SozioNet: networking social science resources

Schmiede, Rudi; Meier, Wolfgang; Heise, Sue; Schumann, Natascha

Preprint / Preprint
Sammelwerksbeitrag / collection article

SozioNet: Networking Social Science Resources

Wolfgang Meier, Natascha Schumann, Sue Heise, and Rudi Schmiede

Department of Sociology,
Darmstadt University of Technology,
Darmstadt, Germany,
sozionet@ifs.tu-darmstadt.de

Abstract. SozioNet[1] forms part of a forthcoming national social science information portal, which is currently being developed by the German Infoconnex[2] initiative. Inspired by successful examples like MathNet[3] or SOSIG[4], SozioNet provides access to freely available web resources with relevance to social science. It is based on a network of social science institutions and scientists, to agree on and establish common metadata standards. 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.

1 Introduction

Infoconnex represents a joint effort to integrate existing information services and ongoing activities related to education, social science and psychology. In addition to the other activities coordinated by the initiative, SozioNet concentrates on freely available web resources with relevance to German social science. It thus complements the already existing information services available to social scientists, including, for example, the Virtual Library of Sociology[5]. The project is funded by the German Federal Ministry of Education and Research and was launched in spring 2002. However, the basic idea for the project already goes back to 1998.

SozioNet has been greatly inspired by such successful projects as MathNet, PhysNet and SOSIG. Although obliged to its predecessors, SozioNet differs from them in several aspects, concerning basic concepts, technology and standards, for example, by the use of ontologies and current semantic web standards like DAML [12] or OWL [4]. This paper presents an overview of the project, the metadata schemes and the standards used in SozioNet. We then introduce the general infrastructure for metadata creation and resource harvesting. Finally, the paper takes a closer look on some of the innovative aspects implemented in two of the core components: the user-interface for metadata creation and the harvesting component. The user-interface offers a personalized wizard for metadata creation with distinguishing features like automatic summarization of existing metadata, and the management of metadata records in persistent collections. It is

[1] http://www.sozionet.org
[2] http://www.infoconnex.de
[3] http://www.mathnet.org
[4] http://www.sosig.ac.uk
[5] http://www.vibsoz.de

T. Koch and I.T. Sølvberg (Eds.): ECDL 2003, LNCS 2769, pp. 245–256, 2003.

build on top of state-of-the-art technologies like Apache's Cocoon and the forthcoming XForms standard. The harvesting component is based on a highly configurable pipeline concept, which reverses the pipelines introduced by Cocoon, while building upon the same component framework.

Unlike other scientific communities, the social sciences have not yet established a (living) pre-print culture. However, many institutions have begun to make working papers, project reports or dissertations available on the web. Other web-sites provide or collect materials dedicated to specific subjects, and increasingly, individual scientists start to publish their work results.

But, searching for these resources can be a rather tedious task: most resources are somewhere hidden on faculty servers, project pages or individual homepages, all implementing their own specific site structure, web design and search facilities. 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 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 DL dedicated to a given discipline. Researchers or students should be able to browse and search resources by domain specific categories and concepts taken from established classifications or thesauri.

At the same time, a strong commitment to open standards is required to ensure that valuable resources can be accessed and reused from contexts outside the community. The RDF Resource Description Framework [7] and related standards provide a well-established foundation for metadata interchange. Moreover, the ongoing efforts to establish standards for the "Semantic Web" will enable individual resources to be connected to a rich web of information objects.

Like MathNet and PhysNet, SozioNet is based on the principle of self-organization: social scientists and scientific institutions are encouraged to publish relevant resources and institutional information on the web, using the metadata standards established by the project. SozioNet provides tools to help authors create semantically rich metadata. The project also implements a general infrastructure to harvest the relevant resources from the web, and to search through these resources with a domain specific focus.

Besides providing access to common resources like working papers or project reports, SozioNet also aims to improve the visibility of social science institutions. Similar to the concepts first introduced by MathNet, institutions are motivated to create a so-called SozioNet page, which summarizes important institutional information on a standardized, additional web page. When browsing through different institutional homepages, users have constantly to adapt to a variety of web designs, site structures or navigation facilities. The SozioNet page will provide an additional, standardized entry-point into the web site, and will be dynamically generated from machine-readable metadata. To summarize, SozioNet aims to

- provide access to freely available web resources with relevance to social scientists
- build up a network of social science institutions to agree upon and establish common standards for the publication of scientific resources on the web

- improve the quality of web searches by concentrating on a given scientific domain
- improve the visibility of social science institutions and the resources and services they provide

Currently, in March 2003, 13 institutions have signed a cooperation agreement to participate in the development of SozioNet. This includes social science faculties as well as independent research institutes not directly affiliated to an university. While some faculties are just starting to make their resources available through SozioNet, there are also many institutions which have been publishing larger volumes of social science materials for a much longer period of time. This includes, for example, the German Youth Institute in Munich or the Social Science Research Center in Berlin. These institutions host a considerable amount of relevant materials. They usually have their own content management and publishing workflow. SozioNet has thus to meet the requirements of both, small university institutes and independant research institutions with pre-existing, possibly large collections of resources.

The remaining sections are organized as follows: The following section introduces the metadata schemes used in SozioNet and the backing ontology model. Section 3 provides a brief overview of the general architecture. Finally, section 4 will have a closer look at two central components: the user interface for metadata creation and management, and the harvesting component.

2 Metadata Schemes

As outlined above, it is vital for SozioNet that all players involved agree on a common basis for the creation of semantically rich metadata. This process can build upon existing standards like, for example, Dublin Core. However, there are additional aspects, which are specific to the social science domain. These domain specific aspects should be made transparent to the end user and to software agents accessing the resource.

All metadata in SozioNet is encoded in RDF [7]. RDF provides a simple language for expressing metadata and is commonly used to embed structured metadata into documents. RDF just defines a basic model and a serialization syntax. It does not specify a vocabulary. Vocabulary semantics, expressed in the form of RDF schemas, have to be defined by the communities using RDF. In addition to the RDF core standard, RDF schemas ensure the validity and data integrity of a given RDF data set, thus defining a specific vocabulary [2].

The benefits introduced by the RDF and RDF schema standards can not be overestimated. In particular, they enable metadata designers to rely on standardized and well-known schemas for the common elements of the metadata model, which can be clearly separated from domain specific vocabularies in a way that is transparent to users and software agents.

In the context of the W3C's semantic web initative, the foundations provided by RDF and RDF schema have been further extended. RDF schema defines modelling primitives for classes, properties, the relationships between classes and properties, and restrictions. The DARPA Agent Markup Language (DAML) extends this basic vocabulary by introducing a still limited set of additional language elements and useful distinctions [12].

It also clarifies many aspects that have been intentionally left open in the RDF schema standard.

From our point of view, DAML ontologies are usually easier to read and understand than comparable schemas written without the help of DAML. Also, tools like Hewlett Packard's Jena [8] offer direct support for this standard, for example, to create instances of a DAML class or to introspect the class hierarchy starting at a given instance. DAML will be superseded by the W3Cs Web Ontology Language (OWL), which is available as a working draft ([4], see also [9] for a discussion of the relationship between OWL and RDF).

SozioNet uses metadata in two main areas: first, participating social science institutions have agreed to enhance common resources like working papers, lecture notes, educational materials etc. with high-quality metadata. Second, RDF metadata is also at the core of the SozioNet secondary homepage, which provides information about the institution, faculty staff, research fields, educational focus, and so on.

Example 1.

```
<rdf:RDF>
  <sn:ResearchPaper rdf:about="http://www.zeitschriftarbeit.de
      /docs/2-2000/wolf.PDF">
  <dc:title>Das Netzwerk als Signatur der Epoche</dc:title>
  <dcq:abstract>Der Aufsatz beinhaltet ...</dcq:abstract>
  <dc:language>DE</ dc:language>
  <dcq:IMT>application/pdf</dcq:IMT>
  <dcq:created>1999-07-14</dcq:created>
  <dc:publisher>Landesinstitut Sozialforschungsstelle
  Dortmund</dc:publisher>
  <dc:creator>Wolf, Harald</dc:creator>
  <dc:subject rdf:resource="http://www.sozionet.org/1.0
    /classification#10220"/>
  <dc:subject rdf:resource="http://www.sozionet.org/1.0
    /classification#1080404"/>
  <dc:subject rdf:resource="http://www.sozionet.org/1.0
      /thesaurus#soziologische_Theorie"/>
  </sn:ResearchPaper>
</rdf:RDF>
```

A (shortened) sample metadata record for a working paper is shown above. For common web resources, most metadata properties are taken from the Dublin Core element set and the Dublin Core qualifier scheme.

A closer look at the example reveals several domain specific aspects. In particular, the SozioNet metadata scheme is backed by an ontology. The ontology defines, for example, different document classes like working paper, dissertation or lecture note. It also models the relationships between different object types and includes references to the classification and the thesaurus. The ontology basically 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.

Another important part of the ontology models the social science classification and thesaurus. Both are well established in German social science and are constantly maintained and enhanced by the Social Science Research Centre in Bonn. For SozioNet, the classification as well as the thesaurus are also defined as ontologies. The metadata record thus links directly to the definition of the corresponding term using an RDF reference in the subject property. This is a powerful feature of the implementation of the ontology.

Since classification and thesaurus are themselves defined in RDF and linked to the metadata record, it becomes possible for a software agent to explore the structure of the classification or thesaurus, and to navigate through the classification terms or descriptors connected to the current item. This is supported by RDF-related tools like, for example, Jena [8].

Currently, the base ontology as well as the classification and thesaurus are defined as DAML ontologies. We are prepared to migrate our metadata schemes to OWL, once this working draft is approaching a stable state.

The base ontology is extended by further object types which are part of the SozioNet homepage. This includes, for example, institutional information, information on faculty members or educational focus. The ontology model for these items has been largely influenced by existing proposals[6]. The draft also uses many properties from the vCard standard for general contact information.

In a future perspective, the different information objects should interconnect to a web of recombinable information items. For example, the creator property in a given metadata record could directly link to a personal description of the author on his homepage. However, implementing such features would require an additional standardization of personal homepages (as in MathNet's Persona Mathematica). As all participating institutions should be involved in the standardization process, we currently concentrate on common web resources. Additional information objects will be included in the future.

Embedding RDF into HTML or XHTML documents is still an open issue. It is a common practice to include the RDF into an XML comment. However, this contradicts the basic concepts of XML. The revised version of the RDF/XML syntax specification follows a proposal of the Dublin Core Initiative [1]: the RDF metadata should be saved into an auxiliary file, which is linked from the HTML document header. SozioNet supports this recommendation. Authors should not directly embed their metadata into HTML. Best practice should be to provide a separate file containing only the metadata.

3 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.

[6] For a list of example ontologies refer to http://www.daml.org/ontologies

3.1 Metadata Creation and Harvesting

As indicated in figure 1, SozioNet provides web-based tools for the creation of metadata records (see section 4). These tools are 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 generated metadata records should either be directly embedded into the source document, or preferably, stored as an auxiliary file in the same server context as the referenced resource. SozioNet gathers the

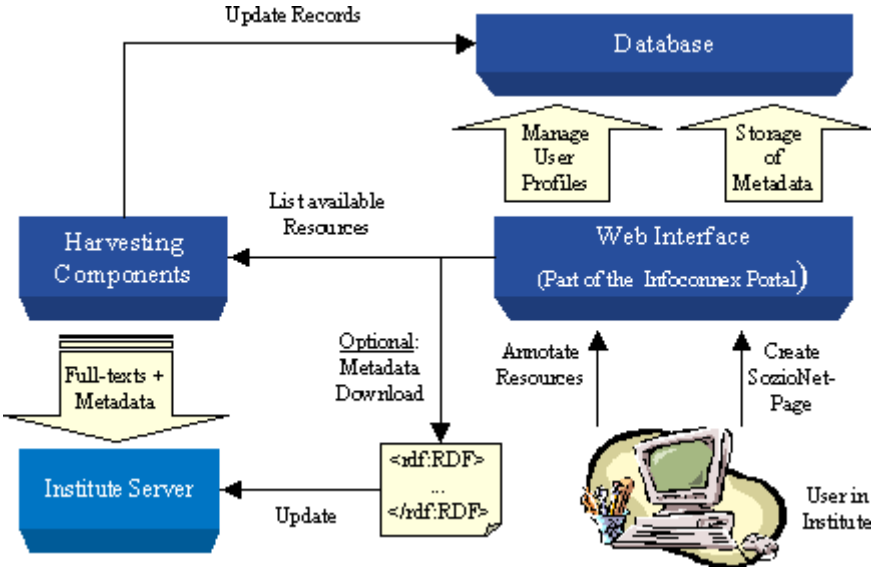

Fig. 1. Architecture Overview

available resources through a harvesting component (see the left side of figure 1), which will be described in more depth below. The harvester will periodically scan through all web-addresses known to the system, extracting metadata and indexing the fulltext-content of the resource.

However, the basic harvesting scenario has some known limitations: first, harvesting an entire faculty web site will necessarily include a large number of irrelevant resources without value to social science. Second, faculty members may not have direct access to the webserver, so updated documents have to be forwarded to the person responsible, which may take some time. Finally, the harvester will not recognize document updates immediately, but only during the next scheduled harvesting run.

SozioNet thus implements a hybrid solution: Metadata records entered through the SozioNet tools are stored in a central database. The harvester is just responsible for updating these records if it finds newer versions on the corresponding web site, and for indexing the fulltext content of the referenced resource. Alternatively, the location of new resources can be entered directly into the harvester's database. These resources will

be included into the scheduled gathering process. The harvester will thus not scan entire web sites, but only those locations which are known to contain relevant resources. This is in concept similar to the strategy implemented by SOSIG, i.e. the URLs in the manual created catalogue are used to initiate a gathering process, which tries to find additional resources to include.

Since several institutions run their own content management system or digital library, the harvester requires additional support to periodically retrieve metadata records from these existing systems.

The OAI protocol is well-suited for this task [6]. Furthermore, OAI data providers are usually rather simple to implement on top of existing systems. In this scenario, the harvester obtains metadata records conforming to SozioNet standards through OAI. It then only needs to access the referenced documents to fulltext-index their content.

3.2 Document Processing and Storage

A native XML database (NXD) is used as storage backend for storing metadata records and (optionally) fulltext contents. As outlined above, the database just represents a temporary data store. The harvester compares the database contents to the metadata descriptions it gathers from the web. If a newer record is found for the same resource, it will replace the one already stored in the db. The database is also used to store other temporary data, e.g. user profiles, thesaurus and classification data.

Using a NXD has a number of benefits: first, it allows the developer to think in XML from A to Z. All components in SozioNet are based on XML, including, for example, the project's web pages, the harvester and the metadata tools. Second, the database provides a schema-less document store, i.e. arbitrary document types can be mixed in the same database collection without requiring modifications to some underlying storage model. For example, RDF metadata records may differ in domain specific parts, while still conforming to SozioNet standards. It is also very likely that our metadata schemes will evolve over time, so documents may follow different versions of the same scheme.

Finally, XML is increasingly accepted in the social sciences not only as a standard for document authoring, but also for the exchange of valuable research data and research results. Since the project builds on XML database technology, SozioNet is able to integrate and search XML based data sets. Participating institutions are encouraged to extend the use of XML on basis of the already existing XML document types.

SozioNet currently uses eXist[7], an open source native XML database system, featuring index-based query processing, tight integration with other open source components (like Apache's Cocoon), simple deployment, and support for a variety of common standards as, for example, the XML:DB API for database access [5]. Particularly useful for SozioNet, eXist also offers query language extensions for index-based keyword searches, queries on the proximity of terms, or regular expression based search patterns. Despite its relatively short history, eXist has already been successfully used in a number of commercial and non-commercial projects. However, since database access is provided through established standard interfaces, SozioNet components do not depend on this particular NXD implementation.

[7] http://exist-db.org

The fulltext content of a resource is fetched by the harvester and indexed in a separate process. As will be described in more depth below, the harvester transforms HTML documents into well-formed XML by passing them to an HTML normalizer. The HTML normalizer closes open tags, adds missing elements or quotes attributes to generate a well-formed XML output stream. The output is then forwarded to one or more transformation steps.

Storing the now well-formed XHTML without modifications does not make much sense: while the HTML standard basically encourages authors to explicate the document structure, most HTML creators tend to misuse the corresponding tags for formatting purposes. For example, section headings or tables are very often used as layout elements and thus loose their structural meaning.

As a result, queries on the document structure - as offered by the XML database - may produce rather poor results when applied to heterogeneous HTML documents. In SozioNet, HTML documents are thus transformed into a simplified XML document type, which tries to preserve the basic document structure, e.g. section headings, paragraphs or links. The same applies to other document formats like PDF. However, PDF contains even less structural information, so the benefits are limited. Thus, as an alternative to storing the document's fulltext-content as XML, the harvester also supports fulltext-indexing of the content by forwarding it to Apache's Lucene.

3.3 Search Facilities

Currently, metadata search facilities are also implemented on top of the XML database. This is not an ideal solution, since RDF/XML is just one serialization form of RDF and it is always possible to represent the same given RDF data set by different XML documents. To be correct, the database just deals with the XML serialization of RDF and not with the underlying RDF data model itself.

To leverage the full power of RDF, one needs a tool which is able to work directly on the basic RDF triple model. Several tools are available, for example, Hewlett-Packard's Jena [8] or Sesame [3]. However, most tools tested in the project lacked some required features, e.g. efficient keyword searches in strings, and there have been some doubts with respect to scalability and performance. Also, the automatic reasoning capabilities built into many tools are not really a requirement at the current stage of development. Thus, for the time being, we decided to offer metadata search facilities on basis of XML. We are still prepared to switch to a specialized RDF search engine if required by future developments.

4 SozioNet Components

This section will have a closer look at some of the core components of SozioNet. However, most components are still rapidly evolving and will have to be covered by separate papers later on.

4.1 Metadata Creation

While some institutions have already defined their own publishing workflow, it seems that most sociologists have no experience with metadata creation. It is thus vital to support metadata creators with proper tools, which do not require any knowledge of metadata formats and only a minimal knowledge of the underlying standards. In SozioNet, we thus developed a reusable interface for metadata editing, called MetaWizard. It offers some outstanding features:

- Based on a server-side implementation of the forthcoming XForms W3C standard, allowing forms to span multiple screens. Uses a model-view-controller architecture.
- Personalized user interface: each user has a home-collection, containing the metadata records he created. Records entered through previous sessions can be revised or removed at any time.
- Existing metadata, e.g. META tags in HTML, is preserved by summarizing the referenced resource at the start of a new editing session.

As described above, entered metadata records are stored in a native XML database. The interface is fully personalized, i.e. every user has his own user profile and home collection, to which created records are assigned. User authentication is done by the database.

Users may either create new records, or delete or edit records they have entered during a previous editing session. Metadata creators are encouraged to download the created metadata and to store it on their institution's webserver, but they are not obliged to do so.

When creating new metadata records, the harvester will try to retrieve the web page at the location specified by the user. This feature has been inspired by the DC Dot service[8]. 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. This summarizing feature works very well, for example, if the web page already contains Dublin Core or other common metadata fields in HTML META tags. Of course, existing RDF metadata conforming to SozioNet standards is also recognized, making it possible to upload hand-edited files. However, if no metadata is found in the document, the results of the summarizing process will be limited: in many cases, only the title and a few fields from the HTTP header, including content language, modification date and mime type, can be extracted (for a more sophisticated approach to automatic metadata extraction from social science resources see [10] and [11]).

The user-interface has been implemented on top of Apache's Cocoon[9]. Cocoon offers a rich application server environment for the development of XML based applications. In particular, current versions of Cocoon come with XMLForms, a server-side implementation of the new XForms standard, which provides a powerful technology for the definition of user interfaces. Contrary to simple HTML forms, XMLForms may span over multiple screens and support a model-view-controller separation. Cocoon handles the session management and the transformation of XML pages into HTML. The developer can thus concentrate on the processing logic, which is defined by the means of so called actions.

[8] http://www.ukoln.ac.uk/metadata/dcdot/

[9] http://xml.apache.org/cocoon

The backing model is implemented as a set of Java beans, which are automatically transformed into RDF/XML by a Java-to-XML mapping tool. This way, we can easily adapt to different metadata models and keep the interface highly configurable.

4.2 Metadata Harvesting

Harvesting web resources is supported by quite a number of different software tools. For example, the open source Harvest software[10] is wide-spread and used in many projects, including MathNet and PhysNet.

Harvest makes a basic distinction between gatherers and brokers. A gatherer is used to recursively scan web locations and to retrieve the resources it finds through a number of supported protocols. The gatherer will forward the pages it finds to so-called summarizers, which are responsible for extracting metadata records, filtering contents, and so on. Based on the gathered data, brokers provide the intended retrieval functionality, usually through external retrieval engines. Though Harvest is highly configurable, SozioNet has slightly different requirements:

1. All metadata will be encoded in RDF. The harvester should thus be able to extract RDF metadata records and, even more important, it should maintain the basic structure of the RDF data.
2. We expect a growing portion of documents to be in XML. 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 easily met by XML-aware summarizers, the first is not so easy to deal with: the original Harvest software stores metadata internally in a format called SOIF. SOIF is a simple, hierarchical format, while RDF records can describe an arbitrary complex graph structure. Converting RDF to SOIF thus implies a possible loss of information and an undesirable reduction of complexity.

Several extensions to Harvest respond to this problematic, replacing the SOIF data store by an RDF triple store[11]. However, additional difficulties showed up during our experiments: for example, the existing, XML-based gatherers failed to process HTML pages which were not valid XML. Also, the installation process was fairly difficult, making it hard to redistribute the software to partner institutions.

As a result, SozioNet is currently using its own harvester, based on a variety of freely available open source tools and backed 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

[10] http://harvest.sourceforge.net

[11] see the CAP7 Gatherer Component developed in the project CARMEN (http://http://www.mathematik.uni-osnabrueck.de/projects/carmen/AP7/), which uses CARA (http://cara.sourceforge.net) as an RDF triple store

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. In the cascaded pipeline, the generator is responsible for reading the source document and for producing an XML stream, which is forwarded to the transformation steps. Depending on the type of input document, different serializers are selected: for example, HTML documents are first passed to one of the available HTML normalizers. The normalizers will try to transform the document into well-formed XML, e.g. by adding missing tags, closing opened elements or quoting attributes.

Transformers manipulate their input XML stream, for example, to extract existing metadata, summarize the document, or copy the stream to one or more sub-pipelines, each having a different function. In the basic scenario, two sub-pipelines are used: one to extract metadata from the incoming XML stream, another to prepare and index the fulltext content of the document. Transformers will often be based on XSL, i.e. most of the summarizing functionality provided by the original Harvest software is replaced by XSL stylesheets. Finally, serializers will consume the incoming XML stream and forward the result to the XML database or the fulltext indexer.

The basic concept is similar to the plug-in pipelines implemented in the Greenstone DL software [13]. However, our pipelines rely entirely on XML processing standards like SAX for streaming XML or XSL for transformations.

Multiple pipelines may be defined in the configuration file. The harvester will select the correct pipeline by looking at the mime-type of the source. Currently, there are pipelines and generators for XML/XHTML, non well-formed HTML and PDFs.

The SozioNet harvester not only benefits from the core paradigm of Cocoon, it is also based on exactly the same component framework, called Avalon. Avalon is the driving force behind the configurable pipelines. It implements a framework based on roles and contracts. Each component has a role, for example, in case of the harvester: generator, transformer or serializer. New components can be added through a configuration file or at runtime, by registering them with a component manager or a component selector. The framework usually cares about the configuration of components and provides a general contract for the methods called in each component.

As a result, the harvester is highly configurable. New components may be added at any time by implementing the Java interface for the desired role and registering the created Java class through the configuration file.

5 Conclusion

The SozioNet project complements and extends already existing information services available to social scientists in Germany. It will provide state-of-the art solutions using current and future web standards, thus ensuring that the resources collected can be maintained and accessed in the future.

References

1. Beckett, D. (Ed.): RDF/XML Syntax Specification (Revised). W3C Working Draft 23 January 2003. http://www.w3.org/TR/rdf-syntax-grammar
2. Brickley, D., Guha, R. V.: RDF Vocabulary Description Language 1.0: RDF Schema. W3C Candiate Recommendation. http://www.w3.org/TR/rdf-schema/
3. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In Horrocks, I., Hendler, J. (Eds.): The Semantic Web – ISWC 2002. Proceedings of the First International Semantic Web Conference, Sardinia, Italy, June 9–12, 2002.
4. Dean, M., Schreiber, G. (Eds.): Web Ontology Language (OWL) Reference Version 1.0. W3C Working Draft 21 February 2003. http://www.w3.org/TR/owl-ref/
5. Meier, W.: eXist Native XML Database. In Chaudri, A. B., Rashid A., Zicaro R. (Eds.): XML Data Management. Native XML and XML-Enabled Database Systems. Addison-Wesley, Boston MA, 2003.
6. Lagoze, C., Van de Sompel, Nelson, M., Wagner, S. (Eds.): The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. http://www.openarchives.org/OAI/openarchivesprotocol.html
7. Lassila, O., Swick, R. R. (Eds.): Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
8. McBride, B.: Jena: A Semantic Web Toolkit. IEEE Internet Computing Online. November, 2002. http://dsonline.computer.org/0211/f/w6jena.htm
9. Patel-Schneider, P. F., Fensel, D.: Layering the Semantic Web: Problems and Directions. In Horrocks, I., Hendler, J. (Eds.): The Semantic Web – ISWC 2002. Proceedings of the First International Semantic Web Conference, Sardinia, Italy, June 9–12, 2002.
10. Strötgen, R.: Meta-data Extraction and Query Translation. Treatment of Semantic Heterogeneity. In Agosti, M., Thanos, C. (Eds.): Research and Advanced Technology for Digital Libraries. Proceedings of the 6th European Conference, ECDL, 2002, Rome, Italy.
11. Strötgen, R., Kokkelink, S.: Metadatenextraktion aus Internetquellen: Heterogenitätsbehandlung im Projekt Carmen. In Schmidt, R. (Ed.): Information Research & Content Management: Orientierung, Ordnung und Organisation im Wissensmarks; 23. Online-Tagung der DGI, Frankfurt am Main, 2001.
12. van Harmelen, F., Patel-Schneider, P. F., Herrocks, I. (Eds.): Reference Description of the DAML+OIL (March 2001) Ontology Markup Language. http://www.daml.org/2001/03/reference.html
13. Witten, I. H., Bainbridge, D., Paynter, G., Boddie, S.: Importing Documents and Metadata into Digital Libraries: Requirements Analysis and an Extensible Architecture. In Agosti, M., Thanos, C. (Eds.): Research and Advanced Technology for Digital Libraries. Proceedings of the 6th European Conference, ECDL, 2002, Rome, Italy.