registriert.

Die übergreifenden Metadatenportale müssen sich mit der Nutzerfreundlichkeit und Performanz von kommerziellen Webdiensten messen. Hierbei liegt auf der Hand, dass die Qualität der Suche maßgeblich von der Qualität der Metadaten und vor allem die Verlässlichkeit und Passung der Metadatenkategorien abhängt, über die sie selbst nur bedingt Kontrolle haben.

Es ist also angebracht, dass sich Forschungsdatenarchive und Forscher bei der Konzeptualisierung von Metadaten und bei der Kuratierung nicht ausschließlich auf die Abbildung des zugrunde liegenden Forschungsgegenstands versteifen, sondern vielmehr Metadaten von ihren intendierten Disseminationskanälen her zu denken. In Zukunft wird es immer wichtiger sein, wie Daten in den jeweiligen Meta-Portalen aussehen werden.



Erwartungen und Veränderungen

Der neueste Ankömmling im Bereich der wissenschaftlichen Datensatz-Suche ist Google mit Google Dataset Search[7]. Dieser Service befindet sich noch im Beta-Stadium und ändert sich täglich. Die grundlegende Ausrichtung und die zugrundeliegenden Technologien sind aber bereits dokumentiert[8] und mehr oder weniger stabil. Google basiert die Datensatz-Suche auf Technologien und Strategien der Websuche auf. Nachdem Google seinen Support für OAI-PMH-Schnittstellen 2008 (Mueller 2008) eingestellt hatte, ist es nicht unerwartet, dass sie, anstatt Metadaten, über spezielle Schnittstellen abzufragen oder eine Registrierung zu erfordern, durch das Crawlen der Webseiten von Repositorien und Portalen erfasst werden. Dabei werden strukturierte Daten, die mit Hilfe der Linked-Data-Technologien JSON-LD[9] oder RDFa[10] und den Ontologien *schema.org*[11] oder *W3C Data Catalog Vocabulary*[12] ausgezeichnet wurden, als Grundlage der Darstellung und wahrscheinlich auch des Suchindexes genommen. Durch diese Technologieentscheidungen wird eine flache Metadatenstruktur, bestehend aus einfachen Key-Value-Paaren und eine allgemeine, fachagnostische Datenfeldsemantik, bedingt. Der Fokus auf HTML-Webseiten und allgemeine Webtechnologien für strukturierte Daten als Hauptschnittstelle für Metadaten bedeutet eine grundlegende Abkehr von den gängigen Praktiken in der wissenschaftlichen Technologielandschaft.

Ausblick und mögliche Antworten

Die aktuellen Veränderungen in der Datensatzsuche bedeuten auf der einen Seite, dass Findbarkeit immer weiter aus der Kontrolle der einzelnen Repositorien in die Hand von Metaportalen und Drittanbietern wie Google und Elsevier[13] geht. Auf der anderen Seite bedeutet der durch Google angedeutete Technologiewechsel eine Stärkung der Webinterfaces der einzelnen Repositorien. Die Webseite ist damit wieder primäre Schnittstelle. Repositorien werden die Semantik ihrer Metadaten stärker von den intendierten Disseminationskanälen her denken müssen und einzeln und in Verbünden Strategien für die nachhaltige Findbarkeit entwickeln.

Fußnoten

1. https://www.narcis.nl/ , 11.01.2019.
2. https://vlo.clarin.eu/ , 11.01.2019.
3. http://search.language-archives.org , 11.01.2019.
4. https://search.datacite.org/ , 11.01.2019.
5. https://www.mpi.nl/ISLE/ , 11.01.2019.
6. https://tla.mpi.nl/tools/tla-tools/arbil/ , 11.01.2019.
7. https://toolbox.google.com/datasetsearch , 11.01.2019.
8. https://developers.google.com/search/docs/data-types/dataset , 11.01.2019.
9. https://www.w3.org/TR/json-ld/ , 11.01.2019.
10. https://www.w3.org/TR/rdfa-primer/ , 11.01.2019.
11. https://schema.org/ , 11.01.2019.
12. https://www.w3.org/TR/vocab-dcat/ , 11.01.2019.
13. https://datasearch.elsevier.com/ , 11.01.2019.

Bibliographie

Blumtritt, J. / Rau, F. (2016): *User-Experience von Spracharchiven: Eine Neubewertung der Interaktion von Archiv und Nutzern*, in: Digital Humanities im deutschsprachigen Raum (DHd 2016), Leipzig. http://dhd2016.de/boa.pdf#page=215 .

CMDI and Metadata Curation task forces of the Standing Committee on CLARIN Technical Centres (2018): *CMDI Best Practices*, Version 1.1.1, https://github.com/clarin-eric/cmdi-best-practices/releases .

Mache, B. / Klaffki, L. (2018): *Das DARIAH-DE Repository. Elementarer Teil einer modularen Infrastruktur für geistes- und kulturwissenschaftliche Forschungsdaten*, in: O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 5(3) 2018, 92-103. https://doi.org/10.5282/o-bib/2018H3S92-103 .

Mueller, John (2008): *Retiring support for OAI-PMH in Sitemaps*. https://webmasters.googleblog.com/2008/04/retiring-support-for-oai-pmh-in.html .

Schaffner, J. (2009): *The Metadata is the Interface: Better Description for Better Discovery of Archives and Special Collections, Synthesized from User Studies*, in: Report produced by OCLC Research. https://www.oclc.org/content/dam/research/publications/library/2009/2009-06.pdf .