

RNA-project: Using things like thesauri and taxonomies in real cases!

Jeroen Wester¹ and Hans Nederbragt²

¹ Aduna, Prinses Julianaplein 14-b, 3817 CS Amersfoort, The Netherlands

² Trezorix, Oude Delft 200, 2611 HH Delft, The Netherlands

Abstract. This is a position paper that presents the work been done in the RNA-project, or the Reference Network Architecture project. The RNA-project investigates reliable, efficient and affordable digital knowledge-sharing solutions and makes use of reference structures, like thesauri and taxonomies, and object descriptions, like those found in collection databases. A number of information collections or use cases have been used in the project, like the Army Museum collection and the collection for natural history of Naturalis. The software component architecture used in the project to create solutions for use cases is based on existing, open source components, like Sesame, Spectacle, Lucene, Solr, Umbraco. The paper ends with a number of lessons learned during the project.

1 Introduction to the RNA-project

Many different organisations in the Netherlands are presently engaged in digital knowledge sharing. Examples include museums, libraries and cultural heritage managers. In the Reference Network Architecture project, RNA-project, we experiment with a variety of solutions to these knowledge sharing problems. The project runs from January 2005 till December 2007.

The RNA-project works on digital knowledge networks, clusters of information sources, like knowledge intensive websites and databases, that have connections and are extensible. The goal is to let users explore these networks and find information in a convenient way. Administrators should be able to publish information in the network. The published information should be connected to the network contents in an automated way. Better: the level of automation should be as high as possible.

Reference networks are a key player in the project. The term refers to reference structures, like thesauri and taxonomies, and to object descriptions, like those found in collection databases. In the project we try to connect a reference network to information sources, like collection databases, in order to create an extensible and dynamic knowledge system where the user can find and search for information in multiple ways.

The project uses existing software components and technologies, like RDF, Umbraco and Sesame, to link pieces of knowledge and make it consistently

searchable without the need for setting up a completely new knowledge sharing system or the necessity for all participants to work in a uniform way.

The RNA-project investigates reliable, efficient and affordable digital knowledge-sharing solutions. We use technologies that have already proven their worth in practice, but which have not yet, or hardly, been combined, or have not yet been applied to the field of (in particular) cultural heritage.

The RNA-project applies these solutions to a number of cases that are representative for the knowledge-sharing issues for which many knowledge organisations are currently trying to find an answer.

In this paper we describe some of the RNA use cases, the architecture we used as the base for placing solutions on, and the lessons learned in building the architecture and working on the use cases.

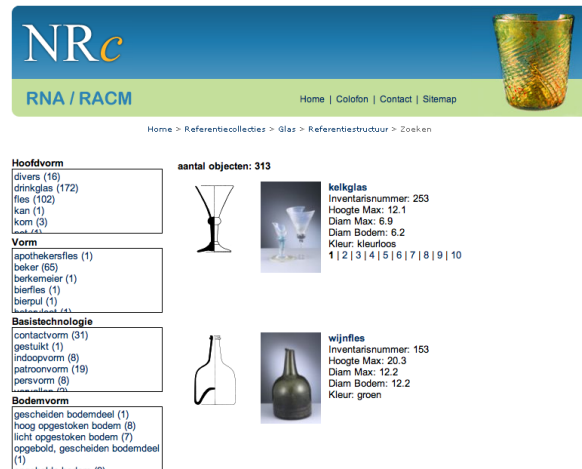


Fig. 1. Example of a faceted navigation interface: Reference collection of glass objects

2 RNA use cases

The RNA project links the uses cases to a number of proposals for solutions that have already proven themselves in actual practice, but have not yet or hardly been related to each other, or have not been applied so far to the field of (in particular) cultural heritage.

- The Legermuseum (Army Museum) has initiated a knowledge site that has made an enormous amount of information available in the shape of articles and object descriptions in a short time.
- Naturalis has become a pioneer in knowledge networking by rearranging and linking its proprietary knowledge and by building large knowledge networks in conjunction with other organisations.

- The Rijksdienst voor Archeologie, Cultuurlandschap en Monumenten (National Service for Archaeology, Cultural Landscape and Built Heritage) is the initiator of the Nationale Referentiecollectie NRc, a digitally consultable knowledge system of archaeological typology.
- A number of cultural heritage institutions in the province of Groningen intends to link regional information and make it available to various target groups.
- The nineteenth-century scientific work of De Clercq and Schmeltz on anthropology forms an early example of object descriptions and reference structures.
- The Dutch Libraries are currently engaged in making extremely diverse practical information findable using modern web technology.

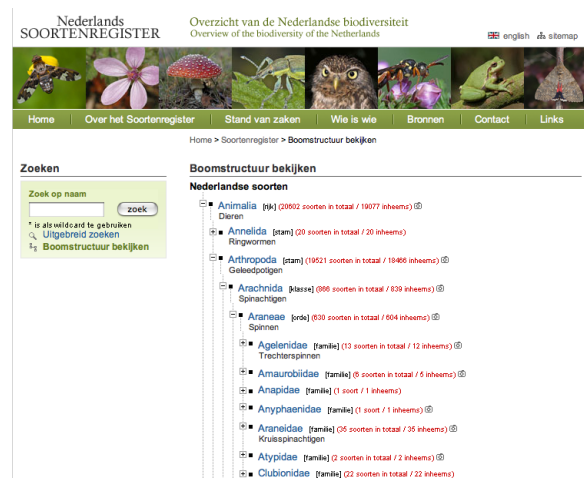


Fig. 2. Dutch Species Register based on the taxonomic thesaurus of Naturalis: RDF/SKOS structure

3 Architecture

This paper shows a software component architecture based on open source components, capable of handling reference structures and object descriptions, that is available and used today. The architecture has been developed by the RNA project partners.

The requirements for the architecture were (1) to support storage and querying of objects, metadata and reference structures, (2) to present objects, metadata and reference structures and (3) to support editing of objects, metadata and reference structures.

High-level overview of the architecture:

- User level. Reference networks as generic and dynamic interface between users and content or collections. Realised with Umbraco [1], Sesame [2], Spectacle [3], Lucene [4], Solr [5].
- Editorial level (internal). Support for management of reference structures. Realised with Umbraco, RNA SKOS-editor [6].
- Editorial level (external). External content production. Realised with (1) Windows Live Writer [7] for content production, and, with use of a RNA plugin [14], for metadata selection and applying metadata to external content like PDF and XLS files, (2) RNA metadata editor [8] for applying metadata to external content.
- System level. Storage of object descriptions and linking of collections and reference structures. Realised with Sesame [2].

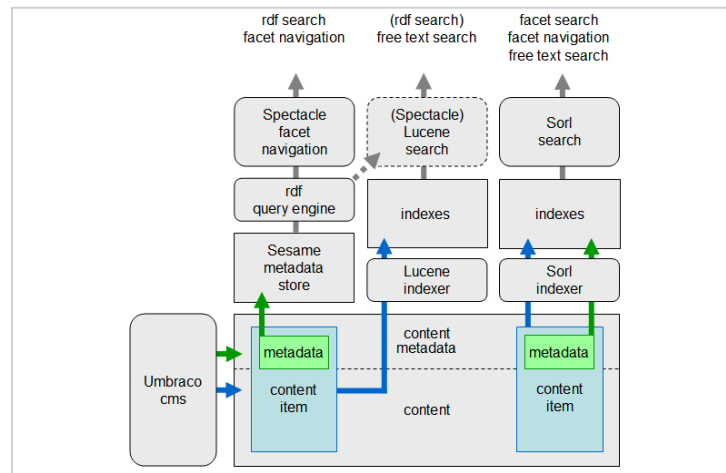


Fig. 3. High level overview of the architecture used in the RNA-project

4 Demonstrators

In the workshop we will try to shortly show the demonstrators listed below. This will be influenced by available time.

- Demonstrator 1. Reference networks used as a generic and dynamic interface between users and content or collections. See the RNA-site for the use of site navigation and faceted navigation [9].
- Demonstrator 2. Supporting internal editorial staff, management of reference structures and content metadata. See the RNA-site for the management of reference structures and content metadata in Umbraco [10].

- Demonstrator 3. Significant Term Extraction. See the RNA website for the use of the STE-tool: a tool that extracts significant terms to suggest classification terms [11].
- Demonstrator 4. The use of Windows Live Writer for convenient content editing and, with use of a RNA plugin, for metadata selection from sets of reference structures [14].
- Demonstrator 5. Alternative use: faceted determination on RACM Glass objects. The use of Spectacle for faceted determination, determination of species with help of facets [13].

5 Lessons learned

5.1 Driven by use cases

Several Cultural Heritage initiatives co-exist in The Netherlands, like the RNA-project, the CATCH-project (Continuous Access to Cultural Heritage) and others. The difference of the RNA-project and the other CH-projects is its aim to create real-life solutions, that are tested in real-life use cases. This practical approach lead to the following insights:

- The use of existing software components for project use cases reduced development time. *Examples* of existing software components used by the project: Sesame (Aduna) for storage and retrieval of data, Spectacle (Aduna) and Solr (Apache) for faceted navigation, Lucene (Apache) for full-text search, Umbraco for content management, Windows Live Writer (Microsoft) for external content production, bypassing a heavy-duty CMS.
- Make an economical, time-saving choice when you pick software components. Better to have something simple that solves most of the problems in short time, than something advanced that solves all problems but takes long time. *Example:* In one use case we needed a semi-automatic classification mechanism, that would suggest terms that match with the contents of a document. The choice between state-of-the-art natural language software that needed training time and a simple statistical tool for term extraction that worked out of the box was not hard. *Another example:* Applying ontology mapping with good results is still a cumbersome affair. In most cases though manual mapping works better at less costs.
- We learned that when you are driven by use cases, it is necessary to have an well organised community of determined content producers and of course users. Without them it is hard or impossible to test the solution. *Example:* The opposite was happening in the the Groningen use case. There were enough ideas and wishes, tools were made to support these ideas, but there was no organisation on the side of the content providers which was strong enough to make things come true.

5.2 The architectural approach

There are two reasons to approach the central problem of the RNA-project, to find reliable, efficient and affordable digital knowledge-sharing solutions, as an architectural problem, rather than a 'information disclosure tool selection' problem:

- Focussing on architecture leads to some form of integration. In other words: we learn to see a chain of components that are needed to solve the problem. This increases the quality of the advice that we can give to future users. *Example:* The RNA-site was build as a coherent collection of open source components aimed at supporting the idea of a dynamic knowledge system in an flexible and affordable way. Before the end of the project this setup was put into actual use by RNA-member Naturalis.
- There is abstraction. In other words: there is not one solution for one specific use case. There seems to be the promise of using this architecture in other use cases, that are not related to this project. An *example* is the use of this architecture for an enterprise search project in a large telecommunication company.

5.3 Conclusion

- The combination of the project's use cases resulted in a high-level approach for the information disclosure chain in the cultural heritage domain. No claims are made here to what extent this approach is general to the domain.
- Combining the input of the different use cases made it easier to come to general models for complex collections than when working in more isolated cases.
- Combining use cases and solutions resulted in generic solutions and applicability. From specific to generic: it was relatively easy to abstract from the solutions generated in the project's use cases. This resulted in generic solutions for broader use. An example is the broader use of the faceted navigation mechanism. At first we used this mechanism to find content by drilling down a large set in small steps, but in a later stage also used as a determination mechanism.
- Combining forces between CH-projects could increase the above mentioned effects. We actively explore the use of RNA-project results in other cultural heritage projects, like CATCH and others.

References

- [1] Umbraco: <http://umbraco.org/>, open source Content Management System based on Microsoft's ASP.NET.
- [2] Sesame: <http://www.openrdf.org/>, Java-based RDF framework.
- [3] Spectacle: <http://www.aduna-software.com/>, open source Java-based faceted navigation engine.

- [4] Lucene: <http://lucene.apache.org/>, a high-performance, full-featured text search engine library written entirely in Java.
- [5] Solr: <http://>, Solr is a high performance search server built around Lucene, with XML/HTTP and JSON/Python/Ruby APIs, hit highlighting, faceted search, caching, replication, and a web admin interface.
- [6] RNA SKOS-editor: <http://www.rna-project.org/>. Reference to this software will be available at the end of 2007.
- [7] Windows Live Writer: <http://get.live.com/betas/>, a desktop application for rich content editing.
- [8] RNA metadata editor: <http://www.rna-project.org/>. Reference to this software will be available at the end of 2007.
- [9] Demonstrator 1: Use of faceted navigation and site navigation, <http://www.rna-project.org/>. Reference to this software will be available at the end of 2007.
- [10] Demonstrator 2: Management of reference structures, <http://www.rna-project.org/>. Reference to this software will be available at the end of 2007.
- [11] Demonstrator 3: Significant Term Extractor, <http://www.rna-project.org/>. Reference to this software will be available at the end of 2007.
- [12] Demonstrator 4: Use of Windows Live Writer, <http://www.rna-project.org/>
- [13] Demonstrator 5: RACM Glass, <http://www.rna-project.org/>. Reference to this software will be available at the end of 2007.
- [14] RNA-plugin for Windows Live Writer, <http://www.rna-project.org/>. Reference to this software will be available at the end of 2007.