International Journal on Digital Libraries manuscript No.
(will be inserted by the editor)

Interactive Context-aware User-driven Metadata Correction in Digital Libraries

David Bainbridge, Michael B. Twidale, and David M. Nichols

This is the authors' version, the final version appears as:

Bainbridge, D., Twidale, M.B. and Nichols, D.M. (2012) Interactive Context-aware User-driven Metadata Correction in Digital Libraries, *International Journal on Digital Libraries*, 13(1) 17–32.

http://dx.doi.org/10.1007/s00799-012-0100-5

The final publication is available at www.springerlink.com

Abstract Personal name variants are a common problem in digital libraries, reducing the precision of searches and complicating browsing-based interaction. The book-centric approach of name authority control has not scaled to match the growth and diversity of digital repositories. In this paper we present a novel system for user-driven integration of name variants when interacting with web-based information—in particular digital library—systems. We approach these issues via a client-side JavaScript browser extension that can reorganize web content and also integrate remote data sources. Designed to be agnostic towards the web sites it is applied to, we illustrate the developed proof-of-concept system through worked examples using three different digital libraries. We discuss the extensibility of the approach in the context of other user-driven information systems and the growth of the Semantic Web.

1 Introduction

All too often in our professional roles as knowledge workers we encounter online information that is incorrect. The memorable incidents tend to be the more outrageously incorrect, but there are other forms of error that while more minor are far more frequent, and hinder our ability to perform our information work.

As the real world examples presented later in this paper attest there is a real weakness in our information systems—in particular our digital libraries—when it comes to the names of people. Existing approaches to identity management include authority control, algorithmic disambiguation and author-driven maintenance. In this paper we describe a system that is orthogonal (yet complementary) to these approaches by allowing users to adapt and enhance the representation of authors across a broad range of the systems they use for information searching.

Inspiration for the work came from an experience of one of the authors a few years ago visiting the local public library with his daughter. In attempting to lookup what books in the *Asterix the Gaul* series by René Goscinny and Albert Uderzo the library had, he became concerned that his author search was not returning all the matches it should, as there were titles missing from the result set that he knew he'd previously had out on loan. Further investigation revealed errors in the author names that were preventing over 70% of the matches to be returned. Surely there was something that could be done to assist users in such situations?

In this paper we describe an application for users to address issues of metadata control in the web systems they use; whilst also capturing potentially valuable usage data. This approach is inspired by open source development and Wikipedia, where users can correct their own view of the data. It aligns the incentives of this immediate correction with the re-use of the data through both crowdsourcing and traditional error-reporting. We view the developed system as part of an exploration of the design space for user-driven data quality manage-

University of Waikato, New Zealand and University of Illinois, USA

ment. We first briefly outline some issues in authority control and identity management. Then we present the system through three worked examples using existing online information systems. We extend earlier work [3] with a description of the software architecture (Section 4), followed by a discussion of the extensibility of the approach in the context of other user-driven information systems and the growth of the Semantic Web.

2 Background

Searching for materials created by a particular author is a common information need, familiar to users of libraries for centuries. It can facilitate both known item searching, and a kind of browsing (discovering what else an author has written). In order to be effective, we need a way of consistently finding all the works created by the same person, even if the exact name string varies due to different conventions of providing the information (e.g., full first names or just initials) [4]. Authors may change their name style or even their surname over time. Name authority control and, subsequently, access control, have emerged as approaches to manage name variants [16, 20, 23]. However, just because there are well-established principles for the use of name authority control it does not mean that the problem is solved. There are various reasons why an end user may struggle to find all the materials in a collection related to the same person:

- Cataloguing is expensive and time consuming and cataloguers are fallible. Even the best libraries with strong processes for name authority control will contain examples of errors.
- Collections built up over decades, or even centuries, will be subject to evolving bibliographic conventions. Maybe the earlier entries were not done so rigorously, or were done in a way different from current practice.
- Collections can be merged with others catalogued less rigorously, or catalogued in different ways.
- The data itself may be harvested from other sources, perhaps multiple sources with insufficient resources for detailed quality checks [14].
- There may not be the time or resources to periodically re-catalogue everything to bring it all up to the current state of the art.
- Translation and transliteration from different alphabets can introduce name variants, as can the use of diacriticals in foreign names not used in the main language of the catalogue.
- The growth of items in collections created by people with surnames originating in certain regions of East and South Asia, where particular surnames are especially common and disambiguation is consequently more difficult [19].

The above reasons particularly apply to well-organized libraries mostly dealing with books, showing why even there, despite the best efforts of librarians, some problems will remain. However, even a comparatively small error rate will have an effect on a searcher's interaction; although the user may not be fully aware of the implications.

Bibliographic databases of journal and conference papers are typically assembled from heterogeneous sources. Any given paper will contain the authors' names, but formatted according to the convention of the particular journal, and different journals have different conventions. Furthermore, the references in the paper also need to be indexed and are also constrained by both formatting and the vagaries of the authors' use of the given citation style [4]. Additionally, existing name authorities focus on authors who have written books [22]; many more researchers write (and co-author) journal articles, conference papers, white papers, technical reports, etc.

The growth of digital libraries such as institutional repositories (IRs) has brought new challenges, such as bibliographic data provided by the authors of the documents in question [26]. Although it might be expected that authors would have a particular interest in both getting their name correct and also keeping a consistent form to maximize impact, Salo [21] has shown why data quality problems in IRs can be particularly acute. To address these issues Xia [26] proposed that authors should be asked to identify variants of their name when depositing work in an IR.

The expansion in the population of authors seen in IRs has led to proposals for new name authority services [11]. The general problem of accurately identifying authors has been the subject of several initiatives, at both national and international levels. Rotenberg and Kushmerick [20] describe the background leading to the Open Researcher and Contributor ID (ORCID) project [10], which has widespread support to produce an industry-wide unique author identifier. Existing similar systems often encourage authors to claim their publications to help in correcting errors and resolving uncertainty [20]. The *iamResearcher* system [27] uses a social media approach to try to provide incentives for authors to identify their own publications. An ongoing question with these approaches is whether scholars have sufficient motivation to keep their entries in an identification registry up-to-date [20].

In contrast to these global solutions, it is now feasible for specialist digital libraries to be produced by small institutions, even individuals. Powerful open source software such as DSpace, EPrints and Greenstone has lowered many of the barriers to DL creation. However, with limited resources, and perhaps minimal access to professional cataloguers, data quality errors are likely to be even greater than in the other kinds of DLs and databases noted above.

The observable name-related errors in many information systems have led to various approaches to automate the process of correctly grouping and assigning names, including: probabilistic profile-based models [24], heuristic-based hierarchical clustering of names [13], weighted clique heuristics on networks of name variants [9], and using external web-based information sources such as online curricula vitae [17]. Highly visible attempts at author disambiguation can be seen in the individual profiles produced by Google Scholar Citations and Microsoft Academic Search. These systems use proprietary algorithms which are supplemented by manual editing from current authors.

An alternative technique has been to merge existing national authority control systems with those in other countries. "Authority records representing the same entity from the world's national bibliographic agencies would be linked and made available on the Internet. Such a VIAF [Virtual International Authority File] ... would permit national or regional variations in authorized form to co-exist, thereby supporting worldwide users' needs for variations in preferred language, script, and spelling" [5]. The possibility of global integrated authority control data has made feasible many potential scenarios of use [7,12]. A key idea has been to use authority records as a building block for the Semantic Web [23].

Although the primary focus of name management is information retrieval, the processing of data about scholarly communication in citation databases is used for other purposes. Errors of identity and attribution have the potential to affect the calculation of individual researcher metrics, such as the h-index, and citation statistics that are used as inputs to university ranking systems [25,20].

Most prior work looks at approaches from the perspective of the DL and its owners. Methods have evolved to address the issue at the time of acquisition of individual items, or the processing of an ingested large dataset. Maintenance and error correction is a continuing task and several systems have provided mechanisms for researchers to assist through 'claiming' their work, merging split identities or separating merged entries [20].

2.1 An alternative approach

The approach we describe in this paper is substantially different from those described above; it is much more modest in its scope, and is currently a proof of concept application aimed at coping with relatively small numbers of errors. Consequently, it should be seen as complementary to other methods rather than as an alternative. It is focused on the moment of use, enabling end-users to cope more effectively with the name issues that they will inevitably encounter. Given that many use contexts will involve searching multiple different DLs it allows for flexibility and a consistent approach across multiple systems.

So how do we differentiate from these very thorough industrial strength approaches? We have a fast and light approach that does not have the thoroughness of these, but is applicable to bibliographic databases created by others. You do not need access to the full dataset. This makes it more oriented to end-users needing to make do with what is already out there rather than to providers improving what they have. Furthermore it is applicable to multiple databases, thus supporting users who search across datasets rather than purely within a single one. At least in the case of academic use of digital libraries this is a common occurrence.

Recent work by Li *et al.* [15] shares an aspect of our approach in using user activity to address name disambiguation; they aim to take explicit user feedback and integrate it with a machine learning algorithm. However, their approach relies on explicit user registration and they do not detail how their users are incentivised to use their system.

3 System Walkthrough

Seeking a way to constructively assist users when they encounter situations such as those described above, we have created a proof-of-concept system called Computer Says No ... Computer Says Maybe ... Computer Says Yes, or CSN for short. The name was chosen to reflect the fact that the approach provides a way for users to go beyond the frustrating—from their point of view—read-only nature of the web pages produced by today's digital library systems, and actually change the content of pages to something more correct and meaningful when problems arise.

Care has been taken in developing the system to be agnostic towards the digital library systems it operates over; moreover, it has been designed in such a way that it does not need the explicit co-operation of the organization providing the digital library to function, meaning the user can use CSN where and when they

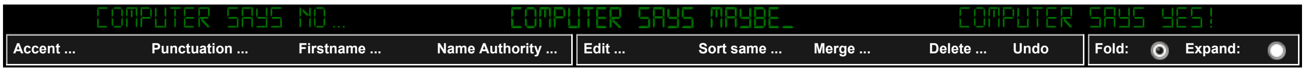


Fig. 1 The default state of the CSN menu toolbar.

see fit. Furthermore, the changes made by the different users using CSN are centrally tracked, and summary statistics are made available to other users of the system (and of course to the system's maintainers). This tracking means that when a user of CSN encounters a piece of information they consider dubious, they can then get an overview of how other people have changed it, or if indeed it has ever been changed at all before.

In the current version of our application (see Section 4 for more details about its implementation) and in the examples below we focus on a few of the problems with name use in order to illustrate the approach and differentiate it from prior work. Inspired by the motivating case of the Asterix books, we focus on the split citation problem and (to a lesser extent) mixed citations.

Once installed in a user's web browser, Figure 1 shows the default state of the menu toolbar that appears at the top of web pages displayed in the browser. From this toolbar an interactive series of transforms to the web page below can be initiated. Notice that the items in the bar are grouped into three areas. Working from left to right, the first area is specifically dedicated to commonly performed transforms to people's names—achieved with a few quick clicks of the computer mouse—such as standardizing the use of punctuation, if middle names are used or not, if first names are initialized or not, and so forth.

The middle area supports more general—but typically more manually intensive—transforms. Through *Edit* CSN allows the user to arbitrarily change any section of text (typically the text of a hyperlinked name) to whatever text they want. This is a useful way to deal with unusual situations such as names intentionally styled not to conform to convention, such as *e.e. cummings*, or to names where little standardization has occurred, such as changing *Moammar Kadhafi* to *Muammar Gaddafi* or editing one of the other 40 or so known variants of the former leader of Libya's name. After such an edit, the user may then use *Sort* to reorder names so the newly changed name is adjacent to the variant they have chosen to standardize on. This would then most typically be followed by using the *Merge* action to join the two names together. However, it should be noted that there is nothing stopping the user continuing with other types of transform offered by CSN, either applied to other names on the page, or indeed the item they have just changed.

Alternatively, through *Delete* the user of CSN can erase names from the page that are superfluous to requirements. While browsing faceted author information by surname, for instance, this functionality can be used to delete *Fox, George* from the list of names to help emphasize the author *Fox, Edward* (assuming the user is seeking articles in the area of digital libraries, or vice versa if looking into the area of grid computing).

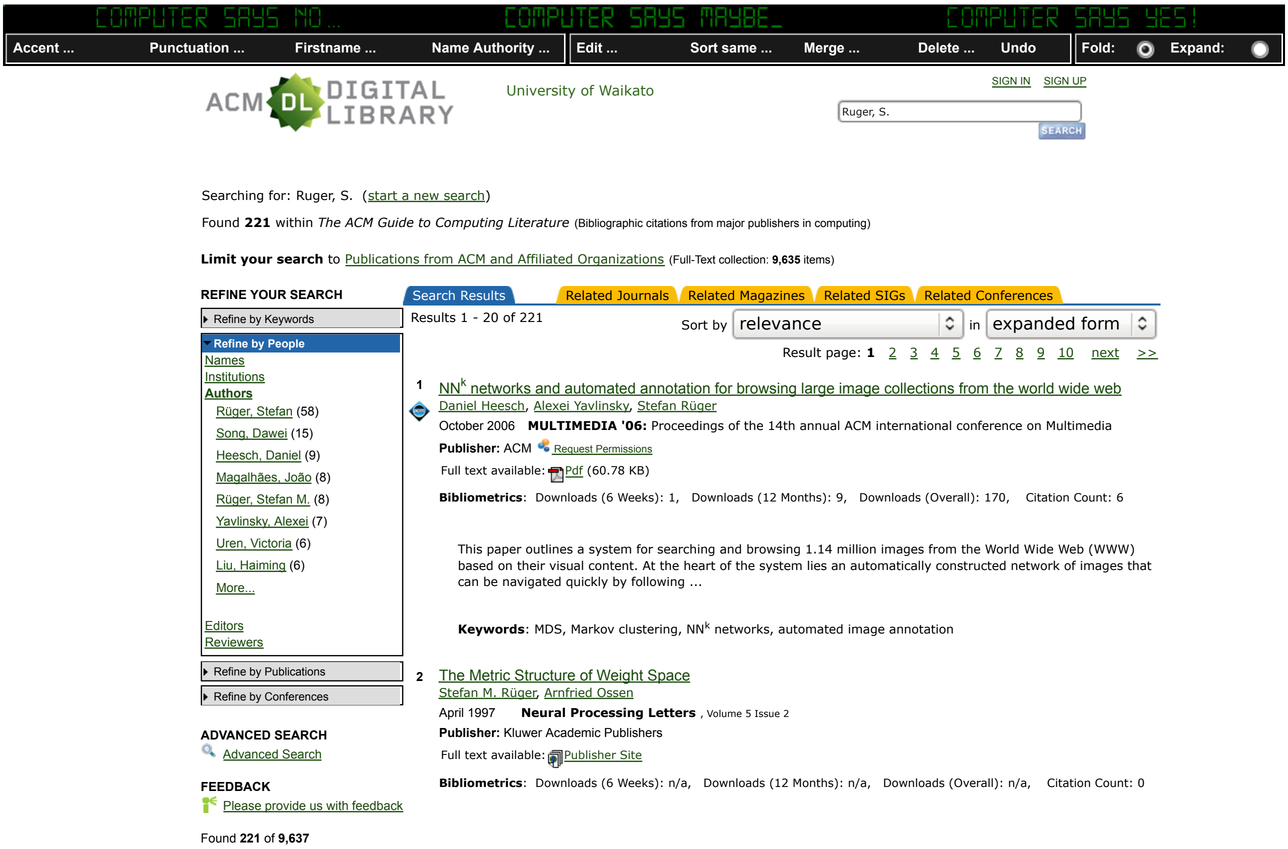
Although the original intention of this feature was for use with faceted names, it can actually be used on any area of a page that the user finds distracting. For example, when perusing the result set returned from a search (it does not even have to be one where the user has used CSN to merge results) we have found using the toolbar's delete functionality an instinctive way to weed out false positives—items in the returned result set where it is often apparent from even a cursory glance of the summary information presented (using the user's domain knowledge of the subject area) which items are not on topic for the formulated query. This pattern of use works well, in particular, in offsetting the problems caused by any errors in algorithmic-based name disambiguation used by the digital library.

The rightmost area to the CSN toolbar controls the mode the software operates in: fold (the default) or expand. The former reduces the text information displayed—for example, removing diacritical marks from characters to assist with the merging of names with versions that do not make use of them. The latter expands the text information displayed—for example, expanding non-accented names in a text search box to include different variants of the name with accented characters, thereby increasing the range of query terms utilized in the search.

3.1 Initial example: single name fold

Having described the general approach CSN takes in overview, we now move on to consider a series of examples that illustrate its capabilities in more detail. As an initial example, we take the task of searching within the ACM Digital Library for papers by the author *Stefan M. Rüger*.

Figure 2 shows the initial screen encountered by the user, with Figure 3 an enlarged version of the *Refine by People* box from the left-hand side of the page that shows the sequence of transforms the user makes to this area using the CSN toolbar.



University of Waikato

Ruger, S.

Searching for: Ruger, S. (start a new search)

Found **221** within *The ACM Guide to Computing Literature* (Bibliographic citations from major publishers in computing)

Limit your search to Publications from ACM and Affiliated Organizations (Full-Text collection: **9,635** items)

REFINE YOUR SEARCH

- Refine by Keywords
- **Refine by People**
  - Names
  - Institutions
  - **Authors**
    - Rüger, Stefan (58)
    - Song, Dawei (15)
    - Heesch, Daniel (9)
    - Magalhães, João (8)
    - Rüger, Stefan M. (8)
    - Yavlinsky, Alexei (7)
    - Uren, Victoria (6)
    - Liu, Haiming (6)
    - More...
  - Editors
  - Reviewers
- Refine by Publications
- Refine by Conferences

ADVANCED SEARCH

Advanced Search

FEEDBACK

Please provide us with feedback

Found **221** of **9,637**

Search Results | Related Journals | Related Magazines | Related SIGs | Related Conferences

Results 1 - 20 of 221 Sort by relevance in expanded form

Result page: **1** 2 3 4 5 6 7 8 9 10 next >>

1 NN^k networks and automated annotation for browsing large image collections from the world wide web
Daniel Heesch, Alexei Yavlinsky, Stefan Rüger
October 2006 **MULTIMEDIA '06:** Proceedings of the 14th annual ACM international conference on Multimedia
Publisher: ACM Request Permissions
Full text available: Pdf (60.78 KB)
Bibliometrics: Downloads (6 Weeks): 1, Downloads (12 Months): 9, Downloads (Overall): 170, Citation Count: 6

This paper outlines a system for searching and browsing 1.14 million images from the World Wide Web (WWW) based on their visual content. At the heart of the system lies an automatically constructed network of images that can be navigated quickly by following ...

Keywords: MDS, Markov clustering, NN^k networks, automated image annotation

2 The Metric Structure of Weight Space
Stefan M. Rüger, Arnfried Ossen
April 1997 **Neural Processing Letters** , Volume 5 Issue 2
Publisher: Kluwer Academic Publishers
Full text available: Publisher Site
Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0

Fig. 2 Searching for *Rüger, Stefan* in the ACM Digital Library: initial screen.

Figure 3a shows the initial view of the *Refine by People* box produced by the ACM digital library. Looking through this list of names, Rüger is listed twice: once without his middle name, and then (four lines later) with it.

As noted in the start of Section 3, the CSN toolbar has a selection of actions associated with it, grouped as follows:

- Accent, Punctuation, Firstname, Name Authority
- Edit, Sort Same, Merge, Delete, Undo
- Fold and Expand

Through the course of this article we will eventually explore all of these through a selection of examples using a variety of digital library systems. For now, for the problem at hand, we will illustrate how with CSN we can direct the ACM Digital Library to recognize the two separate occurrences of Rüger's name as one with the *Firstname* action.

Figure 4 shows the result of clicking on the Firstname action in the CSN toolbar, where the toolbar has expanded downwards to provide a selection of options available within the Firstname action. The instruction immediately below the main heading directs the user to click and drag out an area of interest within the main web page (when the time comes in our worked example, this will be the boxed list of names previously shown in Figure 3a); the next two items control how the names will be changed: the tick next to the item "Firstname" means the whole first name will be left when the change occurs (but any middle names will be removed); changing the tick to be "Only initial" means even the first name will be reduced—down to its initial letter.

Choosing to work with the "Firstname" transformation, in Figure 3b the user has started to drag out an area of interest. Initially moving the mouse cursor around selects individual elements within the page, such as the line *Rüger, Stefan (58)*. Clicking on this item, and then dragging down, causes the selected rectangular area to expand in size. Equally, moving the mouse cursor back up reduces the size of the box again. At this stage of the interaction CSN captures all mouse clicking events, preventing them from propagating any further to elements in the web page. This is so clicking on an item that is hyperlinked, for example, will not cause the browser to navigate away from the current page. When no item from the CSN toolbar is active, then user interaction with hyperlinks on the page proceeds as normal.

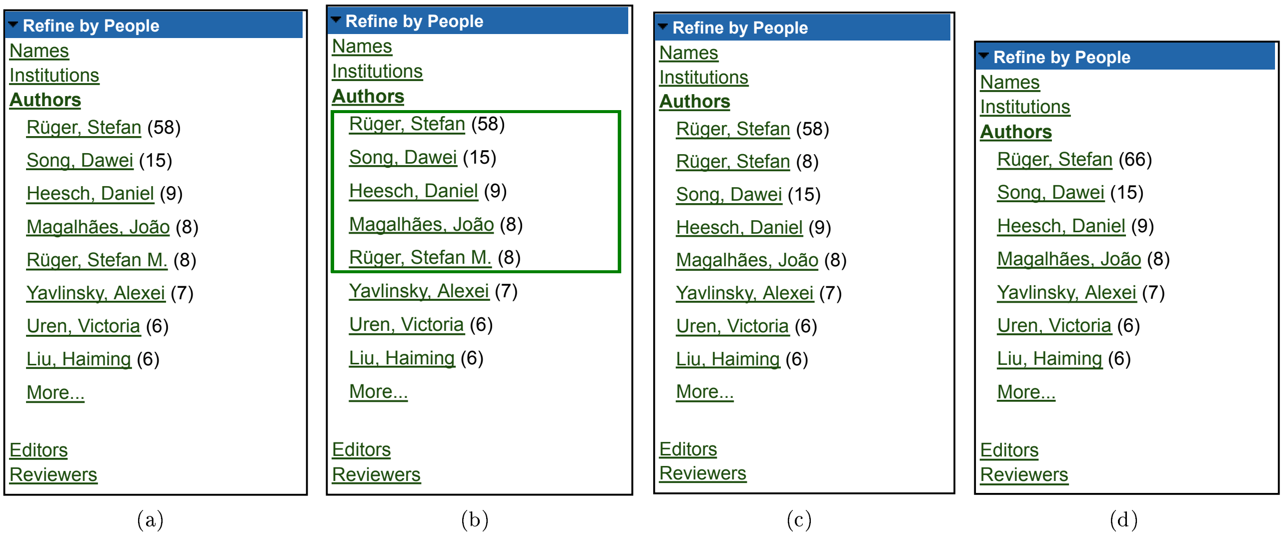


Fig. 3 The different stages of transforming *Rüger, Stefan M.* (a) the initial list of names presented (b) dragging a selected region (c) folding the selected region by firstname (d) merging identical results

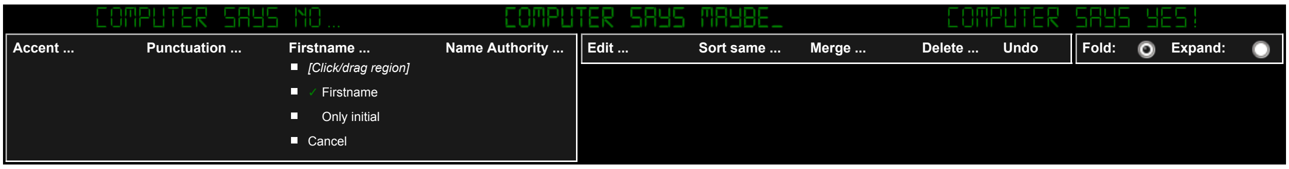


Fig. 4 The options available for the Firstname action.

When the user releases the mouse from a dragging operation, the selected action (Firstname folding in this case) is applied, and any items that are now identical in name are moved next to one another, and the CSN toolbar returns to an inactive state. The merging of identical items does not occur at this point as there are cases where it makes sense to apply further transformation. In Figure 3c we can see the result of this applied to the first three items of the author list in the ACM digital library. Note that the entry *Song, Dawai (15)* has been unaffected by the procedure, it happened to be between the two values we were interested in, and has now moved to be after them.

As this is our first example, we will not consider any more advanced functionality and move straight on to merging the items. In Figure 5 we see the options that result from clicking on the Merge action. Again it is possible to interactively select the area of the web page we wish identical adjacent items to be merged in. There is also a *Previous region* item which, as the name indicates, means the region from the previously selected action will be used. Clicking this results in Figure 3d. The two versions of Rüger's name have indeed been merged, showing a count of $58 + 8 = 66$ matches.

At any stage of this sequence of changes—by way of explanation to the user as to what has happened—hovering over an item that has been changed by CSN brings up a tooltip that captures the history of changes. If the element from the original page already has a tooltip, then the CSN information is appended to it.

Clicking on the newly formed link allows the user to see the result of these merged items (Figure 6). The two result sets corresponding to the searches for the two variants of the name *Rüger, Stefan* (one with a middle name, the other without) in the ACM digital library are brought up simultaneously, and shown side by side. While it would be highly desirable to render the search results as a single list, there are a variety of issues that make this difficult to achieve reliably across a wide range of digital library systems (or even semi-reliably!). We return to this point and discuss it further in Section 5 below.

In terms of the side-by-side frames approach used in CSN, one advantage this has is that it works independently of the digital library system used. Furthermore, to compensate for the lack of a single unifying list, some care has been taken over the formation of the elements that constitute the side-by-side frames. For instance, neither frame is permitted to include a vertical scroll-

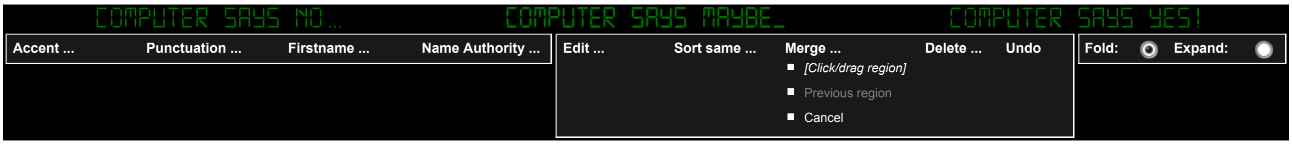
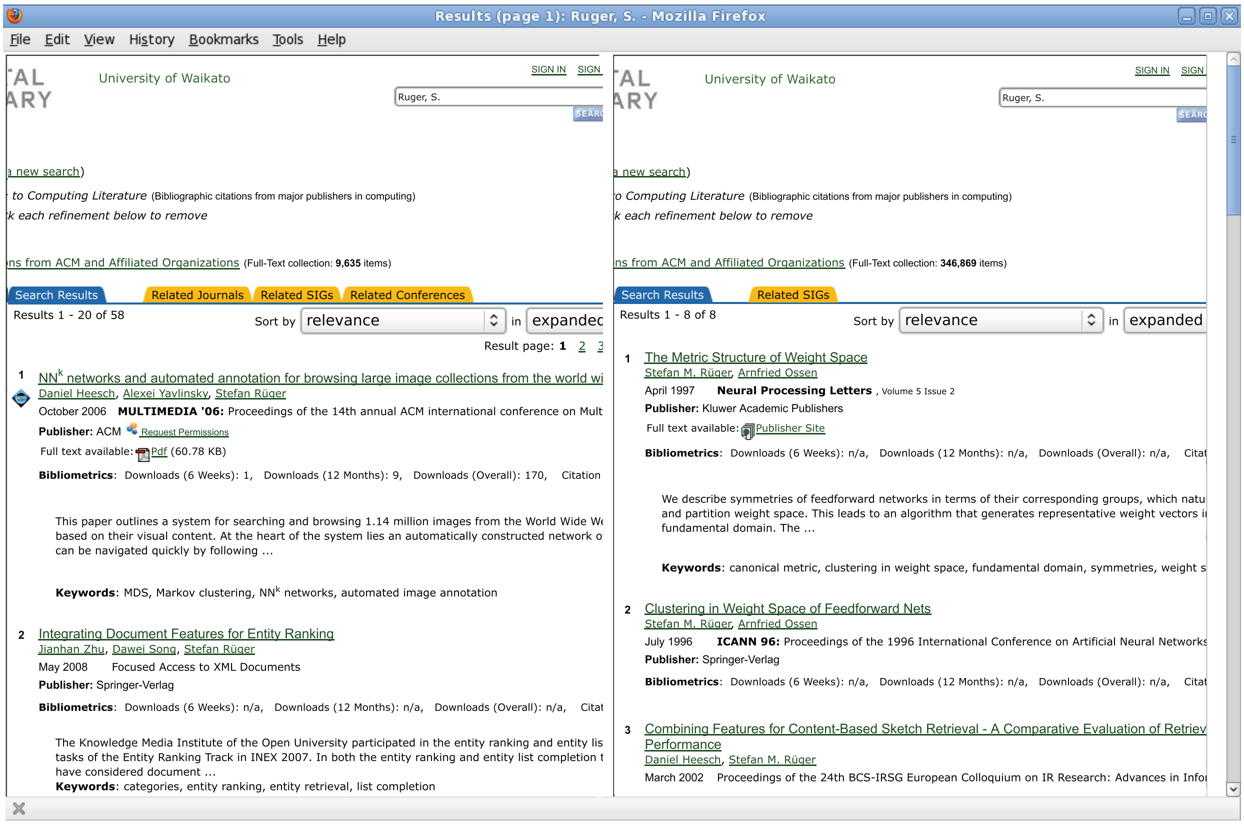


Fig. 5 Options for merging


Results (page 1): Ruger, S. - Mozilla Firefox
File Edit View History Bookmarks Tools Help
TAL
ARY
University of Waikato
SIGN IN SIGN
Ruger, S.
SEARC
Search Results
Related Journals
Related SIGs
Related Conferences
Sort by relevance in expandec

a new search)

to Computing Literature (Bibliographic citations from major publishers in computing)

k each refinement below to remove

ns from ACM and Affiliated Organizations (Full-Text collection: **9,635** items)

Results 1 - 20 of 58

Result page: **1** 2 3

1 NN^k networks and automated annotation for browsing large image collections from the world wi
Daniel Heesch, Alexei Yavlinsky, Stefan Rüger
October 2006 **MULTIMEDIA '06:** Proceedings of the 14th annual ACM international conference on Mult
Publisher: ACM Request Permissions
Full text available: Pdf (60.78 KB)
Bibliometrics: Downloads (6 Weeks): 1, Downloads (12 Months): 9, Downloads (Overall): 170, Citation

This paper outlines a system for searching and browsing 1.14 million images from the World Wide We based on their visual content. At the heart of the system lies an automatically constructed network o can be navigated quickly by following ...

Keywords: MDS, Markov clustering, NN^k networks, automated image annotation

2 Integrating Document Features for Entity Ranking
Jianhan Zhu, Dawei Song, Stefan Rüger
May 2008 Focused Access to XML Documents
Publisher: Springer-Verlag
Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citat

The Knowledge Media Institute of the Open University participated in the entity ranking and entity lis tasks of the Entity Ranking Track in INEX 2007. In both the entity ranking and entity list completion t have considered document ...
Keywords: categories, entity ranking, entity retrieval, list completion

a new search)

o Computing Literature (Bibliographic citations from major publishers in computing)

k each refinement below to remove

ns from ACM and Affiliated Organizations (Full-Text collection: **346,869** items)

Results 1 - 8 of 8

1 The Metric Structure of Weight Space
Stefan M. Rüger, Arnfried Ossen
April 1997 **Neural Processing Letters** , Volume 5 Issue 2
Publisher: Kluwer Academic Publishers
Full text available: Publisher Site
Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citat

We describe symmetries of feedforward networks in terms of their corresponding groups, which natu and partition weight space. This leads to an algorithm that generates representative weight vectors i fundamental domain. The ...

Keywords: canonical metric, clustering in weight space, fundamental domain, symmetries, weight s

2 Clustering in Weight Space of Feedforward Nets
Stefan M. Rüger, Arnfried Ossen
July 1996 **ICANN 96:** Proceedings of the 1996 International Conference on Artificial Neural Networks
Publisher: Springer-Verlag
Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citat

3 Combining Features for Content-Based Sketch Retrieval - A Comparative Evaluation of Retriev Performance
Daniel Heesch, Stefan M. Rüger
March 2002 Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Infor

Fig. 6 Searching for the merged result of *Rüger, Stefan* and *Rüger, Stefan M.* in the ACM Digital Library.

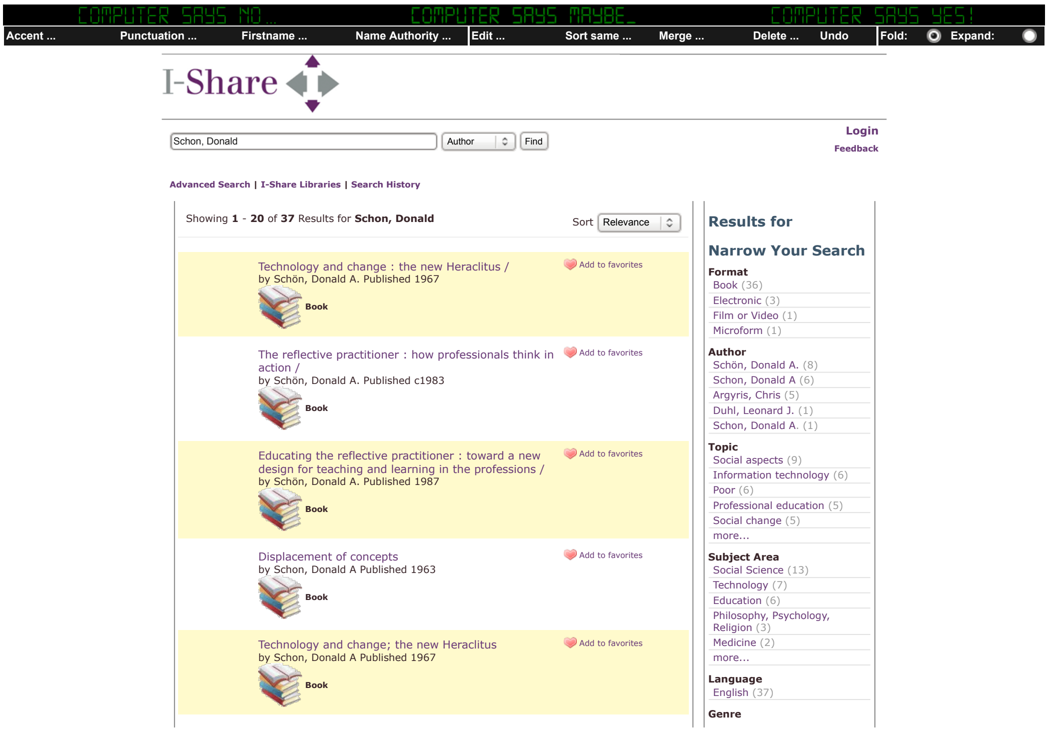
bar. Instead the outer page will add in a scrollbar if needed, and avoids the known user confusion caused by having inner scrollbars within a larger region which may in some circumstances include its own scrollbar. The approach of side-by-side positioning also plays to the trend in display technology for devices to be increasingly wider.

For our example (Figure 6) it is indeed the case that there is a scrollbar located on the right-most side. This is because the search results within the frames exceed the height of the browser's page display area. Changing the position of this scrollbar moves the view of the search results shown within the two frames in unison.

3.2 Combined folding

We illustrate our second example within I-Share, Illinois' statewide integrated academic and research library system. On this occasion we are interested in the author *Schön, Donald A.* and—due to the nature of the errors that occur—this time we will need to combine a sequence of name folding transforms to achieve the desired result. Figure 7 gives an overview to the structure of the I-Share digital library, with the faceted browsing area located on the right, marked out with the heading *Narrow Your Search*.

Figure 8 zooms in on the author block of names to the *Narrow Your Search* sidebar, and shows the sequence of transforms the user makes using CSN. Figure 8a shows the initial names produced by I-Share. In this particular circumstance our author of interest is incorrectly represented three times. While in all three



Schon, Donald

Showing **1** - **20** of **37** Results for **Schon, Donald**

Technology and change : the new Heraclitus /
by Schön, Donald A. Published 1967
Book

The reflective practitioner : how professionals think in action /
by Schön, Donald A. Published c1983
Book

Educating the reflective practitioner : toward a new design for teaching and learning in the professions /
by Schön, Donald A. Published 1987
Book

Displacement of concepts
by Schon, Donald A Published 1963
Book

Technology and change; the new Heraclitus
by Schon, Donald A Published 1967
Book

Results for

Narrow Your Search

Format
Book (36)
Electronic (3)
Film or Video (1)
Microform (1)

Author
Schön, Donald A. (8)
Schon, Donald A (6)
Argyris, Chris (5)
Duhl, Leonard J. (1)
Schon, Donald A. (1)

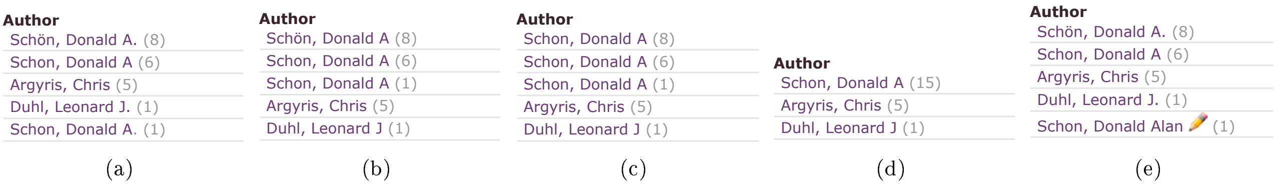
Topic
Social aspects (9)
Information technology (6)
Poor (6)
Professional education (5)
Social change (5)
more...

Subject Area
Social Science (13)
Technology (7)
Education (6)
Philosophy, Psychology, Religion (3)
Medicine (2)
more...

Language
English (37)

Genre

Fig. 7 Searching for *Schön, Donald A.* in the I-Share Digital Library: initial screen



Author
Schön, Donald A. (8)
Schon, Donald A (6)
Argyris, Chris (5)
Duhl, Leonard J. (1)
Schon, Donald A. (1)

(a)

Author
Schön, Donald A (8)
Schon, Donald A (6)
Schon, Donald A (1)
Argyris, Chris (5)
Duhl, Leonard J (1)

(b)

Author
Schon, Donald A (8)
Schon, Donald A (6)
Schon, Donald A (1)
Argyris, Chris (5)
Duhl, Leonard J (1)

(c)

Author
Schon, Donald A (15)
Argyris, Chris (5)
Duhl, Leonard J (1)

(d)

Author
Schön, Donald A. (8)
Schon, Donald A (6)
Argyris, Chris (5)
Duhl, Leonard J. (1)
Schon, Donald Alan ✎ (1)

(e)

Fig. 8 The different stages of transforming *Schön, Donald A.* (a) the initial list of names presented (b) folding the selected region by punctuation (c) folding by accent (d) merging identical results (e) editing content manually with Seaweed

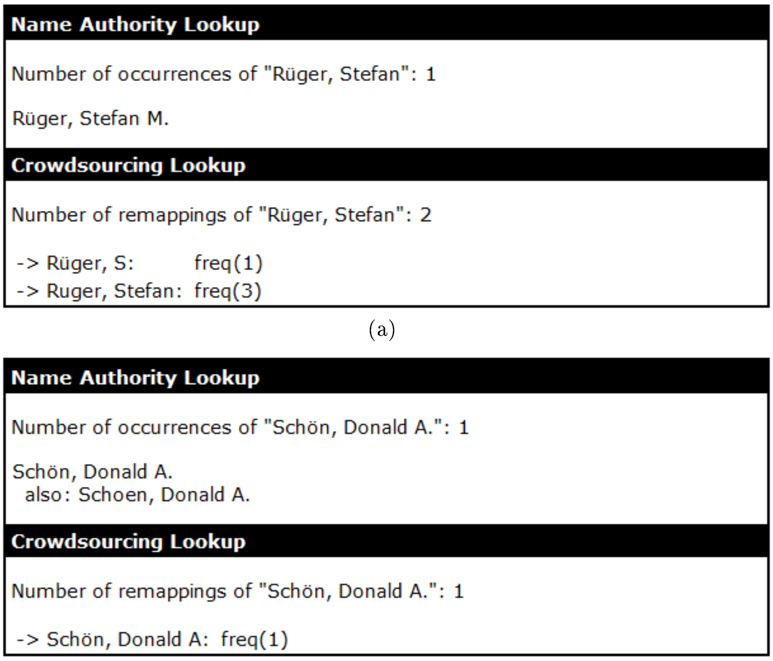
cases both his firstname in full and his middle name as an initial are represented, there are differences too: his surname is spelt once with and twice without an umlaut; similarly the initial for his middle name appears once with a full stop, and twice without.

To compensate for this, first the user decides to use the Punctuation action by selecting a target region. The result of this operation is Figure 8b, where the bottom name has moved up to the third position in the list as it is now identical to the name in the second position. Next the user uses the Accent action, applied to the previously selected region. The list now looks like that shown in Figure 8c. Performing the merge action (again on the previous region) has the desired result of collecting the 15 matching articles by Schön gathered together (Figure 8d).

As a point of technical interest, in the I-Share digital library system the faceted area is implemented through an AJAX-based sub-system, and runs once the main page is loaded. This causes no complication for CSN—from a user's perspective it is quite natural to wait until all the necessary information has come up on the screen, and then start to apply the capabilities of CSN to the live Document Object Model (DOM) that has been formed in the web browser.

3.3 Name authority and crowdsourcing

The CSN toolbar also makes use of a remote name authority search service and crowdsourcing of previous edits people have made using CSN to provide further



Name Authority Lookup

Number of occurrences of "Rüger, Stefan": 1

Rüger, Stefan M.

Crowdsourcing Lookup

Number of remappings of "Rüger, Stefan": 2

```
-> Rüger, S:        freq(1)
-> Ruger, Stefan:  freq(3)
```

(a)

Name Authority Lookup

Number of occurrences of "Schön, Donald A.": 1

Schön, Donald A.
  also: Schoen, Donald A.

Crowdsourcing Lookup

Number of remappings of "Schön, Donald A.": 1

```
-> Schön, Donald A:  freq(1)
```

(b)

Fig. 9 Exploiting name authority metadata and crowdsourcing (a) *Rüger, Stefan* (b) *Schön, Donald A.*

abilities to transform information presented by the digital library. In terms of user interaction, accessing these capabilities differs slightly to our preceding worked examples. Upon activating the Name Authority action, for example, and entering the interactive element selection phase of CSN, hovering over an item for a second or two brings up a popup window like the ones shown in Figure 9, where potential connections between an item in the web page and entities known to the system are made.

Displayed in the upper part of the popup is the result of using OCLC's on-line Virtual International Authority File server (http://viaf.org/) via a SRW/U (Search and Retrieve Web/URL) interface for searching the Library of Congress' Name Authority List. This can yield a variety of information, starting with all the known people that share that name—our two examples happen to be the only entries in the name authority file with that particular combination of first and last name, however there are 31 matches on Edward Fox, for instance; for each person, the canonical version of their name is displayed along with any known alternative variations the person has used when publishing. Figure 10 shows the raw response (all possible matches) to searching for *Schon* as the surname, and *Donald* as the firstname, which is returned in MARC-XML format.

The key information CSN needs to parse out from the returned records is held in the `<mx:datafield>` and `<mx:subfield>` elements. Built-in lookup tables in the CSN code form the basis for mapping the `tag` and `code` attributes these elements contain—with optional indicators for tags (appearing as `ind1` and `ind2`)—into the information that is displayed in the popup window. In Figure 10, for example, the information in tag 100 provides the canonical name for Schön, with tag 400 providing a known variant, Schoen.

Continuing with the two authors from our initial two examples, having hovered over the topmost author name in the Ruger example, in Figure 9a we can see that there is a potential connection between our selected item *Rüger, Stefan* and *Rüger, Stefan M.* For the Schön example, *Schön, Donald A.* turns out to already be the canonical form of his name, and through the known alternatives we glean that *Schoen, Donald A.* is another variant he has published under.

In the lower part of the popup, the result of searching a central repository of changes made by all CSN users is displayed, along with frequency information. This leverages the crowdsourcing part of CSN. The idea of this section of the popup is that, by letting the current user see what transformation others have made when encountering the same lexical name, they can (hopefully) make a more informed decision as to what to do next, based on how frequently certain transforms have been made in the past. This information can be displayed to the user because every time a change is made using CSN the transform is transmitted to a cen-

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="/SRW/searchRetrieveResponse.xsl"?>
<searchRetrieveResponse ...>
  <version>1.1</version>
  <numberOfRecords>1</numberOfRecords>
  <resultSetId>rrglyr</resultSetId>
  <resultSetIdleTime>300</resultSetIdleTime>
  <records xmlns:ns1="http://www.loc.gov/zing/srw/">
    <record>
      <recordSchema>info:srw/schema/1/marcxml-v1.1</recordSchema>
      <recordPacking>xml</recordPacking>
      <recordData>
        <mx:record ... >
          <mx:leader>00000cz  a2200000n  45 0</mx:leader>
          <mx:controlfield tag="001">oca00797874 </mx:controlfield>
          <mx:controlfield tag="005">19940224162208.9</mx:controlfield>
          <mx:controlfield tag="008">
            821110n| acannaab           |a aaa |||
          </mx:controlfield>
          <mx:datafield ind1=" " ind2=" " tag="010">
            <mx:subfield code="a">n  82101313</mx:subfield>
          </mx:datafield>
          ...
          <mx:datafield ind1="1" ind2=" " tag="100">
            <mx:subfield code="a">Scho&#x308;n, Donald A.</mx:subfield>
          </mx:datafield>
          <mx:datafield ind1="1" ind2=" " tag="400">
            <mx:subfield code="a">Schoen, Donald A.</mx:subfield>
          </mx:datafield>
          <mx:datafield ind1=" " ind2=" " tag="670">
            <mx:subfield code="a">Argyris, C. Theory in practice, 1974.</mx:subfield>
          </mx:datafield>
          <mx:datafield ind1=" " ind2=" " tag="670">
            <mx:subfield code="a">His The design studio, 1985?:</mx:subfield>
            <mx:subfield code="b">cover (Donald Scho&#x308;n) p. 4  of cover
                                  (b. 1930; Ph.D.)</mx:subfield>
          </mx:datafield>
         ...
          <mx:datafield ind1=" " ind2=" " tag="999">
            <mx:subfield code="a">33619</mx:subfield>
          </mx:datafield>
        </mx:record>
      </recordData>
      <recordPosition>1</recordPosition>
    </record>
  </records>
  <echoedSearchRetrieveRequest xmlns:ns2="http://www.loc.gov/zing/srw/">
    ...
  </echoedSearchRetrieveRequest>
  <extraResponseData xmlns:ns4="http://www.loc.gov/zing/srw/">
    ...
  </extraResponseData>
</searchRetrieveResponse>
```

Fig. 10 XML response (abridged) from OCLC's Virtual International Authority File SRW/U service for the query, FamilyName=Schon, FirstName=Donald

trally managed information store specifically set up for this task. This can subsequently be searched by author name to yield the desired result—details about how this part of CSN operates can be found in Section 4 below.

Given the experimental nature of CSN, there has not been much time to gather large frequency counts, however the essence of the idea can be seen at work. Figure 9a in the lower portion of the window shows that previously *Rüger, Stefan* has been mapped to *Ruger, Stefan* (i.e., without the accent) three times, and to *Rüger, S* once. As the user moves the mouse cursor within the popup window, the highlighting elements feature continues, and through that they can choose to select whichever region within the popup they like, and this is substituted into the original location in the main web page. Equally they can move the cursor out of the popup window if there is nothing of interest shown, and the window disappears.

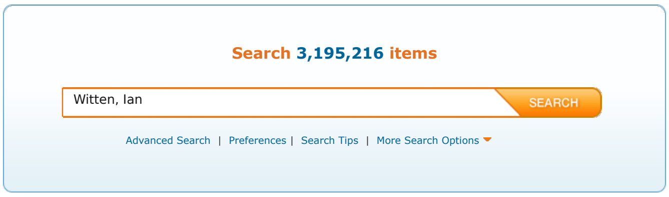
The Name Authority action displays both these categories of information in the popup; when the other actions are active they only show the crowdsource information.

3.4 Editing, deleting, undo

The CSN toolbar also has options for manually editing, sorting and deleting, along with an undo facility. In the case of editing, again the user selects an area of interest: in this case a pencil icon is added against each element in the selected region, signifying that it can be edited directly. Using the technique of Seamless Web Editing we have previously developed [2], these elements can be edited directly—there is no need to reload the page to activate an edit function as in a Wiki or Blog—with operation akin to a word processor. In Figure 8e this editing capability has been activated for the last name in the list. Beyond what can be shown in this figure, if the user of CSN now clicks somewhere within that name a blinking curor appears, representing the position of the current edit point for typing; other editing functions supported include character delete and cursor movements via the arrow keys.

The edit ability adds a “catch all” capability to CSN, allowing the user to apply the full extent of their domain knowledge to change the author names to correspond to what they know is correct. The Sort action reorders the elements so identical elements are adjacent, and when used in tandem with the Edit action, allows the user to manually control what should be merged.

With delete the user can select as before a single element or a range of elements. The action performed once the mouse button is released is to delete these items from the page. This action is a useful way to remove author names that are not of interest to the user. Equally, when browsing the “merged” (side-by-side) set of results, this is a useful feature to remove items that are not of interest. In the case of the author David Bainbridge, for example, there is both a researcher in the digital library field who has no middle name, and another in the field of child psychology. If looking for the DL researcher, in a result set that also erroneously includes matches by the child psychologist, then it is a simple matter to weed them out from the list based on the title metadata information displayed.

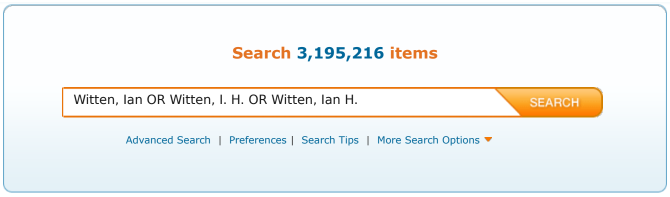


Search 3,195,216 items

Witten, Ian

SEARCH
Advanced Search | Preferences | Search Tips | More Search Options

Fig. 11 Entering *Witten, Ian* as a query term into IEEE Xplore



Search 3,195,216 items

Witten, Ian OR Witten, I. H. OR Witten, Ian H.

SEARCH
Advanced Search | Preferences | Search Tips | More Search Options

Fig. 12 The result of expanding *Witten, Ian* in the IEEE query box using the Name Authority action in CSN

Given these freeform abilities to edit and delete content, it was a natural step to add in an undo feature. In addition to this, CSN allows you to save the page with the accumulated transforms (edits and deletes), leveraging the same centrally managed information store mentioned earlier for crowdsourcing. Visiting the exact same URL again (i.e., when one performs the exact same search) CSN intercedes, and produces the saved version of the page, with the option of reverting to the original version if desired.

3.5 Expanding terms

So far our examples have demonstrated the folding capability of CSN with the radio button in the top right-hand side of the tool bar set to Fold (see Figure 1). For our final example we illustrate CSN’s expansion capability, and show it being applied (for the sake of variety) to the IEEE Xplore digital library.

Changing the radio button in the CSN from *Fold* to *Expand*, alters the behavior of the other items in the toolbar. Now when the user selects an action it replaces the highlighted text with an expanded version. In this

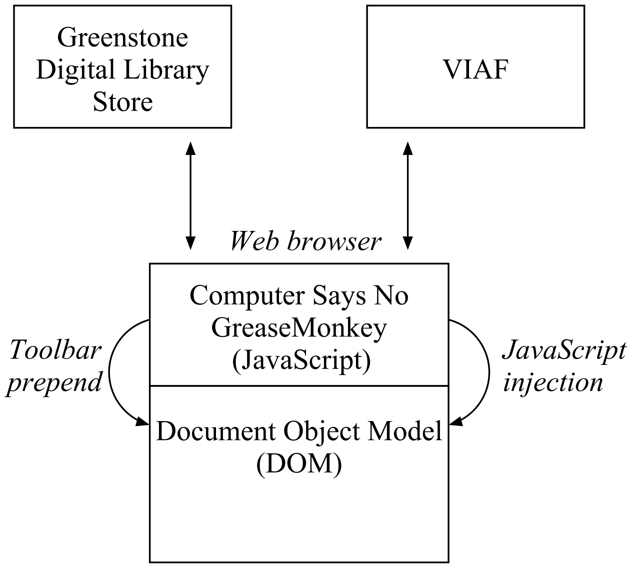

Greenstone Digital Library Store
VIAF
Web browser
Computer Says No GreaseMonkey (JavaScript)
Toolbar prepend
JavaScript injection
Document Object Model (DOM)

Fig. 13 Overview of the software architecture to CSN

mode, rather than apply CSN to faceted lists of names it makes the most sense to apply the action to a text query box where the user has already entered some information, although this is not a constraint—if there are other HTML elements where it makes sense to do this, then they are able to do so.

Visiting the home page to IEEE Xplore, Figure 11 focuses in on the search term box this digital library site provides, where the user has entered the query *Witten, Ian*, but not yet pressed the search button. Instead, as their next step they click on *Name Authority* from the CSN toolbar (with the radio button in the toolbar set to expand), which (like the folding mode) then invites the user to select an area of text to transform. Selecting the text in this query box (*Witten, Ian*) results in Figure 12, where the query term has been expanded to include the known variants *Witten, I. H.* and *Witten, Ian H.* for this author retrieved from the Name Authority service. Pressing the search button now, the user initiates a broader query for articles by this author.

4 Software Design

Figure 13 gives an overview of the software architecture of CSN, which is comprised of three main components:

- a user-script extension to augment a web browser's functionality (lower portion of the figure) which manipulates the DOM of the currently displayed page (under the user's guidance), and accesses two external servers to further enrich the functionality provided;
- a bespoke Greenstone digital library server installation (top left of figure) for crowdsourcing support, used as an information store to record all the text transformations made by the users of the CSN system—from which a summary of transformations can be obtained; and
- a connection to OCLC's Virtual International Authority File web service (top right of figure), used to provide name authority lookup support.

User-scripting is a technique that allows for the functional customization of a web browser. Firefox was the first to support this capability through GreaseMonkey [18], allowing an extension written in JavaScript to operate at a higher level of trust within the browser than is traditionally possible within an HTML web page. The idea proved popular, and has subsequently become supported in browsers such as Chrome, Safari, Opera, and Internet Explorer.

Particularly pertinent for CSN, this higher level of trust allows Cross-Origin Resource Sharing (CORS) in AJAX calls. This means the extension can access the content of the current web page through the JavaScript API for DOM manipulation and at the same time be in contact with servers operating beyond the web site domain that served up the page currently being viewed in the web browser. The ability to mix these two capabilities is what allows the CSN implementation to take an excerpt of text that the user has hovered over (say from *dl.acm.org*) and see if there is a match on the name authority server (*viaf.org*); or for the same excerpt, notice that there is crowdsourced information it can retrieve from the CSN (Greenstone) server on how other people have transformed the text. We will return to these two external services shortly, but let us first consider in more detail how the interface and web browser interaction CSN provides is implemented (the lower component in Figure 13).

The most striking visual feature to CSN is its toolbar. This is implemented as an HTML table, set to be 100% wide, and inserted into the top part of the page being displayed when the document's *onLoad*() event is triggered. The pulldown menus that appear as the result of clicking on items in the toolbar, such as *Punctuation* and *Merge*, are `<div>` elements, and their appearance is controlled by manipulating their CSS style property, `display`.

As noted in Section 3, there are times during CSN's operation—such as during the selection of an area of web content to transform—that any pre-existing interactivity the web page being viewed provides should be temporarily suppressed. This can be controlled through a combination of calls to JavaScript's DOM Event methods *preventDefault()* and *stopPropagation()*.

The interactive selection of elements itself is one of the more intricate parts of the implementation. The key operation here is the ability to identify the smallest, closest binding set of HTML elements that covers the area selected by the user with the computer mouse. Building on a Web browser's event model, which can be set to report the inner-most HTML element that the current (x, y) co-ordinate of the mouse pointer is over, a range of these elements is maintained by CSN. Additional elements may be added, or removed, depending on the movement of the mouse (advancing over new elements or else retracting over existing ones). When the right conditions are met these elements are merged into a single element that is the parent to all the existing elements, and the iterative process of adding and subtracting continues. The complementary operation of dividing a parent element into its child elements was considered, and would certainly lend a certain symmetry to the interaction, but it was found not to be necessary given the way, in general, regions in CSN were selected.

With a region of interest selected, manipulating this to affect changes such as eliminating middle names or accent folding were straightforward to implement. The text in the selected area is simply replaced with the new version. While all strings in JavaScript, like Java, are encoded in Unicode, in contrast to Java, Unicode normalization is not part of the JavaScript language, and so CSN needed to be primed with character encoding lookup tables to provide this particular capability. The necessary Unicode mapping information for this was readily available on-line.

WYSIWYG editing of entries in the web page to make them manually identical before merging them is the catch-all feature in CSN, for any awkward situations where none of the pre-canned folding operations suffice. Here we use a technique we have previously devised called Seamless Web Editing [2], or Seaweed for short, to provide the necessary functionality. A key feature of Seaweed is that no page load is required for the editing to take place, and this distinguishes it from the commonly encountered approaches used in, for example, blogging software and other on-line document editing systems, such as Google Docs. Indeed, for the requirements of CSN, where additional transformations of the document may have already occurred—eliminating middle names, and folding firstnames down to their initial letter—it is vital that the editing operation takes place without a page reload. We know of no other WYSIWYG technique that could be used to do this.

To implement the multiple frames technique for merged queries, where the result sets are shown side by side (as shown in Figure 6 for example), it became necessary in the implementation of CSN to inject JavaScript (in addition to the HTML syntax inserted to produce the toolbar, as discussed above) into the web page being viewed. When two hyperlinked items are merged in CSN, the implementation forms a new hyperlink that calls a JavaScript function, called *multiSearch()*, which takes as one of its parameters an array of URLs formed from the *href* attributes of the hyperlinks being merge. The complication that arises from this is that the JavaScript function *multiSearch()* is defined in the CSN extension, and while a user-script extension has full access to all the content of the web page being displayed, the converse is not true. The solution is to generate a string in the initialization code of the CSN extension that wholly embodies the *multiSearch()* function, wrap it in `<script>` tags and then insert it as an element in to the `<head>` section of the page being viewed. Doing this places the *multiSearch()* function in to the scope of any JavaScript that is initiated from within the web page.

Having described implementation details of the front-line interface to CSN, we now turn our attention to the two back-end services, shown in the upper part of Figure 13: a bespoke Greenstone digital library server installation, and the Virtual International Authority File (VIAF) web service provided by the Online Computer Library Center (OCLC). Contact with these services is initiated by CSN through AJAX calls. In the case of VIAF, information is retrieved from the server (read-only); in the case of Greenstone, information is both stored and retrieved from the server (read/write capable).

The approximate author matching search capability in the VIAF API is used by CSN in two distinct ways: to determine the canonical version of an author's name, and to discover any known aliases. Both capabilities start the same way—issuing a query—but diverge in how they parse and then analyze the returned XML information. The information extracted from these server requests is used to populate the "tooltips" that appear when a user hovers over a name, and in the expanding mode CSN uses it to rewrite the contents of the query search boxes. Being a service that returns Unicode compliant XML content greatly simplifies character encoding issues when processing the data in JavaScript returned through the VIAF service.

The role of the central Greenstone server first and foremost is to provide a location where CSN can save information. It was principally introduced to support the crowdsourcing aspect to the system—where a user can see how others have changed particular names—but it can also act as a repository for saving the final version of a transformed page so that when that individual visits the page again, the page can be restored to how they left it [2].

For the crowdsourcing capability, when a name transformation is made in CSN, the JavaScript system forms a mapping between the original lexical unit to its new version. Simultaneously the same information is transmitted as a time-stamped document to the digital library. Not only does this document contains this mapping as the main body to the document, separate metadata fields are set (such as the original name, and the folded name). Representing the information this way simplifies the task of delivering frequency counts of mappings to the user. Upon initiating a query, HTML syntax is returned, which is processed in a similar fashion to the XML data returned by VIAF, utilizing the parsing classes built-in to JavaScript, from which the summary statistics can be extracted.

5 Discussion

Our aim in this paper has been to show the potential of a range of applications that allow end users to cope with problems with author name disambiguation. Despite the best efforts of DL maintainers, such problems will persist and so it is worth considering the design of features that make it easy for people to cope when they occur (hopefully infrequently) rather than devoting all research and development in the hope of eliminating them entirely. This ameliorative strategic allocation of resources may well be appropriate in other settings where a perfectionist preventative perspective predominates. The approach has various features that can be extended as outlined below.

CSN uses different heuristics to address particular common problems in name disambiguation, including omissions of diacriticals, first and middle name truncation and variable punctuation. Further heuristics can be added, but we have to be careful that these additions do not have undesirable interactions, and that they do not make the user experience overly confusing or clumsy. For instance:

- likely error patterns in entering, transcribing and OCR-ing names;
- considering wider contextual information in the bibliographic record such as co-authorship patterns, institutional affiliations, the title of the article and particular terms in that title, the venue of publication, and the subject area;
- the namespace of all published authors or particular populations derived from census records, phonebooks or other name databases, since greater confidence can be ascribed to variants of a less common name;
- patterns in the writing style of the full text; and
- using online disambiguation information such as home pages, CVs and Wikipedia.

Separately or collectively these can be used to form an opinion about the probability that two records actually refer to articles written by the same person. These same sources of disambiguation information have been proposed and used in various automated approaches (see [22] for an overview). However with a suggestion based approach, we are able to exploit more tentative indicators—provided we make it clear to the end user the underlying reasons for these guesses. A useful metaphor to inspire future design might be the scenario of interacting with a particularly helpful closed stack librarian. You ask for all the material available by a particular author (say D.M. Nichols) and she returns with various items noting “Well here’s what we have, but it looks to me like this pile is by a different person than that pile. I also brought these that might just be by the same person even though they just say they are by a D. Nichols, since they seem to be on the same topic, and here are some others that I’m less sure of.” Such human helpfulness can be contrasted with the robotic literalness of only giving you exactly what you asked for and no hint that you might have got more if you had asked differently.

5.1 Merging result sets

In developing CSN, some time was spent investigating if there was a way to reliably, (or even semi-reliably) merge at the syntax level two issued queries into one, as this would have produced a more straightforward way for the user to view the merged results, and scales when merging three or more queries. Conceptually it should be quite straightforward to combine separate queries to the same digital library with a Boolean OR operation. Unfortunately, features such as faceted search impinge on this idea. In the case of faceted search, in the majority of digital library systems we investigated this operation was implemented as a post-processing filtering step applied to the result set that was returned. Implicit in this approach, then, is an ANDing of terms in cases where more than once facet element is given at a time in the query (as would be the case in our example of two merged authors).

This is why in the work reported here we issued separate AJAX queries to overcome these complications. Retaining the basic approach, a promising avenue to investigate further is to exploit the fact that the two (or more) searches we are seeking to merge are drawn from the same digital library, and therefore the pages returned will have significant sections that are identical HTML: the header and footer, and top-level navigation aids for instance. The sections that are different will help identify where the result sets within the respective pages are, and search specific navigation aids such as next and previous pages, and how many matching terms were found.

There are various other ways that this work can be extended:

- Currently CSN is principally focused on supporting the resolution of split citations. The edit and delete options give very basic support for addressing mixed citations, but more is needed.
- CSN should be tested on a larger set of digital libraries to strengthen the claim for generality.
- Where we can gain participation from DL maintainers, explore how use-based data quality control can be used to inform cost effective centralized updating.
- Important though name authority control is in its own right, very similar approaches can be applied to other bibliographic fields, including title, subject, and publisher.
- It remains an open question whether the approach can also address the particular data quality control problems of institutional repositories.
- The approach could also be applied to other non-bibliographic databases, including digitized cultural heritage collections that also often have distinct data quality problems with various fields.

Many of these possible extensions will need to involve user testing. The ideas of end-user and crowd-sourced feedback are now embedded in many websites but the specific technique of local correction may present conceptual challenges for users.

6 Future Directions

As noted, CSN is a system developed to help explore a design space, not as a standalone application to be implemented as-is. In developing it we have been able to think more deeply about the range of functionalities and use scenarios that might be addressed by particular applications within this area of the design space. We outline some of them here as an illustration of the potential of the approach.

The technical approach of CSN is constrained by both the nature of page content and by the availability of web services such as VIAF. The current version uses VIAF to move from considering page content as text to working with a specific identity record. The growing trend towards more semantic web content, particularly when embedded in web pages, suggests more powerful versions of CSN will soon be feasible. The open linked data movement envisages machine-readable content will be widely available as both stand-alone data (the Web of Data) and embedded in web pages (e.g., microformats within the Web of Documents) [6]. In a Semantic Web environment a CSN-like tool may have immediate access to unique numeric author identifiers (e.g., from ORCID) that have been output as part of a search and browsing interface. Alternatively, a tool such as *Epiphany* [1] could add semantics, such as RDFa links, to enhance plain web pages. The names that a user manipulates in CSN could then be precisely linked to other web services. A user merging name variants would then be operating with a similar level of accuracy to error-correcting facilities within current databases [20]. Semantically-enhanced web content could also allow CSN to better cope with alternative styles of result presentation, such as the facet issue noted in Section 5.1. Linked data has the potential to enable other client-side applications that, like CSN, empower users and also potentially *generate* valuable data.

The CSN approach is, in part, inspired by collective activities such as open source software and Wikipedia where a small (but non-zero) group of end users are inspired (often out of irritation) to fix errors or utility gaps in a piece of software or an encyclopedia. Our approach is somewhat less radical than these; it is not a ‘wikification’ of the metadata. Instead, the system allows the user to correct their *own* view of the data, with the opportunity to pass this correction information on to other users and the owners of the data. This more limited approach (and others like it) will, we hope, open up the possibility of exploring the design space of more participatory error identification and correction.

CSN sits between the extremes of classic read-only databases and the Wikipedia model. The user-centric error-fixing of CSN may be contrasted with much current provision of features for end users to report errors and suggest corrections where the error-correction acts by the end user rely on substantial amounts of altruism. You report an error to benefit all users, maybe not even knowing whether your suggestion will have an impact. This act of altruism is tempered only in the special case of authors reporting errors about their own work. We speculate that this special case of self-interested error correction is consequently a noticeable proportion of all reported errors. Most bibliographic databases do have some feedback channels for error correction [20], but these are usually not well-integrated into the user interface and are sometimes little more than a form of email to the collection maintainer. The Wiki model of Wikipedia allows anyone to edit anything, although there may be checking of the changes made both for accuracy and for vandalism. Wikipedia has a variety of controls in place although there is a strong ethos amongst Wikipedians that these controls should only be applied to particularly problematic, contentious entries [8]. One decision is whether to allow anonymous edits, or require some sort of identity verification as a way of providing a measure of social control of certain anti-social editing actions.

Intermediate models between these extremes allow users, with varying degrees of authentication, to edit or suggest changes to data that may be held about themselves [20]. User, or even author, supplied changes are not perfect and could even introduce more errors. It is for this reason we designed CSN to be less disruptive than Wikipedia-style models, so that users can derive personal benefits from editing activities without having write-access to the core metadata. A further consideration with both edits and suggestions are the costs of managing and assessing this extra flow of information.

A data quality manager for a system with some degree of user feedback (implicit or explicit) is faced with a stream of corrections, merges and suggestions. Prioritizing editing work may involve treating each item of feedback as a 'vote' for an error fix or, more generally, suggest concentrating on areas of higher usage. Where users have been identified then their votes may be weighted according to status [15] or the quality of previous suggestions. Further, errors are unlikely to be equally distributed throughout the database; for example, a group of records imported at the same time may be more likely to share a common problem. We can consider each metadata value in a database to have some probability of error, which is updated based on known errors, shared metadata values, item history, usage-related data, etc. Any database thus has an *error probability field* that is shaped by the nature of the records and the history of the database. As errors are found, possibly from user activity in a CSN-like system, the field updates and maintenance priorities can be updated. Effective ways to display this complex form of error-related information to both users and collection maintainers will be a major challenge for interface designers for digital libraries.

7 Conclusion

In this paper we have moved authority data from server to client, and placed it under interactive control of the user. We have also described how this shift enables user activity to be gathered for crowdsourcing applications to benefit both other users and collection maintainers.

The developed CSN system is intended to be a proof of concept, a single point in the design space of applications intended to deal with the name authority control problem. We want to show that this is a relatively unexplored part of a socio-technical design space, distinct from areas that have been more actively explored. In particular, it is distinct from techniques to normalize records at the time of acquisition (including manual cataloguing processes), and automated processes used to clean up large sets of data acquired *en masse* through importing and harvesting.

The operation of CSN relies on the judgement of the end user and may only be successful with users who have some domain expertise. Even with these significant constraints, it is nevertheless a robust method, able to work with a degree of human-like fuzziness on real-life data, across numerous web-based databases without needing access to the detail of underlying data.

References

1. B. Adrian, J. Hees, I. Herman, M. Sintek, and A. Dengel. Epiphany: Adaptable RDFa generation linking the web of documents to the web of data. In P. Cimiano and H. Pinto, editors, *Knowledge Engineering and Management by the Masses*, pages 178–192. Springer Berlin / Heidelberg, 2010.
2. D. Bainbridge and B. J. Novak. Seamless web editing for curated content. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'10)*, pages 168–175, Berlin, Heidelberg, 2010. Springer-Verlag.
3. D. Bainbridge, M. B. Twidale, and D. M. Nichols. That's 'é' not 'þ' '?' or '□': a user-driven context-aware approach to erroneous metadata in digital libraries. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital libraries (JCDL'11)*, pages 39–48, New York, NY, USA, 2011. ACM.
4. J. Beall. Metadata for name disambiguation and collocation. *Future Internet*, 2(1):1–15, 2010.
5. R. Bennett, C. Hengel-Dittrich, E. O'Neill, and B. Tillett. VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress name authority files. In *World Library and Information Congress: 72nd IFLA General Conference and Council.*, 2006. http://www.ifla.org/IV/ifla72/papers/123-Bennett-en.pdf.
6. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 4(2):1–22, 2009.
7. T. Burrows. Identity parade: building web portals about people. *OCLC Systems & Services: International digital library perspectives*, 23(4):329–331, 2003.
8. P. B. de Laat. How can contributors to open-source communities be trusted? on the assumption, inference, and substitution of trust. *Ethics and Information Technology*, 12(4):327–341, 2010.
9. D. Feitelson. On identifying name equivalences in digital libraries. *Information Research*, 9(4):paper 192, 2004. http://InformationR.net/ir/9-4/paper192.html.

10. M. Fenner. ORCID: Unique identifiers for authors and contributors. *Information Standards Quarterly*, 23(3):10–13, 2011.
11. A. Hill. What's in a name? prototyping a name authority service for UK repositories. In *Culture and Identity in Knowledge Organization: Proceedings of the Tenth International ISKO Conference (ISKO 2008)*, pages 196–202, Würzburg, Germany, 2008. Ergon.
12. M. Kaiser, H.-J. Lieder, K. Majcen, and H. Vallant. New ways of sharing and using authority information: the LEAF project. *D-Lib Magazine*, 9(11), 2003. http://www.dlib.org/dlib/november03/lieder/11lieder.html.
13. A. H. Laender, M. A. Gonçalves, R. G. Cota, A. A. Ferreira, R. L. Santos, and A. J. Silva. Keeping a digital library clean: new solutions to old problems. In *Proceeding of the Eighth ACM Symposium on Document Engineering (DocEng'08)*, pages 257–262, New York, NY, USA, 2008. ACM.
14. C. Lagoze, D. Krafft, T. Cornwell, N. Dushay, D. Eckstrom, and J. Saylor. Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries (JCDL'06)*, pages 230–239, New York, NY, USA, 2006. ACM.
15. Y. Li, A. Wen, Q. Lin, R. Li, and Z. Lu. Incorporating user feedback into name disambiguation of scientific cooperation network. In H. Wang, S. Li, S. Oyama, X. Hu, and T. Qian, editors, *Web-Age Information Management*, pages 454–466. Springer Berlin / Heidelberg, 2011.
16. D. McKay, S. Sanchez, and R. Parker. What's my name again? Sociotechnical considerations for author name management in research databases. In *Proceedings of the 22nd Annual Conference of the Australian Computer-Human Interaction Special Interest Group (OZCHI 2010)*, pages 240–247. CHISIG, 2010.
17. D. A. Pereira, B. Ribeiro-Neto, N. Ziviani, A. H. Laender, M. A. Gonçalves, and A. A. Ferreira. Using web information for author name disambiguation. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'09)*, pages 49–58, New York, NY, USA, 2009. ACM.
18. M. Pilgrim. *Greasemonkey Hacks: Tips & Tools for Remixing the Web with Firefox*. O'Reilly Media, Inc., 2005.
19. J. Qiu. Scientific publishing: Identity crisis. *Nature*, 451(7180):766–767, 2008.
20. E. Rotenberg and A. Kushmerick. The author challenge: Identification of self in the scholarly literature. *Cataloging & Classification Quarterly*, 49(6):503–520, 2011.
21. D. Salo. Name authority control in institutional repositories. *Cataloging & Classification Quarterly*, 47(3):249–261, 2009.
22. N. Smalheiser and V. Torvik. Author name disambiguation. *Annual Review of Information Science and Technology*, 43:287–313, 2009. Medford, New Jersey: Information Today.
23. B. Tillett. Authority control: State of the art and new perspectives. *Cataloging & Classification Quarterly*, 38(3):23–41, 2004.
24. V. I. Torvik and N. R. Smalheiser. Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, 3(3):Article 11, 2009.
25. A. van Raan. Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1):133–143, 2005.
26. J. Xia. Personal name identification in the practice of digital repositories. *Program: electronic library and information systems*, 40(3):256–267, 2006.
27. Y. Yang, P. Singh, J. Yao, C.-m. Au Yeung, A. Zareian, X. Wang, Z. Cai, M. Salvadores, N. Gibbins, W. Hall, and N. Shadbolt. Distributed human computation framework for linked data co-reference resolution. In G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, editors, *The Semantic Web: Research and Applications*, pages 32–46. Springer Berlin / Heidelberg, 2011.