Forschungsdatenmanagement mit Open-Source-Software

C. Willmes, S. Brocks, C. Hütt, D. Kürner, K. Volland und G. Bareth

Einführung

Im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG, http://www.dfg.de) geförderten Sonderforschungsbereich 806 (SFB806, http://www.sfb806.de) wird eine webbasierte Forschungsdatenbank (http://crc806db.uni-koeln.de) entwickelt. Der SFB806 ist ein Forschungsprojekt mit über 100 beteiligten Wissenschaftlern, das sich mit der Ausbreitung des modernen Menschen (Homo sapiens) von Afrika nach Europa befasst.

Die SFB806-Datenbank erfüllt zwei Hauptaspekte: Zum ersten stellt sie eine Möglichkeit zur Publikation und zur Langzeitarchivierung von Daten die im Rahmen des SFB806 produziert werden zur Verfügung. Zum zweiten stellt sie den Forschern des SFB806 eine umfangreiche Forschungsdatenbasis zur Verfügung.

Durch die Forschungsdatenbank des SFB806 sollen die beteiligten Forscher einen intuitiven und integrierten Zugang zu relevanten Daten aus den Bereichen Paläoumwelt, pleistozäner und holozäner Archäologie sowie der Geoarchäologie erhalten. Dazu wurde und wird ein einheitliches Datenmodell [1,2] zur Beschreibung der drei genannten Domains auf der Basis von *Semantic Web Technologie* [3] fortlaufend entwickelt.

In diesem Beitrag wird im Folgenden die Bereitstellung der webbasierten Infrastruktur der SFB806-Datenbank unter Einsatz von standardisierten Datenformaten/Interfaces und Open-Source-Software beschreiben.

Datenmanagement

Die DFG fordert von allen Sonderforschungsbereichen (SFB) die Implementation eines Datenmanagements [4], das die Langzeitarchivierung und Verfügbarkeit der Forschungsergebnisse des SFB für eine Zeitspanne von mindestens 10 Jahren nach Ablauf des Forschungsprojektes gewährleistet. Das bedeutet bei einer maximalen Laufzeit eines SFB von 12 Jahren eine Datengewährleistung von bis zu 22 Jahren. Daten 22 Jahre zugreifbar vorzuhalten ist alles andere als trivial, wenn man betrachtet in welchen Datenformaten und auf welchen Datenträgern vor 22 Jahren (1990 – das Web 1.0 wurde gerade erst von Tim Berners-Lee erfunden) Forschungsdaten archiviert und zur Verfügung gestellt wurden. Diese Erkenntnis legt nahe, Daten ausschließlich auf der Basis von gut dokumentierten bzw. standardisierten Formaten und Interfaces bereitzustellen bzw. zu archivieren.

Aus diesem Grund wird das Datenmanagement des SFB806 zum einen auf der Basis von Open-Source-Software implementiert um die Nachvollziehbarkeit der Systemarchitektur zu gewährleisten und eine höhere Kontrolle über die Implementation des Systems zu haben und ggf. individuelle Anpassungen oder Erweiterung der Software vornehmen zu können. Zum zweiten wird das System auf der Basis von standardisierten Interfaces und Datenformaten implementiert, um die langfristige Nutzbarkeit der Datenbasis zu gewährleisten. Zum dritten werden die Daten in RDF modelliert um die Semantik der Daten klar und eindeutig zu de-

finieren und um die Nutzbarkeit bzw. Integration der Forschungsergebnisse für zukünftige (externe) Forschungen und Anwendungen zu ermöglichen.

Implementation des Datenmanagement

Die von den Wissenschaftlern des SFB806 produzierten Daten werden zunächst im Format in dem der Wissenschaftler die Daten selbst nutzt und verwaltet in einen mehrfach gesicherten Speicherbereich gespeichert.

Der Speicherbereich basiert auf dem verteilten Dateisystems AFS [5], und wird durch das Rechenzentrum der Universität zu Köln (seit vielen Jahren ausfallsicher und datenverlustfrei) betrieben. Die Forscher haben beim Hochladen der Daten die Möglichkeit, Metadaten zur Beschreibung ihrer Daten anzugeben. Hierzu kann der Forscher aus vorgegebenen Vokabularen (Ontologien) zum Beschreiben seiner Daten auswählen. Momentan stehen die internen Ontologien für Archäologie, Palaeoumwelt und Geoarchäologie, die im Rahmen der Doktorarbeit des Erstautors formalisiert werden, sowie die weit verbreiteten Dublin Core [6] und ISO19115 [7] Vokabulare, zur Verfügung.

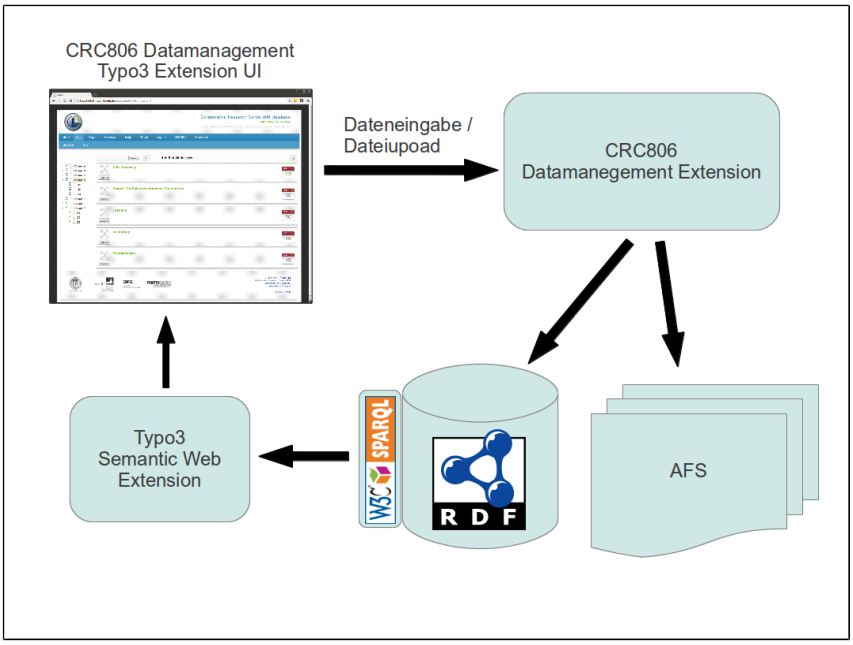

Abbildung 1: CRC806 Datamanagement Typo3 Extension.

In Abbildung 1 ist der Datenupload der SFB806-Datenbank schematisch dargestellt. Der Anwender lädt über das Interface der von uns entwickelten *CRC806 Datenmanagement Typo3-Extension* eine Datei hoch und stellt dazu ggf. Metadaten zur Verfügung. Aufgrund dieser Informationen wird einerseits die Datei im Dateisystem abgelegt und werden andererseits die Metadaten in RDF modelliert. Die RDF-modellierten Daten können über ein weiteres Interface, das die *Typo3 Semantic Web Extension* einsetzt, innerhalb der SFB806-Datenbank Webanwendung über den SPARQL-Endpoint durchsucht und angezeigt werden.

Das Frontend der SFB806-Datenbank ist eine Typo3 [8] basierte Webanwendung, für die eine Extension entwickelt wurde und wird, die das SFB806-Datenamangement (Dateiupload, Dateneingabe, Editieren und Anzeigen/Browsen der Daten) implementiert.

Forschungsdatenbank

Die Forschungsdatenbank dient als zentraler Datenpool für die Forscher des SFB806. Sie umfasst, wie erwähnt, vor allem Daten aus den Domains Archäologie, Geoarchäologie und Paläoumwelt. Die Datenbasis wird ständig erweitert, zum einen durch Integration von Publizierten Datensätzen aus den relevanten Fachdisziplinen und zum zweiten durch Eingabe von Daten durch die Forscher des SFB806.

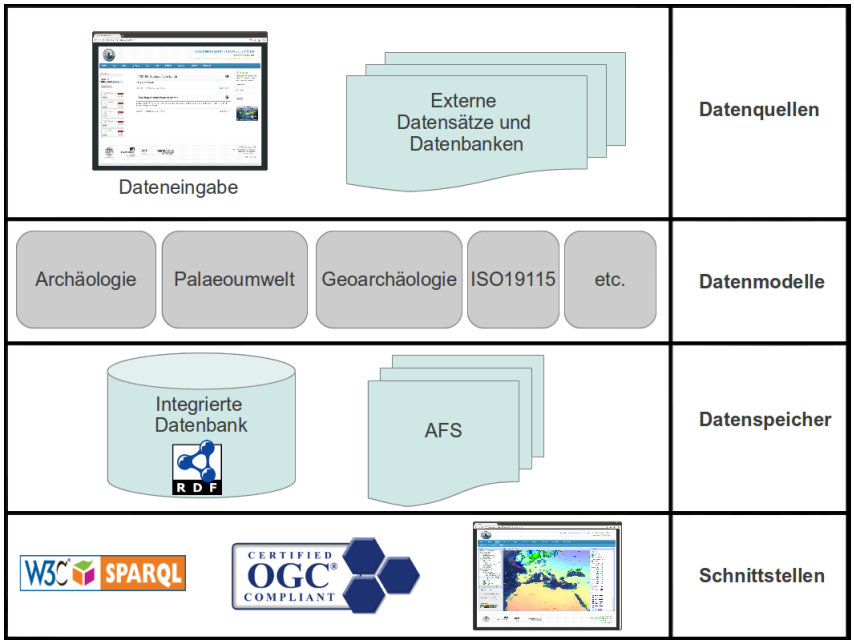

Abbildung 2: Systemarchitektur der Forschungsdatenbank.

Zusätzlich werden die Daten der Forschungsdatenbank fortlaufend mit externen Datenmodellen bzw. Ontologien/Vokabularen verlinkt. Zum eine um die semantische Aussagekraft der Daten zu erhöhen und zum zweiten um die Interoperabilität mit anderen Datenmodellen und Anwendungen zu unterstützen. Diese Technik ist auch unter dem Begriff *Linked Data* [9] bekannt.

Implementation der Forschungsdatenbank

Technisch werden die Daten in einer Graphdatenbank (Sesame [10]) gespeichert. Dazu werden die zu integrierenden Daten in den entwickelten Datenmodellen [1,2] in RDF [11] formuliert um sie dann im Sesame-Triplestore zu speichern.

Zu räumlichen Daten werden i.d.R. nur die Metadaten im RDF-Format gespeichert, der Datensatz selbst wird in einem herkömmlichen GIS-Datenformat (PostgreSQL/PostGIS Datenbank, als ESRI Shapefiles und/oder im GeoTiff-Format) gespeichert um diese über OGC WMS-, WFS- und WCS-Interfaces, und als Dateidownload zur Verfügung zu stellen.

Zum Zugriff auf die Daten der Forschungsdatenbank werden bis dato die folgenden Interfaces bereitgestellt:

- SPARQL-Endpoint
- OGC WMS, WFS, WCS
- GeoExt basiertes WebGIS

Weitere Interfaces sind in Planung (z.B. eine Exhibit [12] Timelinevisualisierung).

Ergebnisse

Bis jetzt sind ca. 20.000 archäologische, ca. 10.000 paläoumwelt und ca. 500 geoarchäologische Datensätze in die Forschungsdatenbank integriert. Diese Daten stehen somit den Forschern zur integrierten Analyse über die genannten Interfaces zur Verfügung. Zusätzlich gibt es eine Datenbank für GIS-Daten, die nicht in die drei genannten Modelle einzuordnen sind (v.a. Höhenmodelle/Bathymetrie und rezente/aktuelle Klimadaten und Umweltdaten), eine Literaturdatenbank und eine Mediendatenbank zur Speicherung aller Datensätze die in keinem der genannten Modelle modelliert werden (z.B. Projektberichte oder Videoaufnahmen von Vorträgen etc.).

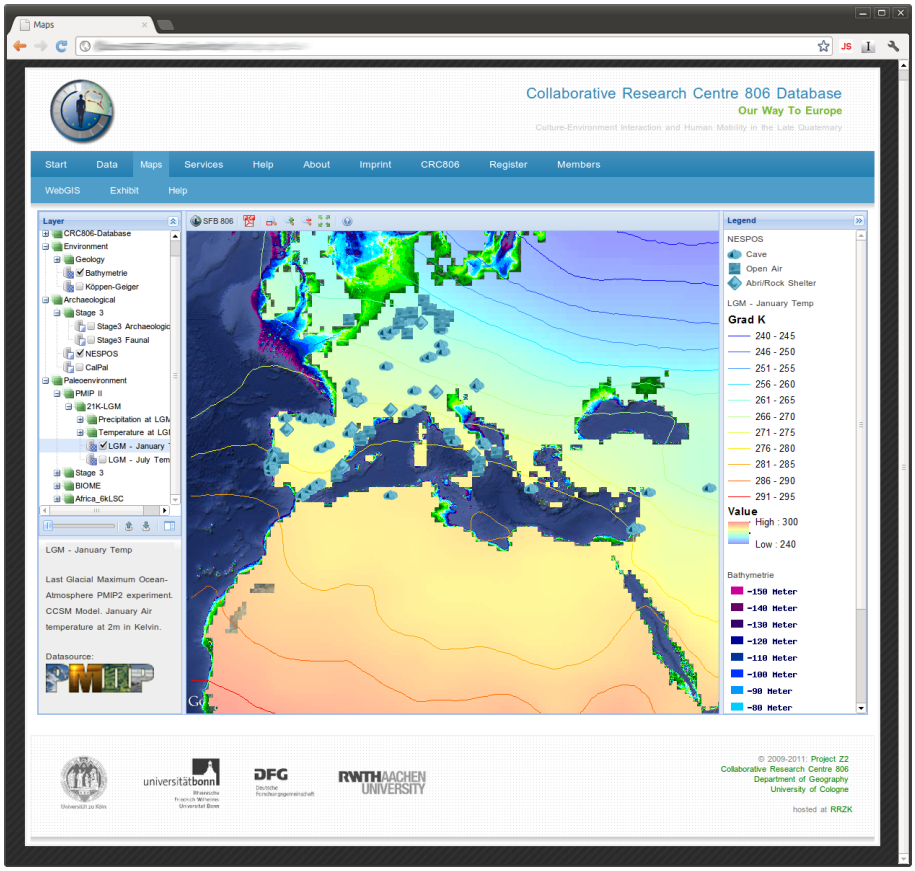

Abbildung 3: Screenshot des WebGIS-Interface.

Des weiteren steht der interessierten Öffentlichkeit ein Großteil der integrierten Datenbasis als OGC Webservices zur Verfügung. Der Zugriff auf den SPARQL-Endpoint ist vorerst nur für Mitglieder des SFB806 vorgesehen, da der Informationsvorsprung, der durch die integrierte Datenbank gegeben ist, zunächst intern für Forschungen genutzt werden soll. Nach Ablauf der aktuellen Projektphase (Sommer 2013) soll der SPARQL-Endpoint auch öffentlich zugänglich gemacht werden.

Fazit & Ausblick

Durch den Einsatz von Semantic Web Technologien ist es möglich, heterogene Datenbanken konsistent und mit überschaubarem Aufwand in einheitliche und eindeutige Modelle zu integrieren. Die Anwendung von Linked Data Methoden ermöglicht bzw. verbessert die Integration der entwickelten Modelle (und auf ihnen basierende Daten) in externe Anwendungen und umgekehrt die Einbindung externer Daten in SFB806-Anwendungen. Ein weiterer Vorteil von Graphdatenmodellen wie RDF ist, dass sie prinzipiell Schemaunabhängig sind. Dies ermöglicht die Weiterentwicklung von Datenmodellen unabhängig von der Anwendungsschicht [13].

Das System befindet sich zur Zeit in der internen alpha-Testphase, der öffentliche Start der SFB806-Datenbank Webseite (http://crc806db.uni-koeln.de) ist für Sommer 2012 geplant.

Kontakt zum Autor:

(Dipl.-Geogr.) Christian Willmes
Geographisches Institut, Universität zu Köln
Albertus-Magnus-Platz, 50923 Köln
+49 (0)221 470 6234
c.willmes@uni-koeln.de

Literatur

[1] *Willmes, C. und Bareth, G.*: A dataintegration concept for an interdisciplinary research database. In: Proceedings of the GI_Zeitgeist Young Researchers Forum, Universität Münster, 2012.

[2] *Willmes, C., Brocks, S., Hoffmeister, D., Hütt, C., Kürner, D., Volland, K. und Bareth, G.*: Facilitating spatio-temporal visualization and analysis of heterogeneoous archaeological and palaeoenvironmental research data. In: Proceedings of the XXII IRSPRS Congress, Melbourne, 2012.

[3] Allemang, D. and Hendler, J.: Semantic Web for the working Ontologist – Effective Modeling in RDFS and OWL. Morgan Kaufman Publishers/Elsevier, Amsterdam, Second Edition, 2011.

[4] Effertz, E.: The funders perspective: Data management in coordinated programmes of the German research Foundation (DFG). In: Curdt, C. und Bareth G. (Hrsg.), Proceedings of the Data Management Workshop 29.-30.10.2010, Kölner Geographische Arbeiten, Heft 90, Universität zu Köln, pp. 35-38.

[5] AFS – Andrew File System: http://de.wikipedia.org/wiki/Andrew_File_System.

[6] DublinCore Metadata Initiative: http://dublincore.org/.

[7] Geographic Information – Metadata ISO19115: http://de.wikipedia.org/wiki/ISO_19115.

[8] Typo3 CMS: http://typo3.org/.

[9] Linked Data: http://en.wikipedia.org/wiki/Linked_data.

[10] SESAME: http://www.openrdf.org/.

[11] Resource Description Framework - RDF: http://www.w3.org/RDF/.

[12] Exhibit - Publishing Framework for Data-Rich Interactive Web Pages: http://www.simile-widgets.org/exhibit/.

[13] Segaran, T., Evans, C. and Taylor, J.: Programming the Semantic Web. O'Reilly, Sebastopol, CA, USA. 2009.