NARCIS: The Gateway to Dutch Scientific Information

Elly Dijk, Chris Baars, Arjan Hogenaar, Marga van Meel

Department of Research Information, Royal Netherlands Academy of Arts and Sciences (KNAW)
PO Box 95110, 1090 HC Amsterdam, The Netherlands
email: elly.dijk@bureau.knaw.nl; chris.baars@bureau.knaw.nl; arjan.hogenaar@bureau.knaw.nl;
marga.van.meel@bureau.knaw.nl

Abstract

NARCIS, National Academic Research and Collaborations Information System, is a project in the Netherlands to build a portal for research information which combines structured research information with information from OAI repositories (publication and other scientific results), websites, and news pages of research institutes. The main goal of NARCIS is to create a central place for searching all these types of data. The research data in NARCIS have been collected via the administrative processes of the different participating institutes within the work flow process. Different techniques are being used to combine the current research information and the research results. 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 the Dutch programme DARE (Digital Academic REpositories) coordinated by SURF. SURF is the higher education and research partnership organisation for network services and information and communications technology (ICT). Partners in the NARCIS project are The Royal Netherlands Academy of Arts and Sciences (KNAW), Netherlands Organisation for Scientific Research (NWO), the Association of Universities in the Netherlands (VSNU), and the Information Centre of the Radboud University of Nijmegen (METIS). The implementing of NARCIS took one year.

Keywords: current research; data sets; harvesting; research information; search facility; web crawling

1 Introduction

In the Netherlands the scientific institutes register current research information and information on research results, like (full text) publications, data sets, models, web publications, and patents. Until the NARCIS project all these types of information were not accessible and searchable at the same time.

NARCIS, National Academic Research and Collaborations Information System [1, 2], is a Dutch project to build a so-called third generation portal. This portal integrates all kinds of types of information from scientific institutes in the Netherlands. NARCIS offers an overall picture of research information and publications of all relevant (university and non-university) research institutes in the Netherlands on the basis of international OAI standards. The portal combines research information systems (structured information) with information from academic OAI repositories via harvesting, and information from websites and news pages via web crawling. The main goal of the NARCIS project is to create a central facility for searching all these data (one-stop shopping).

The Research Information department of the Royal Netherlands Academy of Arts and Sciences (KNAW) took the initiative for this project in the National Platform Data Infrastructure Research Information (DIO). In the NARCIS project KNAW Research Information is working together with the Netherlands Organisation for Scientific Research (NWO), the Association of Universities in the Netherlands (VSNU) and the Information Centre of the Radboud University Nijmegen (UCI) as part of services development within the DARE programme of the Foundation SURF. SURF develops and operates an advanced ICT infrastructure for and on behalf of higher education and research. The aim of the programme DARE (Digital Academic REpositories) is to store the results of all Dutch research in a network of academic repositories, thus facilitating access to them (DAREnet). All the institutes involved do this storage in a similar way, but retain responsibility for and control over their own data. NARCIS is one of the projects of DARE to give better access to information about Dutch scientific activities.

The subsidy of DARE/SURF for the NARCIS project was 83.480 euro; the whole project amounted to 159.160 euro. The NARCIS project did run from September 2004 till September 2005. The results of this project have been approved by SURF in December 2005. There will be a follow-up NARCIS 2 to develop this portal further.

2 The sources of NARCIS

The universities, the Academy and NWO have their own research information systems and their repositories. These systems contain more or less the similar data about research, researchers, research institutes, publications, data sets et cetera (see Fig. 1). Data collection is rather expensive and time-consuming in terms of acquisition and processing the information. Therefore there is much to win by making agreements on who is collecting what data when and where. 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. In this way, minimization of administrative report burden for researchers and institutes as well as registering of data only once can be achieved. Thus, besides the new techniques that are being used in the NARCIS project, the project is also about cooperation of the most important scientific institutes in the Netherlands.

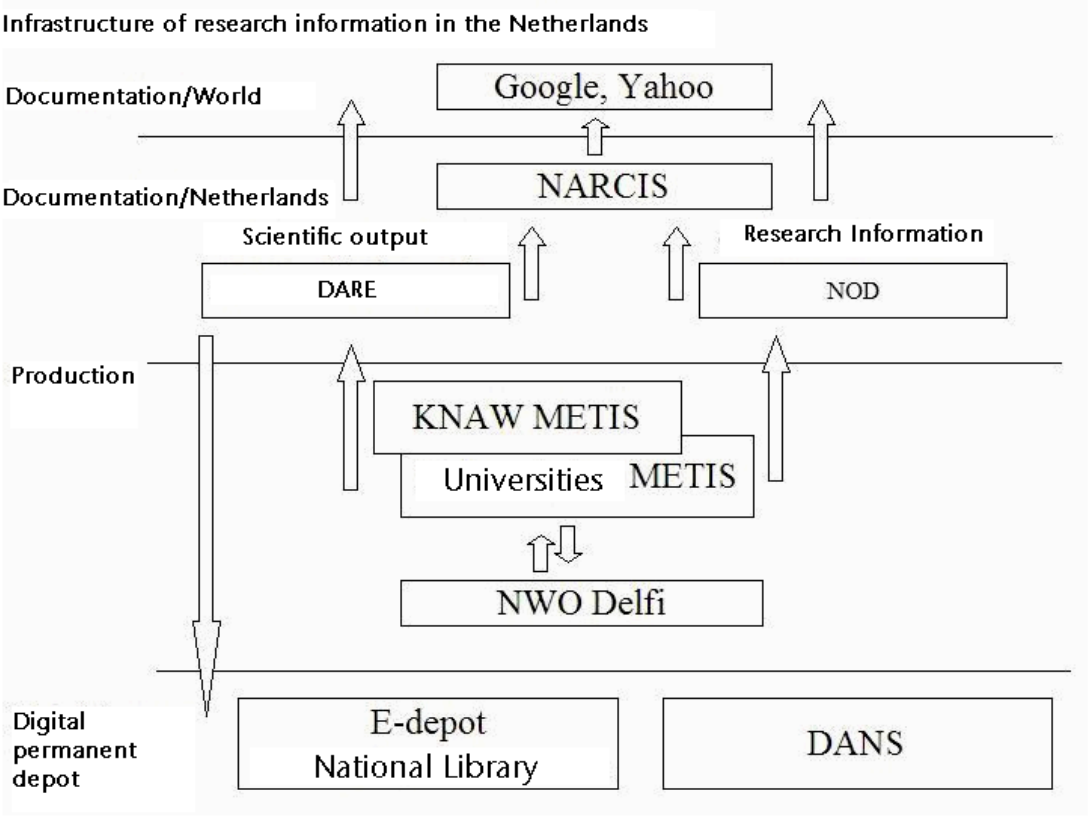

Figure 1: The infrastructure of research information in the Netherlands

NARCIS combines the following information sources:

1. The Dutch Research Database (NOD)
2. Public information from the METIS systems from the universities
3. Research information from NWO/Delfi
4. Information from the digital academic repositories, also available via DAREnet
5. News sites from academic and research institutes
6. Information from outside universities for example data sets of the Dutch organisation DANS

2.1 Dutch Research Database (NOD)

The Dutch Research Database (NOD) is produced by the department of Research Information of the KNAW. This database is the national Current Research Information System (CRIS) and offers an overview of the scientific landscape of the Netherlands. The NOD is a publicly available online database with information on scientific research, researchers and their expertise, and research institutes. The database covers all scientific disciplines and 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. This database is used as the basis for the NARCIS portal.

2.2 Public Information from the METIS systems from the Universities

METIS is the management information system of the universities, containing research information (mostly programmes), the metadata of the scientific output (publications et cetera) and personnel information. Only some of the METIS systems are publicly available and of course not al the data is visible, because there is also private data in METIS.

2.3 Research Information from NWO/Delfi

The Netherlands Organisation for Scientific Research (NWO) promotes scientific research at Dutch universities and institutes through nearly 170 different research programmes and grants. It is the most important funding organisation in the Netherlands. This organisation has its own research information system, called NWO/Delfi. In NWO/Delfi one can find information on research projects and researchers. This information is publicly available at the website of NWO.

2.4 Information from the Digital Academic Repositories, also Available via DAREnet

Since January 2004 all the Dutch universities, KNAW and NWO have academic repositories containing full text publications and other scientific results. The great impulse for these repositories was given by the Dutch programme DARE (Digital Academic REpositories), under auspices of SURF. 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. An example is the website DAREnet that offers a review of the scientific output of Dutch universities. NARCIS is also harvesting these university repositories in order to combine research information from the Dutch Research Database, the universities, and NWO with research results from the repositories.

2.5 New Sites and Scientific Reports on Websites of Academic and Non-academic Research Institutes

Another source for NARCIS is the news on websites from university and important non-university institutes. This information has been made available by making use of web crawling. Also scientific output, like research reports, available on these sites, is a source for NARCIS.

2.6 Information from Outside Universities for Example Data Sets of the Dutch Organisation DANS

In the Netherlands DANS (Data Archiving and Networked Services), an initiative of KNAW and NWO, is the national organisation responsible for storing and providing permanent access to research data from the humanities and social sciences. Parts of DANS are the Steinmetz Archive for the social sciences and the Netherlands Historical Data Archive (NHDA). DANS is also acting as a go-between in the acquisition of large data sets from organisations like Statistics Netherlands (CBS), the Social and Cultural Planning Office (SCP) and the land registry. In NARCIS the metadata information from the data sets of DANS are available and searchable with all other types of information.

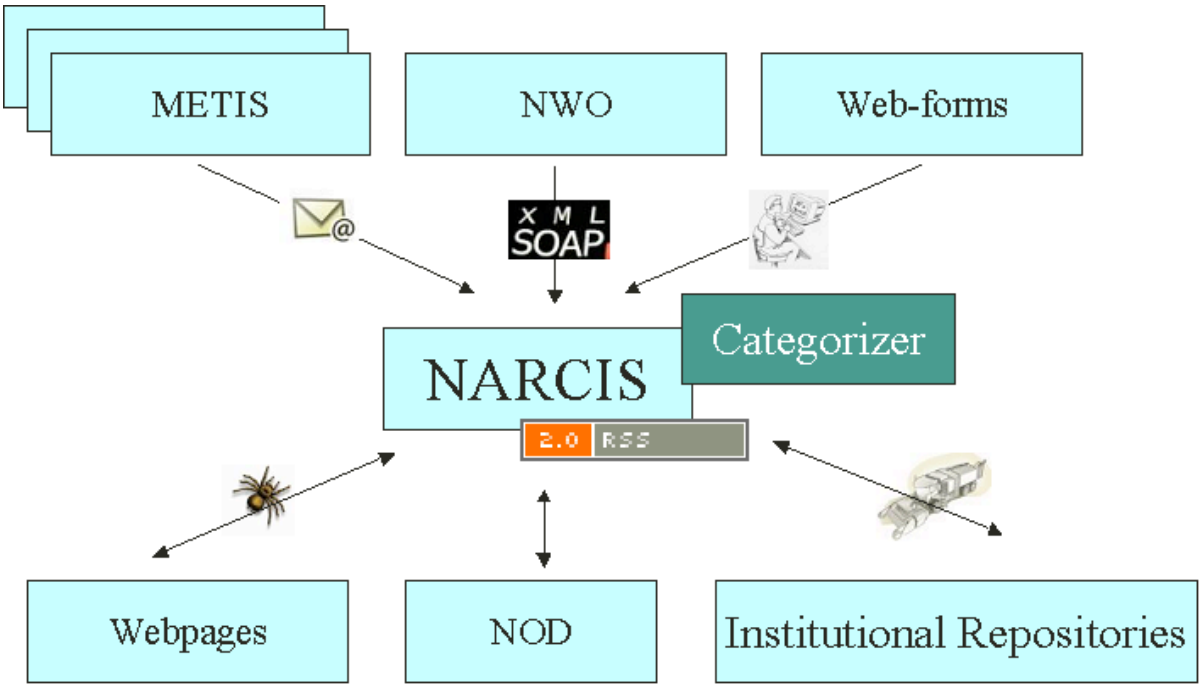

Figure 2: The applied techniques used in the NARCIS project

3 Realization of Sub Goals

To reach the main goal of the creation of a central facility for searching all these data different sub goals had to be realized (see Fig. 2):

1. Linking of different current research registration systems
2. The development of an exchange schema
3. Harvesting of repositories by using OAI-PMH
4. Installation of a spider tool
5. Development of a web tool for automatic categorising
6. Building a portal as a central facility for making available the different types of information

3.1 Linking of Different Current Research Registration Systems

In the Netherlands the 13 universities have their own management information system, called METIS. Arrangements with the producer (University of Nijmegen, Information Centre) were made to enter METIS information (about research programmes and projects) into the information system (VSOI) that forms the basis for the Dutch Research Database (NOD) via an export of the data in a specific format. Universities can send their information by mail, CD-rom or FTP. Through a specific module in VSOI the research information from the universities is easily imported into the Dutch Research Database. This database is one of the sources of NARCIS.

3.2 The Development of an Exchange Schema

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 an XML schema was needed. The implementation of this schema makes it possible to gather automatically current research information (programmes and projects) from the data providers NWO/Delfi and METIS in a uniform way. The definitions of the schema are determined by KNAW, NWO and RU-UCI (METIS). For the NARCIS project XML-SOAP (Simple Object Access Protocol) is being used. SOAP is a lightweight protocol for exchange of information in a decentralized, distributed environment. By introducing XML-SOAP on the NARCIS server, the server of the NWO and the METIS servers, a sufficient data exchange between the local systems is created. When NWO 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. So, there is an exchange of information without

anybody actually having to do something manually. When the first publications of such a research project are available in METIS the metadata of these publications are sent back to NWO and to the repository of the institute. In this way the researcher needs to deliver his publication only once.

3.3 Harvesting of Repositories by Using OAI-PMH

Since a few years the Dutch Universities and other research institutes (NWO and KNAW) developed their own repositories with full text publications and other output of their scientific research. These research results are stored in open archives, which contain XML documents. NARCIS has developed a data service on all these repositories. There is a connection between the NARCIS portal and the different repositories, and through the use of OAI-PMH, Open Archives Initiative - Protocol for Metadata Harvesting, all the documents from the repositories are being harvested and included in the portal. By doing this it is possible to search scientific results at the same time as searching research projects or expertise of researchers.

3.4 Installation of a Spider Tool

To gather information from web pages we have used a different technique: within the framework of the NARCIS project a web crawler, or spider, has been developed. J-Spider, an open source tool, was being used as a plug-in for the NARCIS portal. J-Spider has been adjusted to spider also PDF or RTF format, because many websites contain valuable information in these formats.

Also an interface has been developed to produce spider tasks, which can be tailored to special needs. These spider tasks make it possible to tell the crawler to visit a certain website, or even go to a specific part of that website, and store the full text into the NARCIS index. So these pages with scientific news or full text publications are fully searchable in NARCIS.

For this project the performance of J-spider 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, the pages were no longer downloaded at this stage, but the full text was being indexed and URL was being saved. The result is that there were fewer problems with the memory capacity; as a result 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. With the spider tool it is possible to gather information on websites, for example news items, information on data sets (from the DANS website), web reports et cetera.

3.5 Development of a Web Tool for Automatic Categorising

After gathering of the scientific information in those different ways a tool for documenting and searching the information was needed. The purpose was to give the user of the portal the possibility to make searches not only using words but also by category. Because there are too many items in NARCIS to add categories or thesaurus terms manually, part of the project was the development of a categorizer.

The categories of the classification of the NOD were used as a basis for this tool. This classification, with 259 categories, exists of subject fields and scientific disciplines for all sciences. The categorizer was based on the functionalities of the open source search engine Lucene. The 'Similarity part' of Lucene was being used: finding the documents that look like the training set. These training sets contain each 50 descriptions of current research that are representative for each category. On the basis of similarity with the training set, the categorizer decides in which category or categories the new document belongs. These categories are automatically added to the documents in the portal.

The categorizer tool gives rather good results. However, it depends on the document type how good the categorizer works. For example web pages give a better result than OAI records. An explanation for this is that the size of OAI records is very small. It is desirable that the categorizer gives good results for all the different types of information. In a follow-up project, NARCIS 2, the categorizer will be improved or another categorizer will be chosen.

3.6 Building a Portal as a Central Facility for Making Available the Different Types of Information

The main goal of NARCIS is to create a central facility for searching all the scientific information produced by Dutch universities, research institutes, KNAW and NWO. This facility has been made by developing a bilingual (Dutch and English) research information portal, with the URL www.narcis.info (see Fig. 3) [3, 4]. The portal (a prototype version) contains about 400,000 items and one can find information on current research programmes and projects, researchers (addresses and expertise), research institutes, (full text) publications, data sets, and news items.

Concerning the appearance of the website, people are used to search at simple looking web search engines like Google, and are used to fill in just one or two keywords to find information. This idea has been used as starting point to make a simple looking website with the possibility to search by words. It is also possible to make use of the limiting options in the search bar. One can restrict the search action to the different items 'Persons', 'Organisation', 'Current research', 'Publications (www)', 'Publications"(DARE), 'News', and 'Data sets'. Due to the categorizer it is also possible to search by subject fields or disciplines.

The user can make searches by using Dutch or English words. However, the results will be different depending on the language used, because the research information is available in the language in which is it has been published. For example, the user will not find a publication written in English by using a Dutch keyword.

To give the portal a more personalized character a notifying tool has been added to the portal. The user can make an RSS feed on the content of NARCIS. This feature gives the end-user a possibility to follow the latest developments on the specific subject. Furthermore one can arrange the RSS feed in such a way that one will be informed only on one or some selected types of information (for example exclusively new full text publications).

Figure 3: The homepage of NARCIS after some changes as a result of the first end-user test [www.narcis.info]

4 The Preliminary End-user Test

A few months after the portal has been built, it became time for an end-user test. A small group of end users have tested the NARCIS website in January 2006. This end-user test has been carried out among Dutch information specialists. These specific users for the test were selected as they all have much expertise in the field of information retrieval and they know the situation of the scientific landscape in the Netherlands. Consequently, they would tackle other problems than 'normal' end-users. This quick test gave the possibility to make some first changes to the website. A broader users test, also involving different groups of users, has been planned for spring this year. In this test there were questions about three topics concerning the 'look and feel', and the user-friendliness of the portal:

1. 'Look and feel' of the NARCIS home page
2. Search functionality and limiting options
3. Display of the several record types after searching

The users were asked about these topics via a questionnaire sent by email. It took them about 20 minutes to fill in the form. The users were guided through the site and the questions were partly 'closed' with limited answering possibilities and partly 'open', where one could give some remarks.

4.1 Look and Feel of the NARCIS Home Page

In general the 'clean' appearance of the home page ('looks like Google') is well appreciated and the availability of the RSS-feed is considered as an added value. There are some criticisms as well. Some users made remarks on the length of the explanatory text (too long), and the unwanted necessity of scrolling on the homepage. There were also questions about the option 'now popular'. This option refers to subjects/persons that have raised many search activity during the last month. It should give a notion of the actual topics in research information, but most of the users did not find it relevant. After analyzing the results of this test, this option has already been deleted from the website, with the consequence that there is little need to scroll. The explanatory text on the homepage will be shortened and the language button will be moved to a more visible place.
The results of this test have already led to changes in the homepage. Other issues, like criticism on the presentation of the limiting options, will be tested in the extended end-user test.

4.2 Search Functionality and Limiting Options

The search options and search performance are very well appreciated. After searching one can use the 'search within results' feature at the bottom-site of the page. But the users did not see this bar or did not understand limiting options. Therefore, all the search options will be offered in one single bar at the top of the page.
The users all used the limiting options in the search bar. The other limiting option, by just clicking on the record type mentioned in each individual record (for example click on 'person' and one will see all the persons on a subject), is not understood (see Fig. 4). The way the record type is being displayed makes a user think it is just a type indicator and not a limiting option. This remark will be tested in the broader survey.

4.3 Display of the Several Record Types after Searching

In the results set the different record types are shown with icons. Not all the icons are understood by the user group, so this will be changed (see Fig. 4). Another remark concerns the representation of the record types. In NARCIS the different record types are not be represented in the same manner. In the case of structured information (for instance NOD or OAI record) clicking a record in the results list will lead to a full record displayed within NARCIS. With 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. The first representation is clear, but the users do not understand the second one. Reason for this is the unfamiliarity with this kind of representation. It is clear that it is necessary to explain this typical NARCIS feature. Also the users would like to see information about the volume of the full documents, before they start downloading. This feature will soon be implemented.

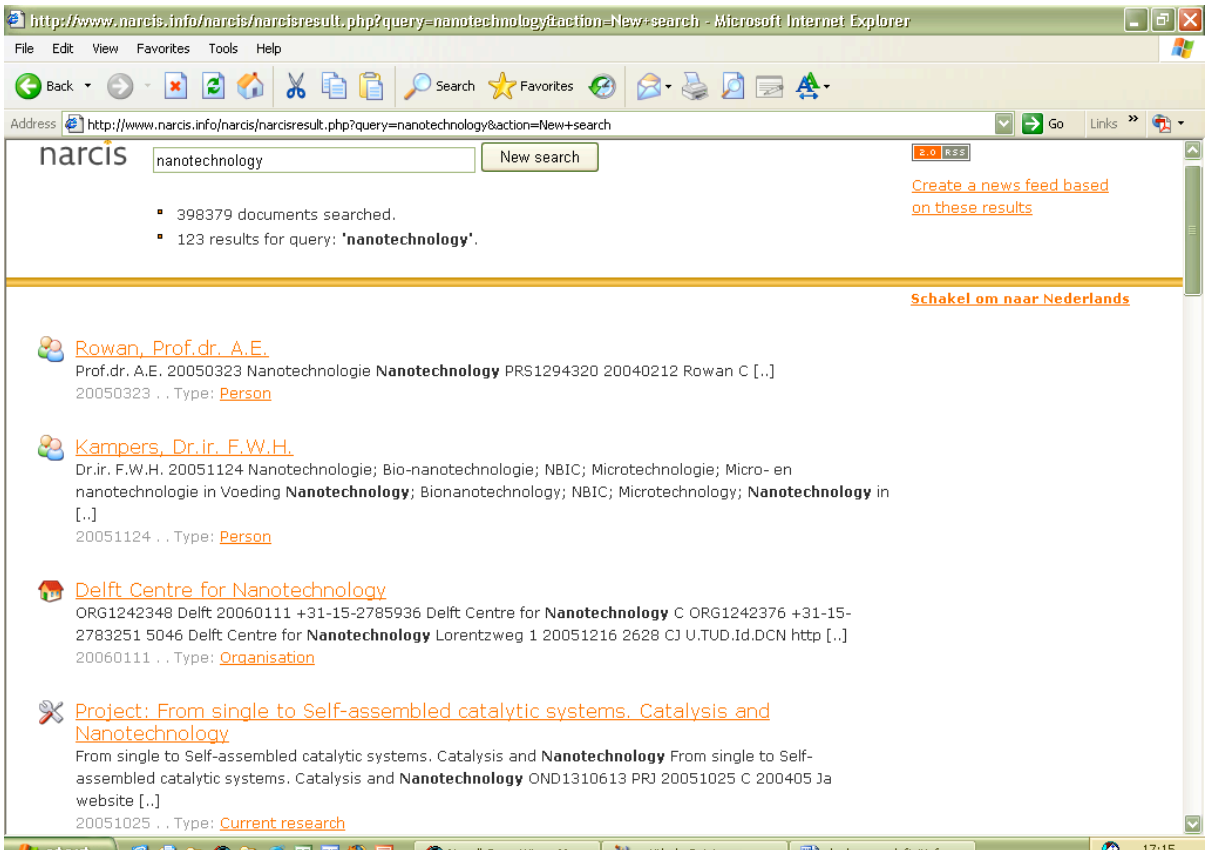

Figure 4: Display of records after searching

5 Conclusions and Further Developments in NARCIS 2

5.1 The Content of the Portal

NARCIS – WWW.NARCIS.INFO - is the entry for scientists, policy makers, intermediary organisations, journalists and the public for obtaining an overview on research in the Netherlands. The portal contains information on research institutes, researchers, and research activities publications (metadata, and full text), data sets (metadata) and news (web pages). The 400,000 items in the portal are searchable at the same time (one-stop-shopping), but one can also search the different information types separately.

5.2 The Topicality of the Portal

The information one can find in NARCIS is up-to-date, because the information is being gathered in an early stage of the registration process of the scientific institutes. In 2006 the automatic exchange of data between the universities and NARCIS will be possible. At the moment the exchange of data takes place by means of a not structured text file. By introducing the XML schema which had been developed in NARCIS the exchange of data will be improved. And due to the developed interface it is quite easy to make new spider tasks, so NARCIS will be adapted to actual scientific developments.

5.3 Quality of Information

One knows that the quality of the information is high, because scientific information specialists select the sources of NARCIS. All information is originating from university and non-university scientific institutes. Thus the results of searching in the NARCIS differ from using a search engine like Google.

5.4 The Volume of NARCIS

The number of items in NARCIS will grow. Within the scope of the Dutch DARE programme the universities put extra effort in the growth of their repositories. Also, in the future it will be possible for scientific organisations to offer their URL for spidering. Of course an editor will review these URLs, before making a new spider task. Another 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 METIS systems 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.

5.5 Improvement of the Used techniques

The web crawling and the integration of the web crawled information in NARCIS is functioning already. The NARCIS crawler is the open source J-SPIDER, which is available via:

http://j-spider.sourceforge.net/other/index.html

However, as explained before, the automatic classification is not yet satisfactory. In a following project, NARCIS 2, an experiment with another, commercial, automatic classification tool will take place. This tool, Collexis, 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. If this test does not work the NARCIS categorizer will be further developed.

5.6 Unique Digital Identifiers

A problem in NARCIS is the identification of the unstructured information. For example, how can one connect a researcher and his expertise originating from the Dutch Research Database with his or her publication in a repository? As a part of the DARE programme unique digital identifiers for authors, objects and institutes will be developed. This will connect all sources and will facilitate the exchange of information within the NARCIS portal.

5.7 End-user Test

A preliminary users test has been carried out among Dutch information specialists. As a result some adaptations are made to the NARCIS website. This ‘new’ website will be tested in a broader users test that will be carried out this spring. In this more extended test about the ‘look and feel’ and the user-friendliness of the search facilities, also the content of NARCIS will also be questioned.

5.8 NARCIS Consortium and Business Plan

A part of the project is the cooperation between the most important scientific institutes in the Netherlands. To continue this relation a NARCIS consortium is formed in which the partners are working together to develop NARCIS further. In June 2006 a business plan will be ready.

References

[1] van MEEL, A. M. NARCIS. National Academic Research and Collaborations Information System: Eindverslag. KNAW, Amsterdam, The Netherlands, 2005, 21 p.

[2] http://www.onderzoekinformatie.nl/nl/oi/onderzoeksprojecten/narcis/eindverslag/Eindverslag_NARCIS.pdf

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

[4] NARCIS project information on the website of KNAW Research Information http://www.onderzoekinformatie.nl/en/oi/onderzoeksprojecten/narcis