

Available from: http://doi.acm.org/10.1145/1952222.1952274

Copyright © ACM, 2010. The definitive version was published in Proceedings of OZCHI (2010).

This is the author’s version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. If your library has a subscription to these conference proceedings, you may also be able to access the published version via the library catalogue.
What’s My Name Again? Sociotechnical Considerations for Author Name Management in Research Databases

Dana McKay  
Swinburne University of Technology  
Library, Institute for Social Research  
John Street, Hawthorn VIC 3122 Australia  
dmckay@swin.edu.au

Silvia Sanchez  
The Hiser Group  
Level 18/35 Bourke Street, Melbourne, VIC 3000, Australia  
silvias@hiser.com

Rebecca Parker  
Swinburne University of Technology  
Library, Institute for Social Research  
John Street, Hawthorn VIC 3122 Australia  
rparker@swin.edu.au

ABSTRACT
Managing names in bibliographic databases so that they have a one-to-one match with individual authors is a longstanding and complex problem. Various solutions have been proposed, from labour-intensive but accurate manual matching, to machine-learning approaches to automated matching which require little input from people, but are not perfectly accurate. Researchers have a particular interest in name management: they are often authors, and receive academic credit based on their work and need correct citation records. However they are also searchers and have an interest in finding all the works by other authors. There has been little work on the tensions between these two needs, nor on how researchers manage their own identities with their choices of name. This paper reports on a study of researchers that investigates both their relationships with their own names, and what they would like from research databases when they are searching for specific authors.

Author Keywords  
Author names, researchers, research databases, sociotechnical aspects of HCI, information seeking, search interfaces.

ACM Classification Keywords  
H.5.m Information interfaces and presentation: miscellaneous

INTRODUCTION
Author name management in bibliographic databases has been a known problem for nearly as long as library catalogues have existed: in 1647 Thomas Hyde grew frustrated with the numerous spellings of Shakespeare’s name and filed them all under a single entry in the catalogue of Oxford’s Bodleian Library (Weinberger 2007). The problems with names became evident in HCI as early as 1989, where a case study demonstrated that searching for names has its own specific problems (Wright et al. 1989). The specific problems relating to author names were first addressed in the digital library field in 1999 (Lawrence et al. 1999) and since then a range of approaches to managing names have been investigated (see for example Snyman et al. 2000; Feitelson 2004; Hong et al. 2004; Laender et al. 2008; Pereira et al. 2009).

Researchers have a dual relationship with the problems with names: they search for names in research databases, but they are also authors in those databases with an interest in ensuring their citations are correct. There is clear evidence to demonstrate that researchers use author metadata when information seeking in research databases (see for example Bishop 1998)). Similarly, it has been evident since 2000 that researchers are interested in ensuring that their citations in research databases are correct (Cruz et al. 2000). This interest in correct citations has only increased with the trend toward research assessment using citation measures (Steele et al. 2006).

Despite researchers’ clear interest in both searching by name and managing their citations, there has been little research into their interaction with names in research databases. Although it is evident from, for example, the institutional repository literature that researchers are concerned with their reputations and like to control their own information (Gadd et al. 2003; Miller 2006), there has been no research into how authors feel about name management as it applies to their own names (or indeed, whether they attempt to manage their names themselves at the time of publication). Similarly, despite a large body of research on name disambiguation (where computer systems or librarians attempt to restore a one-to-one mapping between names and authors), there has been little work on whether it is useful, and if so, how to effectively display the results of automated name disambiguation to information seekers.

In this paper, we present the results of a study that examined researchers’ attitudes to name management (both with respect to their own names and their forays into author searching) and their interface requirements for a digital library presenting disambiguated names.

OZCHI 2010, November 22-26, 2010, Brisbane, Australia.  
Copyright the author(s) and CHISIG  
Additional copies are available at the ACM Digital Library  
(http://portal.acm.org/dl.cfm) or ordered from the CHISIG secretary  
(secretary@chisig.org)  
BACKGROUND ON PROBLEMS WITH NAMES

In this section we will cover the background of our work in three main areas: how academics search for author names, how names lose their one-to-one match with identities in research databases, and existing approaches to name management.

Searching for Names

It is evident from the literature that both lay searchers and researchers find author names useful when using search engines and when retrieving known information (Wallace 1993; Bishop 1998; Jones et al. 1998). It is also evident that in some research collections at least, ambiguity in researcher names means that searchers do not find what they are looking for, even when it is available (Jones, Cunningham et al. 1998). The end point of name management must be research databases that allow searchers to find a complete publication list for any author when they search for that author. Thus far there is no research to demonstrate how to display researcher names in such a way that searchers can find all the work by a given author and interpret the results to understand why results associated with variant names have appeared.

How Names in Databases are Problematic

There are two major problems with the way names appear in research databases: mixed citation and split citation.

Mixed citation occurs when the same name refers to more than one individual in a research database, which, while not exclusively affecting these groups, occurs particularly frequently as a result of transliterating names from Chinese, Japanese and Korean scripts (Qiu 2008).

Split citation occurs when a single author’s publications appear under more than one name, and may occur for a range of reasons (Feitelson 2004) including name changes due to personal circumstances (Hamp-Lyons 1997); variation in transliteration of non-roman names (Sweetland 1989; Qiu 2008); typographical errors (Sweetland 1989); journal requirements for formatting of authors’ names on publications (Puniamoorthy et al. 2008); authors’ deliberate decisions to use name variants (for example to obscure their gender to avoid prejudice in the peer review process (Tregenza 2002)); and simple inconsistent name usage by authors (Jordan 1997).

Approaches to Name Management

Early approaches to name management in databases centered on authority control, where each individual is assigned a ‘master name’ in the database or catalogue, and all other names or name variants are filed under that name (Auld 1982). With the advent of digital libraries, the standard approach to managing names moved away from authority control and toward access control, which treats all an author’s names and name variants as equally valid, and uses a key of some kind (usually a number) to represent an author’s identity (Cruz, Klink et al. 2000).

Considerable research effort has been expended on how best to implement access control in online bibliographic databases. Early work suggested author-mediated access control would be most effective in research databases, since authors have a vested interest in maintaining their own citations (Cruz et al. 2000). While author engagement is probably the least error-prone approach to name management it is well known that author engagement with research databases (particularly institutional repositories) can be hard to generate (Allen 2005; Fried Foster et al. 2005), thus there are considerable social barriers to this approach.

Later work has focused heavily on automated approaches, including SQL-based matching (Hong, On et al. 2004); matching using web-based information to supplement the information held in research databases (Pereira et al. 2009); and (most commonly) machine-learning algorithms (for example Adams et al. 2002; Feitelson 2004; Lee et al. 2005; Laender et al. 2008). While these approaches may require little from researchers, it has not been determined whether the results are more acceptable to researchers (as searchers and as authors) than databases where no name management at all has taken place.

METHODOLOGY

We used a two-stage investigation to understand researchers’ problems with names. The first stage was a series of focus groups designed to elicit researchers’ approaches to: managing their own names; whether or not their search strategies typically included author names; and if they did search on author name, whether they found name ambiguity a problem. We used a grounded theory analysis to understand the results of the focus groups. The second phase of this research used low-fidelity prototyping to determine what researchers would expect from a search interface that presented disambiguated author names.

UNDERSTANDING RESEARCHERS’ PROBLEMS WITH NAMES

We ran a series of five focus groups with a total of 19 researchers recruited from research staff at a small but research-active Australian university. To ensure we understood the issues around author names in a wide variety of research cultures the focus groups were discipline specific and spread over a variety of disciplines. The disciplines covered were: science and medicine, engineering and computing, humanities, design, and business.

In each focus group we questioned researchers about a range of topics including their typical search strategies when looking for information; how they chose a name under which to publish (and what circumstances affected this choice); how they assessed whether a name referred to an author of interest; and how well they felt large bibliographic databases managed names.

During the focus groups, we walked researchers through specific scenarios supported by screen shots of research databases. This allowed us to elicit feedback from researchers about their current user experience with names in research databases.

The data we collected in these focus groups was analyzed using grounded theory analysis (Glaser et al. 1967). Three major themes emerged in the results of the focus groups:
researchers’ relationships with their own names; researchers’ experiences searching for author names; and how name variants could be managed in research databases.

**Researchers’ Relationships with Their Own Names**

Prior to this work, little was known about how academics chose their publication names, including whether or not they tried to exert any control over their names, and what their considerations were when they selected a name for any given publication. This section examines academics’ approaches to choosing a publication name; the ways of and reasons for variation in publication names; and the concept of name-as-brand as it relates to individual authors.

**Choosing Publication Names**

Very few researchers (two out of nineteen focus group participants) deliberately selected the name they would use as their publication name at the beginning of their careers:

‘My author name is [First initial] [Middle initial] [Surname] so it doesn’t tell my gender’

‘I made a decision to always use [name]. Many journals don’t ask for a middle initial, so I put it in the first name field.’

There was considerable regret among the majority of researchers who had not deliberately chosen a name variant, both because it meant they had a number of name variants (and hence split citation):

‘I would like to use [name], but I have all these existing articles with a variety of names’

and because they had not investigated the field to find out if any other researchers shared a name with them prior to selecting their own publication names:

‘There’s this other author with the same name as me at Oxford who’s very prolific and he always makes me look terrible.’

**Variation in Publication Names**

Researchers gave a number of reasons for variation in their publication names, but the reasons fell into three main categories: accidental variation in publication names; variation enforced by publishers; and deliberate variation chosen by individual researchers.

Accidental variation occurred largely where researchers had not made a conscious decision which publication name to use, either at the beginning of their careers or any later stage:

‘I wish I had locked myself into something specific, but you don’t know that when you start out.’

Under these circumstances, authors are not averse to assistance in managing name variation at the time of publication, as one researcher commented:

‘It’d be handy to know how you’ve published in the past.’

Enforced variation occurs where a journal publisher insists on a specific name format (for example initials followed by surname) even though this is not a researcher’s first preference or customary publication name:

‘You have no choice; it depends on what the journal wants.’

In some fields of research, this means that researchers themselves cannot choose a name variant and use it consistently.

The final reason for name variation is deliberate variation on the part of researchers to maintain specific identities:

‘I have two public personas’

‘You may want to distance yourself from earlier research.’

This deliberate name variance has considerable implications for any automated name management approach; it suggests that researchers may actually want their citations split in some cases. Given this want, it is important to balance the needs and desires of researchers as authors with the value information seekers may find in the availability of more complete citation records in research databases.

**Researchers’ Names are like Brand Names**

It was clearly evident in the focus groups that researchers who shared a name with other researchers were well aware of it, particularly if the other researchers were working in the same research field:

‘My name is a problem with research databases. I get credited with all sorts of works that aren’t mine.’

In addition to this awareness, a number of researchers mentioned using a name variant on their publications that was not their preferred or everyday name:

‘It’s not the name you like, it’s what people see you as.’

This concern about other researchers and the considered selection of names demonstrate that researchers see their publication names as brand names. Where researchers are using only a single brand name (i.e. there is no deliberate use of name variants), they may welcome the ability to consolidate their work in research databases.

**Researchers’ Experiences Searching for Names**

We questioned researchers about their experiences searching research databases using author name metadata. We discovered that they do search by author name, but that they recognize a number of difficulties with it. In the absence of any solution to the problem of names, many researchers have either developed specific search strategies for dealing with author information, or developed their own heuristics for disambiguating author names within search results.
Searching Using Author Name Metadata

Our work confirms what the literature tells us (Bishop 1998): researchers are interested in works according to author, and search for these works using author names as keywords:

‘I make my students aware of the top authors in their research fields’

‘I do search on author.’

It is fairly established practice for researchers to look for published work by well-known authors in specific industries or fields.

Name Searching Isn’t Straightforward

Despite author names being useful search metadata, researchers had problems with author searching and complained that it was often difficult to get the results they were interested in or hoped for:

‘Author searching is messy; there are so many misspellings.’

In addition to the more general problems with author search, researchers also demonstrated a clear awareness of the potential for both split citation:

‘Is John Smith the same as John K. Smith? My name appears [two different ways] like this and I find it really annoying.’

and mixed citation:

‘You can’t assume everyone is the same person just because they have the same name.’

Researchers’ Workarounds for Name Searching

Researchers employed two different strategies for dealing with the problems posed by name searching: searching outside bibliographic databases for the author as an individual, and disambiguating names themselves while viewing search results.

Two researchers mentioned searching for authors as individuals; in both cases they mentioned publication listings on individual homepages:

‘I would usually search the university website for a staff homepage.’

All the other researchers in our focus groups mentioned using heuristics to determine whether a given citation in a research database was in fact authored by the person they were searching for. These heuristics relied on supporting information such as research field, affiliation, and co-author. For every strategy mentioned, however, there were both supporters and detractors. This was noticeable when researchers talked about, for example, affiliation:

‘If the affiliation’s the same, it’s probably the same person’

‘Affiliation’s not that useful, people move around so much’

and co-authorship:

‘In my [research] discipline it’s useful because there are only two or three authors per paper’

‘Co-author doesn’t tell you that much; often they’re students and they change each year’

and even research discipline or subject area:

‘I would look at the subject area [of the publication] first’

‘It’s tricky to use subject area when you work in multidisciplinary areas.’

The only metadata that researchers generally held the same opinion of for use in name disambiguation was the metadata traditionally used in authority control, such as middle name and year of birth. They did not find it at all useful:

‘Another academic with the same name as me was born the year before me, so it would look like a mistake if our birth years were listed’

‘I would only rarely know the middle names of ... the top authors in my field’.

What Researchers Want from Name Management

Researchers were universally interested in name management, both as it is applied to their own names and as it would help them when they were searching for author names.

Management of Researchers’ Own Names

When discussing the potential for an automated name disambiguation system that associated all researchers’ publications in a single database with their identity, researchers could see value in this proposal:

‘People could Google you and everything would be in one place’

‘I’d like to have everything listed under one name.’

Despite seeing the advantages in automated name disambiguation, it was very important to researchers that any automated approach be accurate:

‘You’d want to be very sure they were the same person.’

In addition to the concerns about accuracy, some researchers expressed concerns about automated systems aggregating work they simply didn’t want aggregated:

‘It’s up to [the individual researchers] whether they want their publications brought together or not.’

These concerns about accuracy and control run counter to the approach of many suggested solutions to the mixed and split citation problems in research databases, and demonstrate that names have a social value to researchers that should not be overridden by a technical solution to the problem.
Searching Among Managed Names

Researchers were excited about the potential for name management to make author searching easier. The idea that finding all of any given author’s publications in a research database could be simple was very appealing to them:

‘If you’re searching for your favourite author, you want to see everything they’ve written.’

They also stressed that the interaction with author names should be simple:

‘When you click on an author’s name, you want to get all their publications back—you don’t want to have to click twice.’

Even when searching, though, authors had some misgivings about the accuracy of any automated or non-transparent name management approach:

‘You’ll never be 100% sure it is the same person.’

There is a real tension here between simplicity and correctness; as we will demonstrate in the next section what researchers require to verify the accuracy of automated name disambiguation adds to the complexity of user interfaces.

PRESENTING NAMES IN RESEARCH DATABASES

We used low-fidelity prototyping to determine how researchers would like to see author name variants presented in search results interfaces. We conducted four sessions with two participants each. Participants were recruited from the same university as the focus group participants; however to avoid having primed researchers to think about their own names no researcher participated in both the focus groups and the low-fidelity prototyping.

In these sessions we walked the participants through activity scenarios and presented them with paper wireframes of potential interface designs that might be used to display name variants in search results.

In each session, researchers were asked to comment on how they would interpret and interact with a search interface at each step of an author-searching scenario, including deciding how to search; seeing results with ambiguous author names; and seeing results specifically designed to display author name variants in a way that allowed researchers to identify individual authors. Researchers were invited to highlight any strengths and weaknesses of the interface designs presented to them, and to suggest ways to improve the low-fidelity prototypes in order to better present name variants.

The interface elements researchers highlighted as important when displaying name variants in research databases were: specific author searches; the display of search results; and the ability to refine search results.

Specific Search by Author

Researchers explained that the ability to search specifically by author is a useful tool that research databases ought to provide. This functionality seemed particularly useful when researchers were looking for published work by a specific author, as opposed to searching for work on a particular topic.

Researchers explained that limiting the search results specifically to author names would improve the author searching experience. Limiting the number of results that appeared when looking for ‘Robinson’ to ‘just authors named “Robinson”’ was perceived by researchers as a better outcome than obtaining search results that included ‘Robinson’ in the author field or in the title.

Displaying Search Results with Name Variants

Several factors have an impact on researchers’ acceptance of name management when displaying search results: the relevance of search results; the information displayed in the search results; and the highlighting of information relevant to name disambiguation.

Relevance of Search Results

The main concerns researchers had with presenting name variants in research databases were with search results. It was important to them that the search results accurately demonstrated the relationship between their search terms and the name variants returned:

‘If I don’t understand why the system has returned these results, I’m going to think the system is broken and crap.’

It was important to researchers that they understood why they retrieved the results they did, but for significant name changes they admitted this might be difficult:

‘I can’t possibly guess that Simone Miller [a name from the scenario] is the same person as Simone Robinson, but I can guess that S. Robinson is.’

Information Displayed with Search Results

We discussed how researcher name variants should be presented with reference to a low-fidelity prototype that displayed search results that included name variants (see Figure 1). Researchers felt that it would be useful for research databases to present them with the type of information they used in their own heuristics for guessing whether two names referred to the same person. Participants in the low-fidelity prototyping sessions gave us useful feedback about the information categories that needed to be displayed across search results to assist them in understanding disambiguation of author names, for both mixed and split citations. These information categories were: author name; title of the published work; year of publication; type of publication; and subject or research area:

‘Research area is useful to reinforce you haven’t gone down the wrong track.’

Finally, they felt that the system should avoid information overload and present results as concisely as possible:

‘It’s counterproductive to see all these variants. I’d rather just see merged results’
and that where the research database did describe the relationship between names, the terminology should be simple and everyday:

“What does “also authored as” mean? I think it should be “also known as”.”

Highlighting Information about Name Variants

One of the scenarios we used with researchers presented them with author search results that included works published by an author under her maiden and married names (Simone Miller née Robinson).

Researchers explained that if they were searching for work by Simone Miller née Robinson they would want to see all the works by that author, as long as they could tell that the two names did refer to the same author:

‘Show it exactly as it is on the publication and … use a link to pull it all back together.’

The relationship between different versions of an author’s name has to be clearly visible as part of the interface design, so that researchers can be certain that they refer to a single author.

Refining Search Results is Useful

Having the facility to further refine search results in a research database seems helpful to researchers when assessing name variants. While information categories like publication year or subject area might not be directly related to author names, researchers explained that having the capacity to refine search results by some of these categories affected their ability to locate relevant results by a single author.

During our low-fidelity prototyping sessions, researchers explained that faceted search allowed them to refine authors’ names gradually and would be easily understood by most users, regardless of their computer skills (see Figure 1). This finding is well supported by other work on faceted searching, for example (Ka-Ping et al. 2003; McKay et al. 2004; Suominen et al. 2007; Koren et al. 2008).

Researchers also noted that the author facet needed to display all the variants of each author name and to clearly identify the distinctions between authors who shared a name (see Figure 1).

DISCUSSION

Researchers hold views on name management from two perspectives: as it affects them as authors, and as it affects them as searchers.

Name Management for Researchers as Authors

It is evident from our work that researchers have a considerable investment in protecting their research brand—their names. Researchers’ names represent their research records, and as such hold considerable social weight with researchers. In light of researchers’ concerns for their own names, it is perhaps unsurprising that the potential for automated name management is both exciting and perturbing for researchers. While a simple way to manage their name variants would be greatly appreciated, researchers are concerned about the quality of such disambiguation.

For name disambiguation to be truly acceptable to researchers, it must be perfectly accurate, and also subject to their control where they would prefer not to collate their work under a single name for any reason. This runs directly counter to the majority of automated systems, which are neither accurate nor open to the control of individual authors.

In the case of research databases that depend on researcher support (for example institutional repositories Callan 2004; Fried Foster 2005), database managers would be well advised to exercise caution in implementing any automated system for disambiguating names without the support of researchers whose work is included in the databases.
It seems, then, that the only alternative is to use a researcher-mediated name management approach. However, it is evident from the literature that researchers are not particularly interested in administering collections of their work (at least not in institutional repositories (Callan 2004; Mackie 2004; Fried Foster 2005; Kim 2006; Davis et al. 2007)). This means that any researcher-mediated name management would have to be simple to manage and provide considerable value to researchers to be viable.

While it seems that neither automated nor author-mediated name disambiguation would meet the approval of researchers where they are concerned with their own names, it may be that in reality they would find enough value in both having their own publications collated, and making searching for others’ publications easier, that their concerns would pale into insignificance. Whether researchers would find name management in either form acceptable for their own names is an area for future work.

Name Management for Researchers as Searchers

Researchers are well aware of the problems inherent in author name searching, including the impact of mixed and split citations on search results (and, for that matter, their citation counts). They are also acutely conscious that author search mechanisms are frequently inaccurate and cumbersome. Despite the problems with author searching, author information is sufficiently useful to researchers that not only do they persist in searching by author, but they have developed heuristics to help them determine the accuracy and relevance of author search results. These strategies are often based on implicit, social knowledge such as co-authorship patterns, affiliation and research discipline.

In light of the problems they encounter when author searching, researchers are clearly more convinced of the benefits of name management from their perspective as searchers than as authors with publication histories. Accuracy is still important, but less so than when it comes to their own names.

Although researchers are less concerned how others’ names are managed than they are about their own, any interface that presents researchers with combined name variants needs to reassure them that the disambiguation is (at least mostly) correct. We have seen that researchers use their own heuristics to manually disambiguating names, and it is clear that they would like research databases to provide the information they use in these heuristics (such as affiliation and research discipline) even when those databases are already disambiguating. Where research databases present disambiguated names, it is important to researchers that they have some supporting evidence that individual people have been correctly identified, and that in the case of split citation that they can understand what relationship the differing names bear to one another (particularly in cases such as marriage, where the names may be markedly different).

Even though researchers would like research database interfaces to show clearly how names have been managed in their search results, it is clear from our study that they are excited by the prospect of a research database that could produce comprehensive author search results. As such, it is potentially by emphasizing the value of name management during searching that research database managers can overcome the hurdles associated with automated disambiguation of researchers’ own names.

CONCLUSIONS AND FUTURE WORK

In this paper we examined researchers’ experience with names, including how they manage their own names; how they search for names; what problems they face in doing so; and how they feel about name disambiguation.

What Researchers Want from Name Management

We found that through their experiences while searching, researchers are well aware of the problems with author names as they appear in research databases. Despite this, many authors had not made a considered decision about their publication names at the beginning of their careers, and even later in their careers many had not settled on a single variant (though in some cases this was deliberate).

Because researchers are aware of the issues with names, they find the concept of name management both appealing and disturbing. They like the idea of having their publications sensibly collated in a research database, and the possibility of searching for other researchers’ publications more easily is also enormously attractive. Nonetheless, the idea that automated disambiguation might be inaccurate or erode their control over their own publication record is worrying for researchers. This is significant, given that the majority of approaches to name disambiguation described in the literature are automated and somewhat inaccurate. In addition to being concerned about accuracy, researchers are very used to manually disambiguating names as they search, so any digital library that performs name disambiguation on their behalf will need to provide them with the cues they are already accustomed to using to make it clear to them how individuals have been identified. It is evident from our work that any approach to name management must take into account the social importance of name-as-brand, and also the practices of researchers while name searching, or that approach is destined to fail.

Avenues for Future Investigation

Where the best possible medium lies between researchers’ fear of name management and their interest in its benefits is a subject for further investigation. In the case of institutional repositories at least this is a question worth answering, as name management has uses for authors and searches alike but repositories rely on the support of authors to function. Answering this question would provide considerable direction to the still-ongoing search for the best approach to name management in research databases, as most research databases are dependent on the good will of authors.

REFERENCES


Puniamoorthy, N., Jeevanandam, J. and Narayanan Kutty, S. Give south Indian authors their true names. Nature 452, 7187 (2008) 530-.


Wallace, P. M. How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. RQ 33, 2 (1993) 239-53.
