

Amanda Hill

## **What's in a Name? Prototyping a Name Authority Service for UK Repositories**

### **Abstract**

This paper looks at approaches to name authority control in repository contexts and describes the work of the Names project, which has been funded to investigate issues surrounding the identification of individuals and institutions within repositories of research outputs in the United Kingdom.

### **Introduction**

The problem of uniquely identifying authors has been with us ever since books have been catalogued. National libraries have been creating name authority files for authors of books for many years, starting with card catalogues and now maintaining electronic files in MARC format. However, authority files for the creators of journal articles do not tend to 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-referred versions of articles has led to a corresponding rise in the number of authors identified in such systems.

Without having a means of uniquely and unambiguously identifying the creators of materials in repositories, it becomes difficult to be sure whether all the materials related to a particular author will be correctly associated with that individual. Names of authors may be entered in more than one way, or more than one author may have exactly the same name. This article looks at recent attempts to address this problem in the repository environment and goes on to explain the approach that is planned to be taken in the Names project.

### **Background**

The context for the Names project is the work on repositories that has been undertaken in the United Kingdom in recent years, particularly in the rapid development of institutional repositories. One definition for these repositories is:

... a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. (Lynch 2003)

Much of the development of these repositories in the UK has been seeded by funding from the Joint Information Systems Committee (JISC) through a series of initiatives that began with the eLIB programme in 1994 (Pinfield, 2004). As a result of the development of the ePrints software at Southampton which had been made possible by JISC funding, there were, by March 2003, six institutional repositories in the United Kingdom (Day, 2003), containing between them 7,334 records (it should be noted that 7,158 of these were in one repository: Southampton's Department of Electronics and Computer Science, the first such repository to be established in the UK). Such has been the momentum behind the repository movement since then, that in early 2008 the OpenDOAR registry recorded 87 institutional repositories in the United Kingdom,

holding 303,052 records.<sup>1</sup> The number of records held within these repositories ranges from 179,722 at the University of Cambridge to the 26 repositories that are currently holding under 100 records.

Subject-based repositories have been in existence for a longer period than institutional repositories. The longest-running is arXiv.org (covering e-prints in physics, mathematics, computer science, quantitative biology and statistics), which was first made available online in August 1991 (Ginsparg, 1994) and which now holds over 450,000 e-prints.<sup>2</sup> RePEc is a service that covers working papers, journal articles and software relating to economics and this collection now holds over 560,000 items. RePEc itself benefited from JISC funding as part of the eLib programme in the late 1990s (Krichel, 1997).

### **Name authority control in repositories and related services**

As might be expected, issues related to the reliable identification of authors were identified in the subject-based repositories once they reached a significant size.

Searching by authors' names has been among the top search methods by repository users. When a repository grows to substantial size, it is often the case that name variants cause headaches for both the users and repository managers. (Xia, 2006)

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

Figure 1. illustrates this problem. A 'Browse Authors' search has been performed on the University of Cambridge's institutional repository and the author names beginning with 'A' have been returned. It will be seen that the list reflects the differing ways in which authors' names have been entered into the system. Some have been inverted, others have not, meaning that sometimes the 'A' is a forename, sometimes a surname. The name A.K.Bhushan appears in three different forms in the list and each of the names retrieves a different set of documents, though apparently all by the same person.

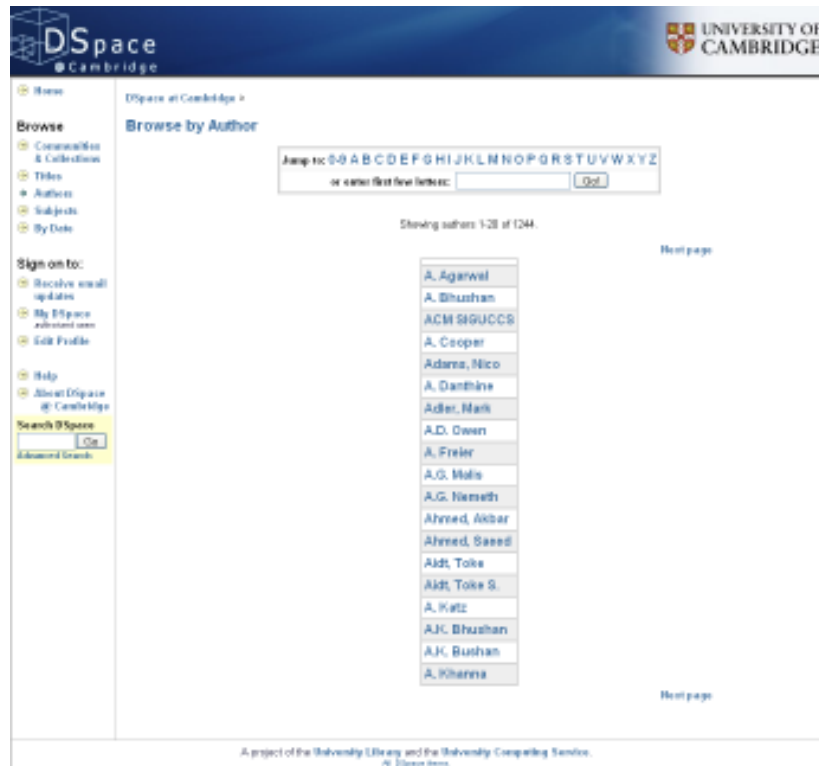
In order to avoid problems of this kind, the RePEc economics repository has introduced a level of authority control for its authors, in order to ensure that individuals are uniquely identified and correctly associated with the materials that they have made available through the services. Authors register with the service (this is also true for arXiv.org) before they can deposit materials and, as a result, RePEc currently holds

---

<sup>1</sup> The OpenDOAR directory of open access repositories is to be found at <http://www.andoar.org/>

<sup>2</sup> [www.arXiv.org](http://www.arXiv.org)

**Figure 1.** Screenshot of browse search on the University of Cambridge's repository



information on 15,000 authors. The service also maintains a directory of economics departments and related institutions, which is linked to the RePEc Author Service to provide affiliation information for individuals. This is how the service was described, in its early stages:

We take a set of digital library data, and we ask authors to tell us which papers they have written. Naturally, this strategy to get the authors involved in producing their own access control data will only work if the authors have incentives to supply such data. Since academic authors are interested in the visibility of their work, they will have good incentives to supply data to a database that is frequently consulted by potential readers. For any database to achieve such a status, it must be relatively large and available at low cost. The RePEc dataset of Economics research is a good candidate. (Cruz et al, 2000)

Variations on the author-registration approach pioneered by the RePEc service have since been adopted by commercially-run systems such as those supplied by Elsevier

(Scopus Author Identifier<sup>3</sup>) and Thomson Scientific (ResearcherID<sup>4</sup>). ProQuest provides a more hand-crafted approach with its Scholar Universe service, which compiles lists of faculty members and associates them with their published works.<sup>5</sup> Faculty are encouraged to add their own details to their profiles. Common to all of the systems is access to a large corpus of existing information about authors and their outputs which can be analysed and used to create associations between the entities concerned. They all also encourage authors to register and update their own lists of publications.

### **The Names project**

With funding provided by the current JISC Repositories and Preservation Programme,<sup>6</sup> the British Library and Mimas, a data centre at the University of Manchester, have been tasked with investigating:

...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, 2007)

The British Library has a long history of expertise in the name authority area and is a contributor to the Name Authorities Co-operative (NACO). The Library of Congress (LC) maintains the LC/NACO authority file, which now contains around seven million records, created by over 500 contributors. The British Library is one of four permanent members of the Policy Committee of the Program for Cooperative Cataloguing which steers the governance of NACO.

Traditional library sources of authority control have proved to be inadequate for the purposes of institutional repositories. Even within the bibliographic world, Calhoun (1996) estimates that around 50% of author names in library catalogues are not represented in the LC/NACO authority file (LCNAF). In repositories the situation is more pronounced. The ePrints UK project<sup>7</sup> and staff at MIT<sup>8</sup> have both investigated the correlation between the Library of Congress name authority file (using services provided by OCLC<sup>9</sup>) and author names held within institutional repositories. The results have not been published, but anecdotal evidence from those involved suggests that fewer than 25% of the authors depositing materials in repositories are also to be found in the LCNAF. It is clear that simply incorporating access to the LCNAF into

<sup>3</sup> Press release announcing the launch of Scopus Author Identifier in June 2006: [http://www.elsevier.com/wps/find/authored\\_newsitem.newsroom/companynews05\\_00484](http://www.elsevier.com/wps/find/authored_newsitem.newsroom/companynews05_00484)

<sup>4</sup> Press release on the launch of ResearcherID in January 2008: <http://scientific.thomson.com/press/2008/8429910/>

<sup>5</sup> <http://www.scholaruniverse.com/>

<sup>6</sup> Information on the programme can be found at [http://www.jisc.ac.uk/whatwedo/programmes/programme\\_rep\\_pres.aspx](http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres.aspx).

<sup>7</sup> This project is described by Ruth Martin in 'ePrints UK: Developing a national e-prints archive', *Ariadne* 35, 2003, available at <http://www.ariadne.ac.uk/issue35/martin/>.

<sup>8</sup> The issue with name authority files is touched upon by MacKenzie Smith, Associate Director of Technology at MIT, in a mailing list message at <http://mailman.mit.edu/pipermail/dspace-general/2006-March/000902.html>.

<sup>9</sup> OCLC's involvement with ePrints UK is described on the project page at <http://www.oclc.org/research/projects/mswitch/epuk.htm>

repository software will not be sufficient for the purpose of uniquely and unambiguously identifying authors of repository materials.

### **Proposed approach**

Initial work on the Names project has been investigating existing activities in this area and gathering requirements from stakeholders of the project. The stakeholders include repository managers, the project's funders and partners, developers of repository software and providers of cross-repository services such as the Intute Repository Search service.<sup>10</sup> These requirements are being used to develop the specification for the prototype that will be produced in the next phase of the project.

The prototype will be using a unique number-based identifier, rather than a controlled form of the name of an author (or an institution) as the primary identity key. It will also need to associate all known variants of the name with that number. In other words, the approach needs to be one of access control, rather than authority control. Such an approach might once have been seen by librarians as "brazenly radical" (Barnhart, 1996), but it makes perfect sense in a context where it is important to record the form of a name as it appears in a published article (to assist retrieval), even though this form may differ across different publications. As yet there is no agreed standard for the form of an author identifier and it appears unlikely that a single world-wide system for assigning a single identifier to an author will be implemented (Tillett, 2007). Therefore the prototype will need to be able to hold information about other identifiers that may be assigned to the same individual in related systems.

There are various sources of data that are available to the project team which will be used to populate the prototype. These include the author names and journal article information in the British Library's Zetoc service<sup>11</sup>, which is an electronic table of contents covering journals and conference proceedings that have been published since 1993, and similar data from UK PubMed Central<sup>12</sup>, a repository for articles published in the life sciences (also maintained by the British Library) and from the metadata supplied to the repositories covered by the Intute Repository Search service mentioned earlier.

One of the activities undertaken by the British Library members of the project has been to analyse a variety of metadata schemas and standards and map the data types to the entities described in the International Federation of Library Association's *Functional Requirements for Authority Data: a Conceptual Model* (FRAD).<sup>13</sup> The schemas that have been mapped include the MARC21 format for authority data<sup>14</sup>, the National Library of Medicine (NLM) Journal Publishing DTD<sup>15</sup>, the Scholarly Works Application Profile (SWAP)<sup>16</sup> and the International Standard for Archival Authority Records, Corporate, Personal and Family names (ISAAR (CPF))<sup>17</sup>. This will make it

<sup>10</sup> The Intute Repository Search service can be found at <http://irs.ukoln.ac.uk/>.

<sup>11</sup> Zetoc is hosted and supported by Mimas and is available at <http://zetoc.mimas.ac.uk/>

<sup>12</sup> Available at <http://ukpmc.ac.uk/>

<sup>13</sup> Available at <http://www.ifla.org/VII/d4/FANAR-ConceptualModel-2ndReview.pdf>

<sup>14</sup> See <http://www.itsmarc.com/crs/Auth0679.htm>

<sup>15</sup> Available at <http://dtd.nlm.nih.gov/publishing/>

<sup>16</sup> Also known as the Eprints Application Profile ([http://www.ukoln.ac.uk/repositories/digirep/index/Eprints\\_Application\\_Profile](http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile))

<sup>17</sup> Available at [http://www.icacds.org.uk/eng/ISAAR\(CPF\)2ed.pdf](http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf)

possible for a variety of different data sources to be adapted for use in the prototype and will ensure that the system will be able to export data in a variety of formats for use by other systems.

The requirements of repository managers are simple: they would like a name authority module that will plug into their existing repository software and provide auto-completion of author names for depositors of materials and for searchers of the system. It is possible for each repository to set up its own authority file for local authors, and functionality of this type is already offered by the latest version of the Eprints software.<sup>18</sup> However, given the level of co-authorship across institutions, a centralised authority file for the UK would avoid duplication of effort. An international file would, of course, be even more useful, but that is beyond the scope of the current project.

Ensuring interoperability with repository systems and other services will be an important part of the development of the Names prototype and any future service that may evolve from it. This will involve the provision of a web service to enable support for auto-completion and may also entail allowing bulk harvesting of the Names data to trusted third parties. Enabling such functionality will also contribute to the wider, world-wide, efforts to unambiguously identify authors and institutions. The prototype's web service will be tested during the course of the project by the Intute Repository Search and UK PubMed Central services to ensure that it meets their requirements for such a service.

The existing name authority services that were examined in the first part of this paper have all recognised that harnessing the knowledge of the authors themselves is an important part of providing a reliable name authority service. Authors are able to tell us about their institutional affiliations (and how they have changed over time) and the variant forms of their names. These may be harder to determine without the involvement of the authors themselves. If the Names service is developed beyond a prototype it will be important to be able to accept information from authors so that the accuracy and quality of the information within the service can be improved and maintained.

Authors will also be able to assist with links across systems by recording the identifiers assigned to them in other systems. Multiple registration requirements across a number of name authority systems may create an element of 'ID fatigue' for authors. If connections can be made across different services, some of this may be avoided. In the Netherlands, a centralised network of author identities has been created by linking data from research information centres to that of the name authority files in the library system (van Spanje, 2007). It is possible in the future that similar links could be forged between a UK Names service and the researcher information held by the Joint Electronic Submission System which is maintained by the Research Councils in the UK. This would enable these funding bodies to easily see how their funding decisions have been converted into published articles, datasets and other repository materials by the recipients of their grants.

## **Conclusion**

The rapid development of institutional and subject-based repositories has brought with it problems of author identification that have been well-known in the library world

---

<sup>18</sup> For details see <http://www.eprints.org/software/v3/>

for many years. The Names project is attempting to bring decades of library expertise together with currently available technologies and data in order to create a prototype name authority service that will help the new repository landscape to solve one of its most pressing problems: the unique and unambiguous identification of creators of repository content. In the future, the involvement of the authors themselves will be essential to maintain and improve the service.

## References

- Barnhart, L. 1996. Access Control Records: Prospects and Challenges. In *Proceedings of Authority Control in the 21st Century: An Invitational Conference*. Available at <http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000003520:000000091779&reqid=354>
- Calhoun, K. 1996. Characteristics of Member-Established Headings in the OCLC Database. In *Proceedings of Authority Control in the 21st Century: An Invitational Conference*. Available at <http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?sessionid=84ae0c5f82404092e15928004ca78c23a9d1325726eb?fileid=0000003520:000000091790&reqid=354>
- Cruz, J.M.B., M.J.R. Klink and T. Krichel. 2000. Personal data in a large digital library. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, 127. Available at: <http://openlib.org/home/krichel/phoenix.a4.pdf>.
- Day, M. 2003. Prospects for institutional e-print repositories in the United Kingdom. ePrints UK Supporting Study no. 1. Available at <http://eprints-uk.rdn.ac.uk/project/docs/studies/impact/>
- Ginsparg, P. 1994. First steps towards electronic research communication. *Computers in Physics* Volume 8, Issue 4, pp.390 - 396
- JISC. 2007. JISC circular 04/06: capital programme, Appendix G, available at [http://www.jisc.ac.uk/fundingopportunities/funding\\_calls/2006/09/funding\\_circular04\\_06.aspx](http://www.jisc.ac.uk/fundingopportunities/funding_calls/2006/09/funding_circular04_06.aspx)
- Krichel, T. 1997. About NetEc, with special reference to WoPEc. *Computers in Higher Education Economics Review*, 11(1), 19-24. Available at [http://www.economicsnetwork.ac.uk/cheer/ch11\\_1/ch11\\_1p19.htm](http://www.economicsnetwork.ac.uk/cheer/ch11_1/ch11_1p19.htm)
- Lynch, C.A. 2003. Institutional repositories: essential infrastructure for scholarship in the digital age. In *ARL Bimonthly Report*. No. 226, 1-7. Available at: <http://www.arl.org/resources/pubs/br/br226/br226ir.shtml>
- Pinfield, S. 2004. eLib in retrospect: a national strategy for digital library development in the 1990s. In *Digital libraries: policy, planning and practice*, ed. J. Andrews and D. Law, Aldershot: Ashgate, pp.19-34. Available at: [http://eprints.nottingham.ac.uk/131/1/elib\\_2003.PDF](http://eprints.nottingham.ac.uk/131/1/elib_2003.PDF)
- Tillett, B. 2007. Numbers to Identify Entities (ISADNs- International Standard Authority Data Numbers). *Cataloging & Classification Quarterly*, 44 (3/4), pp.343-361
- Van Spanje, D. 2007. Digital Author Identification. Presentation at UK Serials Group conference, 17-18 April 2007. Available at [http://dai-uitrol.ub.rug.nl/logboek/FILES/61/DAI\\_UKSG\\_20070417\\_final.doc](http://dai-uitrol.ub.rug.nl/logboek/FILES/61/DAI_UKSG_20070417_final.doc).
- Xia, J. 2006. Personal name identification in the practices of digital repositories. *Program: Electronic Library & Information Systems*, 2006, 40(3): pp.256-267. Available at <http://dlist.sir.arizona.edu/1832/>.