Antrag

AN DIE DEUTSCHE FORSCHUNGSGEMEINSCHAFT
FÜR DIE ANFORDERUNG EINER SACHBEIHILFE

Life

Linked Data for eScience Services

Prof. Dr. Werner Kuhn, Institut für Geoinformatik, Westfälische Wilhelms-Universität Münster
Dr. Beate Tröger, Universitäts- und Landesbibliothek, Westfälische Wilhelms-Universität Münster

ifgi
Institut für Geoinformatik
Universität Münster

ulb Münster

Westfälische
Wilhelms-Universität
Münster

1 General Information - *(Allgemeine Angaben)*

Application for a Grant - *(Antrag auf Gewährung einer Sachbeihilfe)*

Initial Application - *(Neuantrag)*

1.1 *Applicant (1)* - *(Antragsteller)*

First name, surname, title:	Kuhn, Werner, Prof. Dr.
Employment status:	Full professor
Date of birth:	24 August 1957
Nationality:	Swiss
DFG code number of the latest application:	KU 1368/8-1
Institution/ Department:	Westfälische Wilhelms-Universität Münster, Institut f. Geoinformatik (ifgi)
Postal address:	Weseler Str. 253, 48151 Münster
Telephone:	0251–83 33083
Fax:	0251–83 39763
E-mail:	kuhn@uni-muenster.de

Applicant (2)

First name, surname, title:	Tröger, Beate, Dr.
Employment status:	Library Director
Date of birth:	05 January 1961
Nationality:	deutsch
DFG code number of the latest application:	INST 2097/6-1
Institution/ Department:	Westfälische Wilhelms-Universität Münster, Universitäts- und Landesbibliothek Münster (ULB)
Postal address:	Krummer Timpen 3, D-48143 Münster
Telephone:	0251–83 24022
Fax:	0251–83 28398
E-mail:	troeger@uni-muenster.de

The ifgi (Prof. Dr. Kuhn) has the responsibility for the project implementation as well as the management and accounting of the project funds.

1.2 *Topic*

Life: Linked Data for eScience Services

Thema

Life: Linked Data für eScience Dienste

1.3 *Funding Program / Scientific Discipline and Field of Work*

- (Förderprogramm, Fachgebiet und Arbeitsrichtung)

Funding program:	LIS, "Werkzeuge und Verfahren des wissenschaftlichen Informationsmanagements"
Scientific discipline:	Library science, Geoinformatics
Field of work:	Information management and retrieval, information services, workflow management, geospatial data, Linked Open Data

1.4 *Scheduled Duration in Total* - *(Voraussichtliche Gesamtdauer)*

Project start:	1 January 2012
Funded by DFG since:	–
Scheduled duration:	24 months
DFG funding requested:	24 months

1.5 *Application Period* - *(Antragszeitraum)*

Funding is requested for 24 months.

1.6 *Summary*

The overall goal of *Linked Data for eScience Services* (LIFE) is to **facilitate sharing of spatio-temporal information and thus improve interdisciplinary collaboration in science and education**. This approach addresses all kinds of resources, ranging from articles and books over maps to raw data. The **Linked Data** approach will be used as a basis for the university library's eScience services to seamlessly integrate their offerings into both the scientific and the global information infrastructure. These eScience services will enable researchers and students to systematically navigate the dynamic and heterogeneous global network of spatio-temporal information (discovery) and to create the relevant views (access) meeting their information needs. The project particularly aims at **overcoming the information silos** that have been created both in libraries and in the geospatial domain, as the detailed standards for data sharing developed in both fields impede information integration with other data sources. We will develop **workflows that support the semi-automatic acquisition and semantic annotation of geographic information**. We will build on existing standards and extend them with Linked Data interfaces, focusing on a tight integration of bibliographic contents. This will allow for novel user interfaces for retrieving contents through spatio-temporal queries (e.g., "books about medieval Westphalia"). LIFE is a research activity in the *Linked Open Data University of Münster* initiative that fosters exchanging scientific and educational data as Linked Data.[1]

Zusammenfassung

Ziel des *Linked Data für eScience Dienste* (LIFE) Projekts ist es, die **gemeinsame Nutzung von raumzeitlichen Daten in Forschung und Lehre zu vereinfachen**. Dieser Ansatz umfasst verschiedenste Ressourcen, wie beispielsweise wissenschaftliche Artikel, Bücher, Karten, Kartenwerke und Primärdaten. Raumbezogene Daten werden dabei mit Hilfe von **Linked Data** Technologien interdisziplinär recherchierbar und effizient nutzbar gemacht. Diese Technologien bilden die Basis für die eScience Dienste der Universitätsbibliothek, sodass deren Informationsangebote sich nahtlos in die globale Informationsinfrastruktur einfügen, um die wissenschaftlichen Arbeitsprozesse zielgerichtet zu unterstützen. Durch die Integration der verschiedene existierenden Standards mittels *Linked Data* können die bisherigen **Informationssilos geöffnet und zugänglich gemacht werden**. Das hier beantragte Projekt versucht deshalb die bisher erarbeiteten vielversprechenden Linked Data Ansätze praxisorientiert weiter zu entwickeln und ihr Potenzial im Hinblick auf ein innovatives, interdisziplinäres Data Retrieval zu erforschen. Dabei sollen sowohl **nachnutzbare Abläufe und Werkzeuge zur semantischen Annotation von raumzeitlichen Daten** (u.a. Geokoordinaten) als auch die entsprechenden Open Source Recherchewerkzeuge entwickelt werden. LIFE ist eine Forschungsaktivität im Rahmen der *Linked Open Data University of Münster* Initiative zur Bereitstellung von Forschungs- und Lehrdaten als Linked Data.[1]

2 State of the Art - *(Stand der Forschung, eigene Vorarbeiten)*

2.1 *State of the Art* - *(Stand der Forschung)*

In the following, we give an overview of the state of the art in the disciplines relevant to this proposal. In each section, we also point to previous work within the research groups that will be involved in this project.

2.1.1 Cataloguing and digitizing workflows

The standard for cataloguing maps in all German libraries ("RAK Karten", Baader and Poggendorf 1987) was

[1] http://lodum.de

developed in 1987 and is still in use with an amendment[2] from 2010 for libraries of the GBV[3] (Common Library Network). Due to the lack of special retrieval systems for spatio-temporal data (e. g. map based search, hierarchical browsing for locations) most libraries are cataloging just the mandatory metadata fields, but fields for geographical coordinates are only optional and seldom filled in. References to places are established by assigning geographic headings from the "Schlagwortnormdatei" (SWD[4]), a controlled vocabulary. In 2011 the DNB started to enrich the geographic headings of the SWD with geographic coordinates. This service is a promising candidate among other spatio-temporal data services to be integrated into LIFE as one connecting point between library and geospatial sciences.

Digitization systems have been established in many libraries over the past years, based on the international approved metadata standards MODS[5] and METS[6], both maintained by the Library of Congress. A digitization infrastructure is currently to put into operation by ULB Münster. As the applied MODS standard allows the capturing of various relevant metadata, an innovative workflow has to be developed that integrates bibliographic metadata with geo-coordinates. Such a geographic annotation is a fundamental requirement for any efficient geographic information service; however, this approach has not been put into practice at a broader level yet.

2.1.2 Spatio-temporal data in libraries

In bibliographic data management as practiced in libraries, geographic references are most often treated as keywords, along with other thematic information that describes a book's thematic coverage, for example. In most cases, only the place of publication is clearly marked as a reference to a place name, and even these are seldom linked to a gazetteer that resolves that place name to its geographic location (Goodchild and Hill 2008). The same applies for scanned maps, as outlined in Section 2.1.1: the spatial extent of the map is usually not recorded in the metadata, and references to contemporary or historical gazetteers are often missing. Some research projects have tried to tackle these shortcomings over the past years. GEO-LEO[7] is a virtual expert library for the geo sciences that integrates different library catalogues. While it also covers a collection of maps, the system only provides the corresponding metadata; it is not possible to access them from GEO-LEO. Moreover, the collection is limited to the catalogues already integrated, i.e., the user cannot include any other data in the search process to enrich the results. While GEO-LEO integrates data from several sources, it remains an isolated information silo that lacks interoperability with other information on the Web. This also applies for projects such as the historic maps archive at the university library in Bremen.[8] It offers spatial and thematic queries as well as a gazetteer including historic place names, but single items in the collection are not referenceable and can therefore not be interlinked with other data sources. This is despite current activities at various levels that make library data available as Linked Open Data, such as published by the DNB[9], HBZ[10] and the University Library in Mannheim.[11]

Bibliographic content typically originates from different times, especially in an academic environment where the historical dimension is crucial in many fields. There are numerous different types of references to place names in the metadata descriptions of the content. A critical problem of linking the content through place names occurs when place names have changed, or boundaries they refer to have been modified through changes like merges and splits. These changes can concern any places, e.g. municipalities, states or even countries. In Finland, the solution has been to use Linked Data technologies and spatiotemporal reasoning mechanisms to create the Finnish Spatiotemporal Ontology (SAPO; Kauppinen et al. 2008; Kauppinen et al. 2011)[12], meant for offering links between contemporary and historical municipalities. SAPO can be linked to any web-based system to offer query expansion: if a user wants to query using a contemporary municipality name, she can select to expand the query with historical, overlapping municipalities and thus get more search results from around the same area as the original query term.

At the EU level, the Europeana[13] project provides access to the digital resources of museums and libraries, many of which include historic maps and other kinds of spatio-temporal data. Europeana stores spatial

[2] http://www.gbv.de/vgm/info/mitglieder/02Verbund/01Erschliessung/02Richtlinien/01KatRicht/sondktn.pdf
[3] http://gbv.de
[4] http://www.d-nb.de/standardisierung/normdateien/swd.htm
[5] http://www.loc.gov/standards/mods/
[6] http://www.loc.gov/standards/mets/
[7] http://www.geo-leo.de/
[8] http://gauss.suub.uni-bremen.de/suub/hist/index.jsp.
[9] http://www.d-nb.de/hilfe/service/linked_data_service.htm
[10] https://wiki1.hbz-nrw.de:443/display/SEM/Aktuelle+Open-Data-Exporte
[11] http://data.bib.uni-mannheim.de/dokumentation.html
[12] See http://seco.tkk.fi/ontologies/sapo/ for description, and http:/kulttuurisampo.fi/historiallisetAlueet.shtml for an example use in linking content.
[13] http://www.europeana.eu

references as common properties, i.e., by adding a place name such as 'Münster' as a keyword, but without further reference. Maps are stored as images and in most cases not geo-referenced. It is therefore not possible to automatically include these geodata in an integrated retrieval workflow. Within the Europeana Connect[14] project, one of the main objectives is the development of a machine-readable Linked Data representation of the Europeana contents (Isaac 2010). These efforts around a semantic layer for the contents already covered in the archive, and do not include any work on annotation and integration workflows. The research proposed in this project would hence also facilitate larger efforts such as Europeana, and local resources could be interlinked with items in such large digital collections. While we focus on the annotation workflows for professionals in libraries in this proposal, other approaches try to leverage the potential of *crowdsourcing* for interlinked bibliographic databases[15] and in the annotation of historic maps (Simon et al. 2010).

2.1.3 Data exchange in the geoscientific domain

The geographic domain has developed a number of standards for services and data formats to exchange geographic information. The Open Geospatial Consortium (OGC) is the leading standardization organization, consisting of members from industry, government, and academia. The OGC community has developed specifications for accessing maps (Web Mapping Service, WMS), vector (Web Feature Service, WFS) and sensor data (Sensor Observation Service, SOS). These services are accompanied by a collection of specifications that support tasks such as geodata processing or service cataloguing. Most Spatial Data Infrastructures (SDIs) build on OGC services and the Geographic Markup Language (GML), so that these specifications are widely used in practice and have made their way into all major Geographic Information Systems (GIS). Despite the wide usage of OGC standards within the domain, recent developments in the community, such as the GeoSPARQL working group, point to the insight that interfaces are required that make geographic information accessible outside of the domain. This is especially the case for the Linked Data cloud[16] (Bizer et al. 2009), where data from a wide range of domains is interlinked. Although geographic information, such as that provided by the GeoNames gazetteer or LinkedGeoData[17] (Auer et al. 2009) plays a major role in the Linked Data cloud, OGC services do not. Several research efforts (Keßler and Janowicz 2010, Janowicz et al. 2010a,b) hence equip the OGC services with additional interfaces that make the data available to the Linked Data world. Vice versa, Linked Data developers are starting to develop Web applications that place RDF (Resource Description Framework) data with spatial references on a map.[18]

At the geosciences faculty of the University of Münster, the StudMap 14 project[19] provides e-Science services based on OGC services for geographic data for students and researchers. This project features a browser-based map viewer to directly portray geographic data, which is available within the University of Münster's local computer network. Students can view the data and integrate it with their own geographic content. Moreover, they can integrate the available services of StudMap 14 into other software systems, which are used for research and teaching at the faculty (such as ESRI ArcGIS). StudMap 14 is therefore built on interoperable services, which deliver the maps as images to the browser application. The infrastructure of StudMap 14 is based on software available under Free and Open Source licenses, which reduce the costs of the project significantly (compared to an proprietary licensed software stack). In the future StudMap 14 will be enhanced to become a collaborative platform between students to exchange their geographic data (for instance produced during their study projects) with others and to help discover and connect to existing work. Figure 1 shows the current state of the StudMap 14 portal.

[14] http://www.europeanaconnect.eu/
[15] http://openlibrary.org/
[16] http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22.pdf
[17] See http://www.geonames.org/ and http://linkedgeodata.org/ .
[18] http://oegdev.dia.fi.upm.es/projects/map4rdf/
[19] http://gdione4all.uni-muenster.de/joomla/index.php/studmap14

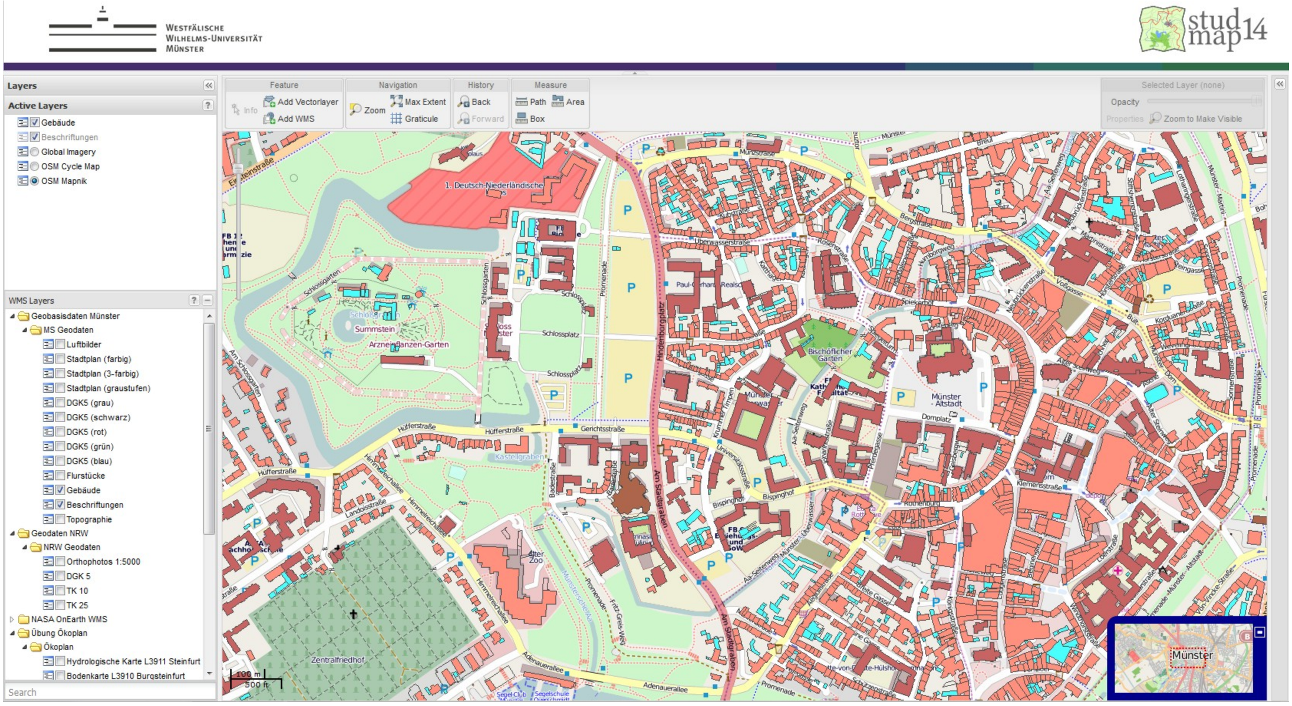

Figure 1: User interface of the StudMap14 portal.

2.1.4 Semantic data integration

Combining data from different sources is often problematic due to the heterogeneity of the data. At the semantic level, this problem is based on the different intended meanings of terms in the data sources' original contexts (Visser et al. 2002). According to Schade (2009), these different contexts lead to naming conflicts (e.g. homonyms and synonyms), scale conflicts (e.g., different units or measurement scales), and precision or resolution conflicts (e.g. if two similar measurements are made with sensors of different precision). The Semantic Web (Berners-Lee et al. 2001) tried to tackle this problem using ontologies (Gruber 1993, Guarino 1998) for the semantic annotation of datasets on the Web. For several reasons – complex ontology languages, lack of easy-to-use ontology engineering tools and unclear chances for success, among others –, this top-down approach was not successful on a broader scale. Instead, the lightweight bottom-up approach of *Linked Data* (Bizer et al. 2009) is increasingly being adopted. This novel approach shifts the focus from sophisticated ontologies to the raw data and loosely structured vocabularies developed within the community, using URIs as identifiers and the simple RDF triples consisting of subject, predicate, and object for statements. By making the previously mentioned URIs resolvable (i.e., by using HTTP URLs) (Berners-Lee 2009), a collection of more than 200 interlinked data sources was made available on the Web over the past years. This Linked Data cloud consists of close to 30 billion RDF triples[20] and covers contents from the sciences, media, libraries, Wikipedia, the public sector, as well as several geodata sources such as Linked Geo Data, Linked Sensor Data, GeoNames, and postcodes. Interlinking different data sources is often done manually on a *per data type* basis or semi-automatically by using link discovery frameworks (such as Silk; see Volz et al. 2009). The goal of this project is to develop workflows and tools that allow librarians to annotate spatio-temporally referenced data in the cataloguing workflow, so that they can easily be interlinked with other sources in the Linked Data cloud.

The Münster Semantic Interoperability Lab (MUSIL) at Institute for Geoinformatics, University of Münster, has been doing extensive research on all aspects of semantic integration of spatio-temporal information for several years. Kuhn (2003) introduced the idea of *semantic reference systems* in analogy to spatial and temporal reference systems. On a structural level, a *semantic datum* (which grounds the meaning of basic terms outside the system) and a *semantic reference frame* (a formally defined framework to which terms can be related to obtain meaning) are needed, which usually come in the form of ontologies (Schade 2009) or conceptual spaces (Gärdenfors 2000). On a functional level, *semantic referencing* (the basic process for relating the terms in a data model to a semantic reference frame) and *semantic translation* (the transformation from one semantic reference system to another) are also needed. Semantic referencing (Scheider et al, forthcoming) can also be seen as the process of annotating data model elements with the semantic reference frame. Semantic reference systems have been studied in detail in the SeReS project

[20] According to http://www4.wiwiss.fu-berlin.de/lodcloud/state/ .

(funded by the DFG under KU 1368/4-1 and KU 1368/4-3). Semantic reference systems play a central role in the International Research Training Group on Semantic Integration[21] (DFG GRK 1498).
The SimCat project (DFG Ra1062/2-1) has developed a semantic similarity measure for concepts expressed in powerful description-logic based ontology languages (Janowicz et al. 2008a). The developed similarity measure has been applied for gazetteer mapping (Janowicz and Keßler 2008), human-computer interaction (Janowicz et al. 2009), as well as ontology engineering (Janowicz et al. 2008b). The second phase of the project (DFG Ja1709/2-2;) put the focus on context-awareness in similarity-based information retrieval by defining a measure to assess the relevance of contextual aspects for a given task (Keßler forthcoming 2011).
MUSIL has worked on semantic enablement for spatial data infrastructures (Janowicz et al. 2010a) in collaboration with the 52°North semantics community[22] before, which also acts as an external partner in this proposal. The approaches developed during this collaboration targeted at opening up the geospatial Web services defined by the OGC towards the Semantic Web and the Linked Data cloud, without breaking the compatibility of the services (Janowicz et al. 2010b, Keßler and Janowicz 2010). The recent developments around Linked Data triggered a number of activities within MUSIL (and the Institute for Geoinformatics as a whole), with the most important being the development of the proposal and coordination of the Linked Open Data University of Münster (LODUM) initiative.
As these diverse activities show, MUSIL has extensive experience in different aspects of geospatial semantics. In this project, we combine our expertise with the university libraries expertise in the bibliographic domain to improve management and retrieval of spatio-temporally referenced information.

2.2 *Publications* - *(Verzeichnis projektspezifischer Publikationen)*

1. Kuhn, W., 2010. Modeling vs Encoding. Semantic Web - Interoperability, Usability, Applicability. Semantic Web Journal, IOS Press, 1(1).
2. Kuhn, W., 2009. Semantic Engineering. In G. Navratil (Ed.): *Research Trends in Geographic Information Science.* Springer-Verlag Lecture Notes in Geoinformation and Cartography: 63-74
3. Kuhn, W., 2005. Geospatial Semantics: Why, of What, and How? *Journal on Data Semantics* (Special Issue on Semantic-based Geographical Information Systems, Spring 2005, LNCS 3534): 1-24
4. Tröger, B.; Lorenz, J.; Klötgen, S.; Przibytzin, H. (2011): Das digitale wissenschaftliche Informationssystem der WWU Münster. In: Fortschritte des integrierten Informationsmanagements an Hochschulen. Wissenschaftliche Schriften der WWU Münster; Reihe XIX, Bd. 2.
5. Vogl, R., Gildhorn, A., Lorenz, J., Wibberg, M. (2010): Integriertes Informationsmanagement an der Westfälischen Wilhelms-Universität Münster – das Projekt MIRO. In: Informationsmanagement an Hochschulen, Springer Verlag, Berlin Heidelberg, 47-62.
6. Vogl, R., Lorenz, J., Gildhorn, A., Schild, Chr.: Unternehmensweite Erschließung von Informationsquellen durch fortschrittliche Suchmaschinentechnologie - Erfahrungen aus dem MIRO Projekt der WWU Münster. In: Praxis der Informationsverarbeitung und Kommunikation (PIK), K.G. Saur Verlag, München, 03/2008.

3 Goals and Work Schedule - *(Ziele und Arbeitsprogramm)*

3.1 *Goals* - *(Ziele)*

The goal of this project is to improve the capture, exchange and integration of spatio-temporal data with a focus on library information systems and geospatial services using Linked Data technologies to satisfy contemporary user needs.
Currently, university libraries are extending their traditional role as providers of bibliographic resources with service-oriented information infrastructures that offer various kinds of data. Among these are spatio-temporal data—different kinds of maps, including scanned historic maps, as well as the underlying geographic information—that are of interest across different disciplines. These data, however, have been neglected so far, due to a lack of appropriate cataloguing and retrieval tools.
Spatio-temporal data are already available in standardized formats through standardized Web services, as defined by the Open Geospatial Consortium. These formats and Web services, however, have been developed for domain specialists and domain-specific applications as data consumers, as they can be found in Spatial Data Infrastructures (SDIs). Because of this strong focus, the standards for geographic information are not used outside the field, which hampers integration of spatio-temporal data with data from other

[21] http://irtg-sigi.uni-muenster.de/
[22] http://52north.org/semantics/

domains. This applies to administrative data (as published in the realm of the *Open Data* movement), business data, data from social networks, and to the scientific content that are in the focus of this project.

This content is available through eScience services. It ranges from bibliographic data, cultural heritage data and medical case databases to data from the social sciences such as interview transcripts. Despite the heterogeneity of these data, they often have spatio-temporal characteristics that allow for integration with geographic information provided according to the above-mentioned standards. Archaeological artifacts, for example, have been found at a specific excavation and are often associated with a place of production; social data can always be associated with the location of the study participants (such as their home town); and books cannot only be associated with the place of publication, but may also contain references to geographic locations. Likewise, many kinds of data carry temporal information. Beyond the inherent creation or update timestamps, temporal aspects such as publication dates (for bibliographic data) or time frames for studies are core elements of most metadata. **The spatio-temporal characteristics of datasets from different domains therefore bears great potential to act as an integrator that connects different datasets to make them queriable and useable together.**

The goal of LIFE is therefore to develop a reusable, open-source collection of workflows and tools to capture, catalogue, manage, search, find, and access spatio-temporal data. A special focus in these workflows is on interlinking the captured data with existing data sources, both in the geo domain (spatial data infrastructures, geo Web services) and in other disciplines (such as found in bibliographic or cultural heritage information systems, for example) through Linked Data. In particular, Linked Data provide geographic information to a wider audience, making them accessible in machine-readable formats. **The Liked Data approach does not only improve discovery and hence potential usage of geographic information, it also adds a level of semantic annotations to the data that the existing services for geographic information are largely missing.** This allows the user to make combined spatio-temporal and thematic queries. The requirements for these semantic annotations will be evaluated in three eScience use cases, which have been established in cooperation with the Institute for Planetology[23], the Institute for Epidemiology and Social Medicine (Section Clinical Epidemiology)[24], and the Institute for Comparative Urban History[25] (all at University of Münster). Based on these use cases (see Section 3.2 for details), which cover different aspects of data capturing, managing, and retrieval, we will work towards the following specific goals of LIFE:

1. **Development of capturing and annotation workflows**. Existing workflows for capturing bibliographic data focus on domain-specific annotation. The interlinking with other related data sources is currently not part of the workflow and carried out manually on demand, i.e., if two or more data sources need to be integrated. To interlink resources automatically, however, it is necessary to *know* that the respective data is available, and where to find it. A dataset that is interlinked with related data through standardized vocabularies from the very beginning facilitates data integration. **We hence focus on an extension of existing capturing workflows (as shown in the three use cases) to integrate existing and related datasets through Linked Data.** The workflows will be generically defined in a high-level modeling language (such as the Business Process Modeling Language), so that they can be easily adapted to new use cases.

2. **Development of annotation and integration tools.** New workflows for annotating data require tools that allow the librarian to efficiently complete the corresponding tasks. The goal is to achieve a maximum degree of annotation and links with minimal effort. **The tools to be developed hence need to guide the user in the annotation process and make use of recommender systems approaches to automatically suggest existing sources to the librarian.** Data sources to be linked are not limited to the context of the university, but should explicitly include external high-quality data sources to make sure that the generated links embed the local data in the Linked Data cloud. These data sources can be found using annotations, keyword search, and especially spatial and temporal co-location. Information on place names will be realized based on existing gazetteers (such as http://geonames.org). The tools to be developed will build on existing link discovery frameworks (such as SILK[26]), indexing tools (Lucence, Solr) and an enterprise content management system (Alfresco). The tools will be made available as open source software to ensure that the outcome of this project is made available for reuse.

3. **Specification and implementation of eScience services.** The project aims at a seamless integration of thematic data in Linked Data sources and spatio-temporal data provided by services in spatial data infrastructures. It is in line with the pronounced requirements for eScience according to

[23] Institut für Planetologie: http://www.uni-muenster.de/Planetology/homepage/homepage.html
[24] Institut für Epidemiologie und Sozialmedizin: http://campus.uni-muenster.de/epi_start.html
[25] Institut für vergleichende Städtegeschichte: http://www.uni-muenster.de/Staedtegeschichte/
[26] http://www4.wiwiss.fu-berlin.de/bizer/silk/

the BMBF[27], which defines it as "approaches towards a knowledge network […] which offer schools, universities, research institutions, companies and administrations individual processes and services for organizing and developing their knowledge." This requires service encapsulations that make the respective data available for the "other side": SDI services providing spatio-temporal data must expose their service descriptions and contents for Linked Data consumers via RESTful services and offer a SPARQL endpoint for querying; likewise, Linked Data sources containing spatial data must be made available through standard geo Web services (Web Map Service, Web Feature Service, Web Coverage Service) to make this information available within Spatial Data Infrastructures. **Encapsulating existing services leverages established forms of data exchange, while adding novel retrieval opportunities and – in the case of offering spatio-temporal data as Linked Data – enhanced semantic data annotations.** The developed software components will become part of the 52° North semantics community software as to ensure long-term availability and further development by the 52° North community.

4. **Design and implementation of retrieval interfaces**. User interfaces for Linked Data most often consist of faceted browsers for data exploration, combined with SPARQL query builders that allow experts to find data that fit specific criteria. The corresponding user interfaces for geo Web services typically consist of interactive map browsers with layer selection and limited GIS functionality. When combining these two worlds, new user interfaces are required that go beyond a plain combination of these two user interface paradigms. Simply combining them would lead to very complex user interfaces that hamper usability. Moreover, both of them lack a proper integration of time as a query constraint. **We therefore suggest developing a guided retrieval process that allows the user to start with spatio-temporal constraints and then refine the results by selection of thematic attributes.** For experts and for applications building on top of the generated Linked Data, an additional application programming interface needs to be defined as one eScience service that allows for any easy integration of the delivered data into other applications. The implementation of powerful spatial query capabilities requires an implementation of the novel GeoSPARQL standard, which is currently being finalized within OGC.

The transfer of the developed tools from the experimental project stage to the regular services of the university library is ensured by the active project participation of all involved parties (library, center for information processing, and Linked Open Data University of Münster team). The prerequisites for an easy integration will therefore be taken into account throughout the development process.

3.2 ***Work Schedule*** - *(Arbeitsprogramm)*

The project work plan is structured as a series of work packages. Their timelines and dependencies are presented in 3.2.9.

3.2.1 WP 0: Project initialization

This work package introduces new staff members to the project group, develops the baseline for the project, and introduces new staff to the internal workflows and the cooperation with other groups. Moreover, it initiates communication with 52n semantics community, the institutes working on the use cases, and the Center for Information Processing.

2 person months.

Steps
1. Hire staff
2. Organize two-day kick-off workshop
3. Set up project website and repositories
4. Establish links for external cooperation
5. Initialize collaboration with project partners

Deliverables

D-0.1 Kick-off meeting report with detailed research agenda
D-0.2 Collection of readings and links relevant for the project on project website

[27] http://www.bmbf.de/en/298.php

3.2.2 WP 1: Use case "comparative urban history"

The Institute for Comparative Urban History owns a large collection documenting urban history, managing their data using monolithic librarian systems. It consists of a bibliography (~150.000 entries, catalogued in Allegro), a collection of maps (~20.000, about 30% catalogued in a MySQL database, partly also with scans), picture postcards (~30.000, catalogued in LitW3) and diapositives (~9.000, not catalogued). The institute receives frequent enquiries for access to this collection for various use cases. However, this information demand can often not be met, because (a) the data is only partly catalogued; (b) the intellectual property rights (IPR) are not cleared; and (c) access is overly complicated because of the heterogeneous systems used for different kinds of contents. The main challenge in this use case is hence to develop workflows that allow for an easy migration from the existing catalogues to a Linked Data representation, while maintaining access restrictions if IPRs are not clear (Linked Data versus Linked Open Data). The management of these contents hence requires a license management that is directly reflected in authorization mechanisms that filter out content that cannot be made available publicly for request from outside the university network. Metadata for these items, however, can still be made available for the general public. In addition to the conversion of existing catalogues, the contents that have not been catalogued yet have to be added to the digital collection. During this process, the contents need to be annotated (see WP4), interlinked (see WP5) and made available through Web services (WP6) and user interfaces (WP7).

5 person months.

Steps

1. Kick-off meeting with urban historians
2. Analysis of existing catalogues
3. Extension of the existing data models
4. Develop requirements specification for annotation workflows and tools
5. Implement and test software components
6. Establish workflows

Deliverables

D-1.1 Kick-off meeting report with detailed research agenda
D-1.2 Requirements specification
D-1.3 Conference publication about the application of Linked Data in urban history
D-1.4 Workflow specification and annotation software (see WP 4 and 5)

3.2.3 WP 2: Use case "cancer research"

The Institute for Epidemiology and Social Medicine conducts cancer research, trying to link spatio-temporal clusters of cancer cases to sources of carcinogenic emmissions or exposition to high concentrations of chemicals in air, water and soil. In addition to the obligatory case data required for this research, a broad range of data sources is required to cover as many possible causes for a cluster of cancer cases as possible. So far, the sources taken into account were retrieved in a non-systematic way, by using sources that were either known from previous research, that had been pointed out by other researchers, and by asking around. The sources found this way had to be inspected for fitness for use one by one, checking their spatial scope and scale, timeliness, and reliability. This unsystematic approach was taken due to a lack of services that offer a structured approach for data retrieval that combines thematic and spatio-temporal constraints, and integrates data offered by different, distributed sources.

The purpose of the inclusion of this use case in the project is to study how spatio-temporal details retrieved, accessed and used in disciplines that have no expertise in geo-web services. The goal is to improve the corresponding retrieval workflow so that such users can browse the Linked Data cloud based on the spatio-temporal and thematic constraints given by their research problem to access relevant journals and books in libraries, or demographic and statistical data. In an ideal case, the retrieved data should be readily available for inspection and further processing in the scientists resepective research environment, such as a statistics program or a Geographic Information System.

5 person months.

Steps

1. Kick-off meeting with Epidemiology researchers
2. Modeling of the current retrieval and usage workflow
3. Modeling of an ideal retrieval and usage workflow from a user's perspective
4. Develop requirements specification for retrieval tools out of the workflow model
5. Evaluate usability of retrieval tools for this use case after development (see WP 7)

Deliverables

D-2.1 Kick-off meeting report with detailed research agenda
D-2.2 Requirements specification
D-2.3 Conference publication about the application of Linked Data in cancer research
D-2.4 Workflow specification and annotation software (see WP 4 and 5)

3.2.4 WP 3: Use case "planetology"

The Institute for Planetology is comparing landforms on the Archipelago of Svalbard with similar landforms on Mars. High resolution aerial imagery (color data, 20 cm/pxl) and topographic data (50 cm/pxl) derived from the stereo channels of six large areas on Svalbard were acquired in 2008 with the High Resolution Stereo Camera – airborne version (HRSC-AX) by a flight campaign led by the German Aerospace Agency (DLR). In cooperation with DLR field work including measurements of various landforms on Svalbard were conducted in 2008, 2009 and 2011 in some of the regions where high-resolution image and topographic data was acquired. In addition, air imagery of the years 1960/61 and 1990 for one region with a comparable ground resolution to the HRSC-AX data set of 2008 was purchased. Furthermore, there exist numerous thematic published thematic maps such as geomorphologic and geologic maps and several hundred published scientific papers mostly dealing about landforms in much defined spatial areas.

The purpose of the inclusion of this use case in the project is to study how large multi-temporal datasets and thematic maps can be linked to spatial-related published maps and literature (e.g., from library catalogues) and recent spatial-related measurements conducted during fieldwork (e.g., own and/or other geodata). The goal is to improve the corresponding retrieval workflow so that such users can browse the Linked Data cloud based on the spatial-temporal and thematic constraints given by their research problem. In an ideal case, the retrieved data should be readily available for inspection and further processing in the scientists respective research environment, such as a statistics program or a Geographic Information System.

5 person months.

Steps

1. Kick-off meeting with planetology researchers
2. Modeling of the current retrieval and usage workflow
3. Modeling of an ideal retrieval and usage workflow from a user's perspective
4. Develop requirements specification for retrieval tools out of the workflow model
5. Evaluate usability of retrieval tools for this use case after development (see WP 7)

Deliverables

D-2.1 Kick-off meeting report with detailed research agenda
D-2.2 Requirements specification
D-2.3 Conference publication about the application of Linked Data in planetology (e.g., ESWC)
D-2.4 Workflow specification and annotation software (see WP 4 and 5)

3.2.5 WP 4: Annotation and integration workflows

The annotation process for new resources in libraries is currently focused on compliance with bibliographic standards. The corresponding workflows need to be extended with the steps required to make the new resource available as Linked Data (see Figure 2). The required extensions comprise two main components: vocabulary selection and integration with other data sources.

Vocabulary selection is a one-off task. In order to assure interoperability with other services that offer bibliographic contents as Linked Data, it is crucial that these services use the same vocabularies for annotation. The main task here is therefore to identify standard vocabularies that have already been established within the community, and to actively participate in the process of establishing standards where they are still missing. The identified vocabularies can then be integrated in the annotation workflow without further input requirements for the user, as they can be realized as straight-forward mappings of existing metadata fields to properties described in the vocabularies (see the implementation in WP 5).

Integration with other sources is a fundamental step to properly embed the annotated data in the Linked Data cloud. Unlike the vocabulary selection, this is a task that needs to be completed for every new resource. The cataloguing workflow therefore needs to be thoroughly analyzed, so that this additional task can be integrated in a non-obstructive way. The integration – adding links to external sources on the Linked Data cloud – will be based on a semi-automatic approach developed in WP 5. The task in this work package is therefore to develop an annotation workflow model that can be seamlessly integrated with the existing workflow. Beyond the actual process, both vocabularies for the links to generate and data sources to link to have to be identified. This part of WP 4 will focus on data sources with a strong spatio-temporal component.

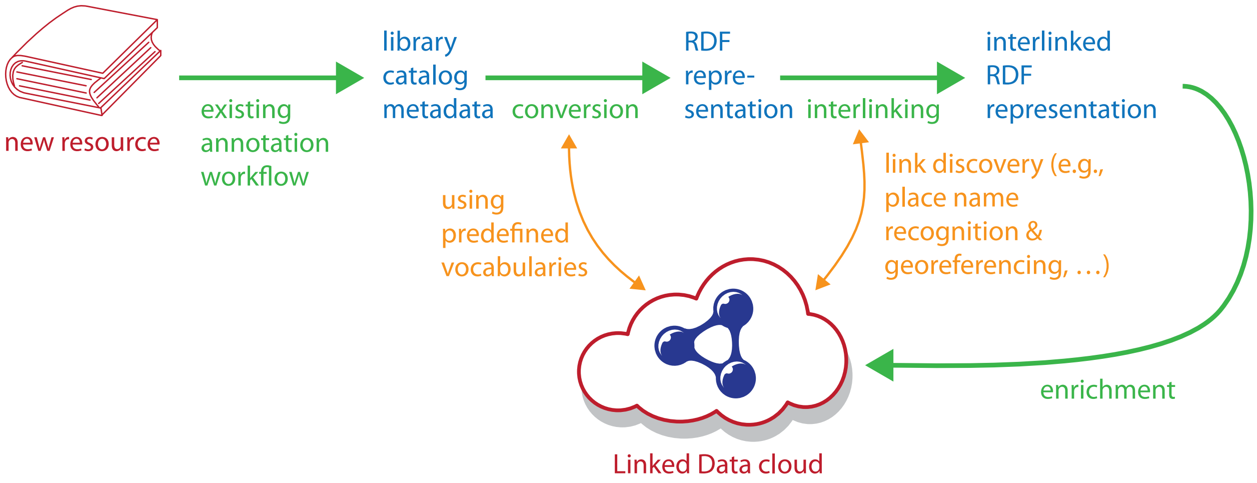
new resource
existing annotation workflow
library catalog metadata
conversion
RDF repre-sentation
interlinking
interlinked RDF representation
using predefined vocabularies
link discovery (e.g., place name recognition & georeferencing, …)
enrichment
Linked Data cloud

Figure 2: High-level overview of the annotation workflow.

5 person months.

Steps

1. Model the current annotation workflow
2. Establish connections to other libraries offering Linked Data
3. Identify vocabularies for mapping of existing metadata
4. Identify external data sources for link recommendation
5. Extend the workflow model with data integration steps (outgoing link generation)
6. Evaluation of extended workflow with library staff

Deliverables

D-4.1 Extended workflow model (e.g., in BPML)
D-4.2 Documentation of identified vocabularies
D-4.3 Conference paper about extended workflow and evaluation (e.g., KMIS)

3.2.6 WP 5: Annotation and link discovery tools

While WP 4 looks at annotation and integration at the level of the workflow, this work package covers the actual implementation of the corresponding tools. The toolset can be divided into two groups, one of them working in the background without any visible changes for the librarian during the cataloguing; the other one requiring user input, hence causing changes to the existing workflow. As discussed in WP 4, existing metadata can be mapped to properties from previously selected vocabularies in the background. The corresponding mapping tool has to be developed in a way that allows for an easy integration into different cataloguing environments. Moreover, the set of properties and vocabularies for mappings have to be kept flexible and easy to configure. Beyond the thematic metadata, the tool needs to support automatic mappings between different spatial reference systems, i.e. between the coordinates of the map at hand and WGS84 coordinates as widely used in the Linked Data cloud.

The integration tools aim at an easy-to use discovery of outgoing links to other potentially relevant data sources in the Linked Data cloud. This requires a selection of data sources to link to, such as DBpedia and GeoNames for that data in the focus of this project. Based on link discovery tools and semantic similarity measurement, the integration tool will then recommend resources from this predefined data pool to link to using *sameAs* and *seeAlso* links. From a user's perspective, the crucial aspect in the development of this tool is a non-obstructive integration into the existing cataloguing tools, so that the extra effort for the user is reduced to a minimum.

10 person months.

Steps

1. Analysis of existing cataloguing tools
2. Abstract software model for annotation and integration tools
3. Open source implementation of both tools
4. Usability testing

Deliverables

D-5.1 Open source mapping tool

D-5.2 Open source link discovery and annotation tool

D-5.3 Journal publication about workflows, toolset and evaluation (extended version of D-4.3; e.g., Semantic Web Journal, Transactions in GIS, B.I.T. online)

3.2.7 WP 6: Service encapsulations

Work packages 4 and 5 are concerned with improved accessibility of bibliographic contents based on spatio-temporal properties. WP 6 tackles the same problem for geo Web services. These services have been defined to improve the accessibility of geospatial information. The corresponding standards, however, are not supported outside of the community working with spatio-temporal information on a daily basis. This is mostly due to a level of complexity that is required for professional geo Web services (such as in spatial data infrastructures), but is seldom required for use cases where spatio-temporal is combined with data from other sources (such as in map mash-ups).

The goal of this work package is therefore to extend existing geo Web service specifications with additional service encapsulations that expose their contents for integration in the Linked Data cloud. We focus here on the Web Map Service (WMS, for raster data) and Web Feature Service (WFS, for vector data) specifications and apply the Linked Data principles to make their contents available for integration with the bibliographic contents exposed as Linked Data in WPs 4 and 5. Exposing the data offered by these services as Linked Data requires a decision on the granularity of the exposition: depending on the use case, one has to decide which is the most fine-grained level that receives URIs as identifiers. For example, in a WFS, one would in most cases want to be able to identify specific features via URIs, whereas a single point of a polygon would not receive its own URI.[28] The corresponding URI schemes as well as the encapsulation translating them to actual WMS and WFS queries, respectively, need to be kept flexible to be useable for any standards-compliant implementation of these service specifications. The goal is to allow users to discover data following the Linked Data principles, and then allow them to retrieve these data in the format that fits their workflow best. This can be either as an RDF representation, or as images or features encoded in GML as delivered by the encapsulated WMS or WFS.

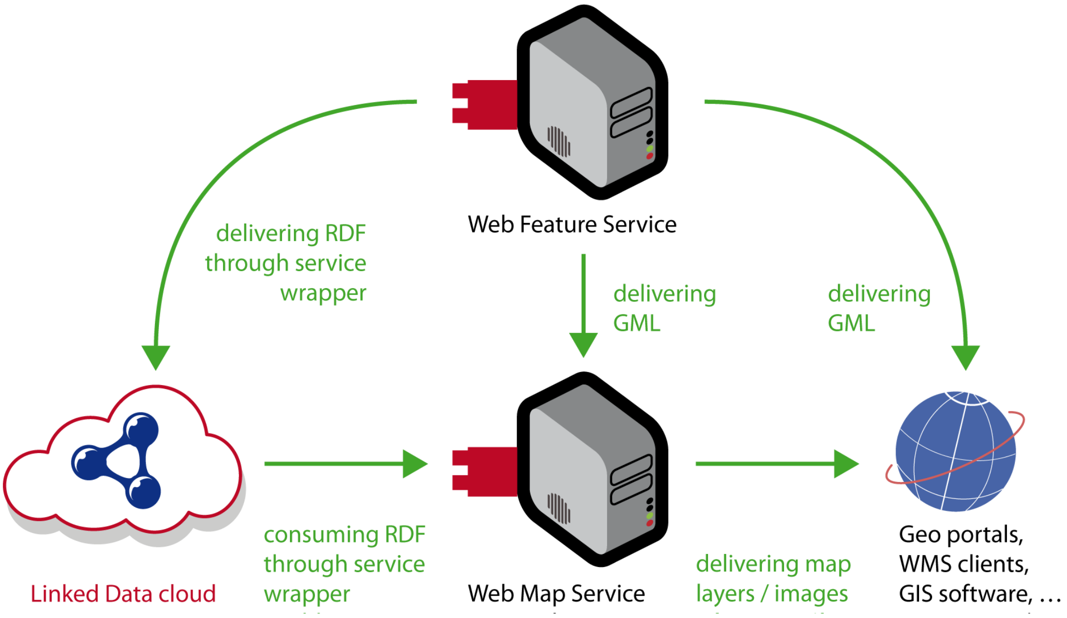
Web Feature Service
delivering RDF through service wrapper
delivering GML
delivering GML
Linked Data cloud
consuming RDF through service wrapper
Web Map Service
delivering map layers / images
Geo portals, WMS clients, GIS software, …

Figure 3: Extending OGC services with service wrappers for Linked Data consumption and delivery.

8 person months.

Steps

1. Decide on reference implementations for the WMS and WFS specifications to use during development
2. Develop URI schemes for the data gather from the three use cases

[28] See also the corresponding outcome oft he geometry working group at GeoVoCamp 2011: http://vocamp.org/wiki/Geometry-vocab

3. High-level model of the service encapsulation; this should be kept generic enough to cover both WMS and WFS, so that it can also be applied to other OGC service specifications later
4. Implementation of an abstract wrapper based on the high-level model
5. Extensions of the wrapper for the actual encapsulations of WMS and WFS

Deliverables

D-6.1 Documentation of URI mapping schemes
D-6.2 Open source wrapper and service encapsulations for WMS and WFS
D-6.3 Conference paper about service encapsulations (e.g., FOSS4G or AGILE)

3.2.8 WP 7: Retrieval interfaces

Retrieval interfaces for Linked Data, such as faceted browsers, focus on thematic relationships between resources and do not support spatio-temporal relationships in an intuitive way so far. In order to make the Linked Data produced in WPs 4 to 6 easy to discover for users, a novel user interface is required that makes use of these spatio-temporal relationships between local resources (library catalogues, historic map databases, cancer case databases, geo databases and services, etc.) and content from the Linked Data cloud. In particular, this novel interface needs to facilitate spatial queries using a map interface that also supports complex spatial constraints that are automatically mapped to GeoSPARQL queries in the background. These need to be combined with thematic and temporal constraints to exploit the full range of information available from the Linked Data representation.

The user interface development will need to go through several iterations with small usability tests in the library as well as the three use cases. In particular, the developed user interface needs to be generic enough to be used across different devices (including mobile devices) and must adhere to common guidelines for accessibility for people with impairments. In order to fulfill these requirements, a browser-based solution should be favored over stand-alone applications that require the user to install software.

8 person months.

Steps

1. Analyze existing user interfaces for Linked Data with a focus on query options for spatio-temporal properties
2. Explore different possibilities for combinations of map-based interfaces and interfaces for thematic queries
3. Usability testing
4. Repeat 2 and 3, compare
5. Finalize implementation based on usability tests

Deliverables

D-7.1 Open source software for user interface
D-7.2 Documentation of usability tests
D-7.3 Conference publication about user interface and usability testing (e.g., GeoViz)

3.2.9 Work phases

	Project months																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
WP 0 Project consolidation	■	■																						
WP 1 Use case “urban history”		■	■	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	■	■	■
WP 2 Use case “cancer”			■	■	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	■	■	■
WP 3 Use case “planetology”				■	■	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	■	■	■
WP 4 Workflow development				■	■	■	■	■																
WP 5 Tool development							■	■	■	■	■	■	■	■	■	■								
WP 6 Service encapsulations											■	■	■	■	■	■	■	■						
WP 7 Retrieval interfaces														■	■	■	■	■	■	■	■			