ARTICLE EXCERPTED FROM:

ISQ INFORMATION STANDARDS QUARTERLY

SUMMER 2011 | VOL 23 | ISSUE 3 | ISSN 1041-0031

TOPIC: ORGANIZATION AND PEOPLE IDENTIFIERS

INTERNATIONAL STANDARD NAME IDENTIFIER (ISNI)

THE OPEN RESEARCHER & CONTRIBUTOR ID (ORCID)

THE NAMES PROJECT

STANDARD ADDRESS NUMBER (SAN)

INSTITUTIONAL IDENTIFIERS (I^2) AND ISNI

NISO
How the information world CONNECTS

Alan Danskin

Amanda Hill

Daniel Needham

The Names Project: A New Approach to Name Authority

ALAN DANSKIN, AMANDA HILL, AND DANIEL NEEDHAM

Unique, unambiguous identification of researchers has become a hot topic in recent years, with a number of initiatives now under way to solve the problem. It is a well-known fact that personal names are not sufficient as a means of distinguishing between individuals: there may be more than one person with the same name and, if only initials are used (as is the case in many bibliographic databases), this problem is compounded.

A person's name may also change during the course of his or her lifetime, or be represented in subtly different ways, creating a need to link all those alternative forms of name together in order to be sure that the materials produced by that individual can be reliably identified and (if desired) collected together.

The Names Project was funded in 2007 as part of the JISC's Repositories and Preservation Programme. It had been recognized that:

> *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.*[Xia]

When the contents of different repositories are aggregated, these problems in retrieving all (and only) relevant materials are compounded. The JISC's call for proposals in September 2006 therefore had a specific requirement for a project which would investigate:

> *…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.*[JISC]

A joint bid submitted by Mimas (at The University of Manchester) and the British Library was successful and work began on the project in July 2007. Early activities included identifying the requirements of the repository community and reviewing the work of existing projects and services in the field of researcher identification and name authority. National libraries have been creating name authority files for authors of books for many years, starting with card catalogs and now maintaining electronic files in MARC format. However, authority files for the creators of journal articles

and other electronic resources often do not 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.

Assessment of research activity has been a significant part of life for researchers employed in the UK's universities since the introduction of the first "research selectivity exercise" in 1985-86.[Day] Research Assessment Exercises (RAEs) have been carried out in 1989, 1992, 1996, 2001, and 2008, with peer-reviewed analysis of publications forming an important part of the process. It seems likely that measurements of the impact of research will be taken into account in the RAE's successor, the Research Excellence Framework (REF). With the increased availability of research materials in repositories and other online locations, it is becoming vital that researchers are reliably associated with their publications in order to show how the outputs of a particular individual are being used and affecting the work of others. Uniquely identifying researchers would assist in this process.

View from Mimas

Mimas is a national data center based at The University of Manchester. The department has a history of providing innovative ways of connecting users with research information and developing technological infrastructure services to support UK academic researchers. Finding a solution to the researcher-identification problem would help to support these objectives.

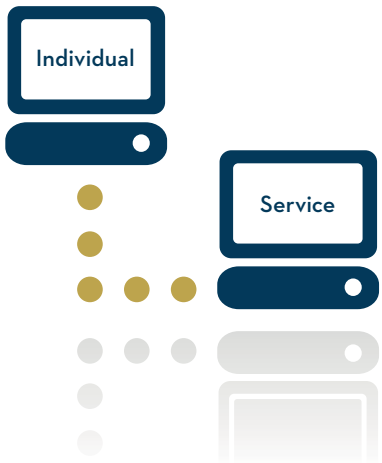
One of the initial aims and ongoing activities of the project has been the design and development of a prototype name authority service for individuals and institutions in order to demonstrate the feasibility of such a system. The original design of the prototype was based upon the project's landscape report, stakeholder requirements-gathering exercises, and consultation with the developer community. Subsequent development of the prototype has been an iterative process, due to the dynamic environment within which it needs to fit and the changing requirements of the varying stakeholders.

The prototype was envisaged as a piece of middleware, comprising of a store of name authority records created using a data model designed by the British Library, and an API through which the records could be queried by external services. In order for the service to be viable, the first thing that was required was a large data set around which the service could be built.

With no available pre-existing set of data pertaining to individuals and their identities, it was necessary to build our own records from scratch. Two approaches to building such a data set were identified: firstly by acquiring access to external data sources containing information relating to individuals and institutions and attempting to automatically disambiguate the unique entities within them, and secondly by providing functionality for individuals and institutions to contribute data to the service themselves. It was determined that the former approach would be more appropriate in the initial phase of prototype development, with the latter being introduced at a later stage.

The prototype was designed to be as flexible as possible in how it acquires and processes external data, which would be accessible in varying formats and provide diverse types of information. For this reason the disambiguation side of the prototype comprises two logical sections. The first is a collection of data source handlers, each of which is tailored to the specific external data source it relates to.

CONTINUED »



TWO APPROACHES TO BUILDING A DATA SET:

1. Acquiring access to external data sources containing information relating to individuals and institutions and attempting to automatically disambiguate the unique entities within them

2. Providing functionality for individuals and institutions to contribute data to the service themselves

Where a match is found with an existing Names record, that record is updated with any new information. Some of the attributes which have proved most useful for matching, aside from entity names, include collaborative relationships, publication title keywords, fields of interest, and institutional affiliation.

This provides the functionality to pull in data from the specified source and then convert it into a collection of internal Names record objects. Data source handlers can be added each time a new external source is acquired. The second logical component accepts these converted Names record objects and attempts to identify and disambiguate the unique entities contained within the source data, first against other records from that source (if necessary), and then against any existing Names records containing information derived from a number of different sources that have the potential to match.

In order to analyze entities identified in different records for potential matches, a disambiguation algorithm examines the various attributes of the entities in question. By using pre-configurable thresholds for the differing matching criteria, the algorithm attempts to determine whether or not the compared entities match according to these rules. Where no match is found, a new record is created for that entity within our own database and assigned a unique persistent identifier. Where a match is found with an existing Names record, that record is updated with any new information. Some of the attributes which have proved most useful for matching, aside from entity names, include collaborative relationships, publication title keywords, fields of interest, and institutional affiliation.

Initially our main external source of data came from Zetoc, the British Library's web interface to its Electronic Table of Contents (ETOC), which is hosted by Mimas. Zetoc provided around 38 million records, containing bibliographic information associated with journal articles and conference papers—a broad basis upon which to build our core record store. Most of the initial development work was carried out using this data. Whilst initial results from the disambiguation process seemed promising, scaling up management and testing of such a large data set proved problematic given the resources we were working with.

Consequently, we began to look at alternative data sources including MERIT data from the 2008 RAE. The JISC-funded MERIT project worked on cleaning up the information submitted to the RAE and the resulting data set, containing information on 45,000 of the UK's top researchers, was a more manageable size to work with. After creating a data source handler to process the MERIT data, we successfully disambiguated the entire data set with a very high level of accuracy, eventually eliminating all mismatched records. In May 2011, following a period of quality assurance carried out by the British Library (described in more detail below), the Names records derived from MERIT were made permanent, and an export of our data was sent to the International Standard Name Identifier (ISNI) initiative, to be matched against their records.

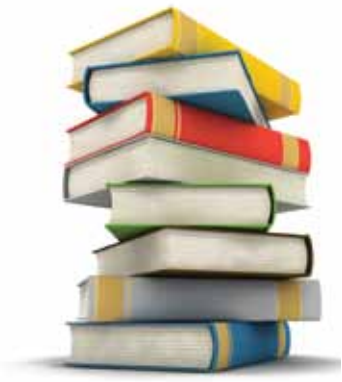
Subsequent to establishing a base set of Names records from the MERIT data, we are now looking at matching the derived records against a subset of the Zetoc data. We are also attempting to match against data exposed by the RDF output of the EPrints repository software, using The University of Southampton's repository as a test case.

In order to provide access to the Names records that had been created during the disambiguation process, it was necessary to design and implement an API through which queries could be made. The API needed to provide facilities to search over the Names records and return results in a flexible way to meet the varying requirements of the stakeholders. Initially SOAP was chosen as the means of providing data search and retrieval; however, following feedback from the JISC developer community during the initial stages of the project, it was decided that a RESTful approach would be more appropriate for the audience that would be using it.

Each Names record is assigned a unique identifier, and this identifier is a resolvable URL as part of the API. The API also provides functionality to search for records using a variety of criteria and returns the results with differing levels of detail and in different output formats. Consequently the API provides a robust and flexible method for searching Names records and can easily be integrated into external systems with minimal effort. This has been demonstrated in both a search interface developed for testing and demonstration of the capabilities of the prototype, as well as an example tool that was written to illustrate the way that external services could use the API to autocomplete a name field in a form with Names data.

The development of the prototype will be ongoing, with the aim of increasing the quality and quantity of records as well as the functionality that the API provides for interacting with them. As part of this work we will be looking at acquiring new data sources to process, as well as refining our disambiguation algorithm to increase the accuracy of results. We will also be reviewing the API and working with external services and repositories to facilitate integration between their applications and the prototype.

The number of new books published in the UK and received by the British Library through legal deposit is about 130,000 per annum; the number of journal articles added to the ETOC system is approximately 2.5 million per annum.



130,000
PER ANNUM

2.5 MILLION
PER ANNUM

View from the British Library

The British Library is the national library of the United Kingdom and is one of the six UK and Irish libraries entitled to receive UK publications under legal deposit. The British Library is a partner in the Names project because control of the names of authors and other contributors to publications is an important and expensive element of cataloging items for the collection.

One of the functions of the library catalog is to enable a user to find "all resources associated with a given person, family or corporate body." [IFLA] To satisfy this requirement it is necessary to identify each individual entity uniquely and to provide links between the variant forms of names by which they are known. In a library context, these functions of identification, disambiguation, and linking are provided by the name authority file. In current cataloging practice the focus is on disambiguation of entries (headings) in a browse index. In a web context, where library metadata has to mix with metadata from other domains, entities have to be explicitly identified to enable joined up services. The Names project has engaged with initiatives developing international identifiers for researchers, including the ORCID Initiative and ISNI.

The way in which authority control is done by libraries is challenged by audience expectations and by the volume of resources that will require authority control. The focus on controlling the authors of printed books no longer satisfies the needs of researchers, who want journal articles, conference papers, data sets, pre-prints, and other resources. The number of new books published in the UK and received by the British Library through legal deposit is about 130,000 per annum; the number of journal articles added to the ETOC system is approximately 2.5 million per annum. Manual processes are not scalable to meet this demand. Automation or semi-automation of authority control processes would enable the British Library to identify individuals in ETOC records and link these identities to existing authority files.

The library has contributed its expertise and metadata to the Names project by contributing to development of the data model, specifying mappings to output formats, and testing of samples of metadata disambiguated by Names.

Testing has been conducted in three main phases, described in more detail below. All of the testing done at the British Library involved evaluation of sample data by catalogers following normal practices used by authority control staff to identify and disambiguate individuals of the same name. Catalogers consulted external sources, predominantly institutional or personal websites, to confirm that identifications made by Names are secure and accurate. The manual review by British Library supplemented extensive testing and validation carried out by Mimas.

CONTINUED »

1

Sample	Number of ETOC records	Number of distinct identities	Identities represented in NACO file (overlap %)
BIRTWHISTLE	140	41	5 (12%)
MCDAID	256	54	3 (5.5%)

Data analysis

The ETOC data numbers approximately 38 million article records. To gain an understanding of the data, samples were extracted for specific personal names. Selecting specific names meant that levels of duplication and disambiguation could be evaluated. The overlap with the LC/NACO authority file was also evaluated and the low incidence of matching influenced the decision to defer loading LC/NACO to Names.

The findings confirmed assumptions about the ambiguity of the data. For example, the name "Birtwhistle, G." concealed seven different identities. The analysis also confirmed expectations that there would be many entries for the same person. For example, "Birtwhistle, G. #2" was associated with sixteen different article records. The low frequency of matching with NACO confirmed that articles receive very little attention from authority control catalogers in other institutions and disambiguation would have value beyond the British Library.

2

	Automated Outputs	Manual Reviewer Findings
Unmatched Identities	48	9 of these could have been matched with data in the records. 4 could only have been matched using information external to the Names processes.
Identities Established	71	—
Mismatches	34	31 of these could have been disambiguated using data in the record. 3 could only be disambiguated using data external to the process.
Correct Matches	96	—
Ambiguous Matches	168	Insufficient data to be certain that matches are correct.
Total Candidate Identities »	298	

Evaluation of disambiguation

A sample file was prepared to enable comparison of automated outputs with manual authority control. The sample consisted of 375 article records associated with the name C. Abbot. Manual review of the sample took approximately four weeks.

The manual review process was time consuming, but very valuable. It highlighted the limitations of the ETOC data for matching, but more positively a problem with the weighting given to subject classification numbers (Dewey Decimal Classification) was identified, which when adjusted improved the results.

Disambiguation of MERIT data 3

Source file size	45,000
Potential Mismatches	33
Actual Mismatches	2
Non-Matches	2257
Non-Matches reviewed	227
Actual Matches identified	7

The MERIT data is a much smaller and better controlled set than the article records and includes institutional affiliations and researcher IDs, neither of which is present in the ETOC records. For this phase of review, the results were filtered to identify possible mismatches or non-matches. Mismatches occur when names of two different individuals are matched incorrectly. The "non-matches" were records for which several individuals with the same or very similar name were identified, but not matched.

There were too many non-matches to review them all manually; therefore every 10th result was reviewed. Mismatches are considered to be more serious than non-matches and all the potential mismatches were reviewed. Only two genuine mismatches were identified. In one of these cases the individuals concerned turned out to be twins who worked at different institutions but had the same initial and family name and were associated with the same paper. One result of this discovery has been the adjustment of the algorithms to prevent matches between individuals with the same name who have collaborated on the same paper.

The MERIT data, deduplicated and disambiguated by Names, provides a core of reliable identities for UK researchers and academics against which other data sets, such as article records, can be matched. The MERIT records have been exported as the first contribution by Names to the ISNI database.

Manual review of the Names outputs highlighted the importance of human inspection of the results. Sampling and filtering created manageable workloads for catalogers, cutting the time to review representative samples from weeks to days. Input from the reviewers has improved the matching and disambiguation algorithms. Future services will require an element of human review to resolve ambiguities and for quality assurance.

Conclusion

Unique identification of researchers is an area of intense interest in the UK and beyond. The Names Project team has aimed to test the feasibility of a service that would provide disambiguation and identification of researchers and make the resulting records available to the wider research community. The work of the project has produced a core set of disambiguated researcher identifiers, accessible through a flexible API, which could be used as the basis of a future name authority service. Plans for further enhancement of the service would include allowing the researchers themselves (or their representatives) to supply information that would improve the accuracy of the data in the system, and further collaboration with related international initiatives such as ORCID and ISNI. | IP | doi: 10.3789/isqv23n3.2011.04

ALAN DANSKIN <alan.danskin@bl.uk> is Metadata Standards Manager at The British Library. **AMANDA HILL** <amanda@hillbraith.com> is an Archival Consultant who manages the Names project on behalf of Mimas at the University of Manchester. **DANIEL NEEDHAM** <daniel.needham@manchester.ac.uk> is Technical Officer for Mimas.

RELEVANT LINKS

British Library. *Value of knowledge: Annual report & accounts 2009/10.*
www.bl.uk/about/annual/2009to2010/performance/performancestats.html#bibliographic

Day, M. *Institutional repositories and research assessment.* Project report. UKOLN, University of Bath, 2004.
opus.bath.ac.uk/23308/1/eprintsuk-rae-study.pdf

EPrints
www.eprints.org/

IFLA. *Statement of International Cataloguing Principles. 4. Objects and functions of the catalogue. 2009.*
www.ifla.org/files/cataloguing/icp/icp_2009-en.pdf

ISNI: International Standard Name Identifier
www.isni.org/

JISC Repositories and Preservation Programme
www.jisc.ac.uk/whatwedo/programmes/reppres.aspx

MERIT project
www.jisc-collections.ac.uk/News/merit-db/

Mimas
mimas.ac.uk/

Names autocomplete form
names.mimas.ac.uk/names/script-test-two/

Names search interface
names.mimas.ac.uk/advanced-search.php

ORCID: Open Researcher & Contributor ID
www.orcid.org/

RESTful: Representational State Transfer
en.wikipedia.org/wiki/Representational_State_Transfer

SOAP specification
www.w3.org/TR/soap/

University of Southampton ePrints
eprints.soton.ac.uk/

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

Zetoc service
zetoc.mimas.ac.uk/