Future Internet **2010**, *2*, 1-15; doi:10.3390/fi2010001

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Metadata for Name Disambiguation and Collocation

Jeffrey Beall

University of Colorado Denver, 1100 Lawrence St., Denver, CO 80204, USA;
E-Mail: jeffrey.beall@ucdenver.edu

Received: 30 November 2009; in revised form: 30 December 2009 / Accepted: 2 January 2010 / Published: 5 January 2010

Abstract: Searching names of persons, families, and organizations is often difficult in online databases because different persons or organizations frequently share the same name and because a single person's or organization's name may appear in different forms in various online documents. Databases and search engines can use metadata as a tool to solve the problem of name ambiguity and name variation in online databases. This article describes the challenges names pose in information retrieval and some emerging name metadata databases that can help ameliorate the problems. Effective name disambiguation and collocation increase search precision and recall and can improve assessment of scholarly work.

Keywords: names; metadata; name disambiguation; collocation; information retrieval

1. Introduction: Name Disambiguation and Collocation

Databases, search engines, and scholarly communication are increasingly paying attention to the need to accurately disambiguate the same or similar names in online databases. Large databases are beset by the problem of a single name being shared by two or more people, families, or organizations. This problem causes search results to be populated with unwanted documents, forcing the searcher to sort through the results and eliminate the unwanted items. Indeed, "… in this day and age, it can be next to impossible to find all the papers written by a given scientific author" [1]. Numerous efforts are currently underway to ameliorate the problem of a single name being used in a database to refer to more than one person or entity. This article describes the name disambiguation problem, describes different methods of achieving effective name disambiguation and reviews the major initiatives now taking place to achieve effective and consistent name disambiguation in online databases. More specifically, this paper looks at the role name metadata plays in disambiguating names in online databases.

Additionally, the paper looks at a related problem, collocating in search results all the variant forms of a single person's, organization's or family's name. Collocation here means grouping together documents by or about a single name regardless of how the name appears in the documents in question. Information retrieval systems are challenged by the different forms a single person's name can take. For example the name *William* can also appear as *Bill* or as the initial *W*. This paper will describe the many different forms a single name can take, the problems these variant forms cause for search engines, and how database search engines can use metadata to resolve the variant name problem.

When a searcher searches for all the works of a particular author, ideally the search results should contain a complete set of only that author's works. In most current full-text search systems, this ideal is rarely achieved. Indeed, the problems associated with ambiguous names and variant forms of names are worsened when one searches across multiple domains and multiple databases. Algorithmic and deterministic use of name metadata are promising and increasingly-popular tools that can help search engines provide complete and precise results in searches that involve names.

The name disambiguation problem and the variant name problem are examples of the homonym problem and the synonym problem [2] in full-text searching. The homonym problem occurs when the same word refers to two or more concepts, such as "dating," which can refer to the social custom (going out on a date) or the process archaeologists use to determine an artifact's age (radiocarbon dating). The synonym problem occurs when more than one term represents a single concept, such as the pair *leprosy*, *Hansen's disease*, which both mean the same thing. Because most current search engines work by matching terms in a search query with terms in online documents, both the synonym problem and the homonym problem yield search results that contain both unwanted and missed documents. In terms of searching names, search results contain items that include names that match the search query but that are for a different entity that only coincidentally shares the same name, or search results exclude documents that contain variant forms of the name that do not match the form used in the original search query.

1.1. Names

In this article, name means "persons and personas (including pseudonyms), organizations, corporate and government bodies and families" [3]. Personal names are names of people, and although this at first sounds simple and straightforward, we will show that there are many peculiarities about names that make consistent name-related information retrieval an ongoing and complicated challenge. In terms of searching, names are searched in two broad categories of search, first as author or other type of contributor, and second as subject. An author search involves looking for a name associated with a resource's creation or authorship, that is, the person or persons responsible for bringing the resource to light, including a writer, illustrator, speaker, *etc*. A subject search involves seeking information about a person or organization as the subject of a document, that is, a document about a person or organization, such as a biography.

1.2. Metadata

This paper will look at several different types of name metadata. Metadata is structured and standardized data that aids in the discovery and management of information resources. Name metadata is one of the essential and crucial elements of metadata, for it is the element that represents the human aspect of information resources. First, name metadata can occur as an element in a metadata record for a resource, that is, a record that functions as a surrogate of an information resource, such as a MARC bibliographic record, a Dublin Core record, or a proprietary format record created and used by a search engine. Second, name metadata may exist in the form of a unique number that represents the name. This article will describe systems that use unique numbers to represent names. Third, name metadata may exist in a name authority record, a data record that records name and other information about a person, corporate body, or family. MARC name authority records are an example of this type of name metadata. Finally, some name metadata may be created on-the-fly and exist only ephemerally, especially in systems that use algorithmic means of name disambiguation.

1.3. Why undisambiguated names are a problem

Smalheiser and Torvik state, “For any work of literature, a fundamental issue is to identify the individual(s) who wrote it, and conversely, to identify all of the works that belong to a given individual” [4]. This identification becomes ambiguous when more than one person shares a name, and those people are represented in a database. For example, if you do a search in Google for David Leavitt, you will retrieve results that are about David Leavitt the American author, a David Leavitt who is a lawyer, and a David Leavitt who was a nineteenth-century banker. Still, search engines such as Google are unable to effectively group its search results by the individuals; they present the names all together, and the filtering is left to the searcher.

In scholarly databases, such as citation and abstracting and indexing databases, forenames are often represented by initials, a practice that increases the name disambiguation problem. Whereas a search engine can easily perceive the difference between “Morris, Michael,” and “Morris, Michelle,” without human intervention or special programming, most search software will be unable to differentiate between these two names when they are represented only as “Morris, M.” and “Morris, M.” Many style guides prescribe using only initials for forenames. For example the *Publication Manual of the American Psychological Association* says, “Invert all authors’ names; give surnames and initials for only and up to including six authors” [5]. The popular database WorldCat.org buys much of its article metadata from the British Library; the product is called British Library Serials, and virtually all of the name metadata in the product uses only initials for forenames. British Library Serials contains tens of millions of records with this abbreviated data.

One of the problems with abbreviated forenames occurs not only with disambiguating names in search results but in the searching itself. In large databases populated with name metadata that includes only initials, searches on a fuller form of the name may not match on documents or metadata that only contains the abbreviated form. For example, a natural-language search on “White, Edmund” doesn’t match metadata recorded as “White. E.” and metadata and documents that only contain the shortened form of the name may not be retrieved in the search. Some metadata databases, such as ISI Web of

Knowledge, convert all searches using the full form of the name to the shortened form. For example, an author search on “Mooney, Chris” is searched by the system as “Mooney, C*” where the asterisk indicates truncation. This procedure has the advantage of including more of the author’s works in the search results, but at the expense of including works by other authors whose names also begin with “Mooney, C.” in the results. This procedure will not work in full-text databases such as Google, because the system searches full-text and not metadata.

In information retrieval, search precision is the proportion of relevant items retrieved to the total number of items retrieved in the search. In the context of name searches, this means the proportion of documents that are by or about the particular person the searcher had in mind to the total number of documents in the search. In some databases, such as in some high-quality abstracting and indexing databases that search only metadata created by humans, precision can often approach or equal 1.0. In full-text databases that rely on word matching, the value is much lower.

Large digital libraries are making research that incorporates data mining possible. Searches for names run against large textual corpora frequently yield voluminous results with low precision. Any system that can disambiguate names in large textual corpora will significantly facilitate research.

1.4. The value of name disambiguation

The greatest value that effective name disambiguation brings to searching is increased search precision. Greater precision in a name search means that more of the results will be about the individual named in the search and not about others who happen to have the same name. Effective name disambiguation saves the time of searchers; it frees them from having to look at each item in the search results to determine whether the item is indeed about or by the person they entered in the search box.

Smalheiser and Torvik list several advantages of name disambiguation in the context of scholarly communication. They describe being able to find a potential research collaborator to tap unpublished information the researcher may have. That is to say, starting with an article or citation, a user, with effective name disambiguation in place, is better able to find the precise individual named in the citation. Additionally, they point out that “Journal editors could assign papers for review more readily by knowing the characteristic publication profile of its reviewers, and conference organizers would similarly benefit from knowing the publication profile of prospective invitees” [4].

Effective name disambiguation also plays an important role in assessment of scholarly activity, such as that assessment carried out by promotion and tenure committees. Such committees might perform two types of search for a tenure candidate, a search of the candidate’s works, and a search of works that have cited the candidate’s research. The first type of search is normally done in an abstracting and indexing database, and the second is done in a citation index. If the candidate happens to have a common name, accurately gathering this information may be difficult or impossible. Ongoing name disambiguation of an individual also enables following the person’s career over time. Ideally, a single search would retrieve all of an individual’s publications (excluding those for which the individual had no role) over the course of the individual’s career. Granting agencies want to follow up on the work of their grantees (or applicants for grants), and if these people have common names, it may be difficult to track their publications or other work. Effective name disambiguation would greatly increase this task’s precision. Cals and Kotz summarize the value of accurate name retrieval in scholarly communication:

> Accurate assessment of scholarly output is important for individual researchers, their institutes, and funding agencies. Tenure, funding, collaboration and recognition often rely on this link. With accurate identification of an individual researcher’s output, individual citation metrics become more valid and reliable than the much condemned impact factors for journals [6].

Scholarly publishing houses also would benefit from effective name disambiguation. A working system, such as one that assigns a unique number to each scholar, would facilitate their work. Such a system would make manuscript processing quicker, would help find the best reviewers (and help prevent conflicts of interest by revealing past collaborations), would help with royalty payments, and would give publishers’ marking departments a more complete record of the author they are assigned to promote [1].

Other domains in addition to scholarly communication benefit from effective name disambiguation. One area is genealogy, where genealogical research is greatly facilitated when ambiguous names are effectively discriminated. Often, the addition of birth and death dates to names in genealogical databases is sufficient to provide the needed uniqueness to personal names.

1.5. Name variation

A single person’s (or organization’s) name can vary or be rendered differently in printed, electronic, and other information resources. Bennett and Williams point out, “The use of widely variant forms of authors’ names without reference or linkage to alternatives causes hardship for searchers. End-users’ search results may be inaccurate or incomplete, resulting in a decrease in the scientific integrity of the research” [7]. Here we describe the main categories or sources of name variation.

Fullness of name. Names frequently vary in fullness. For example, the name *William R. Harrison* can also appear as *W.R. Harrison*, and it can appear as *Will Harrison*. Personal names are shortened by using initials or by abbreviating a name, such as *Geo*. for *George*. Names of organizations are frequently shortened by converting them to either initialisms or acronyms. For example, the *Southern Archeological Society* is sometimes referred to by its initials, *SAS*. However, this shortened form matches the shortened form of the *Scandinavian Airlines System*. Shortening names increases the chances of name ambiguity. People and organizations are more likely to share a shortened form than a longer form of a name.

Different language or script. The name of a single person or organization can vary by language and by script. For example, the following are all different representations of the name of the author Pearl S. Buck:

Pearl S. Buck	[English]
בוק, פירל	[Hebrew]
بـــك، بـــيرل	[Arabic]
賽珍珠	[Chinese]

Additionally, sometimes names in non-Roman scripts are romanized. A name like 毛澤東 could be rendered in Roman script either as *Mao Zedong* or as *Mao Tse-tung*. Moreover, in languages like Chinese that do not use Roman script, there are often varying scripts. For instance, the name for Mao Zedong in traditional Chinese as given above is 毛澤東, but in modern Chinese it is 毛泽东. Search engines will index these names differently, leading to incomplete retrieval. Finally, among different languages, often a person's first name is translated while the surname is not. For example, in a Spanish-language text, *George Washington* might appear as *Jorge Washington*.

Changed name: People change their names for many different reasons. Some of these include marriage, divorce, and gender change, and emigration to another country.

Pseudonym / Nicknames: Many authors use pseudonyms, and people are often known by nicknames. For example, the author *L. Frank Baum* used several pseudonyms, including *Edith van Dyne*. People commonly known by a title may also have only their title – and not their name – in a document. For example, a document may refer to *The Prime Minister* without actually stating his name. This practice also can occur with names of organizations, as in the phrase, "The Church announces …" Finally, people are sometimes known chiefly by their nickname; an example is Cherilyn Sarkisian, who is best known as *Cher*.

Transcription errors: Name variation also occurs due to typographical and other errors and can led to missed retrieval in search engines and databases. For example, the name *Oscar Wilde* has been occasionally mistakenly written as *Oscar Wild*.

Name not present: It's also possible that the name of a person or organization is not present in a document, even though the person or organization is the author or subject of the document. This situation makes it virtually impossible for a search engine to include the document in search results, for algorithms have not reached that level of sophistication. In this case, about the only way for the name to be indexed in the search engine is manually. Specifically, a human-created metadata record that functions as a surrogate for the original resource can have the name included in it.

1.6. Why name variation is a problem

Earlier we described search precision and how it is negatively affected when multiple people are represented by the same name in a database. A similar measure, recall, is negatively affected when the name of a single person is represented in different ways in the same database. Recall is the proportion of relevant items retrieved in a search to the relevant items that exist in the database. Because full-text searching relies on matching, a search on a name will only match the form a searcher enters in a search box. So when multiple names for a single person exist in a database, and a searcher only looks for one form of the name, recall is lowered, because the system will only match on the form entered in the search box. Low recall is a problem for even rare names. This problem is described by Cals and Kotz, who "searched for [the] scientific work of a professor with a rare name from [their] department; papers under seven different names were retrieved because of varying use of initials. Thus people with both common and rare names are hard to find, not to mention researchers with similar names doing similar research" [6].

Shortened forms of name, especially forenames represented only by initials, increases name ambiguity in databases. The shorter a name is, the more likely it is to match other shortened names. For example, the name *C. Mooney* represents both *Chris Mooney* and *Charles Mooney*. The WorldCat.org database is an example of how using variant forms of a single name can decrease recall. This database buys much of its metadata from several different vendors, and the vendors differ on the fullness of name metadata they include. Some include complete forenames when the information is available; others routinely shorten forenames to a single initial. So name searches in WorldCat retrieve a completely different set of results depending on whether the searcher searches for the full forename or just the initial.

1.7. Algorithmic vs. *manual name disambiguation and collocation*

Name disambiguation in databases occurs two ways: manual and automatic. Manual name disambiguation is done in library cataloging. When a new book is processed, the cataloger searches a database of name metadata and identifies the authors and subjects of the work, and adds name metadata to the bibliographic record that matches the authorized forms found in the authority file. If no record for the author exists in the file, the cataloger either creates a new one or just adds the name to the bibliographic record in the form found in the book. Manual name disambiguation is done by humans.

The other kind of name disambiguation is algorithmic name disambiguation. This is the kind of disambiguation done by computers. Whereas manual name disambiguation is deterministic in its nature, algorithmic disambiguation is probabilistic. Still, neither method is error-free. Disambiguation is made more difficult by the practice -- in the past few decades -- of extensive collaboration on research. In fact, it's not uncommon for an article to have as many as ten or twenty authors, often with the forename of each represented by just an initial.

Much algorithmic name disambiguation relies on metadata created manually. This hybrid approach in the end may become the most successful and widely used. In this approach, computers attempt to either disambiguate names of different people or to aggregate variant names in a single person in a given database [8]. This process is informed by manually-created name metadata. The richness and discriminating ability of the metadata powers the algorithmic processes. All online information -- including names -- does a poor job of representing itself and making itself discoverable. Metadata fills this gap.

1.7.1. Manual name disambiguation and collocation

Smalheiser and Torvik argue against manual name disambiguation. They conclude, "Nevertheless, manual disambiguation is a surprisingly hard and uncertain process, even on a small scale, and is entirely infeasible for common names." [5]. Indeed the vast size of the internet makes manual name disambiguation and collocation there virtually impossible. For example, there may be two documents with the same author listed, but there is not enough information to determine with any certainty that the two articles were written by the same person or by two persons who happen to share the same name.

On the other hand, humans can solve some problems and make judgments better than computers can. For example, when one is unsure whether two occurrences of a name represent the same person or two different people, a quick web search, including perhaps an examination of photographs, can often resolve the question. Humans are better than computers at grouping and analyzing variant forms of a single name. The emergence of manually-created, shared name authority files (such as the Virtual International Authority File, to be described below) shows that manual name disambiguation is necessary and sustainable. If algorithmic name disambiguation and collocation were successful, there would be no need for manual approaches to solving these problems. Initiatives that involve manual approaches, such as the application of unique identifiers and shared databases of name metadata, are beginning to proliferate, an indication of the need for and value of manual name disambiguation.

1.7.2. Algorithmic name disambiguation and collocation

Algorithmic name disambiguation is not a feature that can just be switched on in a database. Both Google Scholar and PubMed lack name disambiguation because, “Such software is expensive and time-consuming to develop, and the algorithms are far from perfect” [1]. The medical research domain is one where most algorithmic attempts to disambiguate names are being developed. Indeed, the best algorithmic name disambiguation occurs within a specific domain or field, because it’s easier to create programs that are limited to a specific set of publications, formats (such as journal articles or web pages), and authors. Algorithmic name disambiguation and collocation, like Internet search engines, will always be imperfect. Automatic disambiguation works like relevancy ranking; it is at best a calculated guess. Beall summarizes:

> “Algorithmic failures to achieve quality name disambiguation parallel similar weaknesses in information retrieval systems that rely on full-text searching and probabilistic relevance ranking. These failures demonstrate that artificial intelligence has not advanced as much as we would like. Name disambiguation, like information retrieval, needs a deterministic approach and human intervention to be successful and precise.” [9]

Another effective obstacle to automated name disambiguation is the fact that the vast majority of authors publish only one or a few articles or other titles in their lifetimes. According to Smiraglia and Taylor, “Research from several studies has demonstrated consistently that the distribution of names follows a power law, such that most names occur only once in a file (*i.e.*, most authors have written or edited, *etc*. only one work)” [10]. The fewer articles, web pages, *etc*. by or about a single author there are, the less data the computer algorithms have to use to discriminate among authors with the same name.

Automatic name disambiguation works by gathering as much information about an individual name and then making a guess as to which other matching names represent the same entity and which do not. One of the things that hinders effective automatic disambiguation is changing data. For example, disambiguation programs use data such as author affiliation to group works by an author. But affiliation and other data used for this purpose, such as an author’s email address, often change, and this causes the disambiguation algorithms to fail. For example, if a researcher writes several articles while affiliated with University A, and then moves to University B and continues to publish research, an algorithm would most likely view them as two distinct researchers with the same name, when in fact, only one exists.

On the other hand, once a disambiguation algorithm has been programmed and set up to work on a database, it can be run multiple times at a very low cost. Automatic name disambiguation and collocation may be more effective in data mining, where textual corpora are extremely large and include a large number of names.

1.7.3. Combining manual and automatic approaches

Finally, disambiguation and collocation can be done by computers and informed by manually created metadata. In the end, this hybrid approach may be the most successful, at least for resources on the open Internet.

1.8. What name metadata ought to include

Name metadata ought to exist in discreet name metadata records that are easily machine readable, sharable, and updateable. In the library domain, name metadata exist as “name authority records.” These records are created cooperatively by librarians and shared openly on the internet. With one exception, the standards for these records prescribe creating a main heading that is unique from other names in the database. This uniqueness is achieved by, for example, adding additional data to the heading, such as birth and death dates, and by qualifying or spelling out names represented by initials. An example of such a heading is:

> Mencken, H. L. (Henry Louis), 1880-1956.

The one exception is the practice of creating “undifferentiated name authority records.” These are single records for more than one person that shares the same name. Sometimes it is impossible to differentiate multiple people with the same name because of a lack of information about the person. In this case the Library of Congress advises, “An undifferentiated personal name is called for … as a last resort after all the possible additions to a new personal or to an existing personal name to break a conflict have been exhausted” [11]. The Library of Congress’ rules do not allow for undifferentiated names for corporate bodies; to make these names distinctive, qualifiers, such as location, are added to the authoritative form of the name. Table 1 lists the most useful metadata attributes that could be included in name metadata records to assist with both manual and automatic name disambiguation and collocation.

Table 1. A listing and description of the elements that might be included in a name metadata record. Some of this description is based on information in the publication entitled *Networking Names* by Karen Smith-Yoshimura [3].

Element	Description / Notes
Preferred or authorized form of the name	Referred to as the *heading*
Other forms of the name, including earlier names, nicknames, pseudonyms, shortened or longer forms of the name, name in other languages or scripts, names associated with the person's office (such as Governor of Colorado)	Referred to as *cross references*
Birth and death dates	If available
Gender	
Life events	Includes things such as place of birth and death, place associated with the entity's output, titles held (including titles of nobility), job, nationality, elected or other offices held, military positions held; events associated with the person
Institutional affiliations	University where degrees were earned, places of employment, including dates
Notes	Notes that help identify the entity, e.g. "Author of the Norton anthology" or "Not the same as George F. Smith, archaeologist"
Family	Spouse, parents, children
Works	List of books, articles, art, *etc*. associated with the person; need not be comprehensive
Subject expertise or genres that the person normally creates in	For example: *Writes on supply side economics* or *Illustrator of children's books*.
Languages the person normally writes or creates in or the person's native language(s)	
Brief biography	
Unique identifier	Such as a record number
Dates of the metadata record	When created and edited
Links to contact information	E.g., author's email address

2. Metadata Systems for Name Disambiguation

This section describes the chief metadata systems and databases for name metadata. Name metadata databases fall into one of two categories. They are either native databases, or they are an aggregation of names from several native databases. It is also possible to link individual name metadata records to their counterpart records in other databases [7]. Organizations such as national libraries, database producers, standards organizations, and publishers are putting more effort into controlling and standardizing names in online databases. Some efforts invite participation from the people whose names are actually in the databases – the authors themselves.

2.1. Library of Congress Authorities http://authorities.loc.gov

This database combines name, subject, and title authority records created by the Library of Congress and other cooperating libraries. The name metadata comes chiefly from libraries in the United States, but libraries from other countries, including the British Library, also contribute. The database combines name, subject, and title authority records, but it is possible to search only the name records. These name records, presented in MARC authorities format, include both personal and organization names. There are over 3.8 million records for personal names and over 900,000 records for corporate bodies [12]. The database is open access, but records can be accessed only one at a time. The Library of Congress does offer the complete file for sale. The file is very limited in terms of who can contribute to it. Much of the data comes from the Library of Congress itself; the cooperating libraries must undergo extensive training and review before they are allowed to contribute.

2.2. Virtual International Authority File (VIAF) http://viaf.org/

According to the project’s web page, “VIAF is a joint project of several national libraries, implemented and hosted by OCLC Online Computer Library Center, Inc. (http://www.oclc.org/). The project's goal is to lower the cost and increase the utility of library authority files by matching and linking the authority files of national libraries, and then making that information available on the Web.” The database aggregates name authority records from over 15 participating organizations. Because name headings differ in different databases, the VIAF aggregates the different forms for each heading. So for a single person, there may be three or four name headings listed together. Indeed, voluminous and classical authors may have a dozen or more forms of their names in the database. Country flag icons follow each heading to indicate the source of the heading. Several countries frequently share a single form of the name. The database has numerous add-ons, including cover art that correspond to books published by authors listed in the database, publication statistics, and it makes the name authority records available in both MARC and UNIMARC formats. A “history” tab records all changes to each individual record.

One issue with aggregating name metadata databases such as this one is the accurate grouping of name headings that correspond to the same individual. In this case the work is done algorithmically, and the algorithms are rather conservative. That is, the database will only group different headings together when it is sure that the two forms indeed represent the same person, using co-author information and birth and death date information [13]. VIAF also uses bibliographic information to group two or more forms of a name into a single author. This grouping is done when two sources cite an author for the same title.

2.3. WorldCat Identities http://orlabs.oclc.org/Identities/

This OCLC project is still in beta phase and is an experimental database of names. Designed for a popular audience, the project includes a page for every name in the WorldCat database – over 30 million names. Similar to VIAF, the pages include add-ons or additional features generated from the name and accompanying bibliographic metadata. For example, it lists works both by and about each name. WorldCat Identities also provides Library of Congress name authority records in MARC format for each name, when that data is available. Another feature of WorldCat Identities is a tag cloud created from the subjects associated with the bibliographic data. This feature can be used to differentiate people with the same name who write in different fields. Finally, WorldCat Identities often provides links to Wikipedia articles about the individual authors represented on its pages, whenever such articles are available.

2.4. ResearcherID http://www.researcherid.com/

According to its website, "ResearcherID … is a global, multi-disciplinary scholarly research community. By assigning a unique identifier to each author who participates, ResearcherID provides an invaluable index to accurate author identification and increases recognition of work and collaboration among researchers." ResearcherID was created and is sponsored by Thomson Reuters, the company that produces the ISI Web of Knowledge citation index. ResearcherID has two unique features. First, the name metadata is generated mostly by the authors themselves. The product requires one to establish a user account and to populate the user page with citations to one's own work, including books and journal articles. The researcher pages are all open access. A search page leads searchers to individual pages, but it also provides an institution search, allowing searchers to access citations for all who have signed up for the service from, for example, a single university.

2.5. ContributorID

ContributorID is a planned author identification system that will be made available by the CrossRef publisher cooperative. The project has been promised for some time, and it has been the subject of much speculation and discussion, but as of this writing, it has not appeared.

2.6. International Standard Name Identifier (ISNI) http://www.isni.org/

ISNI is a draft standard from the International Standards Organization. Its purpose is to assign a unique number to personal names that appear as authors/contributors or subjects in print and online publications. The standard will be similar to the ISBN numbers that appear in books, with one exception. One purpose of the ISBN is to uniquely identify various editions of works. For example, a book's first edition gets a different ISBN number than its second edition. One work, then, can have many different ISBN numbers. One author, ideally, will only have one ISNI. Thus the ISBN serves to differentiate among various manifestations of a work and among other works, but the ISNI serves to bring together all instances of authorship with a single number.

According to the project's web site, "An ISNI is made up of 16 decimal digits, the last one being a check character" [14]. Example: ISNI 1422 4586 3573 0476

2.7. Digital Author Identification System (DAI)

http://www.rug.nl/bibliotheek/informatie/digitaleBibliotheek/daikort?lang=en

This system is an example of a national system of author name identification. It is also called DAI or the Dutch Author Identification System. The numbers are automatically assigned to professors and researchers at Dutch research institutes and universities. The numbers follow the pattern and are compatible with the ISNI standard.

2.8. Others

Numerous other systems are emerging; some of these show the potential for success, others do not. Another thing that remains to be determined is whether competition among different systems will help or hinder the emergence of metadata for name disambiguation and collocation. This will be determined by several factors, including the degree to which competing schemes collaborate and share data with each other, whether only a handful of initiatives emerge as dominant and make the smaller ones obsolete, and what model becomes the most popular and interoperable with authors, publishers, and search engines.

Among the others, one that deserves a special mention is Wikipedia. Though not a database of names per se, it does have many articles about persons and corporate bodies, including many for names that are not unique. Wikipedia pays special attention to name disambiguation and has special disambiguation pages that list subjects of articles with the same name, differentiated by what they are known for, such as Roger Morris (American writer) and Roger Morris (Engineer). Table 2 lists some additional name metadata databases.

Table 2. Other systems and databases that help with name disambiguation.

Database Name	URL	Comments
academia.edu	http://www.academia.edu	Requires signup.
arXiv's author identifier system	http://arxiv.org/help/author_identifiers	
FRIDA (Norwegian National Research Database	http://frida.usit.uio.no/	
GEPRIS (the German Research Society's research information system)	http://gepris.dfg.de/gepris	Limited to academics and researchers
International Registry for Authors: Links to Identify Scientists	http://www.iralis.org/	Also Called IRALIS. Based in Spain.
Names Project	http://names.mimas.ac.uk/	Still in development.
People Australia	http://www.nla.gov.au/initiatives/peopleaustralia/index .html	Still in development.
RePEc Author Service	http://authors.repec.org/about	

Table 2. *Cont.*

Database Name	URL	Comments
Researcher Name Resolver (Japan)	http://rns.nii.ac.jp/resolver/search.go?AD=init	In Japanese
Scholar Universe	http://www.scholaruniverse.com	Has a free name search and a proprietary keyword search.
Scopus	http://help.scopus.com/robo/projects/schelp/h_autsrch_intro.htm	Proprietary
Universal Author Identifier System	https://clotho.iml.uom.gr:8443/uai_sys/aboutuai.xhtml	

3. Name metadata and the Semantic Web

Name metadata records are well positioned to become an integral and valuable part of the emerging Semantic Web. Systems such as the Simple Knowledge Organization System (SKOS) are designed to represent standardized or controlled datasets, such as names in a name database. Each name in SKOS would have a uniform resource identifier (URI) that could be used to represent the name in many different, linked systems. SKOS offers a great potential to aid in effective name disambiguation and collocation of variant names.

4. Conclusion

Information retrieval involving the searching of names in databases, especially personal names, is aggravated by the fact that multiple persons or organizations can share the same name, and a single person’s or organization’s name can appear in many different ways in multiple databases. Although both manual and automatic name disambiguation and collocation can resolve these problems, manual name disambiguation does not scale to the size of the Internet, and algorithmic name disambiguation is difficult and expensive to set up and as a probabilistic system only gives a best guess and is subject to error.

Name metadata databases are emerging that offer solutions that work with both manual and algorithmic name disambiguation. Also, emerging standards, such as ISNI, the International Standard Name Identifier, ResearcherID, and databases, such as library name authority files, and the new Virtual International Authority File, are all valuable name metadata resources that will help increase precision and recall in name searching in databases and on the Internet.

Name metadata databases will be incorporated into the Semantic Web, helping to improve the searchability of names on the Internet. Uniform resource identifiers will point to discrete name metadata records and will increase search precision and recall on the Internet and in online databases. Metadata for name disambiguation and name collocation will increasingly improve information retrieval by and about people and organizations on the World Wide Web.

References

1. Enserink, M. Are you ready to become a number? *Science* **2009**, *323,* 1662-1664.
2. Beall, J. The weaknesses of full-text searching. *J. Acad. Libr.* **2008**, *34*, 438-444.
3. Smith-Yoshimura, K. *Networking Names*; OCLC Programs and Research: Dublin, OH, USA, 2009.
4. Smalheiser, N.R.; Torvik, V.I. Author name disambiguation. *Annu. Rev. Inform. Sci.* **2009**, *43*, 287-313.
5. American Psychological Association. *Publication Manual of the American Psychological Association*, 5th Ed.; American Psychological Association: Washington, DC, USA, 2001.
6. Cals, J.W.L.; Kotz, D. Researcher identification: the right needle in the haystack. *Lancet* **2008**, *371*, 2152-2153.
7. Bennett, D.B.; Williams, P. Name authority challenges for indexing and abstracting databases. *Evid. Based Libr. Inform. Pract.* **2006**, *1*, 37-57.
8. Vu, Q.M.; Takasu, A.; Adachi, J. Improving the performance of personal name disambiguation using web directories. *Inform. Process. Manag.* **2008**, *44*, 1546-1561.
9. Beall, J. Cataloguing names the old-fashioned way. *Science* **2009**, *324*, 1514-1515.
10. Smiraglia, R.P.; Taylor, A.G. Letters to the Editor. *Cat. Class. Q.* **2009**, *47*, 760-763.
11. Library of Congress. *Frequently Asked Questions on creating Personal Name Authority Records (NARs) for NACO*; http://www.loc.gov/catdir/pcc/naco/personnamefaq.html#21 (Accessed November 8, 2009).
12. Library of Congress. Prints and Photographs Division. *Authority Files for Cataloging Pictures: Common Choices*. http://www.loc.gov/rr/print/resource/228_authfile.html (Accessed November 14, 2009).
13. Hickey, T. Expanding the concept of universal bibliographic control. *NextSpace OCLC Nwsltr*. 2009. http://www.oclc.org/us/en/nextspace/013/research.htm (Accessed November 14, 2009).
14. International Standard Name Identifier. http://www.isni.org/ (Accessed November 22, 2009).