11-2010

# Survey of Author Name Disambiguation: 2004 to 2010

Sarah Elliott
sarelliott@gmail.com

*Library Philosophy and Practice 2010*

ISSN 1522-0222

# Survey of Author Name Disambiguation: 2004 to 2010

Sarah Elliott

## Introduction

As resources are encoded in various metadata schemes, digital libraries grow, and the internet and metadata encoding in general move toward interoperability, the problems of name and identity disambiguation pose problems in metadata development. Databases and search features must be able to determine whether the person who wrote article A also wrote article B. Searchers may want to call up all items written or created by a particular person. Researchers may need to determine exactly who wrote an article in order to pursue contact that author to propose future collaboration or ask follow questions about the data. Most metadata practices do not easily support name disambiguation and the problem grows as the number of resources and varieties of metadata also grow.

Smalheiser and Torvik offer an excellent description of the four main challenges that impact name disambiguation. First, the same individual might write under more than one name due to “orthographic and spelling variations,” spelling errors, name changes (for marriage, etc.), the use of pseudonyms or pen names (Smalheiser & Torvik, 2009). Secondly, there are many different people with the same name. Perreria, et al. identify this problem as a situation with polysemes, as opposed to the first case, in which there are synonyms (2009). Some names, like John Smith, appear again and again, creating the challenge of distinguishing one author from another. Thirdly, according to Smalheiser and Torvik, metadata, especially in article databases or on blogs, may be incomplete; many times only the initials of the first name and middle name are included in article databases and not the full name. Lastly, many articles are multi-authored and interdisciplinary (2009). The growing trend of interdisciplinary work makes it more difficult to tell whether the John Smith publishing about linguistics is different from the John Smith publishing in biochemistry, whereas in the past, two identities might be safely assumed.

Han, et al. point out that “[n]ame ambiguity can affect the quality of scientific data gathering, can decrease the performance of information retrieval and web search, and can cause the incorrect identification of and credit attribution to authors” (2004). They give an example in Digital Bibliography & Library Project where one author page, which should reflect a single author’s work, actually has citations that belong to three separate people (Han, et al., 2004). With so much as stake in resolving the name ambiguity problem, researchers have been working hard on discovering a solution, especially during

the past few years.

### Human or Machine Disambiguation

There is disagreement about whether information science researchers should focus on manual or automatic name disambiguation methods. Smalheiser and Torvik list two reasons why manual disambiguation is not always possible. First, there is the problem of very large digital libraries that often harvest their metadata from other sources. Since they are mixing records from many places that each might use their own form of name authority, and since they are huge, it is not practical to manually create and fix name authorities for those libraries (2009). Smalheiser and Torvik's second reason is that internet searches will always looks at many more resources than can practically be manually cataloged (2009). Search engines and other methods for organizing and finding information on the World Wide Web must implement workable automatic name disambiguation methods.

In contrast to Smalheiser and Torvik, Veve argues that automated name matching will always fail and so human intervention will always be required (Veve, 2009). Veve explains that "the few endeavors that have tried [name authority control in XML], such as the systems for automated generation of authority control, have only been successful in extracting names from XML records but not in turning them into reliable access points" (Veve, 2009). She says that a person will always need to check whatever name authority records computers automatically generate.

It seems that the question is really whether there is a higher priority to provide probabilistic name disambiguation to all data using automatic methods or whether to provide name disambiguation with a high degree of certainty to a small set of records using manual methods. Research is continuing on both approaches and it is possible that the answer to which method is better depends upon the particular project at hand. Perhaps in some cases, a hybrid approach might be possible.

## Approaches to Name Disambiguation

A variety of institutions and individuals are working to address the problem of name disambiguation with several different approaches. Individuals working in name disambiguation research are from a variety of backgrounds including librarians and information scientists, as well as computer scientists. Projects include local efforts as well as highly coordinated national and international systems. Both manual and automatic name disambiguation methods continue to have new literature published. Ultimately, any system that stores metadata about information resources created by people will run into the problem of author name disambiguation and will need to decide between various methods.

### LCAF

As many writers on the topic of name disambiguation are quick to point out, libraries have been in the name authority business for a very long time. The Library of Congress Authority File (LCAF) contains name authority records that have all been manually created, as part of the Name Authority Cooperative Program (NACO), which has over 400 member institutions worldwide (Van Ryn & Starck, 2005). The process of keeping name authorities in order, both in a national system like the Library of Congress and at a local catalog level is highly labor intensive, but by sharing the work among participating libraries, LCAF has been able to keep up with a substantial and useful portion of author names from

monographic book publications. Since the process is so labor intensive, however, LCAF has not been able to expand to include comprehensive coverage of author names for individual articles, let alone newer information resources like blogs.

## The Names Project

The British Library's Names Project is a response to the growing number of institutional repositories in the United Kingdom and the resulting need for authority control of author names for articles. In 2008, there were eighty-seven institutional repositories in the United Kingdom with a combined total of over 300,000 records (Hill, 2008). Very few names for author articles have authority files in LCAF and the repositories were encountering a problem with having no standardization for names that were entered.

The Name Project is still under development in a prototype stage. It is designed to assign each author a unique ID number that will link to various forms of that author's name. This is considered "access control" rather than "authority control," since no authorized headings will be established. Hill says that while this form of organization may not have made sense in previous generations, 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" (Hill, 2008).

As a step forward toward the Names Project, Joint Information Systems Committee requested a report that compiled a survey of other institutions involved with name authority control. The report includes a very brief description of the type, purpose, methodology used, responsible parties, and references for each institution surveyed. The list of institutions includes both libraries and commercial systems: DAREnet, NACO, National Library of Australia, New Zealand Electronic Text Centre, OCLC, IraLIS, ONESAC, RePEc, Elsevier, ProQuest, and Thomson-ISI. In addition, the last section mentions EthOS, Je-S, the UK PubMed Central, and Zetoc as sources the Names Project might consider for extracting data (Danskin, et al., 2008). This environmental survey for the Names Project is a good starting place for understanding and investigating many different library name authority control initiatives.

## ULAN

One example of library and library related manual authority control comes from the Getty. The Getty has created a controlled vocabulary of names from various Getty projects called the Union List of Artist Names (ULAN) (Paul J. Getty Trust, 2010). As the ULAN website explains, ULAN is useful for finding standardized names for cataloging, as access points for searching for artists, and as research tools, since each record has information about the artist. ULAN was designed for "museums, art libraries, archives, visual resource collection catalogers, bibliographic projects concerned with art, researchers in art and art history, and the information specialists who are dealing with the needs of these users" (Paul J. Getty Trust, 2010).

Each record in ULAN is manually created and maintained. ULAN provides an excellent resource for name disambiguation for the 120,000 artists included. However, like all manually created and professionally maintained name authority files, it cannot help keep up with name disambiguation for the vast numbers of new resources that are created daily.

ULAN and LCAF both work by establishing a preferred heading for each name. ULAN contains a field in every record called the "LC Flag" that

indicated whether the preferred form of a name in ULAN is also the preferred form in LCAF. While most names that appear in both authority files will have the same preferred form, some names in ULAN will have a different preferred form than the entry for the same person in LCAF (Paul J. Getty Trust, 2010). Even in such similar manual name disambiguation endeavors, differences arise and systems that use both LCAF and ULAN may still have problems reconciling names.

## An Efficient Manual Approach from UTL

While some name disambiguation projects create massive authority files like LCAF and NACO, individual libraries must determine how to apply name authority control in areas that go uncovered my major interlibrary projects. In her paper, "Supporting Name Authority Control in XML Metadata," Marielle Veve explains a method that "consists of a simple manual approach to extract and create name access points that effectively reduces time and research efforts by efficiently setting priorities, identifying critical descriptive areas in the digital transcriptions, and identifying the most appropriate biographical resources to consult" (2009, p. 42).

To give a little background for the project Veve explains that the University of Tennessee Libraries (UTL) were faced with creating access points for digitized manuscripts which were encoded with the Text Encoding Initiative (TEI) scheme and which catalogers were going to encode using the Metadata Object Description Schema (MODS). UTL encountered a challenge in creating MODS records for items encoded in their TEI files because names were not in LCAF were cataloged inconsistently by different catalogers. These cataloging inconsistencies made caused difficulties in differentiating name, especially since many people in the collection had similar names, were related, or used nicknames. Catalogers also encountered questions when misspellings of the name occurred in the original documents (Veve, 2009).

UTL found a solution to the authority problem in a workflow that permitted important name authority work to be done without interrupting other cataloging tasks or requiring too much time to be devoted to authority work. Veve details all the steps in this workflow, but here a summary will suffice.

In the UTL workflow, one cataloging librarian, who has experience in creating records for the Name Authority Cooperative Project (NACO), is charged with carrying out all the authority work for the incoming records in order to avoid potential inconsistencies in name choices. This authority librarian organizes the incoming records by collection and then makes lists of all the names in the documents, their variations, and their exact location in the documents. After listing the names and tallying the occurrence of each name, the authority librarian then follows a criteria for deciding which names will be searched in LCAF and receive headings if not found. If the name appears in three or more separate files or if the name belongs to an important historical figure, that name receives an authority heading; other names do not receive any authority heading (Veve, 2009). This practice ensures that cataloging time is not wasted on names that appear only once or twice in the entire collection and so do not really need authority control. The authority librarian is able to devote her efforts to the names that really do require disambiguation.

After establishing which names will receive headings, the authority librarian returns to the text immediately surrounding the name and to finding aids for the collection to find identifying information for that individual. If the librarian needs more information than the text offers, she then searches other sources, including Google Book Search, "Political Graveyard—A Database of Historic Cemeteries (http://politicalgraveyard.com), the state finding aids via the state library or state historical society websites, Genealogybuff.com, the

Biographical Directory of the United States Congress (http://bioguide.congress.gov/biosearch/biosearch.asp),... the Civil War Rosters website (www.geocities.com/Area51/Lair/3680/cw/cw.html)" and the Tennessee Genealogy and History Web (TnGenWeb.org) (Veve, 2009). From these sources, the authority librarian compiles a very brief statement of important facts about the individual and uses these to identify the correct record in LCAF. If the name cannot be found in LCAF, the authority librarian then searches the OCLC Connexion Bibliographic File in both author and subject field to see if anyone else used the name as an access point (Veve, 2009).

Veve argues that this workflow proves reasonably efficient with heavy constraints on time and resources. The criteria used to decide which names require subject headings save time by avoiding time intensive work on obscure name headings. By using a series of biographic and historical tools, the authority librarian quickly identifies the necessary information for the names that do require authority headings. Also, by separating the authority work from the rest of the cataloging activities, the rest of the catalogers are able to work quickly in creating the rest of the MODS record (Veve, 2009). This workflow can serve as a model for those projects where manual name disambiguation is required in a local library; it provides a reasonable compromise between authority control for every name in a collection and a lack of any authority control.

## Harvesting Information from "Personal-Portfolios"

Salo discusses options for getting name authority records for institutional repositories. In her paper, "Name Authority Control in Institutional Repositories," She points out that not only is institutional repository software not sophisticated enough to use an authority file outside the repository, but she discusses where name authority data might come from when software becomes more sophisticated. Salo points out that most name authority files created thus far are for authors of books, while digital repositories contain mostly articles, which means that LCAF will not be much help to institutional repositories. Authority records generated by the institution's faculty records may be helpful, but will not cover all names since many articles are co-authored across multiple institutions (Salo, 2009).

Even though software is not currently sophisticated enough to use such resources, Salo theorizes that some new websites that include author registration might be able to provide data to institutional repositories. She says that "experimental personal-portfolio services" are becoming popular on the web and many of these include some form of name authority control. Some examples are RePEc, CrossRef, Research Crossroads, and Cornell University's VIVO. In addition, the International Standards Organization is developing the International Standard Party Identifier, the Joint Information Systems Committee (JISC) is working on the Names Project, and the Netherland's SURFfoundation is creating a Digital Author Identifier (Salo, 2009).

Salo adds that there are questions about each of these regarding who will maintain quality and how repositories will be able to link to multiple sources. OpenID relies on authors to maintain authority control over their own names, but Salo reasonably asks whether authors will be willing to put in the time to link their names and the names of their coauthors in institutional repositories (Salo, 2009). Hill, however, argues that in some settings, like academic repositories, academic authors have incentives to register their work since they want their research to be recognized (Hill, 2008).

Salo concludes that right now all institutional-repository managers can hope to do is clean up projects on their databases. She recommends several options for future development. Software companies need to design software

will better functions for authority clean up and management. She specifically recommends BibApp (http://code.google.com/p/bibapp/) as a project that could be helpful (2009, p. 260).

## BibApp

BibApp is software that allows people to search for campus experts on a particular topic. BibApp imports citations from other locations. For example, the page in the BibApp wiki called “BibApp Citation Import Mappings” explains how fields from RIS, RefWorks SML, and Medline are mapped into BibApp fields. For example, “a1” and “au” in RIS are mapped into the “name strings, role: author” field in BibApp, while “a2” and “ed” are mapped into “name strings, role: editor.” Salo’s suggestion is that institutional repositories and digital libraries can harvest data from databases like BibApp to help indentify authors (Salo, 2009). Since software has not been designed for a project like this, the practical details have yet to be determined.

## FOAF

Friend of a Friend (FOAF) is a system for encoding information about people so that many types of relevant information are combined in one document. For example, FOAF might encode someone’s name, email address, picture, and personal homepage. FOAF has implemented referring to people by their email addresses, since email addresses are unique identifiers (Graves, et al., 2007). Graves says as FOAF compatible system can look at two FOAF documents and if they both contain the same email address, the computer knows they refer to the same person.

FOAF uses the Resource Description Framework (RDF) to encode relationships between people and information about them. The complete documentation for FOAF can be found through the FOAF webpage. FOAF has proved useful to social networking sites and search engines like Google and Yahoo are beginning to use it.

FOAF offers potential help to the author name disambiguation problem, as it does serve as an identity disambiguation tool. However, it seems to rely heavily on people being willing to encode information about themselves and work would be required to use FOAF in digital libraries that use other metadata schemes. FOAF does not appear to be an ultimate solution for digital libraries, especially since digital libraries frequently have only a last name and first initial to use in their efforts. Even if a FOAF record for that individual exists, if it does not list the article in question, it will probably not be able to help discover the identity of the author.

## Stylometry

One intriguing automatic approach to author name disambiguation is through stylometry or the study of the style with which an author writes. Smalheiser and Torvik mention this field as a possible area of investigation for future research. Stylometry is being studied for other purposes, such as identifying the author in texts where authorship is disputed and identifying unsigned documents on the internet. Smalheiser and Torvik are quick to point out, however, that stylometry works best when only a limited number of authors are being considered and it does not work well for multi-author documents (2009). It is possible that stylometry may prove useful at some point in resolving questions of polysemes, when several possible authors have been identified by other automatic name disambiguation methods.

## Supervised Learning for Computers

Han, et al. present two different ways of training computers to disambiguate names. Both methods assume there is a database of citations that contain properly disambiguated authors which can be used as training data for the program. Both methods also use three types of data: coauthor names, paper title keywords, and journal title keywords. The first method is a naïve Bayes model, which Han, et al. identify as a generative model, where the system can, in a sense, propose what data is likely to exist. This first model works by calculating the probabilities that certain authors wrote papers with other authors, and other such probabilities. Then it calculates how probable it is that two authors with the same name are the same person or different people (Han, et al., 2004).

The second method is called Support Vector Machines (SVM) and is a form a discriminating method, which uses both positive and negative training to teach the computer to make the distinction between authors (Han, et al, 2004). The SVM method was developed previously for other text classification uses. These two methods seems to be important for the development of name disambiguation as the following, more recent research efforts in automatic methods refer to them as a starting place for comparing the success of automatic methods.

## K-WAY

The k-way method of author disambiguation uses vectors composed of co-author names, paper titles, and publication venue titles (Han, et al., 2005). These variables are manipulated as vectors through Laplacian matricies and clustered together. Han, et al. report that the larger the data set of citations they can feed into their formulas, the more accurate their results.

## WAD

Pereria, et al. propose a method for mining internet information, especially from curriculum vitae and personal web pages belonging to ambiguous authors. They call their method the Web Author Disambiguation (WAD) method and it follows three steps. The first step is to collection information from the web. Their queries take two forms: “unquoted author name followed by the word ‘publications’, followed by unquoted work title (e.g.: Denilson Pereira publications Using Web Information Creating Publication Venue Authority Files)” or “unquoted author name followed by quoted work title (e.g.: Denilson Pereira ‘Using Web Information for Creating Publication Venue Authority

Files’)” (Pereria, 2009).

Once the queries have been performed, the second step is to extract information from the documents the queries retrieved. The extraction involves sorting out each citation from the document, which may contain information other than citations, and then weighting the documents when a single citation is found in more than one place. The final step in the WAD method is to cluster the documents, using a twenty-four step algorithm. The clusters of citations that are formed should each represent the work of a single author (Pereria, 2009).

Pereira, et al. tested their method with a sample of one hundred randomly selected documents. The WAD method had a 90% success rate in identifying single author documents when compared with the results that were

derived manually. The authors point out that failures of the WAD method came from three sources: misspellings and other errors that cause incorrect citations, citations existing in only one place, and non-uniform display problems, such as where the coauthor's name is not included on the title page but is included one citation (Pereira, et al., 2009). These problems that hindered the WAD method underscore the types of challenges automatic methods face because of inconsistent metadata.

### Tang, et al.

Tang, et al. contribute to current research on automated methods in their poster paper titled "A Unified Framework for Name Disambiguation." In this paper, they make the rather startling statement that they "intend to make a thorough investigation of the whole problem" of name ambiguity (2008). Their solution involves some rather complex algorithms, but essentially, it looks at five factors in an attempt to resolve name ambiguity: co-conference, co-author, citation, constraints (feedback from users), t-coauthor.

The first few relationships are fairly straight forward. If two papers are published in the same conference or journal, they have a co-conference relationship; similarly, the co-author relationship means two papers have an author with the same name. The citation relationship indicates that one paper cites another paper. The "t-coauthor" relationship works roughly like this (to paraphrase Tang et al.): If Paper 1 is authored by Name A and Name B and Paper 2 has authors Name A and Name C, and then if Paper 3 has Name B and Name C as authors, there is a 2-coauthor relationship. These relationships are then manipulated into a formula, which Tang, et al. claim "significantly outperforms the baseline methods" (Tang, et al., 2008).

### Author-ity Project

The Author-ity project is a method for disambiguating author names in the Medline database. Phase I of the Author-ity project compared vectors composed of "shared title words, journal name, co-author names, medical subject headings, language, and affiliation—as well as two distinctive feature of the author name itself (presence of middle initial and suffix)" (Smalheiser & Torvik, 2009).

Phase II of the Author-ity project added new factors. First names are extracted from articles and from publishers' websites when available. Name variations, nicknames, and email addresses are also included in the weighting factors. The Author-ity project uses pair comparisons of papers and then clustering rather than just clustering because clusters can cause a single author's work to be split among multiple clusters representing different research areas. The Author-ity project will probably be expanded in the future by adding more techniques and factors to supplement those already being use. Already, the Author-ity project has 98 percent accuracy in "assigning a given paper to a given author-individual cluster" (Smalheiser & Torvik, 2009).

## Conclusion

The library and information science community has recognized the problem of name disambiguation and research is moving forward with a number of strategies. Some research is pursing a continuation of manual disambiguation for library project. Other research, like FOAF and author registry sites encourage many people to contribute to providing information to help identify authors. Still other researchers are pursing automatic name disambiguation programs. Many of these use similar factors, like co-author

names and fields of study, but they are all growing more complex. New research will probably find ways to include even more types of information to build disambiguation probabilities.

There are a number of ways this research could be extended. It could be fruitful to survey digital libraries or institutional repositories to learn what a small subset of metadata projects are using in name disambiguation. Another project could look a the degree of interoperability that arises from various name disambiguation projects, particularly looking at how data is stored and which projects are most willing to help others use their data. The possibility of using automatic methods to assist manual disambiguation should also be pursued; workflows for manual methods could save individual authority librarians considerable work by providing probabilities from automatic methods and compiling references from other data sources.

## References


BibApp Citation Import Mappings. Retrieved March 15, 2010 from http://code.google.com/p/bibapp/wiki/CitationMappings

Danskin, A., et al. (2008). A review of the current landscape in relation to a proposed Name Authority Service for UK repositories of research outputs. Retrieved March 22, 2010 from http://130.88.120.172/names/documents/LandscapeReport26Jun2008.pdf

The Friend of a Friend (FOAF) project. Retrieved March 18, 2010 from http://www.foaf-project.org/

Graves, M., Constabaris, A. & Brickley, D. (2007). FOAF: Connecting People on the Semantic Web, *Cataloging & Classification Quarterly*, 43, 3, 191-202. Retrieved March 16, 2010 from http://dx.doi.org/10.1300/J104v43n03_10

Han, H., Zha, H., & Giles, C. (2005). Name disambiguation in author citations using a k-way spectral clustering method, *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, 334-343. Retrieved March 18, 2010 from http://doi.acm.org/10.1145/1065385.1065462

Han, H., et al. (2004). *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, Tuscon, AZ, 296-305. Retrieved March 19, 2010 from http://doi.acm.org/10.1145/996350.996419

Hill, A. (2008). "What's in a name?" Prototyping a name authority service for UK repositories, *Proceedings of the 10th International Conference of the International Society of Knowledge Organization*. Montreal, Canada. Retrieved March 22, 2010 from http://names.mimas.ac.uk/documents/Names_ISKO2008_paper.pdf

Library of Congress Authorities. (2009). Retrieved March 21, 2010 from http://authorities.loc.gov/

Paul J. Getty Trust. (2010). About the ULAN. Retrieved March 16, 2010 from http://www.getty.edu/research/conducting_research/vocabularies/ulan/about.html

Pereria, D., et al. (2009). Using Web Information for Author Name Disambiguation, JCDL 2009, June 15-19, Austin, Texas. Retrieved March 18, 2010 from http://doi.acm.org/10.1145/1651587.1651605

Salo, D. (2009). Name Authority Control in Institutional Repositories, *Cataloging &*

*Classification Quarterly*, 47: 3, 249-261. Retrieved February 9, 2010 from http://dx.doi.org/10.1080/01639370902737232

Smalheiser, N. & Torvik, V. (2009). Author name disambiguation. In B. Cronin, *Annual Review of Information Science and Technology*, v. 43, pp. 287-313. Medford, New Jersey: Information Today.

Tang, J., Zhang, D. & Li, J. (2008). A Unified Framework for Name Disambiguation, WWW 2008, April 21-25, Beijing, China. Retrieved February 9, 2010 from http://doi.acm.org/10.1145/1367497.1367728

Van Ryn, P. & Starck W. L., eds. (2005). *NACO Participant's Manual*. 3rd edition . Washington, D.C.: Library of Congress. Retrieved March 22, 2010 from http://www.loc.gov/catdir/pcc/naco/npm3rd.pdf

Veve, M. (2009). Supporting Name Authority Control in XML Metadata: A Practical Approach at the University of Tennessee, *Library Resources & Technical Services*, 53, 1, 41-52. Retrieved March 15, 2010 from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=36261057&site=ehost-live