

Networking Social Science Resources on the Web: SozioNet

Natascha Schumann¹, Wolfgang Meier², Rudi Schmiede³

^{1 2 3}Darmstadt University of Technology, Department of Sociology, Residenzschloss, D-64283 Darmstadt, Germany

[schumann, meier, schmiede]@ifs.tu-darmstadt.de

Abstract

SozioNet¹ is part of a developing internet portal for the social sciences within the framework of infoconnex², the German information network for Education, Social Science and Psychology. SozioNet provides access to scientific resources freely available on the World Wide Web. Social scientists, researchers, students and others are offered fast and easy access to resources located at institutional servers spread across the web. SozioNet is inspired by successful examples like MathNet³ or SOSIG.⁴

The objectives of SozioNet are not only the integration of freely available scientific resources via the internet but also the creation of a network of social science institutions and the improvement of search facilities by annotating resources with metadata.

The institutions supporting SozioNet agree in using a common metadata schema for the annotation of their documents. We use Dublin Core to describe the resources generally.

For the description of the main topics we utilize the classification and the thesaurus of the German Social Science Information Centre.

Introduction

The project SozioNet is part of the telos⁵ working group located at the Department of Sociology at Darmstadt University of Technology. It is funded by the German Ministry of Education and Research and was launched in spring 2002.

SozioNet is part of a forthcoming national social science information portal within the framework of the German infoconnex initiative. Additionally to existing services offered by

¹ <http://www.sozionet.org>

² <http://www.infoconnex.de>

³ <http://www.mathnet.org>

⁴ <http://www.sosic.ac.uk>

⁵ telos = technology for electronic libraries and the organisation of the Semantic Web

the German Social Science Information Centre⁶ and the Virtual Library of Sociology⁷ SozioNet provides access to freely available social science web resources.

SozioNet is based on the principle of self-organization of social scientists and social science institutions. The participants choose relevant resources, create semantically rich metadata and make these accessible to a broader public.

SozioNet focuses on the registration of relevant resources and information which are distributed among servers of different institutions. These resources have not yet been systematically documented.

SozioNet implements a general infrastructure for the creation of semantically rich metadata, and for the harvesting and retrieval of relevant resources with a domain specific focus.

Many social science institutions have started to make their resources available on the web. They offer increasingly research papers, presentations, lecture notes, articles, masters thesis etc. on their websites. Individual scientists also publish their work results. But it is not easy to find these resources as most of them are hidden on faculty servers, project pages or individual homepages. They all implement their own specific site structure, web design and search facilities, so it may take users a long time to find the information they are looking for. Also, common web-wide search engines are usually based on the algorithmic processing of arbitrary contents and fail to address the needs of a given scientific community.

A digital library that focuses on a specific scientific domain can help to improve this situation, given that the participants involved agree on common standards with respect to formal requirements, metadata sets, metadata quality and classification rules.

Participants

Currently, eleven institutions participate in SozioNet. These include social science faculties as well as independent research institutes not directly associated with an university. While some of them have just started to make their resources available, there are also many institutions which have been publishing larger volumes of social science materials for a much longer period of time. So the requirements differ a lot with respect to metadata creation and accessibility. Those institutions who have not yet implemented a workflow for dealing with electronic resources need support for creating metadata. In SozioNet, we have thus developed a reusable interface for metadata editing, called MetaWizard. This tool can be used without knowledge of the underlying metadata schema or markup languages we use.

Institutions that have already established their own publishing procedures and content management system have to add special interfaces so that the SozioNet harvester may collect their data. Initially, we proposed to collect metadata from our partner institutions via the Open Archives Initiative (OAI) Protocol. However, despite the availability of already existing software tools, most of the institutions preferred to set up a simpler services based on HTTP.

The institutions themselves choose the relevant resources they want to make available through SozioNet. All materials remain under the control of the publishing institution or

⁶ <http://www.gesis.org/IZ>

⁷ <http://www.vibsoz.de>

individual. They are also responsible for keeping the resources available and are also responsible for generating high-quality metadata according to the SozioNet metadata schema. The resources, their maintenance and the rights remain with the publishing institution.

Members of the participant institutions have actively been involved in the development and in the evaluation of the SozioNet metadata schema and the SozioNet tools. Two workshops were organized to discuss problems and new ideas. Furthermore there was an active exchange of information via email.

The following institutions are involved in SozioNet:

- The German Youth Institute, Munich
- Freiburg Institute for applied Social Science
- Friedrich-Alexander-University Erlangen-Nürnberg, Department of Sociology
- University of Göttingen, Department of Sociology
- Institute for Social Scientific Research, Munich
- International University Bremen, Information Research Center
- Mannheim Centre for European Social Research
- Ruhr-University of Bochum, Faculty of Social Science
- Sozialforschungsstelle Dortmund
- University of Dortmund, Business and Social Science Faculty
- Social Science Research Center Berlin

There are some other institutions that are interested in using the SozioNet infrastructure, for example the Institute for Technology Assessment and Systems Analysis, Karlsruhe and the project "Buddhist stone inscriptions" by the Heidelberg Academy for the Humanities and Sciences.

Architecture

The self-organization of social science institutions and scientists plays an important role in the SozioNet concept. Storing resources to a centralized server system would clearly contradict this principle. All materials will thus remain under the control of the publishing institution or individual, who are responsible for making the resource available and are also responsible for the generation of high-quality metadata, using the standards established in the project.

As indicated in figure I, SozioNet provides web-based tools for metadata creation. SozioNet gathers the available resources through a harvesting component. The harvester will periodically scan through all web-addresses known to the system, extracting metadata and indexing the fulltext-content of the resource.

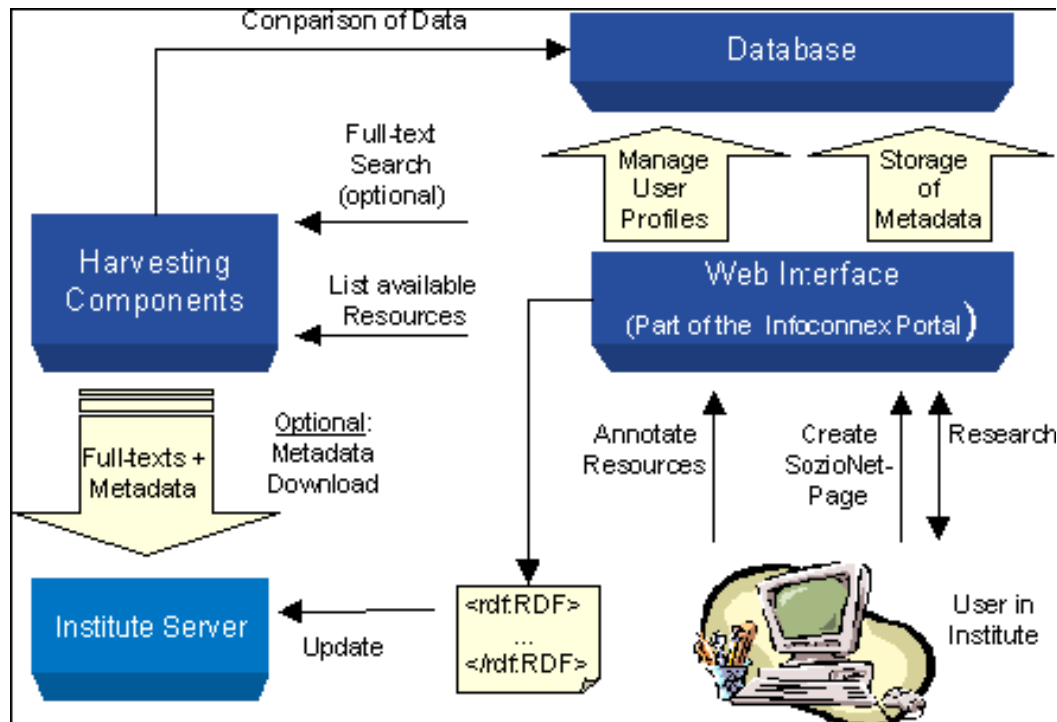


Figure I

Metadata

Due to the overall increase in electronic resources during the last years it becomes more and more difficult to separate relevant documents from those that are not relevant in a specific context. A service that focusses on a specific scientific domain can help to improve this situation, given that the players involved agree on common standards with respect to formal requirements, metadata sets, metadata quality and classification rules. In particular, implementing common metadata standards is a central prerequisite for a digital library to a given discipline. Researchers and students should be able to browse and search resources by domain specific categories and concepts taken from established classifications or thesauri.

In SozioNet we use the RDF resource Description Framework as foundation for metadata interchange. RDF provides a simple language for expressing metadata and is used to embed structured metadata into documents. RDF just defines a basic model and a serialization syntax, but it does not specify a vocabulary.

RDF allows to integrate general metadata, for example Dublin Core as well as additional aspects, which are specific to the social science domain.

Metadata schemes

To ensure a sufficient quality of the service within a specific scientific context, metadata has to be added to the resources. Requirements for a common metadata scheme are not only to focus on the domain specific context of the social science but also to consider international standards.

The SozioNet metadata scheme is backed by an ontology. The ontology is defined in OWL (Web Ontology Language).

All common metadata fields, like author, title etc. were taken from qualified Dublin Core. To describe domain specific aspects, the classification and thesaurus of the social science were included into the ontology. Both are well established in German social science and are constantly maintained and enhanced by the Social Science Information Centre in Bonn, Germany. The thesaurus and the classification are also defined as OWL ontologies.

The SozioNet metadata scheme defines a shared vocabulary for the SozioNet domain, containing concepts which are not covered by simple Dublin Core or other common schemas. Such domain specific schemes are defined in the SozioNet namespace. It defines the general class “publication”, which is refined in subclasses. The following list shows all subclasses from “publication”. The prefix “sn” indicates, that an element belongs to the SozioNet scheme:

- sn:ResearchPaper
- sn:WebPage
- sn:Presentation
- sn:InCollection
- sn:Collection
- sn:Book
- sn:InBook
- sn:Dissertation
- sn:MastersThesis
- sn:Thesis
- sn:LectureNotes
- sn:DataSet

The Dublin Core elements used in SozioNet are as follows:

- dc:title
- dcq:alternative
- dcq:abstract
- dc:creator
- dc:publisher

- dc:language
- dcq:IMT
- dcq:created
- dcq:modified
- dcq:isPartOf
- dcq:isFormatOf
- dc:subject

Using the element "dc:subject" allows to choose one or more terms from the classification, one or more terms from the thesaurus and if necessary to add one or more free chosen keywords. The example in figure II explicates how the general and the domain specific parts are integrated into the SozioNet metadata scheme. The first one is a reference to the classification. The notation 10200 indicates the term "sociology". The second one is a reference to the thesaurus and the last one is a free chosen keyword.

```
<dc:subject rdf:resource="http://www.sozionet.org/1.0/classification#10200"/>
<dc:subject rdf:resource="http://www.sozionet.org/1.0/thesaurus#1200"/>
<dc:subject>sociology</dc:subject>
```

Figure II

The indexing through the classification and the thesaurus is a requirement in SozioNet. Because of the fact that the participants have to index their resources themselves it is necessary to share a common vocabulary that is well known in the community.

Both, the thesaurus and the classification are embedded into the MetaWizard. The way they are integrated and the easy way to handle with contributes to the fact that the participants make active use of it.

Metadata creation

For those participants not familiar with metadata creation the so called MetaWizard was developed. Users do not require any knowledge of metadata formats and only minimal knowledge of the underlying standard. The MetaWizard is based on the XForms W3C standard. The interface is build in XQuery. The MetaWizard is mainly intended for institutions, which have not yet established their own publishing procedures or for individual authors, who would like to create metadata for a limited set of materials.

The MetaWizard offers a personalized user interface. Each user has a home collection, containing the metadata records he created. It is also possible to create different folders. Records entered through previous sessions can be revised or removed at any time. The user can search and browse through his own collection.

The entered metadata records are stored in a database. SozioNet uses eXist⁸, an open source native XML database system. When creating new metadata records, the wizard will try to retrieve the web page at the location specified by the user. If the page is readable, it will be passed to a summarizer to extract existing metadata. The extracted data is then filled into the following web forms and reviewed by the user. Figure II shows one of the web forms.

| | |
|--|---|
| Titel Der Titel des Dokuments (max. 75 Zeichen) | <input type="text"/> |
| Titelzusätze z.B. Untertitel, Kurzfassung, Abkürzung oder Übersetzung des Titels | <input type="text"/> |
| Abstract Kurze Beschreibung des Inhalts des Dokuments | <input type="text" value="SozioNet ist Teil eines im Rahmen von Infoconnex entstehenden Fachportals Sozialwissenschaften, welches die vorhandenen und zukünftigen Informationsdienstleistungen des Faches bündeln wird. Ergänzend zu den bereits bestehenden Informationsangeboten des"/> |
| Autor(en) Format: Nachname, Vorname(n). Zur Eingabe mehrerer Autoren bitte auf das "+"-Symbol klicken. | <input type="text"/> |
| Herausgeber Die fuer die Veröffentlichung des Dokuments verantwortliche Person oder Institution | <input type="text"/> |
| Entnommen aus Ist das Dokument die elektronische Kopie eines Originals, tragen Sie hier bitte Angaben zum Original ein | <input type="text"/> |
| Erstellungsdatum Erstellungsdatum des Dokuments. Bitte Datumsangaben immer in der engl. Standard-Form: YYYY-MM-DD oder nur das Jahr (z.B. 2004) eintragen, falls | <input type="text" value="2003-06-02"/> |

Figure III

After completing a the web forms and reviewing by the user, a metadata record will be created automatically. The user should store this record on the institutes server.

Metadata Harvesting

Harvesting web resources is supported by quite a number of different software tools. For example, the open source Harvest software⁹ is widespread and used in many projects, including MathNet and PhysNet.

Through Harvest is highly configurable, SozioNet has slightly different requirements. First, all metadata will be encoded in RDF. The harvester should thus be able to extract RDF metadata records and should maintain the basic structure of the RDF metadata. Second the harvester should recognize XML documents and pass them directly to the XML-enabled database to preserve the structural information contained in the XML. While the second requirement can be easilier met by XML-aware summarizers, the first is not so easy to deal with. Harvest stores metadata internally in a format called SOIF. It is a simple and hierarchical format. Converting RDF to SOIF thus implies a possible loss of information and an undesirable reduction of complexity. We have thus decided to implement our own harvesting component. It is based on a varity of freely available open source tools and

⁸ <http://exist-db.org>

⁹ <http://harvest.sourceforge.net>

backend by a simple, yet powerful component model. The harvester is entirely based on XML and related standards as, for example, SAX (the Simple API for XML).

The basic idea behind the software could be described as follows: take the core paradigm of Apache's Cocoon and apply it to a harvesting scenario. The core paradigm of Cocoon is the pipeline. Each pipeline starts with a generator, producing an XML stream, which is passed through an arbitrary number of transformation steps. At the end of the cascaded pipeline, a serializer writes the generated XML stream to whatever output format is desired, e.g. HTML or PDF.

Now, instead of transforming the source XML into the desired output format, the SozioNet harvester does it the other way around, i.e. the input could be a PDF, Word, HTML or XML document, but the output will always be XML. This scenario is shown in Figure II.

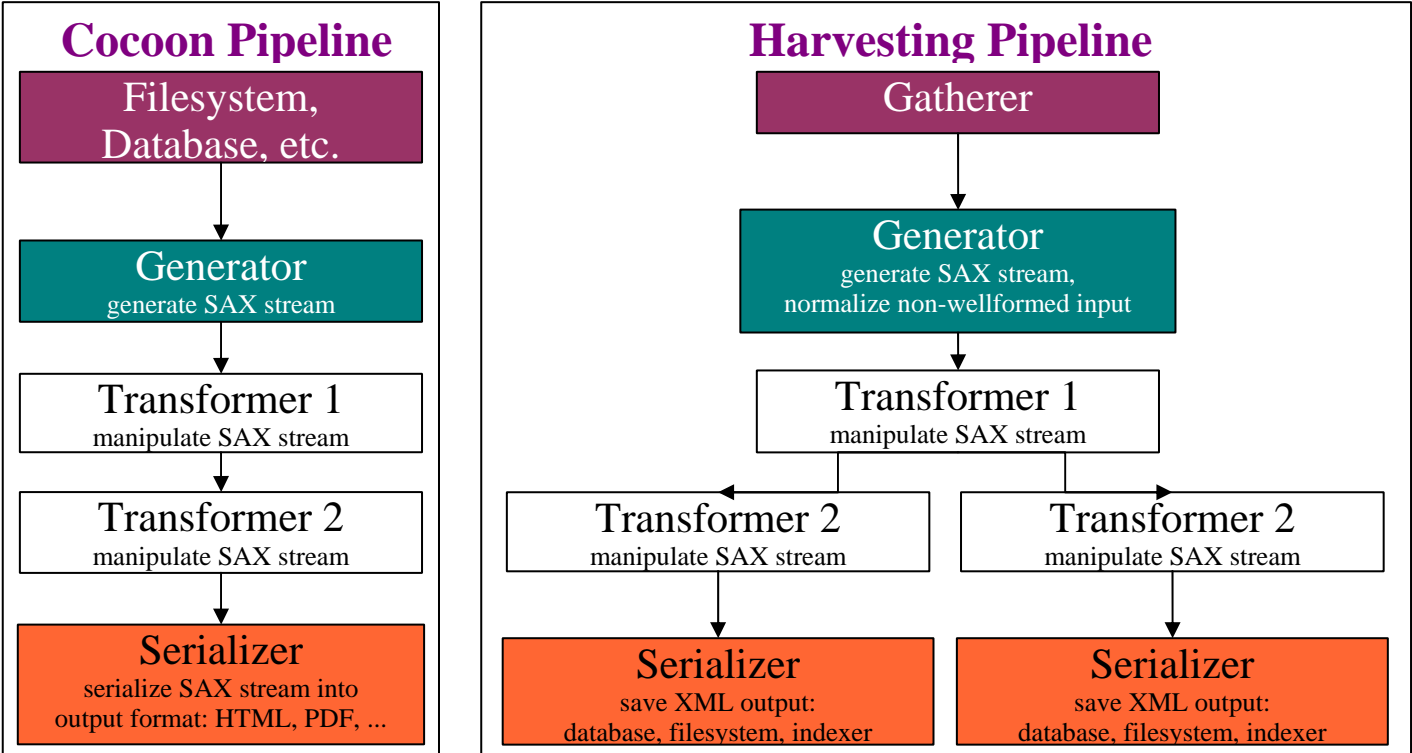


Figure IV

Conclusion

SozioNet provides a general infrastructure for the creation of semantically rich metadata, and for the harvesting and retrieval of free available social science resources. The SozioNet project provides state-of-the-art solutions using current and future web standards, ensuring that the collected resources can be maintained and accessed in the future.

The institutes involved in SozioNet are very engaged to develop and evaluate the service. Even in an international context, there was lively interest in SozioNet. It shows, that there is a general interest in using XML-based standards within the social science.