Published in *Program: Electronic Library & Information Systems*, 2006, 40(3): 256-267.

Personal name identification in the practices of digital repositories

Jingfeng Xia

Introduction

As early as the late 1990s, it was recognised that "names are as important in a digital library as an ISBN number is in a traditional library" (Cleveland 1998). In an ideal situation, personal names are as able to identify digital objects uniquely as they are to identify people. In the real world, however, names are sometimes not unique enough to handle identifications for digital objects. This inability inevitably affects the quality of information retrieval, citation analysis, linking of objects and copyright management.

It is a well-known fact that personal name variants can reduce the precision level of online searches for digital materials. A common practice is that people alternatively use full names and abbreviations. Numerous efforts have been devoted to improving search effectiveness and efficiency in both theoretical explorations and practical implementations (e.g., French et al. 1997). The problems, however, cannot be easily solved because different forms of names may represent legitimate irregularities.

For digital repositories (defined as either a local institutional or subject/discipline-based archive for depositing and providing access to digital contents – http:// en.wikipedia.org/wiki/), personal names, particularly authors' names, are one of the major concerns for precise data retrieval. This is because articles and other textual work are the primary form of eprints in digital repositories, although multimedia data have also been included in certain institutional repositories (Lynch and Lippincott 2005). In the following discussion of the article, 'name' refers exclusively to author names.

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. In previous studies the cases of name misspelling and duplicate data entries were not the major concern (e.g. Bilenko et al. 2003), because they are easily machine detectable.

Name variants usually fall within the range of the following basic forms:

a) Full name vs. abbreviation. This confusion comes from an author inconsistently using his/her names in publications. For example, the same author may have used such forms of her name when she published research articles: 'Susan Elaine Higgins', 'Susan E. Higgins', 'Susan Higgins', or 'S.E. Higgins'. Many digital repositories advise their users to apply the "last name + first name initials" strategy to conduct an

author search as the solution to avoid such a type of name difference. Unfortunately, this search strategy worsens the situation when the second type of name variant exists.

b) Different physical persons sharing the same name, or even similar names. The "last name + first initial" search model will make 'Sandra Higgins' and 'Susan Higgins' like one single person. Running simple tests against large repositories such as arXiv (the open access repository of materials in high energy physics, atmospheric and oceanic physics, mathematics and computer science (http://arxiv.org)) easily demonstrates this problem. When a repository is large enough to return hundreds of articles for an author with the same or similar names, it is easy to frustrate impatient users. Even if patient users are still able to identify the work of a particular author by judging article titles and writing styles, or by manually expanding each article's link, they may be totally confused by the third type of name ambiguity.

c) One physical person with totally different names. This can be the result of marriage status changes as well as other changes. Only when one person is well-known, can s/he be recognised as the same author for different names. For example, the works of Alice Hawthorne and Septimus Winner can be cross referenced as a single author (DiLauro et al., 2001). However, the majority of ordinary article authors will not be so fortunate. This type of name variation makes searching difficult both from the perspective of the repository manager and the normal user.

Name variants not only diminish the precision of data returned in an online search, but they also make citation analysis unreliable to varied degrees. Citation analysis is an important way of evaluating the quality of scientific research, and is widely used to review the contributions of scientists in a given discipline at any research institution. Fellegi and Sunter (1969) pinpointed citation matching problems many years ago. More recent work extensively addressed the problems and attempted to design useful solutions (e.g., Han et al., 2004; Pasula et al., 2002).

Confusion caused by name variants has had varying effects on different types of digital repositories. Roughly defined, there are two categories of digital repositories:

- Institutional
- Disciplinary or subject-based.

Institutional repositories are normally run independently on the server of a university or research organisation, which collects the work of its own employees (Lynch, 2003). Generally speaking, due to a limited number of scholars in each research field inside an institution, name ambiguity has not been a big concern in these digital archives. Naturally, this hypothesis does not take data harvesting from external data providers into account. The majority of institutional repositories are still in their infancy and name ambiguity receives little attention from both authors and readers.

On the other hand, disciplinary repositories have experienced much difficulty with authors' names. Such repositories commonly seek national and international co-operation for self data archiving, which can create confusion. The bigger a disciplinary repository is,

the more scholars tend to deposit their articles, and the more users search its database, thus raising the possibility that authors' names will present conflicts and ambiguities.

Among the most successful disciplinary repositories, arXiv has attracted some 170,000 brief papers in physics, mathematics and computer science, with almost 3,000 new submissions coming in each month (Cornell, 2001). It receives about 2 million visits a week. For 3rd March 2006- one day only - it had 351,628 hits (arXiv 2005). Running an author search using 'Johnson – a popular last name – and adding 'N.' as the first name initial on the arXiv's physics server as a test, as many as 107 articles are returned with several variations of the name: N. Johnson, N.F. Johnson, Neil Johnson, Neil F. Johnson, N.P. Johnson, and R.A. Johnson. The last two authors have only one article for each, while the former four names appear to be the variants of one physical person at Oxford University. Readers will feel overwhelmed if more articles are added in the future with the same form of names but belong to different people.

In the traditional cataloguing process, the variation of author names is controlled by establishing the authority files to represent the official name of the author, which could be used by other libraries, although work on name authorities tend to be more common in the US than in other parts of the world. Variants of a name, including spelling variants, are listed as cross references when the official name is established. Such name authority control is organised in the US by the Library of Congress (LC) and is also maintained by and available through OCLC.

For obvious reasons, this traditional authority mechanism is not applicable to digital repositories. First, the LC name authority file is not only created by LC itself, but is also contributed to by hundreds of the Name Authority Co-operative Project (NACO) participant libraries (Van Ryn and Starck 2005). Over the years, with the contribution of cataloguers all over the world, this file keeps growing to a remarkable size. Now, OCLC has also worked on incorporating Chinese language name authority records into its cataloguing utility through co-operation with Hong Kong University Library. However, there is a lack of a centralised organisation that is able to show leadership at national or international levels to control name authority in the digital environment, although some efforts have been taken toward this direction at various levels, e.g., the work of IFLA's Functional Requirements and Numbering of Authority Records (FRANAR) group (IFLA 2005).

Second, the number of books in cataloguing bibliographic databases is far less than the number of articles published (or unpublished) in journals. Writing an article requires less effort than publishing a book in almost every scholarly discipline. Digital repository managers will have to work with more authors' names without having too much hope for collaboration from other repositories. There appears to be no single digital repository at this moment that has been able to claim successful control of its author data. If there is one in the future, the control may only be limited to the local level. Nationwide, name authority control for digital repositories is still a dream for a long time to come.

Previous efforts

Name ambiguity is a common problem associated with information access and citation analysis for many types of digital collections. Varied strategies have been adopted to solve the problem. An example is the work of the Eisenhower Library at the Johns Hopkins University, which tried to enhance searching in its Levy Music Collection (Choudhury et al. 2000; DiLauro et al. 2001; Warner and Brown 2001).

With more than 29,000 pieces of sheet music in this collection, librarians at Johns Hopkins decided to develop automated ways to reduce the amount of human labour and time involved in the process of disambiguating personal names. An automated name authority control system was created to associate each name with an individual. A series of research methods was applied to identify the authorship of the music works.

Basically, the personal names subset of the Library of Congress name authority file was loaded into a local database. Context from notes fields in the authority file was retrieved to compare to the names in the Levy Collection. The characteristics of matched names were contrasted against the characteristics of unmatched names, which yielded correlations that helped deduce confidence levels for name matches. A confidence level below an established threshold triggered manual processing.

A Bayesian probabilities model was then applied to produce the confidence measure. High confidence was used to modify probabilities until they stabilised. In the analytical process, not only was data from the authority file notes fields accepted as evidence, but this research also relied on publication date and author birth/death dates to disambiguate names. The commonness of a name was also taken into consideration. Eventually, this work confirmed that automation is possible in processing personal names for digital projects.

Similarly, another computer system was built to assist precisely named entity matching in digital collections at the Columbia University Library (Davis at el., 2003). This system used computational linguistic technologies to enrich catalogue records and improve access to scholarly digital image collections.

Facing the same problem of dealing with a large volume of image collections with variation of names, researchers at Columbia were concerned about the accessibility of their online materials. What they did was to measure the frequency of occurrences of names identifying the images. Two approaches were employed in their experimental work:

- the use of a named entity finder,
- the use of authority lists.

The system they devised sequentially removed modifiers in names being searched and repeatedly ‘decayed’ the particulars to make names more general. It recorded occurrences of name variants and adjusted the algorithm to “use high-precision, low-recall matched as seeds for correctly matching more ambiguous terms” (Davis et al., 2003: 127). Then, the tool generated specific keywords to allow itself to optimise searching processes. For example, a project name variant can be decayed sequentially as the following:

- William R. Thorsen House (Berkeley, Calif)
- William R. Thorsen House (Calif)

- William R. Thorsen House
- William Thorsen House
- Thorsen House
- The Thorsen

Turning the discussion from general digital collections to digital repositories, it is easy to find similar efforts for exercising name authority control. In this part of the digital world, the majority of research projects, if not all, deal with the archives of textual materials, namely, articles, reports, book chapters, etc. Research purposes are explicitly set to disambiguate author names in order to facilitate information retrieval, standardise citation analysis, and gather scientific data. Usually, it is those researchers with a background in computer science and engineering, rather than library science, who are actively involved. Naturally, programming and statistics are the most popular research tools.

Among others, Han and his colleagues selected two supervised learning approaches to clarify names in author citations (Han et al., 2004). The first approach uses the naïve Bayes probabilities model. This captures all authors' writing styles by applying positive training citations based upon probabilities. It also emphasises prediction for the most likely canonical names from citation databases. The second approach uses a discriminative model that detects the distinction between different authors from both positive and negative training citations. It uses a distance measure to weigh different citation attributes such as co-authors, article titles, and journal titles.

Both approaches were applied to two types of data: data from publication lists on the Web, and data from the Digital Bibliographic and Library Project (DBLP), an open access database that provides bibliographic information on major computer science journals and proceedings, in Germany with more than 300,000 bibliographic XML records (http://www.informatik.uni-trier.de/~ley/db/index.html). The researchers found that co-author names were the most robust attributes to disambiguate names. Also, journal title words worked better than article title words. Although this research was only experimental, it did produce some insights into the advantages of using statistical methods to resolve name conflicts involved in like studies.

In another research study, scientists investigated a system-oriented solution for name authority control for authors and publication venues (Hong et al., 2004). Beginning by identifying 'semantic' irregularities in names, this research concentrated on three core elements as building blocks:

- linear change – which refers to a name changing from A to B;
- split – which delineates one name entity being split into multiple names;
- merge – which involves multiple name entities becoming one.

Further, the research explored how to use two system functions – *update* and *search* – to support these three core elements. Both *update* and *search* were designed to work on normalising name variants in the SQL query environment. Both tried to catch data changes in database tables and make links between legitimate matches of authors' names.

A test was carried out in an OpenDBLP system at Pennsylvania State University (http://opendblp.psu.edu/). This is a fully DBMS-based system with a Web service-based programming interface. A testing Web site was created on the Pennsylvania State University domain. Some demonstrations of the work are presented in their research article. The results showed that many of the name authority problems were expressed and solved with the research strategies (Hong et al. 2004).

Working with a large scaled disciplinary repository, Cruz et al presented in their article some practical as well as proposed approaches for handling authors' names in the RePEc Economics digital repository (Cruz et al. 2000). The system applied a decentralised archiving model to characterise its maintenance. It provided each participating institution the authority to control the content of its archives, which resided individually on an anonymous Web server and collaboratively with each other. With such a multiply centred scheme design, the process of name normalisation could be possibly implemented, according to the authors, by registering personal information of individual authors in separate servers. Hypothetically, the combination of personal name and the period of his/her registration with a particular digital repository – 'significant time' – would be unique enough to serve as the true identifier.

An internal handle was empowered to link all available personal data. The handle could organise such data into existing resource metadata as collected by RePEc. Being unique and stable, the internal handle combined data from both the handle of the resource and the names of the authors. However, it did not work well in situations where more than one person shared the same name.

The RePEc database had about 1,800 resources that had at least one registered author at the time their paper was written. Therefore, it was realistic to utilise authors' registration data to help identify the relationship between the work and the creators. It only remained questionable if authors updated their registration after their physical affiliations were changed. Yet still, Cruz and his colleagues discovered a potentially important solution with practical significance for name disambiguation in digital repositories.

Similar studies are plentiful. Each has its own concentration and solutions. For instance, Atkins et al worked on reference linking with Digital Object Identifiers (DOIs) "to enable readers to find content on the Internet with a persistent and reliable identifier" (Atkins et al. 2000). Pasula et al. used a relational probability model to handle the issues of identity uncertainty and citation matching (Pasula et al. 2002). Many of these studies have become visible as conference presentations and later published in conference proceedings, mainly in the fields of computer science and engineering.

Future solutions

With the exception of the work of Cruz et al, research projects of this type primarily concentrate on how to solve existing problems instead of how to prevent the problems from happening. The development of digital repositories is still at the initial stage. Even the oldest disciplinary repository, ArXiv, has only a 'hisotry' since 1991. It is important to keep improving the system so that problems are not only caught after data entry, but are prevented at the start. Taking personal names as an example, it is useful to detect the reasons causing

name ambiguities and develop solutions to disambiguate names. However, it is more critical to make suggestions to digital repository applications or to customise existing software, metadata, and databases so that name identifiers can become most unique at the time data are deposited.

Most of the work mentioned above took a research approach rather than a practical orientation because the authors were not typically involved in the management of digital repositories. A manager will be happier with a system that brings into the database an appropriate dataset than with a system that requires additional mechanism to reorganise the incoming dataset. Yet, this request can be potentially satisfied by adding metadata elements and requiring users to input the data at the front-end.

To date, many digital repositories have relied on authors depositing materials themselves in the archive. However, some managers of repositories believe that managed archiving by an intermediary is preferable. In the self-archiving model, authors are encouraged, if not required, to deposit their work into the repository themselves. Despite different types of repository software applications such as DSpace (http://dspace.org) and EPrints (http://eprints.org) having been implemented, all bear a user-friendly interface to let depositors input relevant data as metadata elements and upload actual files in textual or other formats. The metadata elements identify the actual files and facilitate OAI-PMH data harvesting. Some systems employ a flat file strategy that places all metadata files under designated folders. Others use relational databases to store both metadata and eprints files. For example, in EPrints, the 'eprints' and 'users' database tables have a list of metadata fields to store the information about authors and eprints.

Most repositories use author names in metadata as the identifier to retrieve data for the function of author search. Such a single element identifier strategy can optimise query efficiency for online searches, but cannot guarantee query effectiveness. The confusions mentioned above clearly demonstrate the ineffectiveness. A new approach to improve the limitation has been discussed as having a combined identifier that will be able to disambiguate author names. The favorite combination is "author name + birth date" because it is very rare that two persons share the same name and birth date. However, this approach is not particularly feasible as authors tend to be hesitant to provide their birth date information.

Cruz et al (2000) proposed an "author name + registration time" approach to solve the problem, which is one of the few suggestions that has practical value and is workable for data identification. Unfortunately, this model must be ignored by other repositories because RePEc applies a very different organizational policy, totally unique from all other repositories (E-LIS, a disciplinary repository in the domain of library and information science (http://eprints.rclis.org/), has a similar decentralised management, but is also different in other regards). Each contributor author establishes his/her relationship with RePEc by registering to its satellite server, which served one or more institutions. When the author changes physical affiliations, s/he needs to re-register to another appropriate server. Therefore, registration time works well to verify author identities.

For a centralised repository, which is the norm in practice, the registration time approach is not useful since all registration is done through only one server. When a contributor author makes a move to another organization, it is not necessary for him/her to re-register to the

server. Hence, registration time will not serve as uqiue identification. To deal with the problem, Cruz et al have triggered a useful way of thinking: author affiliation data is a good candidate for a composite identifier. Since author registration time is not applicable to many repositories as a good indicator of affiliation, we may propose another type of affiliation, that is, the working or research place of an author (i.e., the organisation where the author is officially affiliated at the time of article publication). It is usually indicated in articles.

It is also necessary to include the publication date of an article as an identifier. This prevents any confusion resulting from a change of workplace for an author. This will also be able to identify indiividua articles. Hence, we have "author name + affiliation + publication date" as the composite identifier for data retrieval. It is extremely unlikely that two physical persons will share the same name and work at the same institution/discipline at the same time of article publication. Moreover, none of these pieces of personal information is confidential. They all appear in publications. Both authors themselves and repository managers can feel comfortable releasing the information.

It must be noted that this solution is only sufficient for distinguishing authors with the same name, but does nothing to solve name ambiguities or to deal with the other two types of name problems: name changes and different spellings of one name. In practice, different spellings of one name are the most common cause for an inaccurate return of an author name search. A possible supplementary solution is also straightforward. Why not let authors add the variants of their own names in repository applications? Most of the current repositories do not provide such an opportunity to help authors. What is required is an additional field that stores this piece in existing author information.

By adding these additional metadata elements, name ambiguities of the three common types will be efficiently solved. In the case where several people, who work in separate institutions, use the same name (Han et al. 2004), their workplace will help maintain their uniqueness. It is more efficient to ask authors to list all variants of their names in publications than to let others figure out the variants. Who can be more discernible than authors themselves for the variations of their own names? Furthermore, it will not cause much of a burden for the author when depositing items, but will save significant time for everybody else.

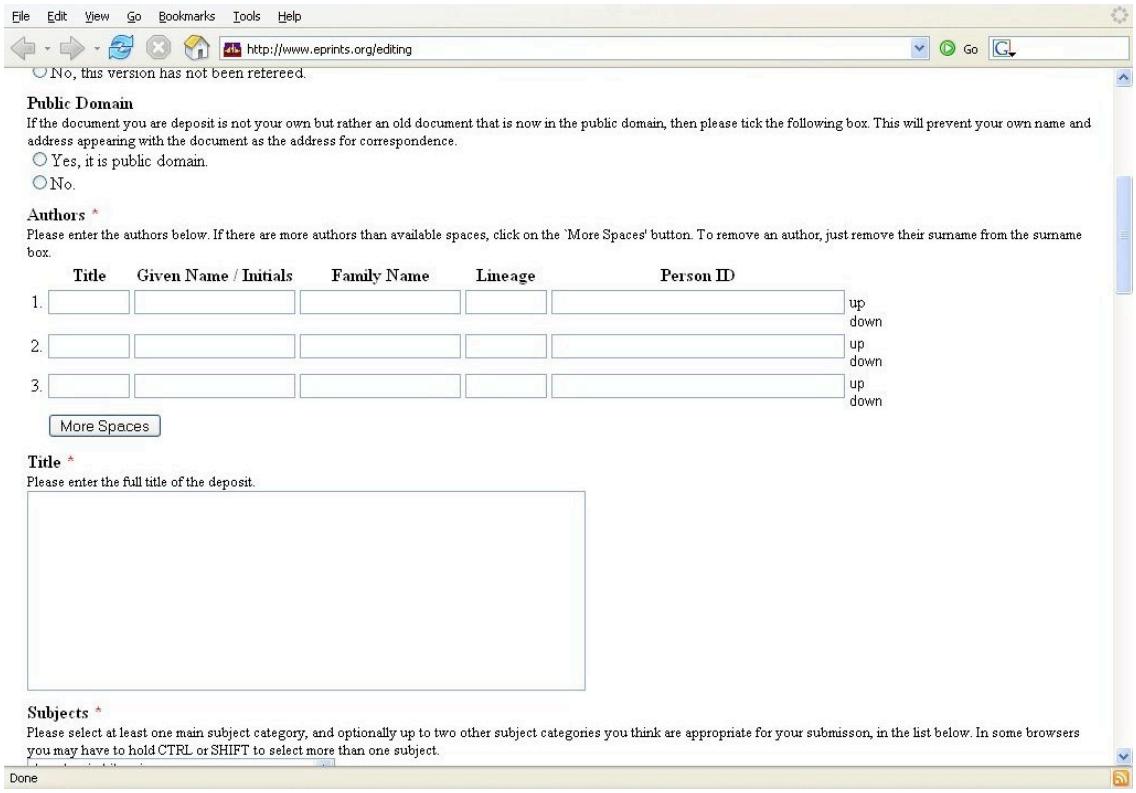
With regard to implementing composite identifications and alternative names, the re-engineering of metadata fields, database tables, and interface appearance is necessary. Most current repositories have adopted Dublin Core as their metadata standard. Yet, the Metadata Encoding and Transmission Standard (METS) and other metadata schemas have also become increasingly popular. It is the task of the developers to analyse the structure of a particular metadata schema being employed so that the proposed metadata fields can be accommodated.

Adding metadata fields to database tables that store metadata involves extra work. The work requires a thorough consideration of the data in order to avoid any potential damage to existing data. For instance, if Eprints is the software used the repository developers must edit ArchiveMetadataFieldConfig.pm (Eprints 2005). Then, the database tables will have to be revised, if not totally erased and recreated. One way to avoid erasing data is through re-engineering the data structure in SQL, which requires great care. Of course, the best way is to make modifications to the application structure – metadata and database –when a new repository is first implemented.

The application interface needs to be slightly customised to support corresponding changes at the backend. Compared to changes to metadata and databases, the customisation can be simple. Additional textboxes need to be added to accommodate data entries for author affiliation and different names of the author. It is important that the affiliation field is set to be "required".

In the Eprints software (a popular repository application developed by Southampton University) the interface presents an 'author' field as a required field that has textboxes in one line for the inputs of an author's name including Title, Given Name/Initials, Family Name, Lineage, and Person ID as shown in Figure 1. Different lines are presented for more than one author of a work. Possible changes to this field can be proposed as one line of textboxes for the name of an author, and an additional line allowing the author to enter his/her alternative name. An expanding button could be designed to bring up more lines if needed. The same mechanism can be applied to other authors of the same work.

Figure 1. EPrints interface for authors to input metadata and deposit eprints archives



The current EPrints interface does not have a place for authors to enter their affiliation information. Therefore, a different textbox would be required to meet this need, asking the author to enter his/her affiliation of publication time, if different at the depositing time. In previous practices, an e-mail address was a required field for the sake of identification (Cruz et al. 2000). This application was questionable, however, because many people have more than one e-mail account, and some accounts are commercial such as Hotmail and Yahoo.

At first glance, the changes seem to be costly. The cost level may increase dramatically for digital databases that have already collected a great amount of data. Furthermore, the changes could possibly damage existing data, which would be devastating for a repository. It is natural, therefore, that managers may doubt the benefits of the proposed changes.

However, the level of difficulties and dangers is by no means as high as it first looks. For existing repositories, application developers need to work closely with repository managers. Complete understanding of the data and data structure is definitely the key to success. Any change to a database will have to be tested repetitively against data in a test environment before it is executed to live data. In the real world, changes of this scale occur in many databases and appear to be safe everywhere. It is normal that when the business of any organization has experienced changes its database may need modifications accordingly. Digital repositories usually have databases with relatively simple structures and may only store a handful of metadata fields and eprints files

For newly deployed repositories, the changes will bring minimum effects on datasets. This, of course, does not mean precautions are not required. As a matter of fact, serious testing will be required as it is for existing databases. The ideal solution is to make all necessary changes before a repository is implemented and there is absolutely no data in databases. Preferably, most developers of software for digital repositories will be really involved in the enhancement process and contribute to creating better systems for users.

No matter what condition a repository is in, it is worth the effort to make good modifications to current digital archiving applications that have room for improvements in their structure. The benefits of perfecting search functions to improve efficiency and effectiveness can outweigh the costs of re-engineering in many cases. The purpose of digital preservation is to maximize the use of digital materials. Well designed retrieval strategies will ensure the success of a repository.

Conclusions

Digital repositories have the mission of providing open access to the general public (Suber, 2005). The attitude of users in exploiting a particular repository is greatly influenced by its quality. Factors which affect a particular repository's quality include its content and the ability of the repository to return appropriate data to users as a result of a request. It is the goal of managers to aim for a quality repository.

Unfortunately, authors' names, because of the existence of variations, have negatively affected the retrieval capability of many repositories. How to optimise search quality should be a primary consideration for repository managers. This paper has highlighted current name search problems in the practice of digital archives. Possible solutions with practical applicability are suggested including:

- use of composite identifiers that combine author name, publication date, and author affiliation

- asking authors to input the variants of their name, if any, at the time of depositing articles.

Unlike previous studies that focused on strengthening existing data, this paper attempts to convince repository managers that emphasis should be on configuring a database appropriate dataset. It is believed that the proposed changes will help disambiguate name variations and thus enhance online searches. If implemented properly, these modifications will add extra metadata fields for unique identification and, consequently, make digital repositories more user-friendly.

Recently, the Joint Information Systems Committee has launched the Digital Repositories Programme attempting to bring together a programme of work relating to digital repositories (http://www.jisc.ac.uk/index.cfm?name=programme_digital_repositories). It tries to enhance the coordination of people from across various domains in the development of digital repositories by facilitating discussions about technical and social issues in digital practices. For example, one of its projects – VERSIONS - has been designed to distinguish versions of an article published by an author. This will represent an area in need of further research.

References

arXiv (2005), A*rXiv Monthly Submission Rate Statistics*, Available at: http://arxiv.org/show_monthly_submissions

Atkins, H., C. Lyons, H. Ratner, C. Risher, C. Shillum, D. Sidman, and A. Stevens (2000), "Reference linking with DOIs", *D-Lib Magazine*, Vol. 6 No. 2, Available at: http://www.dlib.org/dlib/february00/02risher.html

Bilenko, M., R. Mooney, W. Chen, P. Pavikumar, and S. Fienberg (2003), "Adaptive name matching in information integration", *IEEE Intelligent Systems*, Vol. 18 No. 5, pp. 16-23.

Choudhury, G.S., C. Requardt, I. Fujinaga, T.G. DiLauro, E.W. Brown, J.W. Warner, and B. Harrington (2000), "Digital workflow management: The Lester S. Levy digitized collection of sheet music", *First Monday*, Vol. 5 No. 6, Available at: http://www.firstmonday.org/issues/issue5_6/choudhury/index.html

Cleveland, G. (1998), "Digital libraries: definitions, issues and challenges", *IFLA Universal Dataflow and Telecommunications Core Programme.* Occasional Paper 8, Available at: http://www.ifla.org/VI/5/op/udtop8/udtop8.htm

Cornell (2001), "Online physics archive that is transforming global science communication, 'arXiv.org,' is moving from Los Alamos to Cornell University", *Cornell News,* Available at: http://www.news.cornell.edu/releases/July01/ginsparg.archive.ws.html

Cruz, J.M.B., M.J.R. Klink, and T. Krichel (2000), "Personal data in a large digital library", *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

Davis, P.T., D.K. Elson, and J.L. Klavans (2003), "Methods for precise named entity matching in digital collections", *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, Houston, Texas, 125-127, Available at: http://www1.cs.columbia.edu/nlp/papers/2003/davis_al_03.pdf

DiLauro, T.G., S. Choudhury, M. Patton, and J.W. Warner (2001), "Automated name authority control and enhanced searching in the Levy Collection", *D-Lib Magazine*, Vol. 7 No. 4, Available at: http://www.dlib.org/dlib/april01/dilauro/04dilauro.html

Eprints (2005), "Modifying the metadata fields in an archive", *EPrints 2 Technical Documentation*, Available at: http://www.eprints.org/documentation/tech/php/intro.php

Fellegi, L.P. and A.B. Sunter (1969), "A theory for record linkage", *Journal of the American Statistical Society*, Vol. 64 No. 328, pp. 1183-1210.

French, J. C., Powell, A. L., Schulman, E. & Pfaltz, J. L. (1997), "Automating the construction of authority files in digital libraries: a case study", In C. Peters & C. Thanos (eds.), *Research and advanced technology for digital libraries, first European conference, ECDL '97*, pp. 55-71, Berlin: Springer.

Han, H., L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis (2004), "Two supervised learning approaches for name disambiguation in author citations", *Proceedings of the Joint Conference on Digital Libraries*, Tucson, Arizona, pp. 296-305, Available at: http://clgiles.ist.psu.edu/papers/JCDL-2004-author-disambiguation.pdf

Hong, Y., B.W. On, and D. Lee (2004), "System support for name authority control problem in digital libraries; Open-DBLP approach", Available at: http://nike.psu.edu/publications/ecdl04.pdf

IFLA (2005), *Functional Requirements for Authority Records: A Conceptual Model*. IFLA UBCIM Working Group on Functional Requirements and Numbering of Authority Records (FRANAR), Available at: http://www.ifla.org/VII/d4/FRANAR-Conceptual-M-Draft-e.pdf.

Lynch, C.A. (2003), "Institutional repositories: essential infrastructure for scholarship in the digital age", *ARL Bimonthly Report*, No. 226, Available at: http://www.arl.org/newsltr/226/ir.html

Lynch, C.A. and J. K. Lippincott (2005), "Institutional repository deployment in the United States as of early 2005", *D-Lib Magazine*, Vol. 11, No. 9, Available at: http://www.dlib.org/dlib/september05/lynch/09lynch.html

Pasula, H., B. Marthi, B. Milch, S. Russell, and I. Shpitser (2002), "Identity uncertainty and citation matching", *Proceedings of Neural Information Processing System: Natural and Synthetic*, No. 15, Available at: http://www.cs.berkeley.edu/~milch/papers/nipsnewer.pdf

Suber, P. (2006). *Open Access Review*, Available at: http://www.earlham.edu/~peters/fos/overview.htm

Van Ryn, P. and W.L. Starck, eds. (2005), *NACO Participants' Manual*, 3rd Edition, Washington DC: Library of Congress, Available at: http://www.loc.gov/catdir/pcc/naco/npm3rd.pdf

Warner, J.W. and E.W. Brown (2001), "Automated name authority control", *Proceedings of JCDL*, June 24-28, 2001, Roanoke, Virginia, 21-22.