ARROW Project overview: Managing Author Names

The project is a joint undertaking of the Swinburne University, the University of New South Wales and the University of Newcastle.

Project goals

The project is examining the effective management of the names of authors in an institutional repository. Specifically, the project aims to develop a practical toolkit to manage author names in a repository that will assist the effective identification, disambiguation, matching and display of names. The toolkit will consist of guidelines and open source tools.

The project also aims to identify interoperability requirements with other key name projects and services including the Australian Access Federation and People Australia, as well as commercial services such as Scopus, Web of Knowledge and Google Scholar.

Issues being addressed

A repository needs to be able to take account of name variations to match, interoperate and authorise with confidence. It also needs to support name variations for inquiry and reporting.

Author names vary in different contexts and over time. The legal name of a person recorded in a university's Human Resources or financial system may be different to the name they prefer to publish under, are known by in their research community or would like displayed in a repository. There are many university systems that record names with their own rules and variations that a repository may need to interact with. Different publishers use different conventions for names, and a publisher is not necessarily consistent. People's names may change over time such as by marriage or deed poll. Their titles may change as well as their affiliation. Occasionally, even their field of research may change. Many of the services and systems include unique identifiers which enhance matching and discoverability. Existing standards or "rules", such as AACR, where one form of name is authorised based on evidence in publications, produce high quality results but at considerable cost.

The development of repositories provides opportunities to test new ways to achieve the benefits of authority control. A repository needs to be able to take account of name variations and be able to match, interoperate and authorise with confidence. This may be to match a record imported from a citation service to an academic already in the repository, for example downloading citations from Scopus or Web of Knowledge. It may be displaying information from the repository as an historical or current record of the person's research output. It may be feeding other services such as an author profile in a Find an Expert service or providing data to discipline-based repositories. It may be using other services to authenticate users to do particular tasks such as update records, provide access to hidden datastreams, or support users to interact easily with a private workspace in the repository. It may also be assisting other name services to manage names for the institution's academics.

No one name is necessarily the authority. Each variation belongs to a context and a time. For the purposes of inquiry and reporting from repositories, there needs to be ways of gathering the research data and outputs appearing under these name variations.

Finally, a repository needs to be able to support searching of the repository by name variations and to display variant names or supply them to other services. New models of exchanging, interoperating and building name information between repositories and related services are to be examined.

Methodology

Initially the project will investigate and analyse the issues and existing practice with the management and display of author names in relevant communities, including developments in the repository community, local university practice, the Australian Access Federation and the National Library of Australia's People Australia Project. It will work closely with the personal name data from Scopus, Web of Knowledge, and Google Scholar.

Building on this work, the project will develop a practical toolkit of guidelines and open source tools to manage author names in a repository that will be released to the repository community. It will test the use and interoperability of the toolkit with other services such as HR and grant management systems. It will examine and test the toolkit for interoperability with name services of major research citation services.

The project will ensure privacy and policy issues are addressed.

Swinburne Researchers - online profile pages for Swinburne research staff

Project overview:

The Swinburne Researchers project creates online profile pages for Swinburne research staff. A web interface enables the user to browse and search for researchers, and to display researcher profiles.

In the interests of consistency, currency and low maintenance overhead, the profile pages draw largely on information available from existing University systems. Swinburne Researchers sources data from the Swinburne ARROW institutional repository, identity management system, staff directory system, and research management system.

Project background:

The Swinburne Researchers prototype was developed through an ARROW partner project aiming to explore the use of the institutional repository as an information source for other university systems, and to explore the ability to package up institutional repository data in different presentation options, e.g. researcher profile pages.

Project development:

The Swinburne Researchers project is being developed in three phases:

1. Develop a prototype of the system, working only with data derived from existing university systems.
2. Move prototype to live system with existing functionality
3. Enhance profile pages with data not in existing university systems. Develop a web interface for data entry and administrative tasks.

Phase one of the project is complete, and included the following tasks:

- Analysis of data requirements (data attributes, data population, data access authentication, data priority)
- Analysis of system requirements
- Swinburne Researchers prototype database developed and populated with researcher profile data (data sourced from existing university systems – identity management, staff directory, and research management)
- Web interface of prototype to search for and display researcher profiles developed
- Interface to publications in the institutional repository available in the prototype

Project issues:

A number of issues relating to researcher identities were analysed and addressed through the project, such as:

- Identifying and selecting data sources
- Who is a Swinburne researcher?
- Currency of researchers
- Currency of profile information
- Determination of research departments
- Determination of research centres
- Researchers with more than one position / department
- Inconsistent researcher positions
- Inconsistent researcher titles
- Inconsistent preferred names
- Incorrect research profile information
- Research publications for a researcher
- Inclusion of research students
- Inclusion of researchers who are not staff members

USQ Academic Working Environment

Peter Sefton : sefton@usq.edu.au

USQ's Learning Futures Innovation Institute (LFII) is building a platform for academic work, known as the Academic Working Environment (AWE). AWE integrates several open systems:

1. Distributed version controlled content management for courseware and research.
2. A Learning Management System.
3. Portfolios for staff and students.
4. Our institutional repository.

Content

USQ hosts development of the open source 'Integrated Content Environment' (ICE) [1] ; a tool for researchers and educators to write anything from blog posts to papers, to theses or entire courses. In our latest work on ICE we are concentrating on tools to embed data and data visualizations in publications, with globally distributed workgroups. Identity management and access are key to this work.

ICE is now being used for collaborative document creation across institutions and we have major issues with granting access to collaborators. Some of these issues would be addressed by a fully functioning Australian Access Federation (which has not made it to USQ yet) but would not deal with all cases. In the last few weeks we have faced difficulties in the following, having to set up yet another login:

1. International collaboration.
2. Collaboration with unaffiliated researchers.

In addition to supporting Shibboleth [2] the ICE project team are currently investigating using OpenID [3] , which can be used by anyone regardless of whether they are part of a federation. For distributed groups with established trust relationships this would allow us to add users by asking them for an OpenID.

Repositories

USQ repository services is a group working on integration between repositories and other systems, such as access portals and workbench environments such as ICE. This presents interesting challenges in identity management as we are dealing with content that flows between systems where the same individuals have different roles.

One of our research areas is exploring how simple access controls might be implemented on top of a repository using a text-search portal in a way that library or repository staff can very easily create 'slices' through, or views of a repository based on who is browsing or querying it. For example, a thesis portal might have a rule that authenticated staff can view everything, while guests cannot access embargoed theses. Our **first priority** in this work is not to support a standard for access control for the sake of it, but to create an application that can be used by librarians and repositorians using available infrastructure such as LDAP servers for authentication, and when that it working to see how the policies might be expressed in the appropriate standards and shared with other services.

[1] P. Sefton, “The Integrated Content Environment for Research and Scholarship,” *ICE Website*, 2006; http://ice.usq.edu.au/introduction/ice_rs.htm.
[2]“Shibboleth®”; http://shibboleth.internet2.edu/.
[3] D. Recordon and D. Reed, “OpenID 2.0: a platform for user-centric identity management,” *Proceedings of the second ACM workshop on Digital identity management*, 2006, pp. 11-16; http://portal.acm.org/citation.cfm?id=1179532.

University of New England

Researcher Identification Management

Why does UNE need to manage researcher identities?

We want to build on and encourage the willingness of our researchers to include their research outputs in our repository by making sure the deposit process makes use of the best available information we have about them, and that the resources are identified unambiguously as their work and are readily available for use in a variety of contexts.

UNE's existing repository software provides limited identity management capability. Is this a problem, and for whom? Do we care that an identity has different facets and a history? How should we deal with researchers' changes of name, affiliation (internal/external), and the timing of such changes?

The repository is not a research management system, yet it is being required to be part of the University's research management activity. How can we ensure trustworthy linkages between the repository's people and their outputs, and the research management system? How will the identities we have established link to research related systems beyond UNE?

The University has separate but interdependent administrative systems, which can and will change even though the key data won't. While these systems support and help to manage particular parts of UNE's business, they frequently need to use and interpret other systems' data for the purposes of the University as whole. How can we maintain linkages between the repository and these systems over time?

Researcher Identification Management system

A provisional researcher identity management system that facilitates linking and data exchange between the following systems, is being developed by the Research Office and the Library at the University of New England:

- User authentication
- Research Management
- Human Resources
- Student
- Repository
- and other UNE administrative systems as appropriate

The system is essentially a warehouse of identifiers from each system, with a limited set of descriptive data, for an individual.

The key personal identifiers at UNE (staff and student numbers) are not intended for public exposure. The warehouse permits the confidential linking of these key identifiers to the identifiers used in other UNE systems. These latter identifiers may be subject to public exposure, either deliberately or inadvertently, but their value is confined to their host system.

Inclusion in the warehouse of a UNE researcher's national/international researcher identifier(s) is anticipated, as is the inclusion of records and identifiers for non-UNE researchers who will be represented in the UNE repository and other UNE research systems. The inclusion of this data in the repository is presumed.

Simon McMillan
e-publications@UNE Project Manager
28 May 2008

University of South Australia

UniSA's research publications and repository population project

Reasons behind the project

We have developed and are in the process of populating an in house database containing the details of our research publications, and whole of career publications for supported researchers. This database will not be made available outside of the Library (or University) as we are not permitted to make all of the data publically available, however we will be able to provide in house (University) reports of publications with up-to-date citation counts for researchers.

A large component of the project has been data cleanup, correcting errors in reference details and verifying publishing information of references obtained from researchers and the University's Research Master database. The bare citation information will be used to populate the University's research repository. We anticipate that this information will also provide a wealth of useful data for the University regarding research output.

Issues we are addressing

We are struggling with author identity with regard to:

- distinguishing between authors with similar names or where author information is sparse
- being able to identify errors in references lists and verify who wrote what
- being able to provide definitive researcher bibliographies via our research repository (this will also impact on any future reporting)

There is no industry standard as such being used for authority control beyond library catalogues.

The library standard has evolved from the print monograph publishing arena, a fairly contained environment. Even so few libraries are able to expend much resources for authority control which is a very expensive exercise (both in expertise and time).

The world of online databases is much larger than the published monograph world, and aggregation of data from one source to another further expands the problem. In addition journals and conference publications often provide less information about authors (initials rather than full names etc.) than monograph publishing.

How to go about it

Currently we are establishing the extent of the issues and are trying to scope what we want to achieve and how to go about it. Attending the workshop and any other relevant sessions is one way of gathering more data.

Deakin Research Online

The reasons behind the project.

Deakin University launched Stage 1 of its research repository in response to the RQF exercise in 2007. In 2008 stage 2 commenced initially to fulfill the requirement of the ASHER grant and other funding bodies to make the repository and as much of its content as possible available for open access.

How we are going about it.

Deakin chose Fez as a its repository platform as it appeared closely related to the requirements, and was being actively developed and tested. The repository has recently been named Deakin Research Online.

Metadata stored by Research Services for HERDC was exported to spreadsheets and used to create the records for DRO both the top 4 and the body of work of Deakin authors. As the top 4 RQF outputs were identified, they were scanned or downloaded and matched to the metadata then added to a 'dark archive' in DRO for RQF purposes under the provision of the Copyright Act.

In addition we are preparing the repository for self submission by the end of 2008.

The issues we are addressing:

Mandatory deposit and its implications

There seems to be general support at Deakin for mandatory deposit in the repository, however unless it substitutes for providing information to the Research Services database for reporting purposes it would be an additional burden for faculty. To gain maximum benefit new material should be safely deposited with its description, to support validation, preservation and accessibility in DRO at the earliest point, rather than via a periodic metadata load from Research Services with the objects followed up later. So at this stage we are pursuing the research repository as the one place to start the process of depositing and identifying research outputs, expecting to feed the required data to Research Services afterwards. Thus it is this necessary to capture all data elements relating to the item required for reporting, regardless of their value for discovery in the repository. These include Research Output Categories and Weightings, the total number of chapters of a book, and a mechanism to enable easy selection and input of ANZRC codes, old RFCD codes, and sponsors.

Author identification and control.

We need to provide an author key to uniquely match the author in Research Services system, but we expect that repository users would want to browse a list of material by an author "cleanly". We are currently discussing if we will try to provide a single unique disambiguated form of the name for browsing or keep to an ambiguous citation format? This will be different to the undifferentiated display form used in the Deakin Research Bibliography. Fortunately Fez provides a display form and the option of linking it to unique keys through an author table. There is currently no easy way to load author information continuously to the table in an ongoing manner. We have decided not to take the data from the HR database, even with a satisfactory control on the security of such data, as it doesn't preserve historic data about past employees who have published items that have been deposited. As Research Services do preserve historic data by manually

updating from HR reports we will use their database. We will have to add other author records manually as we expect items to be deposited by researchers that may never have been on the payroll. We will need to add Deakin University authors to the Research Services database in order to create their keys to add to our records for reporting purposes. We won't always be sure whether authors are new to Deakin, or still elsewhere, so much checking is envisaged.

P. Scott

29th May, 2008

Identifying Researchers: an outline of the issue at La Trobe University

We haven't solved any of the problems associated with this issue but have in motion three tasks which we hope collectively will go some way to solving the issue.

Why are we attempting this?

The primary goal is to consistently, unambiguously and comprehensively attach research products (or products more generally) to an author (including in combination with joint authors). There are a number of reasons for this, not least claiming credit for the output of institutional authors, protecting authors' rights and more generally understanding the outputs of the institution's staff.

Specifically it would be good to automate the production of bibliographies, lists, collections and reports on products of research with some degree of confidence that the machine will properly associate product with author.

Aiding the authentication and authorisation of individuals to give the right people the right kind of access to the right resources is also essential.

The issues

1. Name authorities have been the Library world's response to similar issues for a long time. The problem however is that many institutions have only minimally maintained authorities because they are time consuming and they need to be meshed within some standard international (LoC) and national (NLA) indexes which don't always suit local emphases or needs. For instance there is limited scope for recording affiliation as an attribute. It can be recorded, in the 678 field for instance, but as an uncontrolled text note. It doesn't lend itself to automated search and linking.

2. Multiple valid IDs for an individual
An individual may have good reason to hold or use more than one ID at any one time or over time.

3. Privacy
Automated access to attributes of an individual opens the risk of breaches of privacy.

Arising from a consideration of these previous two issues then personal control over identity and those attributes that are communicated with each identity in what circumstances also becomes important.

4. Affiliations of IDs/agents
A particular agent may have (will have) multiple affiliations over time and at any one time. The relevant affiliation for a particular action, whether authorship or access to resources for instance, is an important attribute of the ID.

For instance the institution with which a particular action of a particular ID is affiliated may wish to claim credit (money) for that action, may wish to control what is being done in its name or at least be aware of it and may wish to have knowledge of the actions of its affiliates for management purposes. The ID needs to control their described affiliation so this knowledge is gained with the agreement of the agent and ascribed to the correct institution.

Equally a provider of services or resources may wish to control access by a particular affiliation.

In order to meet global and portable nature of ideal IDs affiliation needs to be stored as an attribute of ID or links between institutions (read broadly) and IDs need to be made, maintained and made retrievable by query.

Affiliation brings with it the issue of corporate ID and the difficulties of changing entities and relationships between them. Again a superstructure system is required to map changes over time and relationships so affiliations can be mapped. Trying to map research output even within a university is difficult. Corporate name authorities, if maintained, might point to a solution.

5. Agreed language, eg agent, creator, author, researcher
Given the complexity possible when a real person has multiple roles, IDs and affiliations, we need an agreed language that does not make assumptions such as an ID “is” a person.

How we are going about it

1. We are trying, with our Metadata Services Department, to use the traditional Library Name Authority processes and records.

2. We are working with the University's ICT section on a mini-project funded by AAF to implement Australian Access Federation at La Trobe, leveraging as part of that, the work on identity within that community.

3. We are working with our Research Office and HR to use internal codes and records to control and flesh out identity.

A case study that brings out many of the above issues is the development of a repository for a joint CSIRO-University Research Centre with the attendant access restrictions for view, add, edit etc.

These are all in progress with little to show at this stage but we're keen to be involved in the discussions since the issues seem to us to be central to the core task in hand for the repository.

Elsevier

Identifying Researchers - issues and solutions associated with author and researcher identity management in the digital age:

In May 2006 Elsevier launched Author Identifier on Scopus a major navigation tool with a database containing abstracts and references from over 15,000 peer reviewed journals from over 4,000 international publishers. This is Elsevier's solution with regards to identifying researchers globally.

The Scopus Author Identifier automatically distinguishes between authors with the same name and matches variations of author names using advanced algorithms. The Scopus Author Identifier not only analyses all the variants of the author' name but utilises additional data elements associated with the article such as affiliation, publication history, source title, subject area and co-authors. The algorithms behind the Author Identifier are able to match author names with 99% accuracy and assign a unique identifier number to all authors who have published articles covered by Scopus. The algorithms are able to group at least 95% on average of an author's publications; the remaining 5% of records have too little data points to automatically group them with the others and are therefore presented separately, enabling users to group the documents manually.

During the first release in May 2006 over 20 million author profiles were created (1). Scopus has assigned an "Author Details" page to each author giving users an overview of data associated with that author. It is this page on Scopus that gives researchers the tools to provide feedback on the information that is publicly known about them.
The records on Scopus are a direct result of what has been published and provided by Publishers. Elsevier corrects obvious errors however does not change the content. We engage in enrichment of records for indexing purposes for high search precision and recall.

From Elsevier's perspective a global solution for identifying researchers would be welcomed. This would of course require co-operation from a multitude of academic relations, government bodies and professional organisations and publishers globally. Elsevier actively participates in discussions with industry-wide groups such as CrossRef - as a matter of fact, Elsevier is represented in the CrossRef Contributor ID working group, recently formed to explore such solutions.

1 New Scopus 'Author Identifier' facility - Scholarly Communications Report Page 9, Vol. 10 No.6 June 2006
http://www.atypon-link.com/SCR/doi/pdf/10.5555/scrn.10.6.9

Elsevier (cont)

Requirements for identity control and management

1. Researchers should have control over their identifier, as the use of this identifier shapes what is publicly knowable about them.

That depends on the degree of control. A globally recognized identifier could form the basis for metric evaluation of author output. This implies all the documents assigned should be true and correct without the possibility of excluding one of the documents (for whatever reason) i.e. as the ID could be used for metric output evaluation it should safeguard against manipulation. Therefore control should be limited to the "profile". However, this will more than likely be a trade-off between the community and evaluation board et al.

2. For a given research item or output created in a given context, researchers should be aware of (and have control over) what identifier should be assigned.

Again this comes down to consensus, if the public researcher identifier exists with the individual (from birth – death), it should be transparent to all researchers that they control the use of this public researcher identifier, if they forget to add it to a research item or output created they have effectively left that item/output for someone else to claim.

3. There should be general agreement about what sort of claims about a researchers identity can be made using a researchers own public identifier.

Yes this needs to be defined for transparency and acceptance

4. Researchers should not be able to assert that they are somebody else (or supply somebody else's unique identifier).

Yes, additional security and protection against research theft or altering/damaging data quality, again biggest question how and who manages and governs this.

5. It should be difficult for a third party (such as a repository manager) to incorrectly attribute a work to the wrong person. If incorrect statements are made about a researchers identity, it should be clear who the researcher needs to contact to redress the error and a documented process in place to ensure it is corrected.

Would there be a central body that has access to all the buttons of the control panel as this depends on where this information is stored and who is managing this.

6. Research Identifiers should be global and not limited to gated communities.

Once a global standard has been implemented and recognized as a standard by all, the public research identifier should be considered by all as an altruistic innovation.

7. Research Identifiers should be portable; that is, they should reside with the researcher for the duration of their research career.

Sure, given global consensus.

The metadata issues surrounding name authorities - Katie Blake, ARROW

Describing an object in a digital repository means metadata. This metadata is used for different purposes. In the library catalogue world, metadata is used to describe, retrieve, and display. In the repository world, metadata can have other applications. Use cases include:

- Statistical/reporting (eg for the HERDC).
- Populating author personal pages and dynamic CVs
- Synchronising consistently with aggregators and portals, eg Discovery Service, Google Scholar, NEREUS, and with commercial data services.

A digital object is necessarily a compound object. At a minimum one object will have content plus metadata. There may be multiple "sets" or datastreams of metadata per object. There may be metadata for preservation, structure, and technical information in addition to descriptive metadata such as Dublin Core, MARCXML, LOM, ETD-MS.

Author information has generally been included with the descriptive metadata, which is resource focussed, not author focussed. As repositories are used for many purposes multi-purposed, there is demand from repository managers to include institution specific or application specific information such as:

- Author email address, telephone and fax
- Website and social networking address (like Facebook, Myspace)
- A blog
- Author affiliation as distinct from the aegis under which a resource was published (authors move!)

Such information is not resource-specific, is not required for searching, and may not be for public display. Author affiliation and contact information is necessary for some purposes but it does not sit happily in a bibliographic schema.

New movements in metadata look at separating out this kind of information. SWAP and FRBR suggest an AGENT component.

The questions facing repository managers include:

- Should all author information be included in descriptive metadata?
- Should there be a separate section of metadata along the lines of agent data (adopt FRBR approach)
- How to ensure the right balance between resource oriented metadata and author oriented metadata
- If there are separate sections of metadata (descriptive and agent) how is the relationship expressed within a digital object?