NARCIS
Integrating CRIS, OAI and Web Crawling

Elly Dijk, Arjan Hogenaar, Marga van Meel

Royal Netherlands Academy of Arts and Sciences, Amsterdam
Department of Research Information

Summary

NARCIS, National Academic Research and Collaborations Information System, is a Dutch project to build a portal for research information. This project combines Current Research Information Systems (CRISs) with structured information, with information from OAI repositories (obtained via harvesting), and websites and news pages via web crawling. A major goal is to create a central facility for searching all these data.

In this paper the result of the NARCIS project will be described: a (prototype) portal containing 400,000 items. We will show how and why we have used different techniques for the different sources of NARCIS. Finally, we will present the results of end-user testing.

1 Introduction

The traditional dichotomy between current research information and information on research results (often publications, but also datasets, models, web publications and patents) starts to fade away, also in the Netherlands. There is a need for tools which make all these types of information accessible and searchable at the same time (one-stop shopping).

NARCIS, National Academic Research and Collaborations Information System, is a Dutch project to build a portal for research information which combines CRISs (structured information) with information from OAI repositories (obtained via harvesting), and websites and news pages via web crawling. The main goal is to create a central facility for searching all these data. NARCIS has been developed with the purpose to collect research data via the administrative processes of the different participating institutes within the work flow process. The advantage is that researchers or institutes need to register the data only once. The implementation of NARCIS took one year, which can be considered relatively quickly.

Partners in the NARCIS project are the Royal Netherlands Academy of Arts and Sciences (KNAW), the Netherlands Organisation for Scientific Research (NWO), the Association of Universities in the Netherlands (VSNU) and the Information Centre of the University of Nijmegen (UCI).

The idea for the project has been developed within the DIO-platform (National Platform Data Infrastructure Research Information) and has been realised with a subsidy of DARE. DARE is a large programme in the Netherlands (Digital Academic REpositories) under the auspices of SURF. SURF is the higher education and research partnership organisation for network services and information and communications technology. DARE has been established to develop academic repositories for all the universities in the Netherlands as well as for KNAW and NWO (DAREnet). The subsidy of SURF was 83.480 euro; the whole project amounted to 159.160 euro.

2 NARCIS

NARCIS is the name of a flower (Daffodil) and symbolizes the flowering of the information exchange. The universities, the Academy and NWO have their own research information systems and their repositories. These systems contain more or less the same data. Data collection is not an easy job to do and it is rather expensive and time-consuming. That is the reason why there is so much to win by making agreements on who is collecting which data when and where. Starting point is to collect this information during the work flow process. This process leads to an increasing accessibility, improvement of exchange, linking of data and re-use of available research information. In this way, minimization of administrative report burden for researchers and institutes as well as registering of data only once can also be achieved.

2.1 NARCIS sources

For the NARCIS portal we only use information from existing systems and websites:

The Dutch Research Database (NOD) has been integrated into NARCIS. The NOD is being produced by the department of Research Information of the KNAW. This datbase is the coordinating information source for research information (the national CRIS). The NOD contains information about Dutch researchers and their expertise, information on research institutes and research programmes and projects. This database gives access to university and non-university research information. The NOD is a relational database and the information is highly structured: it offers links between research, persons and institutes. One of the major benefits of the NOD is that it contains information of a higher quality than, for example, unstructured web pages.

METIS is the management information system of the universities, containing research information (mostly programmes), the metadata of the scientific output and personnel information.

The Dutch programme DARE has been established to develop academic repositories for the universities in the Netherlands and for KNAW and NWO. All these institutes now have their own

repositories. These repositories are harvestable by using the OAI-PMH protocol (Open Archives Initiative – Protocol for Metadata Harvesting). This makes it possible to implement services on top of them. A well-known example is DAREnet that focuses on offering a review of the scientific output of Dutch universities. NARCIS is also harvesting these repositories in order to combine research information with research results.

NWO is the most important funding organisation in the Netherlands. It is funding about 15 percent of the academic research. This organisation has its own system which has also been made public via NARCIS.

Completely new in NARCIS is the information from websites of non-university institutes. By making use of web crawling this information has been made available. Scientific output, metadata of datasets, news items, and press reports are also sources for NARCIS.

2.2 NARCIS techniques

In the NARCIS project different open source techniques and protocols have been used, namely: XML & XML-SOAP, OAI-PMH (harvesting), categorizing, web crawling, RSS and bulk upload.

2.2.1 METIS

In the Netherlands all 13 universities have their own management information system, called METIS. We have made arrangements with the producer (University of Nijmegen, Information Centre) to enter METIS information into our system via an export of the data in a specific format. Universities can send that information by mail, CD-rom or FTP. Through a specific module in our database we can import this research information easily into the Dutch Research Database, which forms the basis of NARCIS.

2.2.2 XML schema

An XML schema has been developed for the exchange of information between the Netherlands Organisation for Scientific Research (NWO), the METIS-systems of the universities and the NARCIS system of the Academy. This schema is CERIF-based.

We used XML-SOAP (Simple Object Access Protocol). SOAP is a lightweight protocol for exchange of information in a decentralized, distributed environment. It is an XML based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined data types, and a convention for representing remote procedure calls and responses.

By introducing XML-SOAP on the NARCIS server, the server of the Netherlands Organisation for Scientific Research (NWO) and the METIS servers, a sufficient data exchange between the local systems is created. When the Netherlands Organisation for Scientific Research approves a research proposal, the information goes directly into their database. At the same time it goes automatically into the METIS-system and into NARCIS portal. When the first publications of

such a project are available in METIS the metadata of these publications are sent back to Netherlands Organisation for Scientific Research and to the repository of the institute. In this way the researcher needs to deliver his publication only once.

2.2.3 Web crawling

To gather information from web pages we have developed a web crawler, or spider. We use the open source tool J-spider as a plug-in for the NARCIS portal and developed a simple interface to produce spider tasks, which can be tailored to special needs. The web crawler is able to spider websites or parts of websites.

Many websites contain valuable information in PDF or RTF format. J-Spider has been adjusted to spider also PDF or RTF format.

Also the performance has been improved. Originally, J-Spider was using much memory capacity during crawling of information because it retained all the pages and the structure of the website. That resulted, while spidering large websites, in an 'out of memory' report and in the termination of the spidering. As a solution, we currently no longer download the pages at this stage. We index the full text and save the URL.

The result is that we have fewer problems with the memory capacity; so many pages can be spidered in a quick and efficient way. After the spider task, a report with the results is automatically sent to the administrator.

2.2.4 Repositories

In the last years all the Dutch Universities and other research institutes developed their own repositories with publications or other output of their scientific research. These publications are stored in open archives, which contain XML documents. The documents are in Dublin Core format and mostly there is a reference to the full-text publication in PDF or in other formats. NARCIS has developed a data service on all these repositories. So there is a real connection between the NARCIS portal and the different repositories, and through the use of OAI-PMH, Open Archives Initiative -Protocol for Metadata Harvesting, we harvest all these documents and include them in the portal.

2.2.5 Categorising

After all the gathering we also need a tool for documenting and searching the information. With almost 400,000 items we cannot add categories or thesaurus terms manually. Therefore, part of the project was the development of the categorizer.

The categorizer was based on the functionalities of the open source search engine Lucene. The motivation to use Lucene is that it is one of the most used open source search engines. We used the 'Similarity part' of Lucene: finding the documents that look like the starting document. On the basis of similarity, the categorizer decides in which category or categories the document belongs.

The classification codes of the NOD were used as a basis for this tool. This classification exists of subject fields and scientific disciplines for all sciences. There are 259 categories in total.

Scientific information specialists have selected 50 descriptions of current research that are representative for each category. These training sets are used to categorize other research projects. At the moment, in NARCIS, the categorizer tool gives rather good results. However we are not fully satisfied yet. One document type (for example web pages) gives a better result than other document types (for example records). An explanation for this is that the size of OAI records is very small.

2.2.6 Notifying

And last but not least, it is possible to make an RSS feed on the content of NARCIS. This feature gives the end-user a unique possibility to follow the latest developments on the specific subject. The possibility to activate an RSS feed occurs after a search is done, so the RSS feed is based on the search terms given by the end-user.

2.3 The result: www.narcis.info

We have developed a bilingual (Dutch and English) research information portal (www.narcis.info), which contains about 400,000 items (prototype). NARCIS contains information on research, researchers, research institutes, (full text) publications, datasets, and news. All the information that one can find in this portal is of high scientific quality. One can search all this information in one time, but it is also possible to search in the different items 'Persons', 'Organisation', 'Current research', '(Full text) publications', 'News', and 'Datasets'. Besides it is possible to make use of an RSS-feed to keep up with the latest information. One can also search further by an automatic generated category. Searches can be made by inserting both Dutch and English words. However, the results will of course be different depending on the language used.

Figure 1: NARCIS homepage : www.narcis.info

3 The end-users test

Although a lot of organisations have been involved in the development of NARCIS, its usefulness and benefits can only be assessed by the users themselves.

Anticipating on a broader users test planned in spring this year, a preliminary users test has been carried out among Dutch information specialists. On purpose we have selected these specific users. We had selected these specific users for the test, as they all have much expertise in the field of information retrieval. Consequently, they would tackle other problems than normal end-users.

The main goal of the test was to get a quick view on the opinion of these experienced users on the following topics:

1. Impression and functionality of the NARCIS home page
2. Search functionality
3. Limiting options

4. Display options of the several record types

3.1 Impression and functionality of the NARCIS home page

Most users like the clean appearance of the home page ('looks like Google'), but there are some criticisms as well. The most arguable issue is the option 'now popular'. This option refers to subjects/persons that have raised many search statement during the last month. It should give the end-user a notion of the major recent topics in research information. Apparently, even these experienced users find it difficult to estimate the value of this particular service.

The availability of an adaptable RSS-feed is considered as an added value, but a direct link to its functionality is advisable.

Minor issues discussed are the length of the explanatory text (too long), the presentation of the limiting options, and the necessity of scrolling on the homepage (one should try to avoid this).

3.2 Search functionality

In general, search options and search performance are very well appreciated. There is some need for explanation of symbols appearing in the records within the results set.

The searching within results appears to be a good feature. We have intended to create a 'Google look-alike" view of the results list. In this view, the bar with limiting options is only available at the bottom-site of the page. Obviously, the Google 'search within results' feature is not well-known amongst Dutch information specialists. Therefore, it is worth considering offering all the search options in one single bar at the top of the page.

3.3 Limiting options

The users all used the limiting options in the search bar. The other limiting option (by just clicking on the record type [examples: 'person'; 'current research'] mentioned in each individual record) is not understood, not even by these information specialists. The way the record type is being displayed makes a user think it is just a type indicator and not a limiting option.

3.4 Display options of the several record types

Key feature of NARCIS is the access to quite divergent record types. This makes it very difficult to represent all these records in the same manner. Therefore we have decided to use two ways of representing them:

a) In the case of structured information (for instance NOD or OAI record) clicking a record in the results list will result in a full record displayed within NARCIS

b) In the case of unstructured information (for instance data set or web publication record) clicking a record within the results list will lead to the opening of an external webpage within the NARCIS page.

Where route a) is clear, users get confused by route b). Reason for this is the unfamiliarity with this kind of representation. Therefore there is an explicit need for some elucidation of this typical NARCIS feature.

4 Future developments in NARCIS II

4.1 Automatic categorization

As already mentioned within the framework of the NARCIS portal we have developed a tool for automatic categorization, but the results were not satisfactory.

In the beginning of 2006 we will be looking at two different solutions. First we are running a test with a commercial tool for categorization. This tool makes a so-called unique fingerprint of the different categories and the items in the portal. By comparing the fingerprints the software can decide which documents belong to which category.

The other possible solution is the further development of our own tool, developed for NARCIS. On the basis of similarity the Lucene search engine compares documents with a training set. We expect that much effort is needed to make it work.

4.2 Repositories

Since all the 13 universities in the Netherlands use METIS as their CRIS system, an output can be made from the METIS-system to fill the repositories with the metadata and the full text publications. Thanks to this simplification the amount of publications in the repositories is expected to increase.

4.3 Unique identifiers

In bringing CRIS systems and repositories together, there is a tremendous problem in matching names. This is necessary for the exchange of data, but also if someone wants to retrieve a list of publications of a certain researcher. M.A. Smith can be known as M. Smith, Martin or Martin A. Smith and there are even more possibilities. In a small investigation we found people with more than 10 alternatives for one single name.

At the moment there are some experiments to introduce a national DAI, Digital Author Identification. This unique number will make it possible to retrieve all the alternative names of a person and connect the information. It also gives the opportunity to connect the CRIS systems with the repositories. For the exchange of data between NWO, universities and the KNAW there is already a thesaurus with the matching of these alternatives.

4.4 Curriculum Vitae (CV)

One important development will be the inclusion of CVs of Dutch researchers. Within the framework of the HARVEX project (Harvestable Excellence, another DARE project), a feature will be provided for researchers to publish their own CV on the internet. The system will use the personal information and publications of the CRISs and publish it on the internet.

At the time HARVEX will be ready, we will be able to introduce a new type of information on the NARCIS portal, namely CV, and spider all the CVs on the internet.

4.5 XML exchange of data

In 2006 we hope to realise the exchange of data between the universities and NARCIS. At the moment we can exchange data by means of a not structured text file. We aim to introduce the XML schema which had been developed in NARCIS for a better exchange of data.

4.6 Broader users test and adaptations of the website

A preliminary users test has been carried out among Dutch information specialists (See 3. The end-users test). As a result we will make adaptations to the NARCIS website. This 'new' website will be tested in a broader users test that will be carried out this spring.

5 Conclusions

- A research information portal has been developed: www.narcis.info.
- This portal contains about 400,000 items (prototype).
- We have managed to combine different information types in the portal. The portal contains information on research institutes (profiles, addresses, projects, and publications), researchers (expertise, addresses, research projects, and publications), research activities (research projects, and programmes), publications (metadata, and full text), datasets (metadata) and news (web pages).
- Due to the developed interface it is quite easy to make new spider tasks, so we can adapt to actual scientific developments.

- One-stop-shopping: In the portal it is possible to search all the different kinds of information at the same time or to search the different sources separately.
- The introduction of the RSS feed within NARCIS makes it possible for the researcher to follow the latest issues on a subject.
- It is very important that by the far-reaching co-operation between the partners in this project the researchers and institutes do not have to spend costly time doing double (or more) administrative work, such as making descriptions of current research, publication lists, etc.
- The web crawling and the integration of the web crawled information in NARCIS is functioning already. However the automatic classification is not yet satisfactory. In a following project, NARCIS 2, we will develop this tool or experiment with another automatic classification tool.
- The XML exchange schemas (CERIF based) are working well and will be introduced in all METIS systems. It has the potential to become the common scheme for the exchange of research information.
- The future development of unique digital identifiers for authors, objects and institutes, part of the Dutch DARE programme, will connect all sources and will facilitate the exchange of information within the NARCIS portal.
- A NARCIS consortium is formed in which the partners are working together to develop NARCIS further.
- De NARCIS crawler is the open source J-SPIDER, which is available via http://j-spider.sourceforge.net/other/index.html.

6 References and contact information

6.1 References

Meel, A.M. van (2005): NARCIS, National Academic Research and Collaborations Information System: Eindverslag. Amsterdam: KNAW

Background information on the NARCIS website: http://www.narcis.info/narcis/background.htm

NARCIS project information:
http://www.onderzoekinformatie.nl/en/oi/onderzoeksprojecten/narcis

6.2 Contact information

E.M.S. Dijk
Royal Netherlands Academy of Arts and Sciences - KNAW
Department of Research Information

PO Box 95110
1090 HC Amsterdam
www.onderzoekinformatie.nl
E elly.dijk@bureau.knaw.nl

A.T. Hogenaar
Royal Netherlands Academy of Arts and Sciences - KNAW
Department of Research Information
PO Box 95110
1090 HC Amsterdam
www.onderzoekinformatie.nl
E arjan.hogenaar@bureau.knaw.nl

A.M. van Meel
Royal Netherlands Academy of Arts and Sciences - KNAW
Department of Research Information
PO Box 95110
1090 HC Amsterdam
www.onderzoekinformatie.nl
E marga.van.meel@bureau.knaw.nl