

Automatic Metadata Generation – Use Cases

Person-related metadata

Amanda Hill

Hillbraith Ltd.

June 2009



This report is made available under a Creative Commons licence:
http://creativecommons.org/licenses/by/3.0/deed.en_GB

Automatic generation of person-related metadata

Being able to reliably link people with the work that they have produced is becoming increasingly important; for institutions, funding bodies and for the individuals themselves. As more materials are generated in electronic forms and are shared online in different ways, it is vital to be able to identify organisations and people in a reliable, unambiguous way and to automatically be able to insert such information into descriptions of those materials.

A recent report from the JISC-funded MetaTools project succinctly sums up this point:

Resource discovery metadata is a crucial component of the lifecycle of digital resources. Standardised metadata is a powerful tool that enables the discovery and selection of relevant digital resources quickly and easily. Poor quality or non-existent metadata on the other hand is equally effective at rendering resources unusable, since without it a resource is essentially invisible within a repository or archive and thus remains undiscovered and inaccessible. Unfortunately, with digital resources being produced in ever-increasing quantities, finding the time and resources necessary for ensuring metadata of appropriate quality is created is becoming a more and more difficult task.¹

The ability to be able to use an authoritative source to identify people and organizations (and to then be able to re-use that data automatically) is a goal that has been under active discussion in a wide variety of domains in recent times, in contexts including research assessment, submissions of materials to digital repositories and resource discovery.²

In the UK academic community, assessments of research excellence use citations of journal articles by researchers as one means of measuring impact. The interim report of the pilot project for the new Research Excellence Framework shows that analysis of citations of selected papers by authors, combined with expert review, was the preferred method of assessment of the three methods used in the pilot.³ Reliably identifying authors and matching them to their works are important elements of this process (and ones which some of the pilot institutions found difficult, due to a lack of data in the right forms).

Funding agencies are also interested in being able to reliably link researchers with their funding sources (particular grants), their institutions, and the published outputs of grants. There are related requirements in the realm of reprographic rights organisations and in book and journal publication, where it is necessary to link information about works and authors in various ways.

Many UK academic institutions are now building digital repositories of their staff's research outputs, partly as a 'shop-window' to promote the work performed in them. One of the barriers to submission of materials has been identified as the 'keystroke problem'⁴: individuals do not want to spend time entering the necessary information to identify their work. This reluctance might be overcome if the amount of typing involved can be reduced by automatic population of the fields in the various forms (including individual and institutional names). An experiment undertaken by Emma Tonkin and Henk Muller showed that most participants believed that filling in the necessary information was faster when a form had been automatically pre-populated, even though this was not actually the case.⁵

¹ Malcolm Polfreman and Shrija Rajbhandari, 'MetaTools Final Report', October 2008, p.4, <http://www.jisc.ac.uk/media/documents/programmes/reppres/metatoolsfinalreport.pdf>

² For example:

http://www.crossref.org/CrossTech/2007/02/crossref_author_id_meeting.html

³ HEFCE, 2009, 'Interim report of the REF bibliometrics pilot exercise' available at http://www.hefce.ac.uk/pubs/rereports/2009/rd13_09/rd13_09.pdf

⁴ See Dorothea Salo's view on this at <http://cavlec.yarinareth.net/2007/11/02/less-cognitive-load-faster-deposit/>

⁵ Emma Tonkin and Henk L. Muller, 'Semi automated metadata extraction for preprints

Libraries have been using name authority files as a means of assisting users with retrieving bibliographic works for many years. National libraries usually maintain these files, based upon individuals and organizations which have published books. In the UK, the British Library used to maintain its own name authority file (the British Library Name Authority File), but in 1993 it was agreed to merge this with the United States Name Authority File. A number of libraries around the world are now members of the Name Authority Cooperative Program (NACO) of the Library of Congress's Program for Cooperative Cataloguing. The combined name authority file is now known as the Library of Congress/NACO Authority File (LC/NAF). The LC/NAF is extensive, holding over 3.8 million personal and 900,000 corporate names.⁶ However, many creators of electronic resources, cultural heritage materials and journal articles will not also be authors of books (and changes in policy in libraries mean that authority records are not nowadays created as a matter of course: only if there is a risk of confusion with a name). As a consequence, library authority records will not exist for many individuals who might need to be identified as creators of or contributors to digital resources.

In repository systems that rely on depositors to supply information about creators, the quality of metadata is highly variable. Names of creators may be entered in direct order, or may be inverted (which is usually the preferred form of information professionals⁷). Initials only may be supplied, or full first names, or one first name and a middle initial. As a consequence, retrieval is affected:

The naïve user of an institutional repository will swiftly find that the absence of name authority control inhibits retrieval of items by a single author. Should a user arrive at a specific item and desire to see more items by the same author, clicking on the author's name will lead only to results for that particular name spelling or variant...⁸

Having a way of uniquely identifying individuals and institutions and of automatically inserting the resulting identifiers into databases and forms would help in a wide range of different scenarios. Here we have been discussing living individuals, but there are also requirements within universities for uniquely identifying people who have lived in the past: archivists within universities in the UK (and in the wider archival community) are very interested in this area and do not currently have a system in place to create and share information about the creators of archival materials. Museum and gallery professionals have similar requirements for identification of individuals connected to their objects and collections.

archives', *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital libraries*, June 16-20, 2008, Pittsburgh, PA, USA, pp. 157-166.

⁶ Statistics taken from the Library of Congress Authorities help page at <http://authorities.loc.gov/help/contents.htm>

⁷ See, for example, the 2007 recommendations of the National Research Discovery Service Metadata Guidelines of the National Research Discovery Service Project, National Library of New Zealand, available at <http://www.natlib.govt.nz/catalogues/library-documents/national-research-discovery-service-metadata-guidelines>

⁸ Dorothea Salo, 'Name authority control in institutional repositories', *Cataloging and Classification Quarterly* 47:3/4 (April 2009)

Scenarios

In this section a range of scenarios are outlined that illustrate the potential uses of automatically-generated information relating to people.

Title	Easing the task of submitting information
Narrative	<p>Juliet, a post-doctoral researcher, is ensuring that all of her publications are available online by placing copies of them in her university's digital repository. For each paper that she saves in the repository, a number of the fields in the associated deposit form are automatically completed. These include the title of the paper and the names of the authors. Juliet is able to review this automatically-created information and check that it is correct. Fields such as the title of the journal may be missing, but an autocomplete function that dynamically queries a service such as the ROMEO journal copyright policy service⁹ may be used to ensure that they are entered consistently.</p> <p>This scenario assumes that repository systems are able to interpret textual information (including author names) that is held in deposited materials and to extract it as structured data. Experimental work by Emma Tonkin and Henk Muller has shown that there is some potential in this approach.¹⁰ A further stage in this process would be to analyse the data and match the authors' names in them against known names in an internal or external name authority service, thereby assuring that an author is reliably linked to their work.</p> <p>The autocomplete of journal names using the ROMEO service is a feature in the current version of the EPrints software (http://www.eprints.org/software/v3/). This software also has an autocomplete function that uses existing information in a repository (or could be connected to other internal or external name authority services) to suggest author names.¹¹</p>

Title	Finding all works by a researcher
Narrative	<p>Graham is a science journalist working on an article about the development of the theories of Mary Grant, a researcher on climate change. In order to do this, he wants to find paper written by the researcher over the course of her career. The researcher has changed her name twice during her lifetime and has worked in a number of different institutions.</p> <p>Graham searches the Intute Repository Search service for Professor Grant's name. A list of possible matches is returned, with some disambiguating information such as field of interest or institution names. With this information, Graham can determine which of the names denotes the Professor Grant he is looking for. He follows a link in that entry to retrieve information about Mary Grant's papers that have been deposited in a number of different repositories in the universities where she has</p>

⁹ A machine-to-machine interface to this service is available:

<http://www.sherpa.ac.uk/romeo/api.html>

¹⁰ Emma Tonkin and Henk Muller, *op .cit.*

¹¹ <http://www.eprints.org/software/v3/>

	<p>worked.</p> <p>This scenario assumes that, on deposit, works by the researcher are associated with a unique identifier that is linked to her names and to the institutions in which she has worked. This is not currently possible, but work under way in the Names Project is developing a database of unique identifiers which could be used as such a source of information for repositories and for the Intute Repository Search service.¹²</p>
--	---

Title	Improving access to repository materials
Narrative	<p>Brent is an institutional repository manager who wants to improve the quality of the information in his database. He submits the repository's OAI base URL¹³ to an external, web-based service. This service analyses the metadata that is associated with records in the repository and produces a report for Bill. This report lists any apparent problems with the metadata, such as duplicate records, records that have missing titles, authors and dates, or authors whose names have been entered inconsistently in the metadata. Bill is then able to amend the data and improve the discoverability of the materials inside the repository.</p> <p>This scenario assumes there is an external service that can perform the analysis and reporting functions described here. This scenario could also be achieved if the repository software being used was able to generate reports of this kind. The KRIS (Kiwi Research Information Service) and MAT (Metadata Analysis Tool) services developed in New Zealand in recent years are examples of such services.¹⁴ The KRIS tool is used internally within a project consortium to check metadata¹⁵, while MAT was made available on the web for other services to use. It appears that the MAT functionality may now be part of the Greenstone Digital Library software, although this has not been verified. The web-based service is not currently available.</p>

Title	Creating authority files
Narrative	<p>Rebecca is an archivist, responsible for creating name authority records which are associated with individuals, families and organisations mentioned in archival finding aids. These are contributed to a national service based at The National Archives. She is working on cataloguing the papers of the recently-deceased historian John Trant and needs to create</p>

¹² This scenario is based on one (5.2.1) outlined in the Names Software Requirements Specification, available at

http://names.mimas.ac.uk/documents/Names_Software_Requirements_11Jul2008.pdf

¹³ This is the web address that identifies the part of the repository that is used to harvest records from it, according to the Open Archives Initiative's protocol for metadata harvesting. See <http://www.rsp.ac.uk/usage/harvesters> for a fuller explanation.

¹⁴David M. Nichols, Gordon W. Paynter, Chu-Hsiang Chan, David Bainbridge, Dana McKay, Michael B. Twidale, Ann Blandford, 'Experiences in Deploying Metadata Analysis Tools for Institutional Repositories', *Cataloging and Classification Quarterly*, Volume 47, no. 3/4, 2009

¹⁵ The reports generated are available at <http://nzresearch.org.nz/index.php/reports/indexMetadataQualityReports>

¹⁶ International Council on Archives, *ISAAR(CPF): International Archival Authority Record for Corporate Bodies, Persons and Families*, Ottawa, 1996:

an authority record according to the International Standard for Archival Authority Records (ISAAR-CPF).¹⁶

Rebecca consults a central name authority service for researchers and discovers that there is a record for Trant. The service provides an option of exporting its data in the Encoded Archival Context (EAC) format, enabling Rebecca to pre-populate a number of fields in her record. She completes the record with additional detailed biographical information and sends it to The National Archives, with the link to the record in the name authority service.

This scenario assumes there is a name authority service which covers researchers in the UK and which is capable of exporting information in different formats, including EAC. It also assumes that The National Archives is actively collecting name authority files. The Names Project prototype is being developed as a source of information about researchers and their institutions. It supports a range of output formats.¹⁷ The National Archives maintains some authority files which have been generated from the indexes of the National Register of Archives, but the organisation is not currently accepting contributions from other archives.

[http://www.icacds.org.uk/eng/ISAAR\(CPF\).pdf](http://www.icacds.org.uk/eng/ISAAR(CPF).pdf)

¹⁷ This scenario is also based on one (5.6.1) in the Names Software Requirements Specification.

Tools, Standards and Key Issues

There are a number of sources of names available online, ranging from controlled vocabularies such as that maintained by the Library of Congress, to user-generated descriptions of individuals like those to be found in Wikipedia. Many of the existing online sources of names and standards relating to them were described in some detail in the Names Project's *Landscape Report*.¹⁸ Rather than list them in the same amount of detail here, some of the key issues have been identified and are illustrated with examples of some of the main resources.

From the perspective of automated metadata generation within UK universities, there are a number of problems to be faced with many of these sources which mean that none of them are currently entirely suitable for use in the scenarios described above. One of the key issues is the availability of the information in a form that is suitable for incorporation into other systems. The final report of the MetaTools project emphasised the importance of making automated-metadata generation tools available as web services:

Researchers, application developers and cataloguers should benefit from well-defined and implementation-independent interfaces that enable metadata generation tools to be plugged more easily into their applications and processes.¹⁹

In order for a source to be useful, it also needs to be likely to actually hold the names that users would expect to see there. Clearly, some resources are going to be more useful than others, depending on the context of the end user. The Library of Congress/NACO name authority file is useful for book cataloguers as it contains controlled forms of the names of people who have authored books. In the digital repository arena, however, there is not a high correlation between names of the authors of academic papers and the names of book authors, meaning that many researchers are not represented in the file. In addition, this file is available through Library of Congress subscription-based cataloguing tools, but is not available through a machine-accessible web service interface (although it can be searched on the web at <http://authorities.loc.gov/>).²⁰

Other tools are available that do have machine access and are available for free download. These include DBpedia²¹ and Freebase²². DBpedia is a database of structured information that has been extracted from Wikipedia and which contain links to other resources on the web (books in Project Gutenberg²³ and placenames in GeoNames²⁴, for example). This includes over 213,000 entries on people. Wikipedia has 'notability guidelines' for people ([http://en.wikipedia.org/wiki/Wikipedia:Notability_\(people\)](http://en.wikipedia.org/wiki/Wikipedia:Notability_(people))), which mean that individuals will only be included if they have been the "subject of published secondary source material", but there is overlap with people working in the academic community (see, for example, the Wikipedia entry for chemist Peter Atkins at http://en.wikipedia.org/wiki/Peter_Atkins, which becomes http://dbpedia.org/page/Peter_Atkins). The datasets are available for download and can also be accessed a query interface and as linked data (<http://wiki.dbpedia.org/OnlineAccess>).

Freebase is, like DBpedia, comprised of structured content created by its users and

¹⁸ Alan Danskin, Anne Dixon, Michael Docherty, Amanda Hill, Richard Moore, 'A review of the current landscape in relation to a proposed Name Authority Service for UK repositories of research outputs', September 2007-June 2008, <http://names.mimas.ac.uk/documents/LandscapeReport26Jun2008.pdf>

¹⁹ Malcolm Polfreman and Shrija Rajbhandari, *op.cit.*, p.22

²⁰ LC/NAF name authorities are not yet available through a machine-accessible service but the Library of Congress Subject Headings have recently been made available for machine-searching and download. The Library of Congress plans to make other vocabularies and authorities in the future, which may include the name authority files: <http://id.loc.gov/authorities/about.html>

²¹ <http://dbpedia.org/>

²² <http://www.freebase.com/>

²³ http://www.gutenberg.org/wiki/Main_Page

²⁴ <http://www.geonames.org>

harvested from other sources (including Wikipedia). It does not have the 'notability' constraints of Wikipedia and currently holds information on over 1.3 million people. The data is made freely available through an API and can also be downloaded in a variety of formats (<http://download.freebase.com/datadumps/>). With significantly more people included in Freebase, it is possible that this might be a more useful database for person-related data than DBpedia. One important point to bear in mind with sites that are comprised of user-contributed data is their reliability. The entries for the chemist Peter Atkins in both DBpedia and Freebase attribute to him the writing credits for the *Hellraiser* series of films (which were written by another man of the same name).

Reliability is also important in regards to access: when the Library of Congress Subject Headings were first made available as linked data (at <http://lcsb.info>), a number of other services made use of the information. The lcsb.info service was later turned off, leaving those other services without access, until the replacement service was made available on 1 May 2009.²⁵ In order for other services to make use of a source of names, their developers and administrators need to be assured that the source will continue to be maintained.

There are other sites which more directly match the needs of universities in terms of their contents. Included in these would be the lists of authors that are being created by a number of publishers: Thomson Reuter's ResearcherID²⁶, Elsevier's Scopus Author Identifier²⁷ and ProQuest's ScholarUniverse²⁸. These services generally make use of the bibliographic databases owned by publishers and encourage researchers to maintain their own profile information.²⁹ These services are all only fully available to those who subscribe to the products of the publishers (although in the case of ScholarUniverse basic details are available to everyone through the website) and none of the services make their data freely available for machine interaction. The data held in the ResearcherID service is now made available for download as a subscription service³⁰, and institutions can upload data free of charge.

Educational institutions maintain their own lists of employees, some of which may be available within those institutions for use by internal services. Data Protection concerns often prevent institutions from being able to share such resources more widely. Commercially-run services such as those provided by publishers usually require individuals to agree to share or withhold their personal details in order to legally share them. Privacy concerns can be a major stumbling block when trying to obtain information from potential data providers, even when information is already in the public domain.

There are a number of initiatives under way to try to resolve some of the problems around the unique identification of organisations and individuals. In the UK, JISC is already funding the Names Project to develop a pilot name authority service which is aiming to uniquely identify researchers and their institutions.³¹ Internationally, OCLC is looking to develop its Worldcat identities service into an Identities Hub which would bring together information on individuals and organisations from a variety of external sources.³²

In the area of standards development, there are a number of existing name authority standards detailed in the Landscape Report mentioned earlier. Standards currently under development include the International Standard Name Identifier³³ (currently a draft ISO Standard, ISO 27729, which is a number that can be used for "any entity that is or was either a natural person, a legal person, a fictional character, or a group of such entities, whether or not incorporated") and the NISO Institutional Identifier, which is being defined

²⁵ <http://lcsb.info/comments1.html>

²⁶ <http://www.researcherid.com/>

²⁷ <http://info.scopus.com/etc/authoridentifier/>

²⁸ <http://www.scholaruniverse.com/>

²⁹ Peter Atkin's page in ScholarUniverse seems to have been edited by him, for example: <http://www.scholaruniverse.com/profiles/people/8A1730727F0000010044A23153C5CE40?q=firstname:peter+lastname:atkins>

³⁰ <http://isiwebofknowledge.com/researcherid/webservices/rid-dl-faq/>

³¹ <http://names.mimas.ac.uk/>

³² <http://www.oclc.org/programs/ourwork/renovating/leveragevocab/idresource.htm>

³³ <http://www.isni.org/>

now as a means of identifying corporate bodies and their constituent parts.³⁴ The future use and availability of information associated with these standards is not yet clear.

³⁴ <http://www.niso.org/workrooms/i2>

Recommendations

In order to support the scenarios described above (and many other related use cases), the key priorities that emerge are the need for comprehensive, reliable and freely machine-accessible sources of information about people and organisations. Tools to help improve existing metadata sets are also important. It is therefore suggested that JISC look at work in the following three areas:

1. The possibility of developing tools for the automatic analysis and improvement of metadata in existing digital repositories and services, along the lines of the KRIS and MAT systems that were developed in New Zealand.
2. Encouraging liaison with funding councils and institutions to see if there is scope to align approaches to name authority in digital resources area (such as that currently being undertaken by the Names project) with the requirements of institutions and funding organizations. The aim should be to have a comprehensive source of unique identifiers for active creators of digital content for the UK, which is accessible both to people and to machines. This data would then be available for a variety of different purposes.
3. Funding a pilot name authority service for UK historical figures, in collaboration with The National Archives and the wider cultural heritage community.