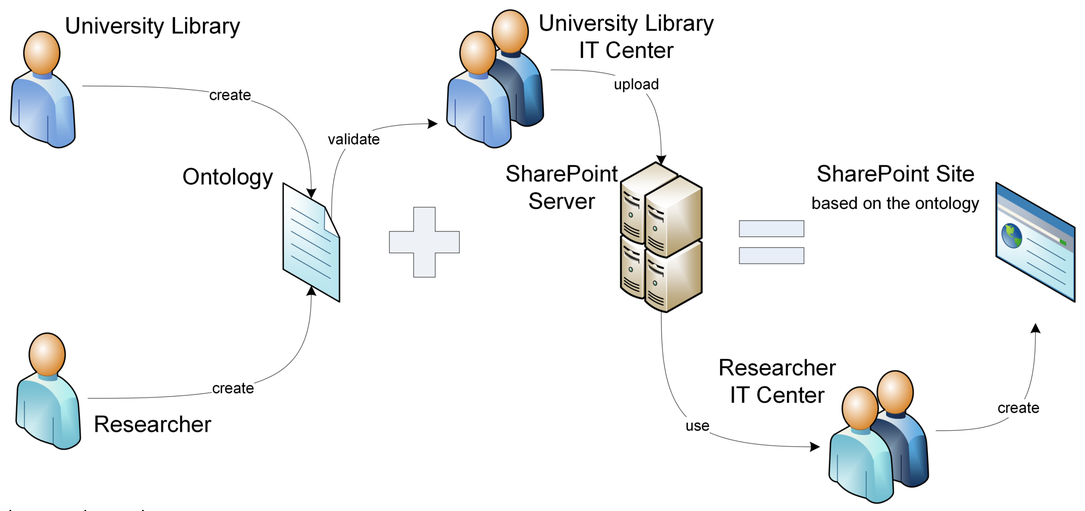
Ontology based semantic data management for pandisciplinary research projects

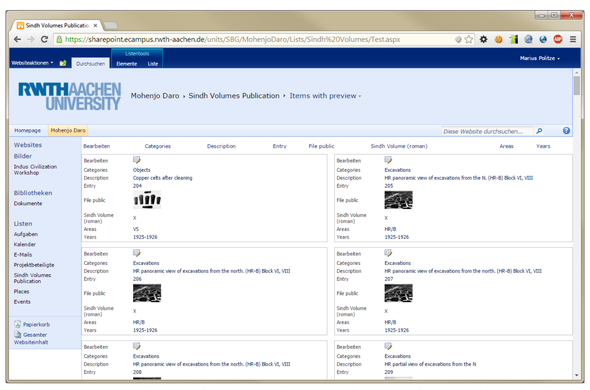
Introduction

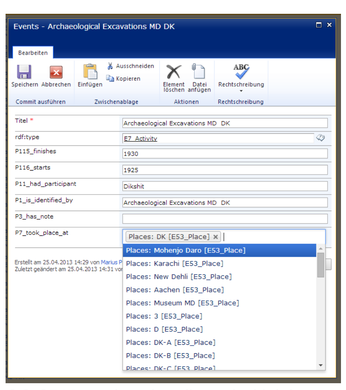
ProjektRepository, initially funded by the German Research Foundation, is a web based pandisciplinary repository for research projects that shall become a central component of cooperation in scientific projects. It is developed by IT Center of RWTH Aachen University on basis of Microsoft SharePoint as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

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 this support for researchers in the field of e-science shall have an equivalent standing as already existing e-learning applications.



Creation Process of a semantic research repository



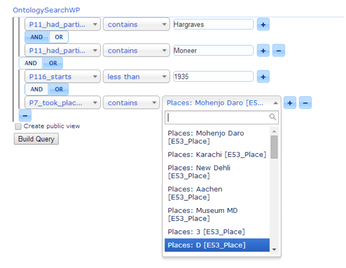


Different Semantic user interfaces: Overview of tagged items (left), metadata fields derived from the CIDOC CRM ontology in edit mode (right)

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 thesauri and subject catalogues.

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.

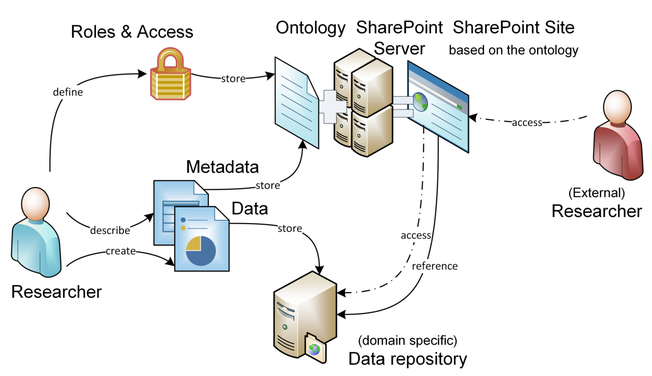


Ontology based search interface

Connecting Non-Semantic Systems: Repository Proxy

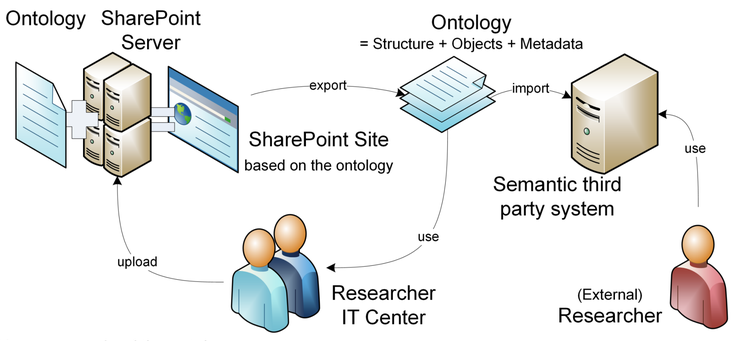
The repository proxy allows to propagate the features offered by ProjektRepository to remote repositories and thus gives the user the possibility to add semantic metadata but store the actual data in the domain specific repository. However this is not limited to metadata but also gives user the possibility to benefit from the role based access control and collaboration and sharing features. 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.



Overview of the Repository Proxy functionalities

Connecting Semantic Systems: Ontology Export



Usage scenario of the 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 organizations which in turn forms the basis of modern interdisciplinary research.

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.

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 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.

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 automation 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. 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:

1. technical means to read and parse the file format
2. metadata fields available in the file type need to be mapped to 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. 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.

Automatically retrieved metadata (title, preview and extend) and preview image of a ShapeFile

M. Politze, B. Decker

Telefon: +49 241 80-29720
E-Mail: politze@itc.rwth-aachen.de

IT Center
RWTH Aachen University
Seffenter Weg 23, 52074 Aachen, GERMANY
www.itc.rwth-aachen.de
Telefon: +49 241 80-29720
Fax: +49 241 80-22134