



Names Project

Final Report

Amanda Hill, Daniel Needham, Alan Danskin

July 2009

Acknowledgements

The first phase of the Names Project was funded by JISC as part of the Repositories and Preservation Programme between July 2007 and February 2009. The project partners were Mimas at the University of Manchester and The British Library.

Executive Summary

The Names Project began in July 2007. It was funded to investigate requirements for a name authority service for UK repositories. Prototype name authority software has been developed as part of this work and a number of connections have been made with UK stakeholders and with international projects working in a similar space.

Prototype development

The prototype¹ has been developed using an iterative approach due to the shifting nature of requirements and exploratory findings. An initial software requirements specification was derived using the outcomes of the requirements gathering phase, followed by design and development work which has been running in parallel, with input from external developers and stakeholders helping shape its course.

Initial prototype work has focused on several main areas.

- 1) A database has been created, required to store name authority records, based on the entities defined in the Data Analysis and *Functional Requirements for Authority Data*² (FRAD) mappings outcomes.
- 2) A back end data collection and disambiguation application is under ongoing development, to acquire data from a variety of sources and identify unique entities within them with which to populate the database.
- 3) A web interface is under ongoing development, working with external partners, to provide machine to machine access to the database, with the creation of an API to provide easy, standardised, flexible querying of the service.
- 4) A web based human search interface has been developed to allow human searching of the names records, and also aid in testing of the prototype.
- 5) A client script has been developed in conjunction with Cranfield University, in order to prototype automated methods of externally retrieving data from the Names service for use in other applications.

All of the above work is still ongoing.

Stakeholder engagement

The name authority area is of interest in a number of different domains, some of which are actively working on solutions to the reliable identification of individuals and institutions. During the project's lifetime the project team have been in contact with UK funding councils and name authority service developers from Australia, New Zealand and the USA, as well as the UK repository community (which is the principal target audience for this project).

Note on terminology

For concision, the term "author" has been used throughout the document. For the purposes of this report this usage should be understood to include:

- names of any party associated with the resource, irrespective of role, e.g. author, composer, contributor, publisher, editor, sponsor
- names of different types of entity, such as names of persons or names of institutions
- names used to identify the subject of a resource

¹ Available at <http://130.88.120.172:8080/axis/index.jsp>

² *Functional Requirements for Authority Data: a conceptual model.* / Edited by Glenn E. Patton. IFLA Working Group on Functional Requirements and Numbering of Authority Records (FRANAR). München: K.G. Saur, 2009. IFLA Series on Bibliographic Control, vol. 34. ISBN 978-3-598-24282-3

Background

The problem of uniquely identifying authors has been with us ever since books have been catalogued. National libraries have been creating name authority files for many years, starting with card catalogues and now maintaining electronic files in MARC format. However, authority files for the authors of journal articles and unpublished materials, including many electronic resources, do not always exist in library systems. The increasing use of subject-based and institutional repositories to hold working papers, reports, research data, and pre-refereed and post-refereed versions of articles has led to a corresponding rise in the number of authors identified in such systems.

Without having a means of uniquely and unambiguously identifying those involved in the creation of materials in repositories, it becomes difficult to be sure whether all the materials related to a particular person will be correctly associated with that individual. Names of authors may be entered in more than one way, or more than one author may have exactly the same name.

Issues related to the reliable identification of authors were identified in the subject-based repositories once they reached a significant size:

Searching by authors' names has been among the top search methods by repository users. When a repository grows to substantial size, it is often the case that name variants cause headaches for both the users and repository managers.³

Where repositories contain relatively few items, the problems associated with loss of precision (the ability to retrieve only the items created by a particular individual) and loss of recall (the ability to retrieve all the items for that individual) may not be particularly noticeable and may be managed by the intervention of repository administrators. When repositories are large (or when the contents of different repositories are aggregated), or where machine to machine operation is required, these issues become more prominent. If author names are not controlled, then a search for a particular name will only retrieve items which match the query's form of the name exactly, creating a loss of recall. If more than one author has the same name, then precision will also be affected, with irrelevant material being returned for a search.

A number of JISC-funded reports and scholarly papers have identified a name authority service as a desirable element in repository infrastructure.⁴ As a consequence, the Repositories and Preservation Programme call in September 2006 had a specific requirement for an investigation into:

...the potential for the development of a Name Authority Service and factual authority for digital repositories, to support cataloguing, metadata creation and resource discovery in the repository environment.

The proposal for the Names project was submitted in response to this part of the call. Work began on the project in July 2007 and the project ended in February 2009 (although work has continued since then, as the project has received continuation funding from the JISC).

³ Xia, J. 2006. Personal name identification in the practices of digital repositories. *Program:Electronic Library & Information Systems*, 2006, 40(3): pp.256-267. Available at <http://dlist.sir.arizona.edu/1832/>.

⁴ For example:Chapman, A. & Russell, R. 2006. *JISC Shared Infrastructure Services Synthesis Study: A review of the shared infrastructure for the JISC Information Environment*Jones, C.et al. 2008. *Report of the Subject and Institutional Repositories Interactions Study*Salo, D. 2009. Name authority control in institutional repositories. *Cataloging and Classification Quarterly* 47:3/4 (April 2009)

Aims and Objectives

The agreed aims and objectives for the project were described in the project plan in this way:

The Names project will investigate the requirements of the UK's repository community for a name authority service and will develop a demonstrator. This process will involve the analysis of existing name authority solutions and the identification of sources of relevant data. The team will determine the types of metadata which will be required for the service and decide whether existing standards are suitable for use within the prototype.⁵

Methodology

The project was divided into two broad phases. The first phase focused on gathering information about existing name authority services and standards, and on scoping the requirements for a service that would meet the needs of the UK repository community. In the second phase of the project, work began on building a prototype to demonstrate one possible approach to building a service that would meet the identified requirements.

The first phase of the project called upon the team to review a range of services and standards. These included library authority resources, subject repositories' name authority solutions and publishers' name authority services. The information thus gathered was published in a Landscape Report that was jointly authored by staff at the British Library and at Mimas. Requirements of the project's stakeholders were gathered through a combination of meetings and email correspondence in the early part of 2008.

A significant area of work in the scoping phase was Alan Danskin's identification and analysis of the necessary data elements for a name authority service. This involved looking in detail at the International Federation of Library Association's *Functional Requirements for Authority Data* and mapping various existing name authority standards to the elements so identified. This was published as the Data Analysis Report.

The second phase of the project began with establishing the software requirements for the prototype. This involved reviewing the stakeholder requirements and data analysis documents and drawing up a specification based on these. This phase continued with the development of the prototype, which is described in more detail in the Implementation section below.

Implementation

The work on the Landscape Report had suggested that the best approach towards developing a useful name authority service would involve a combination of pre-populating a database with information from existing journal article and conference paper details and then allowing individuals to manage and improve the records that had been so created. The project partners are responsible for providing the Zetoc service,⁶ which contains information about millions of journal articles and conference papers, so it seemed sensible to use a subset of data from Zetoc to populate the prototype.

The Zetoc data include author names, article titles, subject classifications and keywords. Additional information has been taken from the Wellcome Trust's grantees database from the UK PubMed Central service. The project has developed a disambiguation algorithm based on these information elements as a way of generating name data for the prototype. This algorithm groups together

⁵ Names Project Plan, October 2007. Available at http://names.mimas.ac.uk/documents/Names_project_plan_v4_Oct07_web.pdf

⁶ Available at <http://zetoc.mimas.ac.uk/>.

information from multiple records that appear to relate to the same individual and generates an authority record which holds that information.

Initially a database was established based on the entities identified in the Data Analysis Report, to create a structure that was capable of storing unique records, as well as allowing for their easy search, manipulation and deletion as required. In order to allow for the provision of foreign names the database uses the UTF-8 character set.

Once the database had been established work began on the back end data collection and disambiguation application. Here work began on a central application, developed in Java, that used collection and disambiguation algorithms tailored to different data sources to gather the data and process it to populate the database.

A process for assigning unique identifiers for records within the system has been developed and an initial author management system has been created which will allow individuals to manage the data that relates to them within the prototype.

The initial data sources we are working with include:

- 1) Zetoc – using the SOAP interface to gather data
- 2) UKPMC – using data provided in an excel extract
- 3) NACO records – provided in XML format.
- 4) Open access listings – using data provided in an excel extract.
- 5) HESA institutional identifiers – using an external library to parse their online identifier listings.

In order to debug and test the effectiveness of the disambiguation process it was necessary to create the web interface in parallel. A JSP search interface was created that receives queries via the request query string, defined in an application programming interface to provide machine access to the records. This API allows for output in a variety of formats, including HTML, comma-separated values, MARC-XML, JSON and a Names-specific format. The current version of the API has been documented and is publicly available at <http://130.88.120.172:8080/help.html>.

On top of this API an html interface has been built to allow easy human search and examination of the records, this search tool is available at <http://names.mimas.ac.uk/prototype/>. Similarly test client applications have been built, in conjunction with Cranfield University, in PHP, JSP, and Javascript using Ajax, to test remote retrieval of data from the Names prototype and its utilisation within external applications. An example of the test script can be found at <http://names.mimas.ac.uk/script-test/>.

This development work is continuing in the second phase of the project.

Outputs and Results

The scoping activities undertaken in the first phase of the project resulted in the publication of a number of reports which are all available from the project's website.⁷

The Landscape Report describes the external context for name authority work in the repository area, looking at a range of existing commercial and repository solutions to the name authority problem and describing the related standards that are being used by libraries, archives and repositories. The subsequent Requirements Report analysed information from the project's various stakeholders to determine the necessary features for a name authority service aimed at repositories. Those requirements were carried through into the Software Requirements Specification, which also incorporated a number of usage scenarios for a name authority service.

A number of presentations on the project have been given, with various members of the project team involved. Some of these have been aimed directly at repository developers, for instance the meeting in Bath organised by the IE Demonstrator Project and the Common Repository Interfaces Group on 6

⁷ Reports are all at <http://names.mimas.ac.uk/documents/>.

June 2008.⁸ A more formal paper was given at the ISKO 2008 conference in Montreal.⁹ All the project's presentations are available from the website.¹⁰

As has already been explained, the work undertaken in the Data Analysis Report formed the basis for the structure of the Names prototype's database. The second phase of the project has been focused on building this prototype and testing the functionality of its API with colleagues at Cranfield University.¹¹

Outcomes

The initial objectives set for the Names project have been met and the prototype is now being extended (in a new phase of the project) into a pilot system which will be developed, in consultation with the repository community, into an increasingly useful tool.

The reports generated by the project team have been favourably received.¹²

A consequence of the work undertaken by the project has been Involvement in a number of international collaborations related to name authority and repository initiatives. These include representation on the Networking Names Advisory Group for OCLC's Identities Hub¹³ and on the NISO Institutional Identifiers group.¹⁴ Meetings have been held with colleagues at the British Library who are involved with work on the International Standard Name Identifier (ISNI).¹⁵ The project was also represented at the International Repositories Workshop held in Amsterdam in March 2009, where progress was made on establishing requirements for an interoperable identification infrastructure for repositories.¹⁶

The issue of unique identification for researchers has been an active topic, particularly amongst bio-scientists, during the course of the Names project.¹⁷ The project manager has joined a Linked-In group on Unique Identifiers for Researchers set up by Cameron Neylon as part of this discussion, to raise awareness of the work that we are doing.

Conclusions

The unique and unambiguous identification of researchers in order to track grant outputs and reward their work is increasingly being recognised as an important area of work. The prototype developed by the Names project demonstrates one possible solution to some of the problems in this area. Its approach of seeding the database with existing name data drawn from Zetoc and other sources shows that it is possible to generate a considerable corpus of name authority data automatically. This data can then be shared with the wider community and its reliability and completeness improved upon by others.

⁸ http://www.ukoln.ac.uk/repositories/digirep/index/CRIG_DRY_Workshop

⁹ The paper is available in the JISC IE Repository at <http://ie-repository.jisc.ac.uk/154/>.

¹⁰ <http://names.mimas.ac.uk/documents/>

¹¹ The prototype is available at <http://names.mimas.ac.uk/prototype/>.

¹² Coverage includes Lorcan Dempsey's blog: <http://orweblog.oclc.org/archives/001445.html> and the CrossRef blog: http://www.crossref.org/CrossTech/2007/10/the_names_project.html. The Landscape Report has also been translated into Japanese by the National Institute of Informatics in Japan (<http://namesproject.wordpress.com/2008/09/08/landscape-report-now-available-in-japanese/>).

¹³ <http://www.oclc.org/programs/ourwork/renovating/leveragevocab/netgroup.htm>

¹⁴ <http://www.niso.org/workrooms/i2>

¹⁵ International Standard Name Identifier: <http://www.isni.org>

¹⁶ <http://www.ukoln.ac.uk/events/ir-workshop-2009/>

¹⁷ See, for example: <http://blog.openwetware.org/scienceintheopen/2009/01/20/a-specialist-openid-service-to-provide-unique-researcher-ids/>; <http://www.gen2phen.org/researcher-identification-primer/author-names-and-authorship-scientific-publications>; <http://bjoern.brembs.net/news.php?item.493>;

Implications

The Names project prototype is a starting point for the development of a more comprehensive name authority system for use by repositories and other online services. Interest in the project has been high within the repository community, with many repository managers and developers keen to find a solution to problems associated with unique identification of creators of repository materials. The continuation of the Names project into another phase will allow the project team to build upon the achievements of the first phase in collaboration with the wider repository community.

References

Chapman, A. & Russell, R. 2006. *JISC Shared Infrastructure Services Synthesis Study: A review of the shared infrastructure for the JISC Information Environment*

Common Repository Interface Group DRY (Don't Repeat Yourself) meeting wiki:
http://www.ukoln.ac.uk/repositories/digirep/index/CRIG_DRY_Workshop

Hill, A. 2008. What's in a Name? Prototyping a name authority service for UK repositories', paper presented at ISKO conference, Montreal, August 2008: <http://ie-repository.jisc.ac.uk/154/>

International Standard Name Identifier (ISNI): <http://www.isni.org/>

Jones, C. et al. 2008. *Report of the Subject and Institutional Repositories Interactions Study*

Names Project Plan, October 2007. Available at
http://names.mimas.ac.uk/documents/Names_project_plan_v4_Oct07_web.pdf

Names Project Prototype: <http://names.mimas.ac.uk/prototype/>

NISO Institutional Identifiers Working Group: <http://www.niso.org/workrooms/i2>

OCLC Networking Names Advisory Group:
<http://www.oclc.org/programs/ourwork/renovating/leveragevocab/netgroup.htm>

Salo, D. 2009. Name authority control in institutional repositories. *Cataloging and Classification Quarterly* 47:3/4 (April 2009)

Zetoc service, <http://zetoc.mimas.ac.uk/>.