Originalveröffentlichung in: Griesbaum, Joachim/Mandl, Thomas/Womser-Hacker, Christa (Hrsg.): Information und Wissen: global, sozial und frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (ISI 2011). Hildesheim, 9.—11. März 2011. – Boizenburg: Werner Hülsbusch, 2011. S. 62-73. (Schriften zur Informationswissenschaft 58)
DOI: *https://doi.org/10.5281/zenodo.4134504*

Komponenten-basierte Metadatenschemata und Facetten-basierte Suche

Ein flexibler und universeller Ansatz

Reinhild Barkey, Erhard Hinrichs, Christina Hoppermann, Thorsten Trippel, Claus Zinn

Seminar für Sprachwissenschaft und SFB 833 – Universität Tübingen
Wilhelmstr. 19, D-72074 Tübingen
E-Mail: vorname.nachname@uni-tuebingen.de

Zusammenfassung

Wenn man verschiedenartige Forschungsdaten über Metadaten inhaltlich beschreiben möchte, sind bibliografische Angaben allein nicht ausreichend. Vielmehr benötigt man zusätzliche Beschreibungsmittel, die der Natur und Komplexität gegebener Forschungsressourcen Rechnung tragen. Verschiedene Arten von Forschungsdaten bedürfen verschiedener Metadatenprofile, die über gemeinsame Komponenten definiert werden. Solche Forschungsdaten können gesammelt (z.B. über OAI-PMH-Harvesting) und mittels Facetten-basierter Suche über eine einheitliche Schnittstelle exploriert werden. Der beschriebene Anwendungskontext kann über sprachwissenschaftliche Daten hinaus verallgemeinert werden.

Abstract

The content description of various kinds of research data using metadata requires other than bibliographical data fields that are alone not sufficient for this purpose. To properly account for research data, other metadata fields are required, often specific to a given research data set. Consequently, metadata profiles adapted to different types of resources need to be created. These are defined by building blocks, called components, that can be shared across profiles. Research data described in this way can be harvested, for example, using OAI-PMH. The resulting metadata collection can then be explored via a unified interface using faceted browsers. The described application is in the area of linguistic data, but our approach is also applicable for other domains.

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-108748

1 Beschreibungsprofile für Klassen von Ressourcen

Wissenschaftliche und andere Publikationen werden in der Regel mit strukturierten Beschreibungen, Metadaten, versehen, wie z.B. mit bibliografischen Angaben zu Autoren, Publikationstitel, Verlagshaus und Erscheinungsjahr, sowie mit einer Klassifikation oder Verschlagwortung. Diese Metadaten erlauben das Auffinden von Publikationen innerhalb von (Bibliotheks-) Katalogen. Auf diese Weise kann auch innerhalb einer wissenschaftlichen Arbeit auf andere publizierte Arbeiten verwiesen werden. Gleichzeitig helfen Schlagworte, verwandte Arbeiten grobkörnig zu gruppieren.

Für Druckerzeugnisse hat sich als Beschreibungssystem eine Kernmenge von Datenkategorien für Metadaten etabliert, die Dublin-Core-Kategorien (Hillmann, 2005). Viele dieser Kategorien sind für Forschungsprimärdaten nicht relevant oder nicht aussagekräftig, um durch die Beschreibung einem möglichen Benutzer einen hinreichenden Eindruck zu geben, um was für eine Ressource es sich überhaupt handelt. Unterschiedliche Klassen von Ressourcen benötigen dabei unterschiedliche Beschreibungsebenen. So sind etwa für die Sprachtechnologie Informationen zu Audioformaten von Aufnahmen wichtig, wohingegen für Textkorpora eher der Zeichensatz eine Rolle spielt, für lexikalische Ressourcen die Struktur der einzelnen Einträge, für Fragebogenauswertungen die Größe der Stichprobe und Methode, etc.

Aus diesen Beispielen wird deutlich, dass die benötigten Beschreibungsdimensionen für Ressourcen stets vom Ressourcentyp abhängen, auch wenn für Archivierungszwecke allgemeine bibliografische Kategorien für alle Typen Anwendung finden können. Daher ist es notwendig, basierend auf einem Klassifikationssystem für Ressourcen und möglichen Prototypen Beschreibungsmuster zu definieren, die in Abhängigkeit vom Ressourcentyp auf die jeweilige Ressource angewendet werden können. Diese Beschreibungsmuster bilden *Profile* für Metadaten.

2 Komponentenbasierte Metadatenbeschreibungen

Profile für unterschiedliche Ressourcentypen sind nicht überschneidungsfrei, weil bestimmte Beschreibungsebenen, wie z.B. bibliografische Informationen, häufig von verschiedenen Ressourcentypen verwendet werden. Somit können auch Beschreibungen unterschiedlicher Ressourcentypen Ähnlichkeiten aufweisen. Um die Wiederverwendung von gemeinsamen Datenkategorien und Beschreibungsstrukturen sowohl bei der Erstellung als auch bei der Interpretation von Beschreibungen zu gewährleisten, wurde ein System für Metadaten entwickelt, bei dem zusammengehörige Datenkategorien und -strukturen zu *Komponenten* zusammengefasst werden. Komponenten sind dabei zunächst Mengen von beschreibenden Datenkategorien. Diese wiederum können selbst zu größeren Komponenten kombiniert werden, um schließlich für einen Ressourcentyp als ein Beschreibungsprofil Verwendung zu finden. Damit werden Komponenten als Bausteine für Profile verwendet, wobei die gleichen Komponenten innerhalb verschiedener Profile enthalten sein können.

Im Rahmen des EU-Projektes CLARIN (www.clarin.eu) wurde zur systematischen Verwendung von Komponenten ein Metadatenschema, die *Component MetaData Infrastructure* (*CMDI*, siehe Broeder et al., 2010, siehe auch htttp://www.clarin.eu/cmdi), entwickelt. Neben einer Beschreibungssprache für Profile und Komponenten enthält diese Infrastruktur dazu auch weitere Werkzeuge, sowohl Editoren als auch Analysewerkzeuge. Diese operieren unabhängig vom Ressourcentyp auf bestimmten Datenkategorien.[1] Bestehende Metadatenstandards wie Dublin Core (Coyle und Baker, 2008), OLAC (Simons und Bird, 2008) oder der TEI-Header (TEI P5, 2007) können als Profile oder auch als Komponenten dargestellt werden, sodass ein Komponentenmodell mit Profilen als Obermenge bestehender Metadatenschemas angesehen werden kann. So werden die bibliografischen Informationen in den Metadaten einer Ressource für Archiv- und Bibliothekskataloge ver-

1 Die Implementierung hätte dabei auch mittels XML-Namespaces erfolgen können, dies allerdings zu Lasten einer erhöhten Komplexität, da potenziell die volle Ausdrucksmächtigkeit von XSchema zur Verfügung gestanden hätte. Die vorliegenden Werkzeuge dagegen basieren zwar auf XSchema, operieren aber auf einer Teilmenge davon und enthalten Restriktionen, die zu einer leichteren Handhabung führen.

wendbar. Andere Datenkategorien dagegen, wie z.B. die Angabe von Annotationstypen bei linguistischen Korpora, werden von allgemeinen Kataloganwendungen ignoriert, aber von spezialisierten Suchmaschinen oder Diensten verwendet.

Um auch institutionsübergreifend die Verwendung gleicher Komponenten und Profile zu ermöglichen, wurde im Rahmen von CMDI die *Component Registry* veröffentlicht. Dabei handelt es sich um ein Verzeichnis, das zentral Komponenten und Profile sowohl zur Weiterverwendung in Institutionen und Projekten als auch zur Validierung konkreter Instanzen zur Verfügung stellt. Die Komponenten erhalten dort einen persistenten Identifikator (*Persistent Identifier* oder *PID*, siehe ISO 24619), auf den sowohl von anderen Komponenten als auch Instanzen verwiesen werden kann und der über ein Handle-System zu einer URL aufgelöst wird.

Innerhalb der Komponenten werden die Datenkategorien mit einer Referenz auf bereits standardisierte oder im Standardisierungsprozess befindliche Datenkategorien verwendet, die in einem Verzeichnis definiert und nachhaltig dokumentiert werden. Bei diesem Verzeichnis für Datenkategorien handelt es sich um *ISOcat*, das aus dem Bereich der Sprachressourcen der *International Organization of Standardization* (*ISO*, siehe ISO 12620:2009, siehe auch http://www.isocat.org) stammt. Die Referenz auf in ISOcat definierte Datenkategorien innerhalb der Komponenten ermöglicht es, dass Datenkategorien von unterschiedlichen Erstellern von Metadateninstanzen in gleicher Weise verstanden werden. Außerdem können Probleme wie nicht der Definition entsprechende entfremdete Verwendungen der Datenkategorien (d.h. *Tag Abuse*) eingedämmt werden.

In den Komponentendefinitionen von CMDI können zudem kontrollierte Vokabulare angegeben werden. Diese können ebenfalls dazu beitragen, das Problem des *Tag Abuse* zu minimieren, da Datenkategorien durch das kontrollierte Vokabular formal auf ihre Konsistenz geprüft werden können. Gleichzeitig gibt es auch Freitextfelder wie Zusammenfassungen und Beschreibungen, deren Inhalt nicht genauer reglementiert wird. Der Gebrauch von Datenmodellen ist nach Maßgabe der zugrundeliegenden Schemasprache möglich. Im Rahmen des CMDI-Datenmodells ist dies mit der Verwendung von XSchema sehr weitgehend umgesetzt worden, angefangen von Datumsformaten bis zu regulären Ausdrücken für Zeichenkettendefinitionen.

Abbildung 1 stellt eine Anwendung des Komponentenmodells für linguistische Korpora dar. Die Komponenten (durch Rechtecke repräsentiert) können dabei selbst weitere Komponenten enthalten. So ist z.B. die Komponente

zur Annotation innerhalb der Komponente zur Erstellung der Ressource (*Creation*) eingebunden. Komponenten sind dabei unterschiedlich komplex (illustriert durch die Dimensionen der Rechtecke), können aber in verschiedenen Profilen und Komponenten erscheinen. Beispielsweise wird die in der vorliegenden Implementierung verwendete Komponente *GeneralInfo* in fast allen Profilen für unterschiedliche Ressourcentypen eingebunden. Tochterkomponenten und Datenkategorien können nebeneinander in Komponenten erscheinen (in der Abbildung: Datenkategorien mit durchgezogenen Kanten).

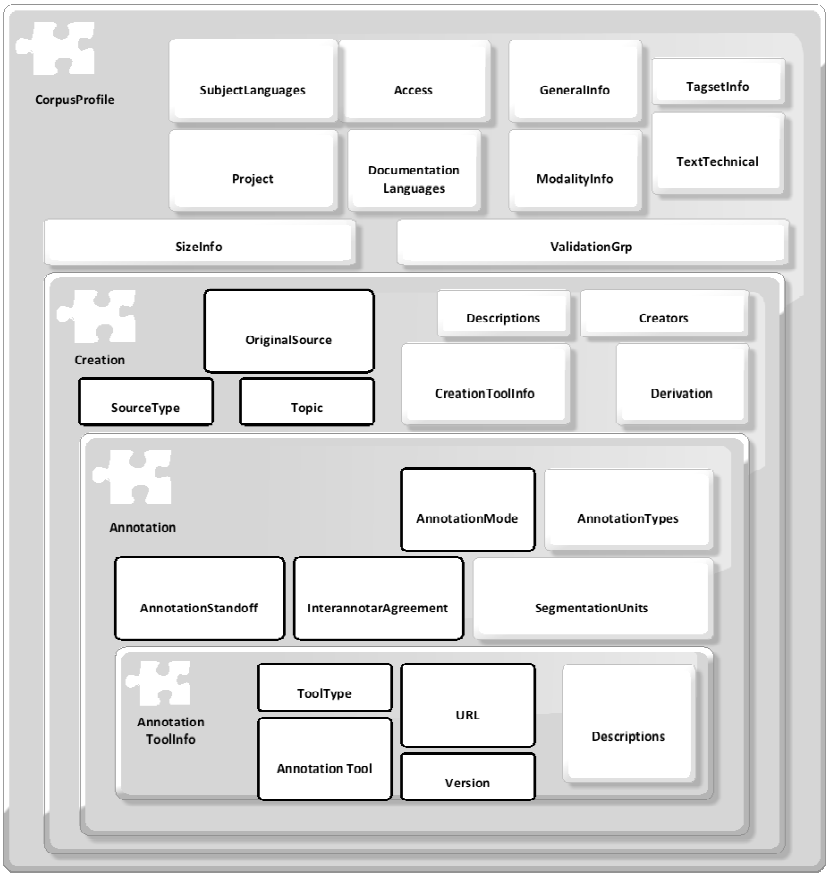

Abbildung 1:
Komponentenmodell für Metadaten zur Beschreibung von linguistischen Korpora.

3 Verbreitung von Metadaten über OAI-PMH

Für den Austausch von Metadaten in Archiven hat sich im Rahmen der Open Archive Initiative ein Containerformat etabliert: das *Open Archive Initiative Protocol for Metadata Harvesting* (siehe *OAI-PMH*, 2008). Obwohl OAI-PMH im Wesentlichen zum Austausch von Dublin-Core-Daten für Katalog-informationen zwischen Bibliotheken gedacht war, erlaubt es dieses Protokoll dennoch, z.B. mithilfe von *Namespaces*, auch weitere Metadatenformate einzubinden. Auf diese Weise können detaillierte Metadaten nach dem Komponentenmodell über einen OAI-PMH-Server bereitgestellt werden. Die einzige Voraussetzung dabei ist, dass es – möglicherweise zusätzlich zu diesen spezifischen Metadaten – bibliografische Metadaten nach Dublin Core gibt, wenn der verwendende Service auf Dublin Core Metadaten beschränkt ist. Services, die von Dublin Core unabhängig sind, benötigen diese Abbildung nicht. CMDI Informationen können damit direkt in OAI-PMH-Containern eingebunden und verteilt werden.

Da ein Komponentenmodell detailreicher ist und in Dublin Core die Datenkategorien optional sind, gibt es immer eine verlustbehaftete Abbildung dieser Komponenten-Metadaten nach Dublin Core. Um eine vollständigere Abbildung von den detailreicheren Metadaten auf Dublin Core vorzunehmen, ist eine profilspezifische Anpassung nötig. So kann eine Person, die in einer Projektleiterkomponente einer Ressource erscheint, in Abhängigkeit vom Ressourcentyp in Dublin Core als Herausgeber oder als Autor aufgefasst werden. Dies kann automatisiert beim Bereitstellen auf dem OAI-PMH-Server erfolgen, sodass keine redundante Dateneingabe erfolgen muss.

Die über OAI-PMH-Server bereitgestellten Metadaten können automatisiert mit Crawlern und Webservices erfasst werden. Zur Zeit werden Metadatenbestände unter anderem von den folgenden sprachwissenschaftlichen Institutionen semiautomatisch erfasst und ausgewertet: MPI Nijmegen, Universität Leipzig, Bayerisches Archiv für Sprachdaten, Universität Stuttgart, Universität Tübingen, Berlin-Brandenburgische Akademie der Wissenschaften und linguistische Sonderforschungsbereiche der DFG.

4 Verwendung von komponentenbasierten Metadaten für die Facetten-basierte Suche

Ein wesentliches Problem bei der Weiterverwendung von Forschungsprimärdaten in anderen Kontexten und der Überprüfung von Ergebnissen anhand der Daten – was zum Beispiel durch die Deutsche Forschungsgemeinschaft gefordert wird (DFG, 2009) – ist neben der Langzeitarchivierung auch und gerade die Auffindbarkeit der Daten (Rehm, et al., 2010). Dies umfasst zunächst nicht den Zugang zu den Forschungsprimärdaten, sondern die Auffindbarkeit ihrer formalen Beschreibungen, wie sie beispielsweise in Bibliothekskatalogen für Schriften vorliegen. Dabei stellt die große Variation von Metadatenkategorien in Abhängigkeit von den Klassen von Ressourcen eine Herausforderung dar. Volltextsuchen über die Metadaten sind für diesen Zweck nur bedingt hilfreich, da sie die in Datenkategorien und Metadatenstrukturen implizit enthaltenen Informationen nicht auswerten. Auch klassische formularbasierte Suchen, die oft als „erweiterte Suche“ bezeichnet werden, sind durch die Variabilität der Metadatenschemas mit unterschiedlichen Komponenten stark eingeschränkt, weil sie nicht alle Varianten berücksichtigen können, ohne zu umfangreich und unübersichtlich zu werden.

Um diese Probleme zu vermeiden, kann man ein Facetten-basiertes Suchsystem (siehe Hearst, 2006) einsetzen, das alle Datensätze eines Datenbestandes mithilfe von Ausprägungen wohldefinierter Facetten beschreibt. Dazu werden einem Datensatz in der Regel mehr als eine Kategorie (Teilmenge) zugeordnet. Die Abbildung zwischen Facetten und Metadatenfeldern diverser Metadatenprofile wird dabei durch den oben beschriebenen Komponenten-basierten Ansatz enorm vereinfacht. Dies liegt daran, dass eventuelle Ambiguitäten in der Lesart durch Referenz auf das Verzeichnis von Metadatenkategorien (www.isocat.org) leicht aufgelöst werden können.

Der Benutzer eines *Faceted Browsers* erhält bereits zu Suchbeginn eine Facetten-basierte Übersicht über den gesamten Datenbestand. Abbildung 2 veranschaulicht dies am Beispiel des im Projekt „Nachhaltigkeit Linguistischer Daten“ (NaLiDa, http://www.sfs.uni-tuebingen.de/nalida) entwickelten *Faceted Browsers*, der einen Zugang zu sprachwissenschaftlichen Forschungsprimärdaten auf der Basis von komponentenbasierten Metadaten erlaubt. Sichtbar sind in der Abbildung die Facetten *origin* (Quelle eines Datensatzes), *modality* (Modalität der Ressource), *resourcetype* (Ressourcentyp), *country* (Ursprungsland), *language* (Sprache der Ressource) und *or-*

ganisation (Institution, an der diese Ressource entstanden ist) sowie ihre Facettenausprägungen und die Anzahl der Datensätze, die mit den jeweiligen Ausprägungen beschrieben sind. Durch die Auswahl einer Facettenausprägung (z.B. die Facette *resourcetype* mit Ausprägung *corpus*) setzt der Nutzer einen Filter, der den Suchraum entsprechend verkleinert. Die ausgewählten Datensätze (Anzahl 4499) werden so wiederum umgehend mithilfe der verbliebenen Facetten beschrieben, sodass der Nutzer gezielt durch Suchräume navigieren kann. Auf diese Weise kann ein Nutzer etwa alle Ressourcen identifizieren, die zugleich aus einem bestimmten Korpus stammen und einer bestimmten Sprache zugeordnet werden. In diesem Suchkontext fächert der Faceted Browser die ausgewählten Ressourcen u.a. bezüglich der Ausprägungen der Facette *genre* auf. Nutzer können so ihre Suche nach einem deutschsprachigen Korpus mit Dialogdaten oder Diskursdaten verfeinern.

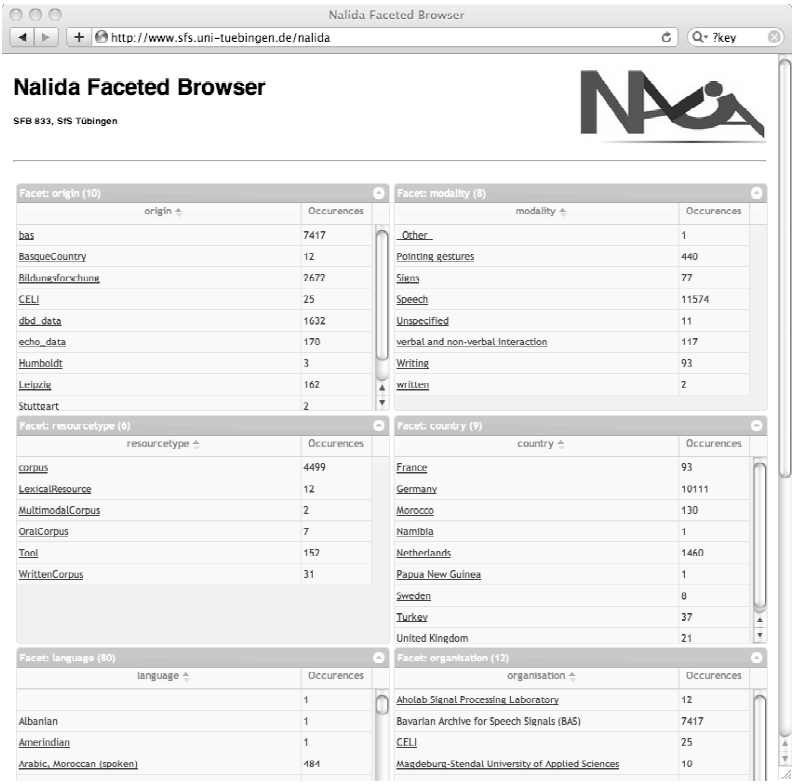

Abbildung 2: Der NaLiDa-Faceted-Browser zum Metadaten-basierten Zugriff auf Forschungsprimärdaten in der Sprachwissenschaft.

Für die sehr großen und heterogen Datenbestände in der Sprachwissenschaft ist die Einführung *bedingter Facetten* hilfreich. Bedingte Facetten sind solche, die nur für bestimmte Typen von Ressourcen relevant sind und die dem Nutzer erst nach Vorauswahl von einigen allgemeinen, sogenannten *unbedingten* Facetten angezeigt werden. Somit erlauben sie eine feinkörnigere Suche in Teilräumen von Metadatensätzen. Beispielsweise wird die bedingte Facette *genre* mit ihren Ausprägungen *discourse, poetry, story-telling, etc.* nur angezeigt, wenn Datensätze vom Ressourcetyp *corpus* weiter exploriert werden sollen. Werden vom Nutzer hingegen Datensätze vom Ressourcentyp *tool* ausgewählt, wird ihre weitere Exploration durch die Einführung der bedingten Facette *tooltype* (mir ihren Ausprägungen *spell checker, POS tagger, named entity recognizer, etc.*) erleichtert.

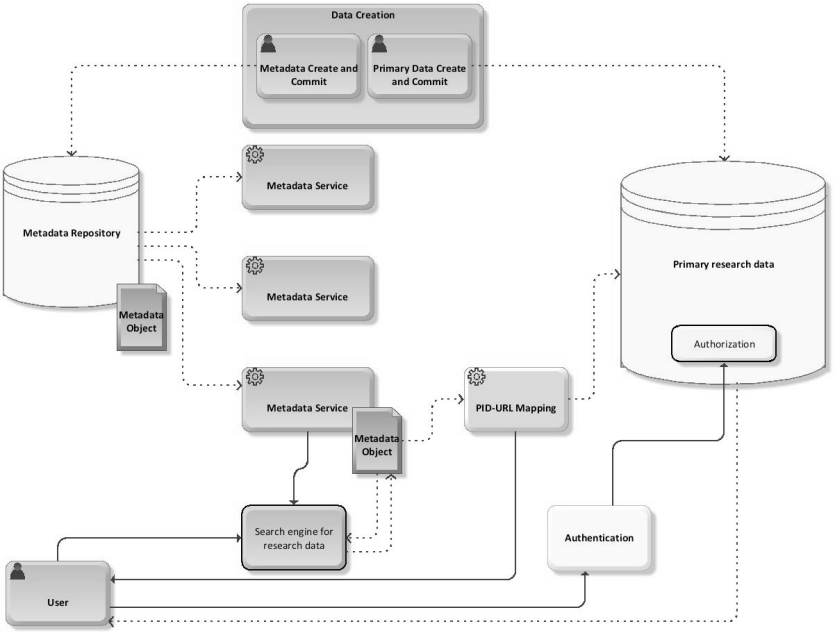

Abbildung 3:
Architektur eines Metadaten-basierten Zugangs zu Forschungsprimärdaten.

Diese Benutzerschnittstelle ist unabhängig von der Datenhaltung der Forschungsprimärdaten, erlaubt aber über die persistenten Identifikatoren (PIDs) der Forschungsprimärdaten die Verbindung zwischen beiden. Abbildung 3 illustriert die Architektur eines Systems, in dem Metadaten und Primärdaten

konzeptionell getrennt sind und unterschiedlichen Rechteverwaltungs- und Zugangssystemen unterstehen. Die *Metadaten-Objekte* sind in einem *Metadata Repository* offen zugänglich und damit lesbar. Sie können von unterschiedlichen Services verwendet werden. Der Zugang zu Primärdaten erfordert dagegen sowohl die Authentifizierung als auch die Überprüfung der speziellen Rechte eines Benutzers, die Autorisierung. Die Suche und die Auflösung von persistenten Identifikatoren auf URLs können dabei wiederum unabhängig als Service realisiert werden.

5 Zusammenfassung und weiterführende Arbeiten

In diesem Beitrag haben wir die Grundzüge von Komponenten-basierten Metadatenmodellen skizziert und aufgezeigt, wie flexibel ein solches System auf unterschiedliche Ressourcentypen angewendet werden kann. Dabei erweist sich ein Faceted Browser als hervorragendes Werkzeug, um erfahrenen Nutzern wie auch Anfängern einen einheitlichen Zugriff auf Kollektionen von Metadatensätzen zu geben. Die Einführung bedingter Facetten sorgt zudem dafür, dass Navigationselemente dynamisch und kontextsensitiv bereitgestellt werden und bringt somit Nutzern eine zusätzliche Unterstützung zur schnellen und strukturierten Exploration großer Datenmengen.

Nach der ersten Implementierung eines Faceted Browsers auf der Basis von CMDI-Komponentenmetadaten für unterschiedliche Korpora, Lexika und computerlinguistische Werkzeuge, sollen in einem nächsten Arbeitsschritt Profile für weitere Ressourcentypen geschaffen und die Inhaltsmodelle von Komponenten überprüft und bei Bedarf angepasst werden. Diese Komponenten sind ferner über die Component Registry zur Weiterverwendung bereit zu stellen.

Ein wichtiger, bereits initiierter Schritt besteht darin, das Komponentenmodell selbst und eine Implementierungssprache für Komponenten in den Standardisierungsprozess im Rahmen der ISO einzubringen. Dies soll dazu führen, dass langfristig und transparent Dienstleistungen für die Forschung und Ressourcengemeinschaft aufgebaut werden können.

Zur Erweiterung des Systems wird außerdem versucht, weitere Archive und Daten produzierende Projekte mit einzubeziehen und gegebenenfalls bei

der Erstellung von Metadatenbeispielen für ihre Datentypen zu unterstützen. Dies soll dazu führen, dass Forschungsprimärdaten langfristig zur Weiterverwendung, Referenz und als Forschungs- und Ergebnisbeleg verfügbar sind.

Literaturverzeichnis

Broeder, D.; Kemps-Snijders, M.; Van Uytvanck, D.; Windhouwer, M.; Withers, P.; Wittenburg, P.; Zinn, C. (2010): "A Data Category Registry- and Component-based Metadata Framework". *Proceedings of the 7th conference on International Language Resources and Evaluation*.

Coyle, K.; Baker, Thomas (2009): *Guidelines for Dublin Core Application Profiles.* Dublin Core Metadata Initiative, 2009-05-18.
http://dublincore.org/documents/2009/05/18/profile-guidelines/

DFG (2009): *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. Deutsche Forschungsgemeinschaft, Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss für Informationsmanagement, 2009.
http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

Hearst, M. (2006): "Design Recommendations for Hierarchical Faceted Search Interfaces". ACM SIGIR Workshop on Faceted Search.

Hillmann, D. (2005): *Using Dublin Core – The Elements*. Dublin Core Metadata Initiative, 2005-11-07.
http://dublincore .org/documents/2005/11/07/usageguide/elements.shtml

ISO 12620:2009: *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources.* International Organization of Standardization, Genf.
http://www.isocat.org

ISO/DIS 24619:2010: *Language resource management -- Persistent identification and sustainable access (PISA)*. International Organization of Standardization, Genf. http://www.isocat.org

OAI-PMH (2008): *The Open Archives Initiative Protocol for Metadata Harvesting.* Protocol Version 2.0 of 2002-06-14, Document Version 2008-12-07.
http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Rehm, G., Schonefeld, O., Trippel, T. Witt, A. (2010): Sustainability of Linguistic Resources Revisited. *Proceedings of the International Symposium on XML for*

the Long Haul: Issues in the Long-term Preservation of XML. Balisage Series on Markup Technologies, vol. 6 (2010). doi:10.4242/Balisage/Vol6.Witt01

Simons, G.; Bird, S. (2008): *OLAC Metadata*. Open Language Archive Community, 2008-05-31. http://www.language-archives.org/OLAC/metadata-20080531.html

TEI P5 (2007): TEI Guidelines. Text Encoding Initiative, 1. November 2007. http://www.tei-c.org/Guidelines/P5/