

ONTOLOGY BASED SEMANTIC DATA MANAGEMENT FOR PANDISCIPLINARY RESEARCH PROJECTS

M. Politze, B. Decker
IT Center RWTH Aachen University
(politze, decker)@itc.rwth-aachen.de

Abstract

ProjektRepository, initially funded by the German Research Foundation, is a web based pandisciplinary repository for research projects that shall become a central component of scientific cooperation in scientific projects at the university. It is developed by IT Center of RWTH Aachen University on basis of Microsoft SharePoint as a widespread standard product for web based communication and collaboration. The product in turn is extended by several features concerning the tagging and formal retrieval of data. These features make use of ontologies to define the structure of the repository.

Keywords: virtual research environment, e-science, semantic repositories, ontology

1. Introduction

Research today is increasingly interdisciplinary and it thrives on collaboration with partners at the home institution, at home and abroad as well as between universities, research institutions and industry. Research projects are therefore far beyond the classical model of bilateral cooperation and design themselves as virtual organizations. For their work researchers need a secure space which allows them to cross domain and organizational boundaries within the project context and to exchange and enrich research artefacts. In addition, mechanisms for crossing the project and organizational boundary are required to be able to build on other research results and to make their own results accessible to the public.

2. Project Goals

The main goal of ProjektRepository is to build a web based platform that offers a low-threshold service to share, store and retrieve research data among different groups of researchers from a variety of fields. This service is integrated into the IT infrastructure offered by RWTH Aachen University. Finally, once fully developed, this support for researchers in the field of e-science shall have an equivalent standing as e-learning applications that already exist at the University. BISCHOF et al. (2009) and DECKER et al. (2012) describe the general goals of the project and its placement at the university.

Our approach to accommodate the context of cooperative interdisciplinary research is to combine librarian methodologies for organizing und structuring artefacts like classification and taxonomies with semi structured approaches like tagging. The product should provide a work-flow that enables researchers of different disciplines to structure their research data and map it to an ontology. The ontologies themselves define the structure of the metadata stored and allow multiple enhancements to the currently available structures offered before. Thus the product offers a user interface to the semantic web. Search and collaboration can profit from the formal definition of the ontology. This can be extended to knowledge sharing in the semantic web and through semantic search engines.

The described product therefore offers the following features:

- An interface to store, retrieve and update the ontologies used in the different projects.
- A functionality that adds the structure provided by the ontology to an existing or a new repository.
- Ways of retrieving data using the structural information provided by the ontology.
- An interface that allows multiple repositories to exchange data and metadata
- Enrich existing data repositories with structured metadata.

3. Metadata Definition

The first step to set up a semantic repository is to formally define the structure of the metadata that is used to describe the stored objects of a certain domain. This structure is defined by the means of an ontology. It is important that the researchers are not left alone while creating the structure for their repository. For ProjektRepository a process was established that supports the researchers during the initial set up of their repository. Figure 1 shows an outline of this creation process. In first step the researcher has to formally define the structure of the metadata for the specific domain of research. Despite from creating a new ontology for each and every repository it is highly recommended to adopt already existing ontologies. To support this process a knowledge engineer is needed who has an overview of existing metadata standards in a variety of domains. Due to their experience with different metadata schemas the university library provides this support for ProjektRepository. In the second step the ontology is imported into the repository. After checking the ontology for logical flaws or inconsistencies the repository is instantiated. This process is currently supported by the IT Center to guarantee the successful creation of complex structures. However it is possible to provide this functionality as a self-service functionality to the researchers.

Until now ProjektRepository has been used with a variety of custom and predefined ontologies such as: Dublin Core (POWELL et al. 2007), INSPIRE (European Commission 2008), ICD 10, ICD O 3 (World Health Organization 2010), CIDOC CRM (International Organization for Standardization (ISO) 2006) and LIDO (International Council of Museums 2010)

In terms of the ontology artefacts stored in the repository are seen as individuals. They are therefore subject to some of the properties defined in the ontology. The basis for the formal tagging process are the classes defined in the ontology. To define relations among the instances object properties are used. Based on the domain and range defined in the ontology object properties are restricted to a limited number of instances. Further description of the instances can be done by using data properties. These properties assign a concrete value, like a text or a number to an instance. The full instantiation process of a semantic repository based on an ontology is shown by POLITZE (2012).

Suppose the repository contains an artefact described by the URI 'http://example.com/file'. The user wants to assign the value 'Sample File' to the metadata property 'Title'. Since 'Title' is essentially a free text field it is defined as a data property in the ontology. Thus by assigning this value to the metadata property the assertions in the ontology (A) are extended such that

$$A = \{ \dots, (http://example.com/file, "Sample File"):Title \}$$

One of the most common application of metadata is to assign values from thesauri or subject catalogues. In an ontology every possible value is then defined as an individual of a certain class. This individual is then connected to the individual representing the data in the ontology using an object property. Regarding the previous example the user wants to assign a value to the 'has_Datatype' metadata property. The possible values for 'has_Datatype' are defined in a catalogue containing the values 'Raw', 'Preprocessed' and 'Result'. In the ontology this catalogue also extends the set of assertions

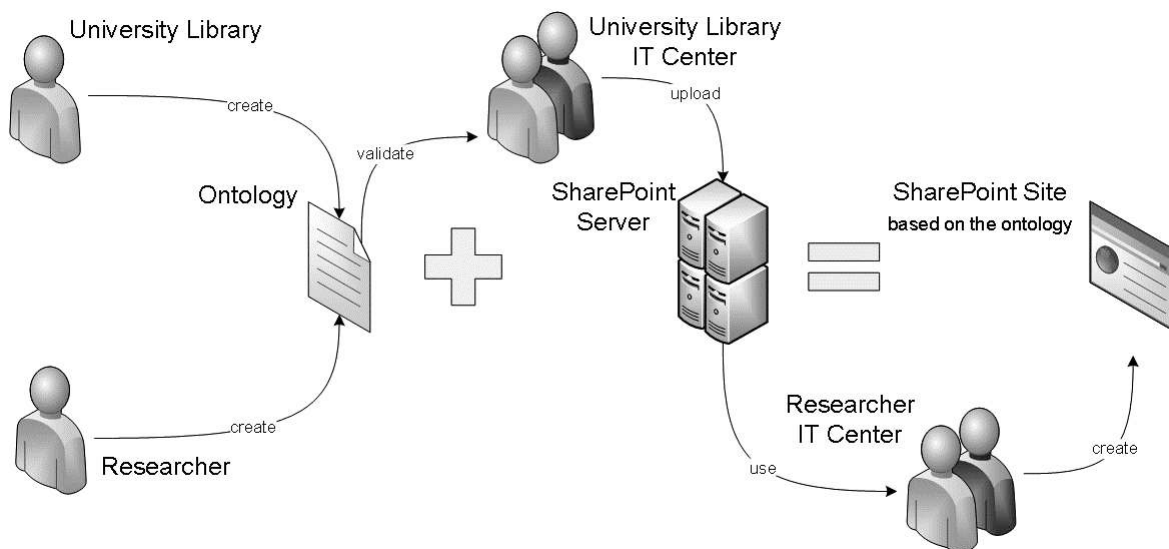


Fig. 1: Creation process a semantic research repository

$$A = \{ \dots, (Raw, Datatype):Type, (Preprocessed, Datatype):Type, (Result, Datatype):Type \}$$

Once a values is assigned to the ‘has_Datatype’ metadata property is also added to the set of assertions in the ontology.

$$A = \{ \dots, (http://example.com/file, Raw):has_Datatype \}$$

Apart from this traditional way of tagging multiple individuals representing data artefacts may also be connected by using object properties. This allows to describe relationships between stored artefacts. Continuing the example above the user adds a file identified by the URI ‘http://example.com/file2’ that is the result of some computation that used the first file. The user can add a metadata property ‘is_result_of’ that describes this relationship between files. Again the assertions of the ontology are extended to store these information such that

$$A = \{ \dots, (http://example.com/file2, http://example.com/file):is_result_of \}$$

4. Semantic User Interface

The integral part of the repository is the interface that the researchers use to add the semantic information to their data. This interface hides all technical details about the definition of the metadata and their structure and its definition in the ontology. It gives the user easy access to some features of the ontology such as access to thesauri and subject catalogues but also to an ontology enhanced search.

While Figure 2 displays an overview of stored artefacts on the left, in the right image the user interface is shown that is used to edit metadata properties. It is rendered as a simple HTML form with the different fields being generated from the object and data properties defined in the CIDOC CRM ontology. While data properties are rendered as simple text fields object properties offer more enhanced features like auto completion while entering values for the property. The possible values for auto completion are generated from individuals from the initial ontology and already existing individuals in the repository that match the range of the object property. In terms of metadata this allows to dynamically extend the set of possible subjects in a catalogue using class relations and reasoning features offered by ontologies.

Of course not only the creation of metadata is supported by the system but also the retrieval of already tagged artefacts was enhanced using the information from the ontology. The search interface can thus be used to formulate complex queries against the stored individuals. Advanced reasoning can then be used to retrieve individuals using the structural information from the ontology. The user interface depicted in Figure 3 provides simple access to some of the reasoning features and helps users to define their queries without the need of deeper knowledge about ontologies. As in the case of the edit form also the search form provides auto correction when searching in object property fields. More advanced search features can be acquired using semantic search engines like SWOOGLE or WATSON by making the defined ontology and the tagged individuals available to these services using the ontology export feature explained in section 7.

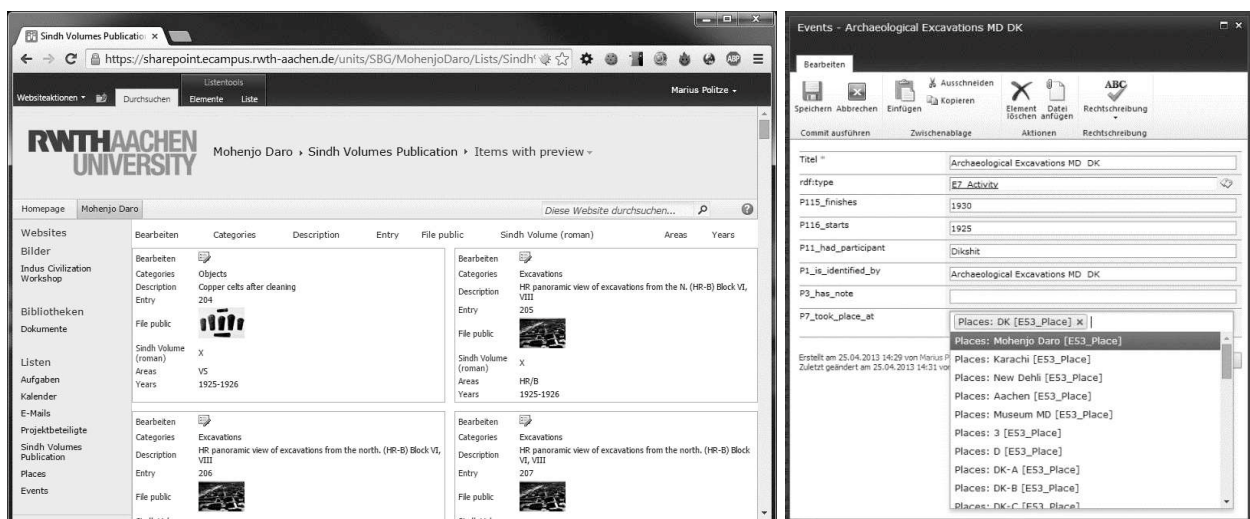


Fig. 2: User interface showing some metadata fields in display mode with a preview image (left) and metadata fields derived from the CIDOC CRM ontology in edit mode (right)

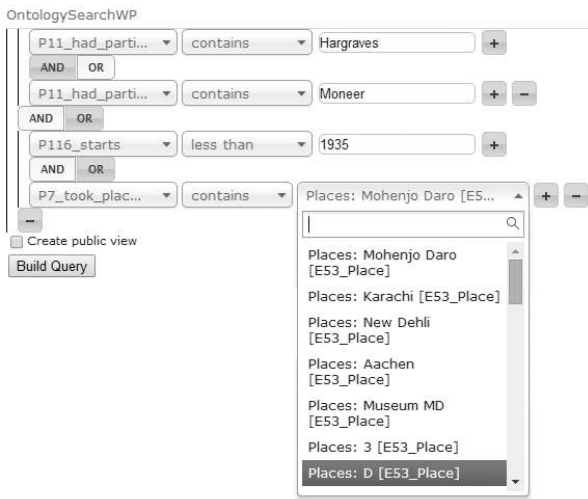


Fig. 3: Ontology enhanced search interface

5. Automated Retrieval of Metadata

Instead of entering the metadata manually into the system some of the metadata can be retrieved automatically from the files. The atomization of metadata retrieval from certain file types can greatly enhance the overall quality of metadata as well as the users' acceptance of the system since less metadata has to be entered manually in the system. Depending on the file type this process is especially useful for technical metadata such as image resolution, number of pages in a document or the modification date.

Even though this process is very valuable it requires a specialized implementation to retrieve technical from certain file types. Essentially it requires two steps: (1) The technical means to read and parse the file format need to be implemented in the repository and (2) the metadata fields available in the file type need to be mapped to the properties for the semantic description. While the first step is mostly a technical issue, the second steps requires an in depth analysis of the supplied metadata in the file type as well as a specific mapping to the ontology used in the semantic repository.

For ProjektRepository this automated retrieval of metadata was implemented for two file formats: Jpeg and ShapeFiles. Figure 4 shows an example of automatic extraction of metadata from a ShapeFile that is commonly used in geo sciences. Apart from some technical metadata like title and the geographic extend the system generates a low resolution preview image of the file that can be viewed in the browser. Both instances of this single artefact are then linked together using specific object properties.

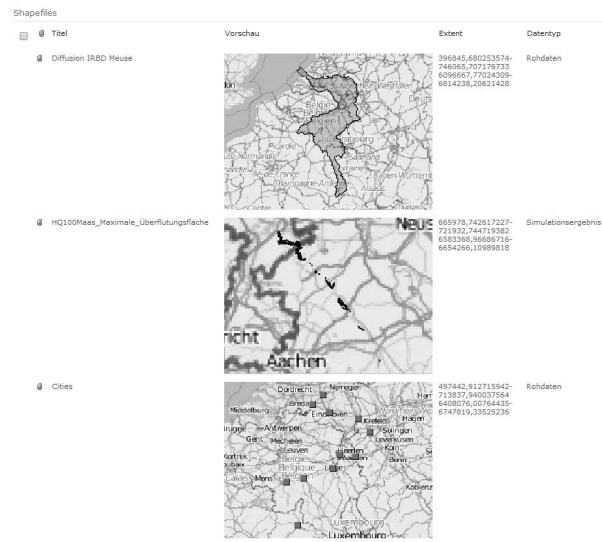


Fig. 4: Automatically retrieved metadata (title, preview and extend) and preview image of a ShapeFile

6. Repository Proxy

In many domains research data is already stored in existing repositories that meet very specific needs of the researchers of this domain. These needs may be the handling of specific file types, very large or a vast amount of files, a user interface that works according to some domain standard or the ability to access the data with domain specific tools. A repository spanning multiple disciplines and domains cannot always satisfy all these needs especially as they may be contradictory. On the other hand these domain specific systems often lack support of defining a metadata structure as well as semantic description of the stored data.

The repository proxy allows to propagate the features offered by ProjektRepository to remote repositories and thus gives the user the possibility to add semantic metadata but store the actual data in the domain specific repository. However this is not limited to metadata but also gives user the possibility to benefit from the role based access control and collaboration and sharing features. As displayed in Figure 5 the repository proxy will only store the metadata in the semantic system. To avoid copying the original files they will remain in the domain specific repository.

Like the automatic retrieval of metadata this is a specialized process that is tailored to certain domain specific repositories. This is especially the case when some metadata is already stored in the domain specific repository and therefore has to be manually mapped to the semantic metadata model. Nevertheless most modern data repositories offer machine readable interfaces that allow to programmatically link them to the semantic repository.

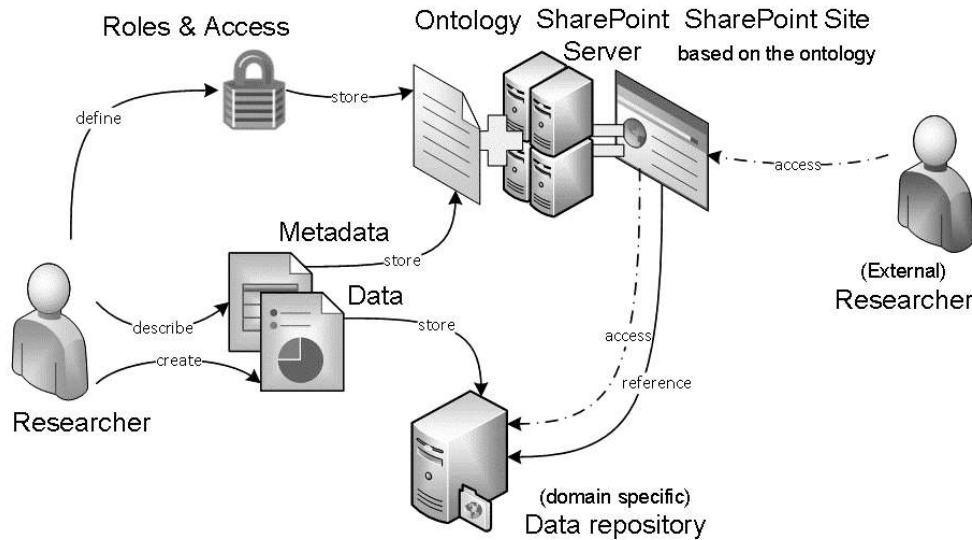


Fig. 5: Overview of the Repository Proxy functionalities

tory.

If the domain specific system is capable of semantic description of the data this process can even be completely automated and thus can be done as needed by the researchers themselves. This behaviour when linking two semantic systems is done by exporting the describing ontology and then importing it into the proxy system.

7. Ontology Export

Reuse, sharing and linking of repositories is one of the most important steps towards building a good and reliable knowledge base within the home institution as well as across multiple organisations which in turn forms the basis of modern interdisciplinary research. Figure 6 shows the different scenarios that become possible using the ontology export that contains the structure, the stored data and metadata within the repository.

To allow an easy integration with other semantic systems such as semantic search engines but also other semantic data repositories the metadata and references to the stored artefacts can be exported as an ontology in the OWL file format and can then be imported by other systems. Using the W3C standard OWL (W3C OWL Working Group 2009) as an export format promises compatibility to current and future semantic systems. Of course the ontology export can also be used to create new or link existing repositories within ProjektRepository.

Even though artefacts and their description need to be included in the ontology export they are not embedded directly into the export file. This is due to the fact that there may be a large number of files each several hundreds of megabytes in size and would thus make the export hard to transfer between remote systems. Instead of embedding the artefact completely in the export it is referenced by a URL within the source system that allows to access the file contents. Accessing the file with

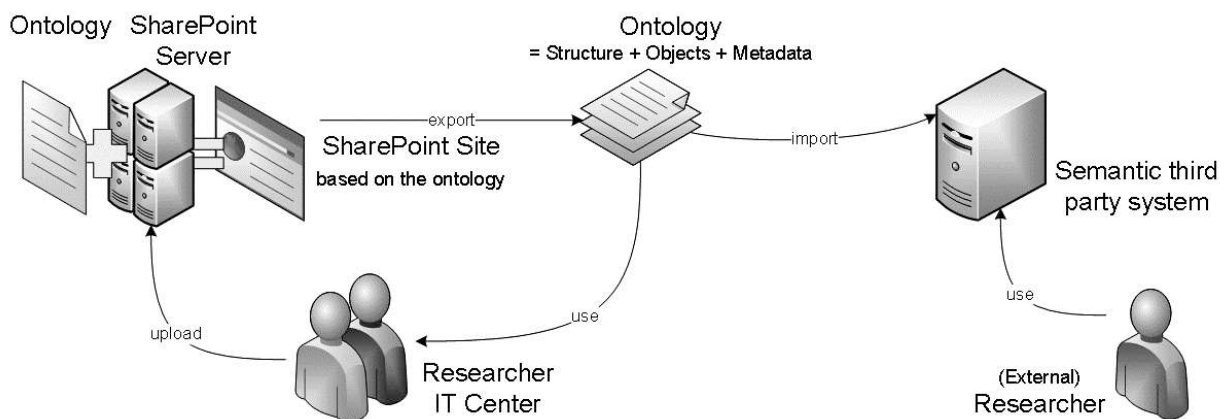


Fig. 6: Usage scenarios of the ontology export

the URL in ProjektRepository also allows the system to authenticate the user and apply role based access control. Access rights to the actual file contents can thus be divided and allows to browse files by metadata without transferring the actual file.

8. Future Work

The intermediate versions of the product are continuously integrated into the system from the start of the project. The product is designed in cooperation with four groups of researchers of RWTH Aachen University: the University Library, the Institute of Pathology, the Department of History of Urbanization and the Institute of Hydraulic Engineering and Water Resources Management.

The current version will be further improved to provide a low-threshold access method to create and maintain semantic data for a variety of researchers. Our future challenges before finally introducing the product are:

- Build a more seamless integration of ontology creation and handling in the user interface
- Provide a lower-threshold to access the semantic search and tools.
- Raise usability and acceptance to implement it for more research groups at RWTH Aachen University.
- Further integration in the existing IT infrastructure.

References

- BISCHOF, C., EICH, U., KNÜCHEL-CLARKE, R., JANSEN, M., SCHÜTTRUMPF, H. (2009): ProjektRepository ein pandisziplinäres Repository für Forschungsprojekte als Komponente einer niederschweligen webbasierten Kooperationsinfrastruktur, German Research Foundation grant application, Aachen, Germany.
- European Commission (2008): Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata (Text with EEA relevance)
- DECKER, B. & GEBHARDT, M. (2003): ProjectRepository – From Heap to Hoard, EUNIS Congress 2012, June 20-22, Villa Real, Portugal.
- POWELL, A., NILSSON, M. NAEVE, A. JONSTON, P., BAKER, T (2007): DCMI Abstract Model. <http://dublincore.org/documents/abstract-model>. 2014-10-30.
- International Council of Museums (2010): LIDO (Lightweight Information Describing Objects): Making it easier to deliver information to portals. <http://www.lido-schema.org/documents/LIDO-Handout.pdf>. 2014-10-30.

- International Organization for Standardization (ISO) (2006): ISO 21127:2006 Information and documentation - A reference ontology for the interchange of cultural heritage information.
- POLITZE, M. (2012): Automated Ontology Mapping of Tagged Data in a Pandisciplinary Repository for Research Projects, master thesis, Maastricht University.
- W3C OWL Working Group (2009): OWL 2 Web Ontology Language Document Overview (Second Edition). <http://www.w3.org/TR/2012/REC-owl2-overview-20121211>. 2014-10-30.
- World Health Organization (2010): International Statistical Classification of Diseases and Related Health Problems – Malta.

Contact information

Marius Politze
RWTH Aachen University
IT Center
Seffenter Weg 23
52074 Aachen
Germany
politze@itc.rwth-aachen.de
+49 241 80 29720