JISC

Project Document Cover Sheet

Project Information			
Project Title	Names Project		
Start Date	1 December 2011	**End Date**	31 July 2013
Lead Institution	The University of Manchester		
Project Manager & contact details	Amanda Hill Hillbraith Ltd. Amanda@hillbraith.com		
Partner Institutions	The British Liibrary		
Project Web URL	http://names.mimas.ac.uk/		
Programme Name	*Digital Infrastructure: Directions*		
Programme Manager	Verena Weigert		

Document Name			
Document Title	*Final Report*		
Author(s) & project role	Amanda Hill, project manager		
Date	July 2013	**Filename**	Names_Phase_Three_Final_Report_Jul13.pdf
URL			
Access	☐ Project and JISC internal	☑ General dissemination	

Document History		
Version	**Date**	**Comments**
1.0	15 Jul 2013	Initial draft, AH
2.0	16 Jul 2013	Incorporating comments from DN
3.0	22 Jul 2013	Added UI screenshot, minor edits
4.0	29 Jul 2013	Added BL perspective

Names Project, Phase Three

Final Report

Alan Danskin, Amanda Hill, Daniel Needham

July 2013

Table of Contents

Acknowledgements

Phase Three of the Names Project was funded by JISC as part of the Digital Infrastructure: Directions Programme.

The project partners are Mimas at the University of Manchester and The British Library.

Executive Summary

The Names project team have developed a set of high-quality data on 50,000 UK researchers and their institutions, with approximately 20% of UK researchers now represented. They have made that data available to the International Standard Name Identifier service, ISNI, and have linked Names records to their ISNI counterparts.[1] The team have freely shared their data, their experience, their expertise, and their software with others and have formed an important part of the researcher identification exploratory landscape in the past six years.

The approach taken by Names shows that the problems associated with the identification of researchers in legacy systems can be resolved through a combination of automated matching and manual quality assurance. The combination of this approach with the automatic allocation of identifiers to new researchers offers a possible solution to the researcher identification problem in the future.

Phase Three of the Names Project continued work which began in July 2007. The original project was funded to investigate requirements for a name authority service for UK repositories which would uniquely and persistently identify individuals active in research. A prototype identifier system was developed as part of this work and this continued to be improved and have data added to it during Phase Two. Phase Three comprised two separate projects: the Option Appraisal and System Enhancement work which ran from 1 December 2011 to 31 December 2012 and the extension of the project from 1 January 2013 to 31 July 2013.

Work in this phase of the project fell into the following categories:

Options Appraisal report

This report was delivered to Jisc in November 2012 and presented a range of possible future uses for the Names system and data, including options for integration with the international initiatives of ISNI and ORCID. It was discussed at a meeting in December 2012.

System enhancement

Between 1 December 2011 and 31 July 2013 the project team continued to work on adding new data to Names, refining the Names interface and algorithms and adding functionality to the Names web interface. Some 3,750 new individual identifiers were added to the Names system during this reporting period, bringing the total number of individuals with unique identifiers to just short of 50,000.

System architecture revision

During this phase a significant process of refactoring has been carried out on the back end software that is used to disambiguate individuals and manage the Names database. In order to make the entire system more reusable the disambiguation and database components have been separated, making them entirely independent, and meaning that they can be repurposed in a number of ways in other systems. These components have been shared on a source hosting system.

Stakeholder engagement

The Names project has been represented at a number of meetings and conferences during this reporting period. Updates on the project's progress continue to be shared through the project's blog and Twitter feed and the team have continued to work with colleagues at ISNI to ensure that individuals identified in Names are allocated international identifiers.

[1] http://isni.org/

Background

The Names project was initiated by Jisc in 2007. The project team at Mimas and the British Library was tasked with investigating requirements for a name authority service for UK repositories which would uniquely and persistently identify individuals and institutions active in research in this country. A prototype name authority system was built and additional funding since 2009 has extended the prototype into a pilot system which has created identifiers for more than 50,000 individuals and organisations. Information within the Names system is freely available to other systems through a flexible API.

As part of the development of the pilot system the project team have refined a disambiguation process which combines automatic matching of new data sources against existing records with a manual quality assurance process undertaken at the British Library. This has resulted in a high-quality set of information on UK researchers, the majority of whom were included in the 2008 Research Assessment Exercise (and who can therefore be seen as the 'cream' of UK academic researchers).

Additional information on researchers and their scholarly outputs has been received from a range of UK institutional repositories, including Robert Gordon's University, the London School of Economics, the University of the West of England, the University of Huddersfield and the Open University. The Names team have worked together with institutional repository managers at these institutions to identify their affiliated individuals and to improve repository data.

The Names team have been involved in the work of a number of external committees related to the identification of researchers and research institutions, including NISO's I[2] (Institutional Identifiers) and OCLC's former Networking Names Advisory Group and current Registering Researchers in Authority Files Task Group. They have also been involved in providing information to Jisc's Researcher ID Task and Finish Group and have published a number of reports and articles on researcher identifiers.

International situation

A number of countries have developed national name authority systems for researchers[2] and Names is the UK's approach to investigating such a system. At the same time as Names was under development, international initiatives also began to emerge to solve some of the same problems. ISNI evolved from national libraries' long-standing work in this area, while ORCID[3] developed from publisher efforts to identify authors of journal articles.

The two international systems have very different approaches to identifier assignment: ORCID relies on a researcher or the researcher's institution to claim an identifier, while ISNI uses existing sources of data and registration agencies to establish an identity and then assign an identifier. ORCIDs fit well into the workflows of publishers who are dealing with a known individual. ISNI records generally carry more comprehensive information about individuals than ORCID ones (most of which contain only a name and an identifier) and are therefore more useful for third parties such as cultural heritage institutions who may need to identify a person or organisation with whom they are not in direct contact.

[2] The report *National Approaches to Researcher Identifier Systems* is available from http://ie-repository.jisc.ac.uk/567/1/Report_on_approaches_taken_in_national_researcher_name_authority_initiatives.pdf
[3] http://orcid.org/

Aims and Objectives

The agreed aim for this phase of the project was described in the first project plan in this way:

> *To produce for JISC a robust, evidence-based options appraisal on the potential roles for a bibliographic-oriented UK national researcher identifier service in the context of the recommendations of the work of the UK Researcher Identifier Task and Finish Group*
>
> *To continue to enhance the existing Names pilot system by increasing the quantity and quality of records held within it, informed by the options appraisal report*

The extension to this phase aimed to:

> *…bring Names to a point where its main elements can be re-used by other services in a semi-automated manner, while also holding open the possibility that those elements could move forward into a service of some kind.*

Methodology and Outputs

The project was divided into four main areas of work, reflecting the objectives listed above and also including engaging with the project's stakeholder communities.

Options Appraisal Report

The project team completed the Options Appraisal Report in November 2012 and presented it at a meeting at the Jisc offices in December of that year. The report was a result of discussions with colleagues at ISNI and ORCID about the role which a national system like Names could play in relation to international name identification services. A number of possible ways forward were presented in the report, from basic maintenance of the current Names data through to a service which could liaise between UK institutions and ISNI and ORCID.

System enhancement

New data has been added to Names from four institutional repositories in this period:

1) University of Huddersfield
2) University of the West of England
3) London School of Economics
4) Open University

In addition to processing this data and assigning unique identifiers to the individuals represented in those repositories, the project team have been working to ensure that the existing Names data is represented in the international system being developed by ISNI. Working in conjunction with ISNI and colleagues at the British Library, ISNIs have been assigned to individuals currently uniquely identified within Names, with the exception of the Open University data which has only just been added (and which will be processed in the coming months by ISNI). Colleagues working on the ISNI data recognise that the disambiguated Names records are of a high quality[4] and are easy to incorporate into the ISNI database.

A further enhancement to the Names web interface has been to alter the ISNI identifiers so that they now link directly to individual records in the ISNI system.

It was planned that the organisations within Names which do not already have ISNIs would be

[4] Quality of metadata is measured in a number of ways, including: completeness, provenance, accuracy, consistency and accessibility. See http://blog.law.cornell.edu/voxpop/2013/01/24/metadata-quality-in-a-linked-data-context/ for further analysis.

assigned them in this period. However, plans within ISNI for identifying organisations have since altered: ISNI will be working to assign identifiers to the institutions currently identified by Ringgold IDs.[5] ISNI have undertaken to supply these new ISNIs back to Names for the institutions already identified in the Names system, deprecating those already assigned by ISNI. These ISNI identifiers will also be used by the ORCID system.

System architecture revision

The revised Names System Architecture is made up of a number of independent components that can be used on their own, integrated into other applications, or used in conjunction to replicate a Names service.

These components represent the culmination of the development work carried out to build the Names prototype over the last five years. The revision has been performed to allow the system to be more portable in case of re-use or redeployment. In addition several improvements have been made to the database architecture to improve speed of response and maintenance.

All of the various components are written in Java. They will be fully documented by the end of July 2013 and made available on bitbucket.org. The components are:

1) Disambiguator

A library that provides functionality for normalising metadata pertaining to individuals or institutions from a source data set, and then returns match scores between these normalised records.

2) Database Manager

A library that acts as an interface between a Names Database instance and another component.

3) Match Service

A web application that accepts a list of individuals and their metadata, and returns a table of match scores for these different individuals. Uses the Disambiguator library.

4) Names User Interface

A web application providing search functionality and Names identifier resolution over a Names Database instance. Uses the Database Manager library.

[5] Press release from Ringgold: http://ringgold.com/pages/pr_270613_ORCID.html

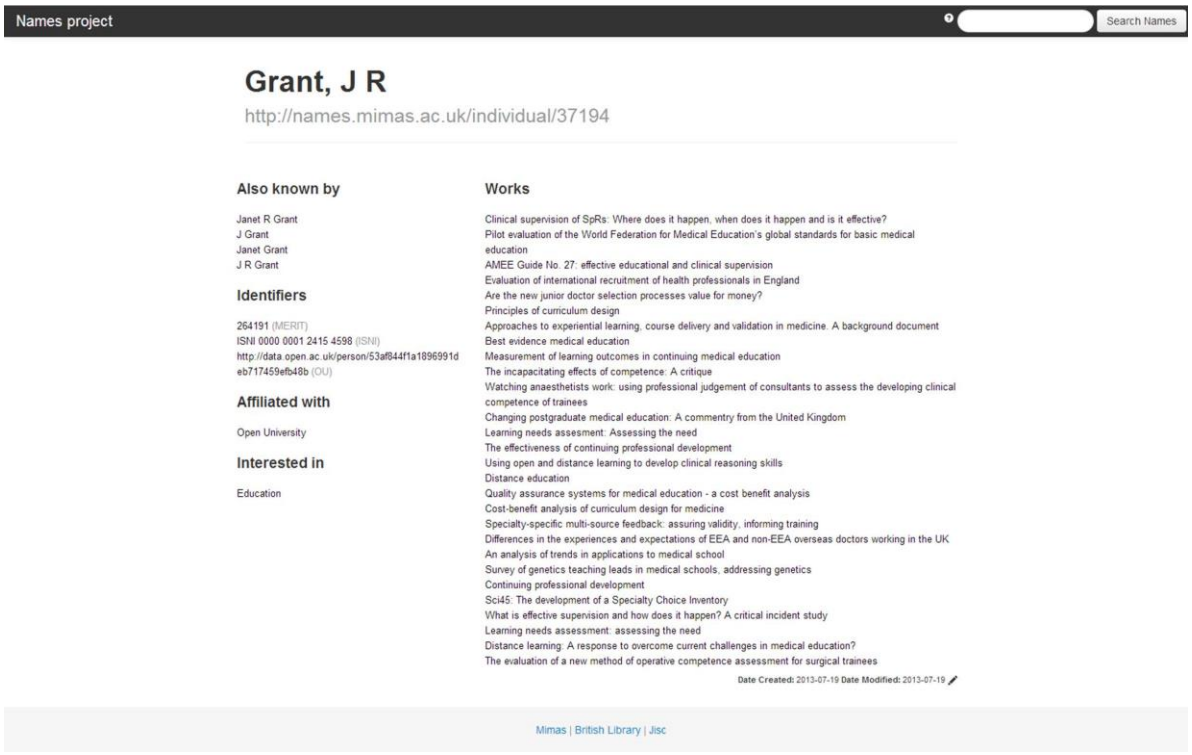

Figure 1: New Names user interface

In addition to these components a number of example applications have been written to demonstrate how the various system components can be used. These will also be available on bitbucket.org. The team plan to replace the current Names web interface with the new Names User Interface so that potential users will be able to see how it appears. The new interface will not have the ability for users to submit information about themselves to Names, as there will be no staff working on Names after the end of July to process such records. It would be possible to reinstate the submission form in the future.

Stakeholder engagement

The Names team has continued to work with colleagues from the International Standard Name Identifier initiative (ISNI). All researchers identified in Names, with the exception of the Open University individuals added in the last month, have now been assigned ISNIs. The Open University individuals will be assigned ISNIs in the coming weeks. Amanda Hill represents Names on the OCLC Registering Researchers in Authority Files Task Group which is preparing reports on issues surrounding author identification.[6]

The project has continued to share updates through the JISC-REPOSITORIES JISCmail list, through its blog,[7] through Twitter, and via presentations. Meetings attended and presentations given during this period include the following events:

1) Knowledge Exchange Summit, March 2012[8]
2) Open Repositories conference, July 2012[9]
3) GrandIR Technical meeting on author identifiers and ORCID, September 2012[10]
4) ORCID launch and outreach meeting, October 2012[11]

[6] http://www.oclc.org/research/activities/registering-researchers.html
[7] http://namesproject.wordpress.com/
[8] Report on Day 1 at: http://namesproject.wordpress.com/2012/03/14/knowledge-exchange-digital-author-identifier-summit/, Report on Day 2: http://namesproject.wordpress.com/2012/03/23/knowledge-exchange-dai-summit-day-2/
[9] Report and links to videos: http://namesproject.wordpress.com/2012/07/12/open-repositories-2012/
[10] Report and videos: http://namesproject.wordpress.com/2012/09/06/grandir-technical-meeting-on-author-identifiers-and-orcid/
[11] Report: http://namesproject.wordpress.com/2012/10/17/orcid-outreach-meeting/

The British Library's view

(This section written by Alan Danskin, Metadata Standards Manager at the British Library.)

The Library has well established workflows for disambiguation of persons and corporate bodies associated with books, but these are not scalable to cover authors of journal articles nor do they cover persons or corporate bodies associated with archival material and various types of manuscript material.

Specific objectives:

1. To enhance the capability for research funding bodies to identify outputs of their funding in the research literature by clear identification of researchers.
2. To develop tools to automate authority control processes
   a. Automated deduplication of names
   b. Automated disambiguation
3. To obtain retrospective control of article records (ETOC)
4. To develop capability to improve resource discovery between sectors through improved authority control.

These objectives have been partially realised by Names.

1. Names has identified 50k UK researchers and has obtained ISNIs for these researchers. This is a valuable resource for future research, but without an ongoing service, it will quickly lose currency and value.

2. Names has developed and tested algorithms for matching names. These have been quality assured by the British Library. The tools for deduplication and disambiguation are effective. Two criteria were used to evaluate quality of the algorithms.

Mismatches: Conflation of the identities of two or more individuals. Mismatching is a serious problem as it results in misattribution and renders the data unreliable. The matching algorithms were designed to reduce the risk of mismatches.

On average, 5% of names processed were identified as potential mismatches, requiring human review. Quality Assurance by British Library found that the range of actual mismatches for files processed was in the range 0%-3%. The largest number of actual mismatches found was 5 out of 786 (0.64%). Changes were introduced which reduced this to 2 out of 786 (0.25%). We believe the level of risk associated with potential mismatches is acceptable and that routine manual review of the potential mismatches would not be required.

Non-matches: Failure to identify duplicate identities. Names reported potential non-matches for each file processed. The percentage of actual non-Matches reported ranged from 0-55%. The variation is (inversely) related to the quality and richness of the incoming data. We believe that the algorithms are tuned as efficiently as possible and any relaxation would risk increasing the number of mismatches. The high rate of potential non-matches is a concern. The cost of QA would constitute a substantial element for cost recovery should any service be offered in future. The variability points to differential pricing models based on quality of data.

3. Three ETOC samples, totalling 201k records (including for 90k Crystallography and 30k Palaeontology) were evaluated using Names. The match rates were disappointing and there was a high rate of potential non-matches; attributable to deficiencies in the ETOC data. Coupled with capacity limitations of the Names demonstrator, it was decided to contribute the ETOC data directly to ISNI. In April 2013 British Library contributed 90 million records to ISNI.

4. The Library specified the data model (and mappings) which underlies Names. The data model is based on IFLA Functional Requirements for Authority Data and provides a rich attribute set for matching and deduplication. The data model is richer than is justified by the data sources Names has worked with to date, but ensures that Names is flexible and extensible to meet future needs and could support integration with archives and museums as well as institutional repositories and libraries.

The Library also engaged with the JISC Researcher Identifier Working Group during 2012-13. There was an exchange of information, but no agreement on a common approach. Both parties did agree that ORCID and ISNI should work together so that their strengths complement each other. From the perspective of the Library, the solution for identification of researchers must be compatible with broader requirements for identification of persons and bodies, including those that are no longer active and across different media.

In conclusion, Names has delivered a good demonstrator with the potential to offer automated services to HEIs and other repositories.

Outcomes

The objectives for this phase of development of the Names Project have been met. The data in the pilot system has been converted into internationally-available data in the form of ISNIs, ensuring that about one fifth of the UK's current researchers (and all of the researchers whose work was submitted into the 2008 Research Assessment Exercise) are now represented in the ISNI system. Options for integrating the Names data into ORCID were presented in the Options Appraisal report: these have not yet been taken up by Jisc.

Converting the Names system into modular components and making them available through a code-sharing site will ensure that the development achievements of the project team can continue to be used by other services and organisations in the future.

Conclusions

1) The Names project has demonstrated that automated or semi-automated solutions can be applied to bulk-process complex authority control tasks traditionally undertaken by cataloguers on an item by item basis. This approach offers the potential to extend authority control to types of resource, such as journal articles, which have previously been neglected on grounds of cost.

2) The quality of the outcome is directly affected by the range and quality of the metadata available. Publishing conventions, such as use of initials rather than full names, hinder accurate identification and comprehensive disambiguation of individuals. Human intervention is still necessary, but filtering enables the human intervention to be focused on ambiguous and anomalous identities.

3) Retrospective author disambiguation is complex and costly, even when partially automated and should be regarded as the solution to a legacy problem rather than the preferred way forward. The Names database and the components of the Names system are resources which can be used by other services to improve their own efficiency.

4) Integration between national systems such as Names and international services like ISNI is possible, with the national system offering the opportunity of liaising with institutions to feed data into the international level and with the potential for saving the research community the fees for institutional membership for ORCID and registration agency costs ISNI. Further investment in Names would be required to establish an automatic updating mechanism between the Names system and ISNI and/or ORCID.

5) The major achievements of the Names project have been the development of the disambiguation algorithm and the quality assurance process for the resulting data. These have enabled the creation of a useful set of information in the Names database which offers free and flexible access to its contents. By making the database structure, the data, and the disambiguation algorithm available through a code-hosting service, it will be possible for other services to make use of these elements in the future. It should be noted that the quality assurance expertise provided by the Names project team is not something that can be made available externally.