

Creating an Academic Information Domain: a Dutch Example

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

Department of Research Information,
Royal Netherlands Academy of Arts and Sciences (KNAW)

Summary

The KNAW presents the outcome of different projects which have resulted in the first steps towards an Academic Information Domain (AID) in the Netherlands. All Dutch academic repositories, the content of Current Research Information Systems (or CRIS's) and the datasets available are brought together in one portal.

When creating an AID, problems such as interoperability and the relation between the data within the different systems have to be solved.

The relation between the data of different systems will be made by implementing DAI (Digital Author Identification). The DAI project, in which each individual researcher is given a unique number, will make it possible to create links between the different kinds of information and Persistent Identifiers are to be introduced to guarantee sustainable access to the data.

1 Introduction

The mission of the Royal Netherlands Academy of Arts and Sciences (KNAW) is to ensure the quality of scientific research in the Netherlands. The Academy also promotes the open accessibility of scientific information. The main task of the Academy's Research Information department is to be the national focal point for research information in the Netherlands.

For this purpose the department has been producing the Dutch Research Database (the NOD), which is the national Current Research Information System. This database contains information about current research, researchers and research institutes. The department is also responsible for the DAREnet website, which gives open access to Dutch full-text publications, and NARCIS, a portal containing scientific information from the Netherlands. In NARCIS the information from the NOD and DAREnet is combined, and extended with scientific news items and datasets.

Within Knowledge Exchange (KE)¹, a collaboration of DFG, JISC, DEFF and SURF, a new concept has been developed for connecting information from CRIS's (Current Research Information Systems) and OAR's (Open Access Repositories): the Academic Information Domain (AID). In this paper the focus will be on the realisation of a Dutch Academic Information Domain. As can be seen in Figure 1, the Academic Information Do-

¹ <http://www.knowledge-exchange.info/>

main can be distinguished from the Personnel Information Domain or the Financial Information Domain².

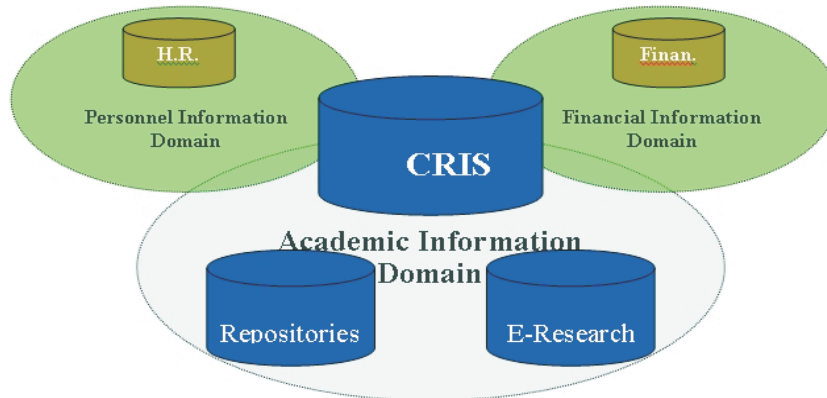


Figure 1: The Academic Information Domain

CRIS's are information systems containing an extensive set of metadata covering various aspects of research information. They are in use in universities, research institutes, and/or governmental bodies and were initially developed for administrative purposes. An OAR contains, for example, full text publications and research data (in the so-called e-Research repository³). Many of the repositories are managed at a single-institution level.

NARCIS can be seen as a first step to the Dutch Academic Information Domain with information from the national CRIS, the NOD, full text publications from the academic repositories, and research data in the fields of the humanities and social sciences from DANS.

To develop the Dutch Academic Information Domain, which links information from the academic Current Research Information Systems and the Open Archive Repositories, Digital Author Identifiers (DAI) and Persistent Identifiers (PI) have been introduced into the NARCIS.

2 Three different worlds

In order to create a Dutch Academic Information Domain three different worlds have to be joined in one portal. A clear distinction can be made between the area of research information, the output of research, such as publications, and that of research data.

Research information focuses on describing programs and projects, the researchers and the organisational units. Research information offers a lot of detail on the level of the institution. Typically, these descriptions are made by research administrators, mostly operating

2 Razum, M.; Simons, E. and Horstmann, W. (2007) Exchanging Research Information. Strand Report of the Institutional Repositories workshop, February 2007.

3 Jeffery, K.G. (2007). Technical Infrastructure and Policy Framework for maximising the benefits from Research Output. Proceedings ELPUB2007 Conference on Electronic Publishing, Vienna, Austria, June 2007.

separately from library staff, and are stored in an institutional Current Research Information System (CRIS).

Research publications consist of books, journal articles and conference papers that have been published as a result of research. This information has an international character and the freely accessible part of it is available via Institutional Repositories (IR). The description of research publications has always been a task for library staff. Thanks to the OAI-PMH protocol, contents of IR may be harvested by service providers, so that access on a national level can be offered (DAREnet, The Netherlands; IRI, Scotland; HAL, France).

Research data, or e-data, are the raw data resulting from research and may be used to write books, articles and so on. In a way, research data may be considered as a typical class of research publications. But there are major differences: research data are normally published according to less fixed (international) rules. It is also not always to the benefit of the researcher to make research data publicly available. Therefore it is not unusual that a researcher keeps the data in his/her own laboratory.

2.1 CRIS's and the Dutch Research Database (NOD)

The Dutch Research Database (NOD)⁴ is produced by the Research Information department of the KNAW. This database is the national Current Research Information System (or CRIS). The NOD is a database which is publicly available online with information on scientific research (research programmes and research projects), researchers (with their working addresses and their area of expertise), and research institutes (with profiles). 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.

The NOD contains information on 36,000 researchers of whom are 7,500 professors, and 9,100 researchers with expertise. One can find 37,000 descriptions of research programmes and projects, of which 20,000 are current. There is also information on 3,200 research institutes.

The NOD is fully accessible by search engines such as Google. This has led to an average of 170,000 unique visitors, and 240,000 visits per month. About half of the users access the NOD via a search engine.

2.2 Open Access Repositories (OAR) with full-text publications

In the Netherlands the four-year DARE programme was launched in 2003. DARE is short for Digital Academic REpositories. The mission of the DARE programme was to get better access to results of publicly funded academic research in the Netherlands and it allowed authors to post their publications in an institutional academic repository.

DARE started with a budget of almost six million euros. All the universities in the Netherlands participated, as well as the Academy, the Netherlands Organisation for Scientific Research, the most important funding organisation, and the National Library. In the first year of this programme an infrastructure of institutional academic repositories was set up, based upon the Open Archives Initiative Protocol for Metadata Harvesting.

4 <http://www.researchinformation.nl>

In 2004 and 2005 the focus was on populating the repositories. Within a year, almost 49,000 objects had been uploaded, not only metadata but also full text. In May 2005 the second milestone was reached: Cream of Science. Cream of Science is the showcase of top Dutch research. All DARE partners selected ten of their prominent academics whose complete publication lists were stored in a special OAR set, with as much full text as possible. One of the prime aims of Cream of Science is to open up top quality content to the scientific community and society at large, and make it more easily and electronically accessible. Another aim is to demonstrate that scholars are willing to post their materials to a repository. In the end over 200 authors were willing to cooperate.

Another project was Promise of Science. The aim of Promise of Science was to set up a national doctoral e-thesis gateway and populate this gateway with 10,000 full-text e-theses before the end of 2006. The total annual production of doctoral theses is around 2,500 items, which equals 5% of the formal scholarly output in the Netherlands.

All these initiatives led to a national infrastructure of OAR's. In addition the different projects act as a catalyst for the process of getting the repositories filled with content.

At this moment (February 2008) together the repositories contain nearly 150,000 full-text scientific publications and research output from all the Dutch universities, some scientific institutes, the KNAW and the NWO. Cream of Science contains 46,000 scientific publications written by 229 prominent Dutch scientists. About 60% of these can be accessed as full text. Promise of Science gives access to over 16,000 doctoral e-theses from all the Dutch universities.

2.3 E-Data in EASY

EASY⁵ (Electronic Archiving SYstem) gives access to a large number of research datasets, or so-called e-data. EASY is maintained by DANS – Data Archiving and Networked Services – a KNAW institute, which is supported by the Dutch Research Council (NWO).

EASY provides an infrastructure for the deposit and permanent accessibility of e-data, especially in alpha and gamma science. The datasets relate mostly to research that has already been completed and published, and the authors have made their basic data available to other researchers. Data and documentation can be downloaded free of charge. For some datasets the depositor's permission is necessary before data can be downloaded.

3 Joining three worlds in NARCIS

It is a real challenge to join three different worlds in one system. There are many technical problems to be solved, as well as organisational problems. Issues such as formats, interoperability, presentation and compatibility need to be addressed. This part of the paper will focus on some of these aspects.

5 <http://easy.dans.knaw.nl/dms>

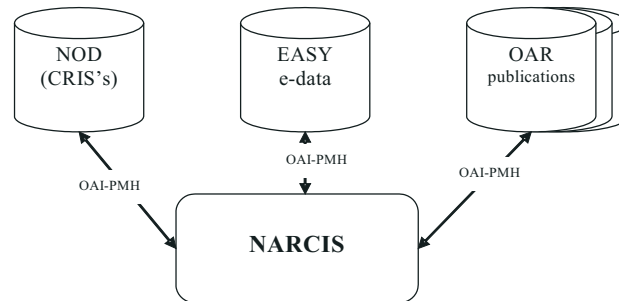


Figure 2: Joining three worlds in NARCIS

3.1 Organisational aspects

In every project organisational aspects are very important. In projects that are aiming to create an Academic Information Domain on a national level technical issues are very important, but organisational aspects are of equally importance. The NARCIS project could benefit from the organisational network created during the DARE programme.

To carry out this programme a national network of different working groups in which every institute participated was organised. DARE is followed by the SURFshare⁶ national programme (2007-2010), coordinated by the SURFfoundation. Through the Surfshare-programme, the SURFfoundation wishes to establish a joint infrastructure that advances accessibility as well as the exchange of scientific information. Within the framework of SURFshare the different working groups established during the DARE programme have been maintained, and the participants see each other regularly to make appointments, discuss problems, etc.

Every institution has participated in the strategic working group where future plans are considered. The important technical working group talked about technical issues such as changes in format, protocol implementation, software, etc, which need to be discussed for the creation of uniformity. In another working group organisational matters are discussed and agreements about the quality and size of the content are made. The participants in this working group ensure that the deadlines for implementing new technical issues or filling the repositories will be reached. Without these working groups the developments of the last years would never have been accomplished.

It is fair to say that, without good organisational structure and agreements about metadata and technical issues, it is impossible to create an AID on national level. Even on the level of just one academic institution these organisational aspects are very important.

3.2 Technical issues

In the following section the most urgent issues such as interoperability, formats and presentation will be addressed.

⁶ For more information about the SURFshare programme, coordinated by the SURFfoundation, see the website: <http://www.surfoundation.nl/smartsite.dws?ch=ENG&id=5463>

3.2.1 Interoperability: OAI-PMH protocol

To gather the information from the different systems the OAI-PMH protocol is used. There is much discussion about the best techniques to facilitate interoperability. In ‘CERIF – Information Retrieval of Research Information in a Distributed Heterogeneous Environment’⁷ the authors gave requirements for information retrieval from distributed heterogeneous systems: ease of implementation, flexibility, support of open standards, effectiveness, and solving problems such as semantic and structural interoperability.

Since the DARE community had already implemented the OAI-PMH protocol and most of the above requirements were met by this protocol, it was a small step to implement the protocol for all systems included in NARCIS.

For NARCIS there are many advantages to using the OAI-PMH interface:

- a) new repositories or systems can be added to NARCIS very easily;
- b) only the metadata is harvested, not the complete database or repository;
- c) it is possible to harvest heterogeneous databases and repositories with one protocol; only mapping with the index needs to be changed when a different XML schema is used.

Interoperability is a very complex matter. Through OAI-PMH, NARCIS builds an index of the metadata stored in all Dutch OAR, the Dutch Research Database and e-data. If the end user wants to see an item, the actual data is retrieved from the different systems at that moment.

3.2.2 Metadata formats and CERIF

Metadata format for OAR & e-data

The original repositories all started off using the OAI-PMH protocol to achieve interoperability. The protocol was developed in 1995 and defines a limited set of HTTP requests a harvester can send to a repository. The OAI-PMH protocol requires the usage of the Dublin Core (DC) metadata format for describing objects in the repository.

DC is very restricted; it only contains 15 metadata elements, but this restrictiveness has also its advantages. It allows easy exchange of metadata between repositories and harvesters.

Unfortunately, DC hinders the further progress of developments, for instance when an exchange of more complex metadata is needed.

Since 1995, newer metadata formats have been developed, offering more possibilities. Within the Dutch infrastructure the need was felt for a more advanced XML schema to describe compound objects. After much debate, the project participants chose the Metadata Object Description Schema (MODS)⁸ for describing bibliographic metadata elements. The main reasons for choosing MODS were:

- a) User-friendliness
- b) A simple extension mechanism
- c) The availability of mapping from MODS to DC

7 CERIF – Information Retrieval of Research Information in a Distributed Heterogeneous Environment, Andrei Lopatenko UM, Anne Asserson UiB, Keith G. Jeffery CLRC

8 For more information see: <http://www.loc.gov/standards/mods/>

There was an especially strong need for an extensible schema in the Netherlands because of the introduction of a totally new element: the Digital Author Identifier or DAI (see section 4). The DAI is used to relate all kinds of variants of a specific author name to one single identifier. Of course, this typical Dutch element is no part of standard metadata formats or schemas. MODS offers the possibility to simply add this element.

Metadata format for CRIS

For the harvesting of the CRIS (NOD) records, a CERIF-based format is used. Only the fields that are put in the index of the system are harvested, thus only those fields that are searchable within NARCIS. If someone wants to see the complete record, the information is retrieved directly from the CRIS. So the CERIF-based XML format is used only for those fields to be stored in the index.

3.2.3 Presentation

Presentation on the website can be a real problem, especially consistency in the presentation of different content types. A portal like NARCIS has many content types such as current research, persons, organisations, publications and datasets. One advantage of the NARCIS portal is the possibility for one-stop-shopping: it is possible to search all types at the same time. On the other hand this presents a problem for some advanced searches.

For instance, for the presentation of publications there is always a need for ranking by date, rather than instead of ranking by relevance. But for persons and organisations ranking by date makes no sense and is not even possible.

Another problem comes with the Dutch open access policy. Not all publications but only full-text open access publications are stored in the OAR's. This results in incomplete lists of publications from certain researchers, while everyone expects a full list and wonders why the list is not complete.

4 Digital Author Identifier (DAI)

Not only in the Netherlands but also in other countries, the CRIS's and OAR's of the institutions are not yet related to each other. The reason for this is the fact that in most cases the CRIS's are maintained by the departments responsible for research administration, while the OAR's are maintained by libraries. The e-data are often not stored at all; in the Dutch case, they are stored in a separate system maintained by the specialised institute DANS.

To connect the information in all these systems, the first step will be to assign a Digital Author Identifier (DAI) to researchers in the Netherlands. Research data, publications and research descriptions may all have the name of the creator, author or project leader in common, but often the names of authors are not consistent in the different information systems. Also within one specific information system there can be many author name variations.

As part of the SURFshare programme (the successor of the DARE programme) a project to import the unique names of authors/researchers into the academic CRIS (the so-called METIS) was started last year. Within the OCLC-PICA library system a thesaurus of author names (40,000) with corresponding DAI's has been created. All the Dutch universities, the KNAW and the NWO have matched the DAI's and author names in their own CRIS. This work was finished at the beginning of 2008. The next step will be the implementation of the

DAI in the repositories of the participants. The DAI will also be imported in the NOD, in the database of DANS (EASY) and in the e-depot of the National Library. This implementation project will be conducted by the Academy in co-operation with the University of Utrecht and SURFfoundation.

NARCIS connects and gives access through the DAI

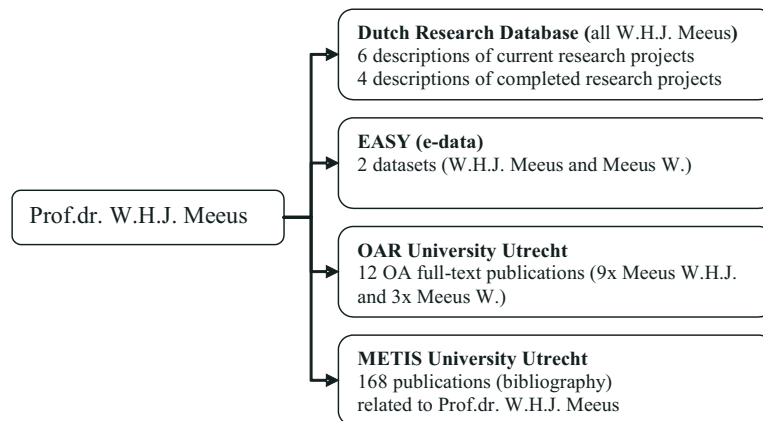


Figure 3: Examples of DAI variants

Because the DAI has to be harvested together with the other metadata elements, the SURFshare participants had to agree on a standard in the XML. The participants decided to use the Metadata Object Description Schema (or MODS). This has been developed by the Library of Congress' Network Development and MARC Standards Office. Unfortunately, the guidelines for the MODS are not yet ready. To take advantage of the MODS schema at an early stage, the participants will use Mini-MODS, a small subset of MODS.

5 Persistent Identifiers and sustainability

By using the DAI, the relationship between research descriptions and research results (datasets and publications) becomes visible, but sustainability is not yet guaranteed. For instance, objects may shift from one repository to another or even may disappear. In less dramatic circumstances, objects may be removed from a university repository, but will continue to be available in the e-depot of the national library. In these situations, the user of systems like NARCIS will be confronted with lost information (broken links), unless an automatic referral to the new location of the object is offered.

Persistent Identifiers (PI's) are necessary to develop such a referral system. Using PI's, objects will no longer be identified by the unstable Uniform Resource Locator (URL), but by a stable Uniform Resource Name (URN). The advantage is clear: when an object shifts to another repository, its URL will change, but its URN will remain the same.

Assigning URN's to objects is, however, not enough. A new service is needed that 'links' the URN to the actual location (URL) of those objects. This is called a 'resolver service'. In

the Netherlands, the DANS institute has actually developed a resolver. Thanks to the DANS resolver, an object is always traceable when it moves from one location to another. It is anticipated that the URN will take over the role of the URL in scientific citations. This will ensure that scholars will in principle always be guided to the object they want to refer to.

There are many different systems of assigning PI's. In Holland, the URN:NBN system has been chosen. This system is a concept of CDNL and CENL and has further been developed by the German National Library. In the concept NBN stands for National Bibliography Number. The fact is that the National Library is responsible for the distribution of URN's.

An example for such a number is: URN:NBN:NL:UI:28-15123. In this example, the string URN:NBN:NL:UI:28 refers to objects from the University of Twente and 15123 is the unique object number.

6 Conclusions and further development

6.1 Conclusions

With the present infrastructure and systems there is a good foundation to build on. At the moment it is safe to draw the following conclusions:

- a) To create an AID, organisational problems are often overlooked, and may be even more difficult to solve than technical problems;
- b) Although the implementation of a Digital Author Identifier (AID) seems an easy task, in practice it is complicated and far more work than assumed;
- c) Commitment from all parties involved is extremely important to creating the necessary infrastructure;
- d) Persistent Identifiers are necessary for sustainability and guaranteed access to the right information.

6.2 Further developments

6.2.1 Enhanced publications

Of course, these developments in interconnecting research related information take place on a European level too. The most promising project in this respect is DRIVER-II. DRIVER stands for Digital Repository Infrastructure Vision for European Research. The project is an initiative of 11 European countries, with the University of Athens as its co-ordinator. The project has received a grant from the European Commission (7th Framework). The KNAW and all Dutch universities are participating in DRIVER-II, united within a so-called Joint Research Unit (JRU).

The main goal of DRIVER-II is the realisation of an infrastructure that will result in a European confederation of modern repositories, in which non-textual objects (images, datasets and presentations) can also be deposited. As a spin-off, DRIVER-II aims to construct 'enhanced publications'. An enhanced publication consists of a traditional publication (for instance, a report) and the objects interrelated to these publications. Here, one may

think of a presentation, a dataset or a project description. The relationship with the development of the Academic Information Domain is clear.

6.2.2 Metadata store

At the moment the OAI-PMH protocol only stores harvested XML in the index, but there is a need for storing the metadata physically. The implementation of a metadata store gives advantages in situations where a new index process is needed. It also gives the opportunity to create a repository on a national level since all the metadata are available within the portal.

6.2.3 Validation tool

Among the participants in NARCIS and other owners of repositories, there is a need for validation of the DIDL 2.3 format. Within the framework of DRIVER-II a validation tool was developed. For the Dutch situation only the validation rules have to be adapted to DIDL 2.3.

The tool must be implemented in a way that is user-friendly. The web interface will be made where one needs this to enter the base-URL and an email address. After validation a status report will be sent to the given email.

6.2.4 Tag clouds and keywords

Searching large collections of information always brings the problem of categorisation and the assignment of keywords. One of the problems with automation of this process is the lack of sufficient content. During other experiments we found that titles and short abstracts are not enough for automated classification, but due to a combination of different collections, such as METIS and OAR's, the amount of content has increased enormously. Through the DAI it is possible to gather information from all these collections. In a preliminary experiment information from publication, descriptions of research, information from expert databases, etc, is being used to generate a tag cloud with expertise. More work needs to be done to generate the expertise precisely.



Figure 4: Tag cloud of expertise

Acronyms

- AID: Academic Information Domain
- DAI: Digital Author Identifier
- DANS: Data Archiving and Networked Services
- DARE: Digital Academic Repositories (programme)
- DEFF: Danmarks Elektroniske Fag- og Forskningsbibliotek

DFG: Deutsche Forschungsgemeinschaft
EASY: Electronic Archiving System
JISC: Joint Information Systems Committee
KB: National Library of the Netherlands
NARCIS: National Academic Research and Collaborations Information System
KNAW: Royal Netherlands Academy of Sciences
NOD: Dutch Research Database
NWO: Netherlands Organisation for Scientific Research
SURF: Dutch higher education and research partnership organisation

References

- Dijk, E.; Baars, Chr.; Hogenaar, A.; Meel, M. van (2006). NARCIS: The Gateway to Dutch Scientific Information. Paper presented at the Elpub 2006 Conference, Banskó, 14-16 June 2007.
http://elpub.scix.net/data/works/att/233_elpub2006.content.pdf
- A. Hogenaar (2007) 'Joining three worlds: research information, research data, and research publications', International Internet Librarian Conference, London, 8-9 October, 2007.
- Theo van Veen (2007). Persistent Identifiers. Information Professional, 7/8 2007.
- Razum, M.; Simons, E. and Horstmann, W. (2007) Exchanging Research Information. Strand Report of the Institutional Repositories workshop, February 2007.
- Jeffery, K.G. (2007). Technical Infrastructure and Policy Framework for maximising the benefits from Research Output. Proceedings ELPUB2007 Conference on Electronic Publishing, Vienna, Austria, June 2007.
- CERIF – Information Retrieval of Research Information in a Distributed Heterogeneous Environment, Andrei Lopatenko UM, Anne Asserson UiB, Keith G Jeffery CLRC
<http://www.ub.uib.no/avdeling/fdok/cris/Taskgroups/CERIF-IRofRIinaDistributedHE20020503.rtf>

Contact Information

Chris Baars
Department of Research Information
Royal Netherlands Academy of Arts and Sciences (KNAW)
PO Box 95110
1090 HC Amsterdam
The Netherlands
chris.baars@bureau.knaw.nl
www.onderzoekinformatie.nl