erschließen - verknüpfen - finden: Forschungsdaten im Digitalen Wissensspeicher

Czmiel, Alexander
czmiel@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Grabsch, Sascha
grabsch@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Jürgens, Marco
juergens@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Maiwald, Anke
maiwald@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

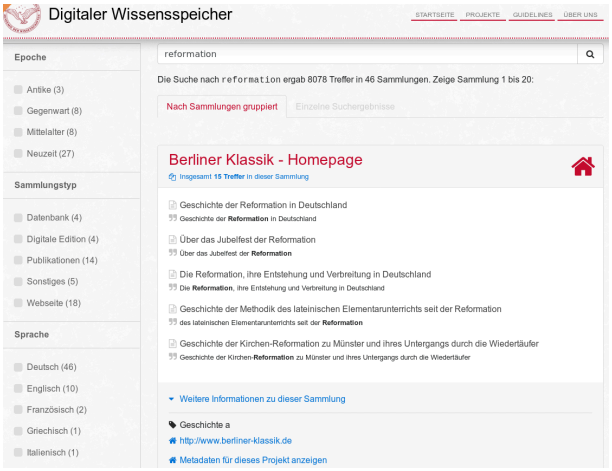
Willenborg, Josef
willenborg@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Deutschland

Die Auffindbarkeit und Sichtbarkeit digitaler Forschungsergebnisse in den Geisteswissenschaften leiden immer noch unter dem Mangel an einer zentralen Plattform für den wissenschaftlich fundierten Zugang, nicht nur zu den Metadaten, sondern auch zu den vollständigen digitalen, semantisch erschlossenen Forschungsdaten. Insbesondere die wichtigen Ergebnisse der geisteswissenschaftlichen Grundlagenforschung an den Akademien haben zu wenig Bekanntheit in den Fachcommunities und der breiteren Öffentlichkeit. Die Landschaft digitaler Forschungsergebnisse, wie Digitaler Editionen, Repositorien oder Forschungsdatenbanken, erscheint, trotz des vermehrten Einsatzes von Standards, technisch fragmentiert und wenig überschaubar. Dazu kommt eine mangelnde Vernetzung der verschiedenen Sammlungen, insbesondere institutionen- und fächerübergreifend, die zu dem aktuellen Zustand einzelner, isolierter digitaler Projekte geführt hat.

Im Rahmen des DFG-geförderten Projektes „Digitaler Wissensspeicher“ wurde an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) seit 2012 ein zentraler Zugang für sämtliche digitale Forschungsdaten und Ressourcen der Akademie geschaffen. Hauptziel war dabei die vollständige Erfassung und Volltext-Indexierung der technisch sowie inhaltlich äußerst vielfältigen und heterogenen Ressourcen der BBAW. Gestützt auf einen Volltextindex (Apache Lucene) und ein anhand der Anforderungen der Akademie entwickeltes Metadenschema (basierend auf dem Metadatenstandard OAI-ORE) wurden über 230 Sammlungen aus 142 Projekten mit insgesamt mehr als 1 Mio. digitalen Ressourcen im Volltext und mit Metadaten erfasst. Durch den Einsatz von Sprachtechnologien (u. a. Donatus) ist eine morphologisch normalisierte Suche möglich. Über Text-Mining-Tools, wie DBpedia-Spotlight, werden die erfassten Ressourcen semantisch angereichert und vernetzt. Eine weitere Verknüpfung erfolgt über die manuell erfassten Metadaten zu den einzelnen Projekten und deren digitalen Sammlungen. Dies ermöglicht z.B. eine automatisierte Zuordnung semantisch ähnlicher Projekte.

Die durch das Textmining gewonnenen semantischen Annotationen ermöglichen eine Vielzahl weitergehender Nutzung. Beispielhaft wird dies auf der Website des Wissensspeichers anhand einer Kartenvisualisierung, die auch verschiedene Filtermöglichkeiten anbietet, demonstriert: die von DBpedia-Spotlight innerhalb der Ressourcen erkannten Orte werden projekt- und sammlungsübergreifend auf einer Karte referenziert. Sie bilden die Grundlage für einen visuellen, explorativen Zugang zu den indexierten Einzelressourcen.

Die Beschreibung der Metadaten erfolgt mittels eines flexiblen OAI-ORE-basierten Metadatenschemas, das die Abbildung aller, teilweise komplexen, Forschungsvorhaben der BBAW ermöglicht. Ähnlich den technisch niedrigschwelligen Anforderungen an die Volltext-Indexierung besteht auch für den Bereich der Metadaten die Möglichkeit, mit wenigen Pflichtfeldern weitere Sammlungen und Forschungsprojekte in den Wissensspeicher aufzunehmen. Die Metadaten zu allen Projekten werden über eine SPARQL-Schnittstelle für die weitere maschinelle Nachnutzung, z.B. die Integration in die Linked-Open-Data-Cloud bereitgestellt.



Der Digitale Wissensspeicher ist unter der Adresse http://wissensspeicher.bbaw.de für den öffentlichen Zugriff erreichbar. Damit bietet der Wissensspeicher eine exemplarische Lösung für einen Katalog heterogener, digitaler geisteswissenschaftlicher Ressourcen, die zentral aggregiert und über eine Volltextsuche verfügbar gemacht werden. Dass es sich dabei nicht nur um digitale Ressourcen der BBAW handeln muss, liegt auf der Hand. Damit auch Sammlungen anderer Projekte und Institutionen von der Entwicklung des Wissensspeichers profitieren können, wurden Guidelines mit strukturellen und inhaltlichen Mindestanforderungen, die Ressourcen und Metadaten für die Aufnahme erfüllen müssen bzw. Schnittstellenbeschreibungen für vom Wissensspeicher unterstützte Formate bereitgestellt. Damit werden niedrigschwellige Zielvorgaben für die (technische) Qualität der in den Wissensspeicher aufzunehmenden Ressourcen mit ihren Metadaten formuliert. Zu den unterstützten Formaten gehören u.a. HTML-Websites, XML-basierte Ressourcen, PDF-Dokumente und institutionelle Repositorien sowie jede Form von Daten, die in der vom Wissensspeicher definierten Form eines XML-Exports ausgeliefert werden können (z.B. relationale Datenbanken oder NoSQL-Datenbanken).

Das große Spektrum unterschiedlicher Erscheinungsformen digitaler Ressourcen in den Geisteswissenschaften kann somit vom Wissensspeicher verarbeitet und integriert werden. Durch die projektübergreifende Volltextsuche in allen verzeichneten Sammlungen sowie die miteinander vernetzten Metadaten verbessert der Digitale Wissensspeicher die Auffindbarkeit, Sichtbarkeit und damit die Nutzbarkeit von Digital-Humanities-Projekten und digitalen Forschungsergebnissen.

Chen, Ko-le, Marian Dörk und Martyn Dade-Robertson: „Exploring the Promises and Potentials of Visual Archive Interfaces.“ In IConference 2014 Proceedings, 735–41. iSchools, 2014. https://doi.org/10.9776/14348.

Glinka, Katrin, Christopher Pietsch und Marian Dörk: „Past Visions and Reconciling Views: Visualizing Time, Texture and Themes in Cultural Collections.“ digital humanities quarterly 11, Nr. 2 (27. Februar 2017). http://www.digitalhumanities.org/dhq/vol/11/2/000290/000290.html; abgerufen am 14.12.2017.

Horch, Andrea, Holger Kett und Anette Weisbecker: „Semantische Suchsysteme für das Internet. Architekturen und Komponenten semantischer Suchmaschinen.“ Stuttgart: Fraunhofer Verlag, 2013.

McMurry, Julie A., Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Mélanie Courtot u. a.: „Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data.“ PLOS Biology 15, Nr. 6 (29. Juni 2017): e2001414. https://doi.org/10.1371/journal.pbio.2001414.

Voß, Jakob: „Was sind eigentlich Daten?“ LIBREAS. Library Ideas, Nr. 23 (2013). http://libreas.eu/ausgabe23/02voss/; abgerufen am 14.12.2017.

Ward, David, Jim Hahn und Kirsten Feist: „Autocomplete as Research Tool: A Study on Providing Search Suggestions.“ Information Technology and Libraries 31, Nr. 4 (11. Dezember 2012). https://doi.org/10.6017/ital.v31i4.1930.

Woutersen-Windhouwer, Saskia (Hrsg.): „Enhanced Publications: Linking Publications and Research Data in Digital Repositories.“ Amsterdam: Amsterdam Univ. Press, 2009.