That's ‘é’ not ‘þ’ ‘?’ or ‘□’: A User-driven Context-aware Approach to Erroneous Metadata in Digital Libraries

David Bainbridge
University of Waikato
Hamilton, New Zealand
davidb@cs.waikato.ac.nz

Michael B. Twidale
University of Illinois
Urbana-Champaign, USA
twidale@illinois.edu

David M. Nichols
University of Waikato
Hamilton, New Zealand
dmn@cs.waikato.ac.nz

ABSTRACT

In this paper we present a novel system for user-driven integration of name variants when interacting with web-based information systems. The growth and diversity of online information means that many users experience disambiguation and collocation errors in their information searching. We approach these issues via a client-side JavaScript browser extension that can reorganise web content and also integrate remote data sources. The system is illustrated through three worked examples using existing digital libraries.

Categories and Subject Descriptors

H.3.7 [**Information storage and retrieval**]: Digital Libraries

General Terms

Design and Experimentation

Keywords

Selective Web Editability, Name Authority Control, Crowdsourcing

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 all the examples presented later in this paper attest to, which are taken from real-world digital libraries and were not hard to come by, it does not take much investigating to discover there is a real weakness in our information systems—in particular our digital libraries—when it comes to the names of people. Despite considerable effort in developing and maintaining authority control systems many users experience significant problems in interacting with systems' representations of people. In this paper we describe a system to allow users to adapt and enhance the representation of authors in the systems they use for information searching.

Some of the inspiration for the work (and the title to this paper) 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 a proof-of-concept system for users to address issues of metadata control in the system they are using. We first briefly outline some issues in authority control. Then we present the system through three worked examples using existing online information systems, and conclude with a discussion of possible future extensions.

http://doi.acm.org/10.1145/1998076.1998084

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 (whether a full first names is provided, or just the initial, and similarly for whether and how any middle name is provided) [2]. 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 [10].

However, just because there are well established principles for the use of name authority control 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 obtained from other sources, perhaps multiple sources with insufficient resources for detailed quality checks [9].
- There may not the 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 originating in certain regions of East and South Asia where particular surnames are especially common [13].

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 be have an effect on a searcher's interaction; although the user may not be fully aware of the implications. When data is automatically gathered from diverse sources error rates are likely to be much higher [9].

Bibliographic databases of journal and conference papers are typically assembled from heterogeneous sources and have more limited resources to name information. 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 [2]. Additionally, existing name authorities focus on authors who have written books; many more researchers write (and co-author) journal articles. "Most cases to date in which author names have been disambiguated have tended to involve manual curation. For example, librarians have traditionally carried out authority control on book collections" [15].

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. 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 [14] has shown why data quality problems in IRs can be particularly acute. In the case of IRs the expansion in the population of authors has led to proposals for new name authority services [6].

Specialist small scale digital libraries are now feasible to produce by small institutions, even individuals. Powerful open source software such as Greenstone lowers 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 [17], heuristic-based hierarchical clustering of names [8], weighted clique heuristics on networks of name variants [5], and using external web-based information sources such as online curricula vitae [11]. 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" [3]. The possibility of global integrated authority control data invited many potential scenarios of use [7, 4]: "We can open up the valuable information within our authority records to users worldwide ... [and be] a building block for the infrastructure of the Semantic Web and beyond" [16].

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. Our approach is substantially different from these; 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 the methods above 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.

In the current version of our application and in the examples below we just 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 primarily focus on split citation and (to a lesser extent) mixed citations. Again, for simplicity, we begin with the use of diacriticals in Latin alphabets.

3. WALKTHROUGH

Seeking a way to constructively assist users when they encounter situations such as those described above, we have created a prototype system called Computer Says No ... Computer Says Maybe ... Computer Says Yes, or CSN for short.

Central to the approach is providing a way for a user 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 a page itself to something more correct and meaningful. They could then even save it in their own CSN "web-space" and draw upon it later on if they so wished. The operations of the different users using the system are also tracked, and summary statistics are made available to other users of the system, and of course to the system's maintainers. This way 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.

The system is implemented entirely in JavaScript so the capability can be added to any web browsers that support GreaseMonkey [12], such as Firefox and Chrome. Having activated CSN it appears as a persistent toolbar across the top of the currently loaded web page. From this toolbar an interactive series of transforms to the web page can be activated.

3.1 Initial example: single name fold

As an initial example, take the task of searching within the ACM Digital Library for papers by the author *Stefan M. Rüger*. Figure 1 shows the initial screen encountered by the user, with Figure 3a an enlarged version of the *Refine by People* box from the left-hand side of the page. Looking through this list of names, Rüger is listed twice: once without his middle name, and then (two lines later) with it.

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

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 as one with the *Firstname* action.

Figure 2 shows the result of clicking on the Firstname action in the CSN toolbar, where the toolbar has expanded 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.

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 (41)*. Clicking on this item, and then dragging down results in the selected rectangular area expanding 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 items that is hyperlinked, for example, will not cause the browser to navigate away from the current page.

When the user releases the mouse from their 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. The merging of these 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 (12)* 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 4 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 $41 + 8 = 49$ 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 5). The two frames shown side by side correspond to searches for the two versions of the name the ACM digital library has for Rüger. 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!).

In terms of the split frames approach used in CSN, one advantage this has is that it works across all digital library systems tested. 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 split-frame search. For instance, neither frame is permitted to include a vertical scrollbar. 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 monitor display technology development which is for the devices to be increasingly wider.

For our example (Figure 5) 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

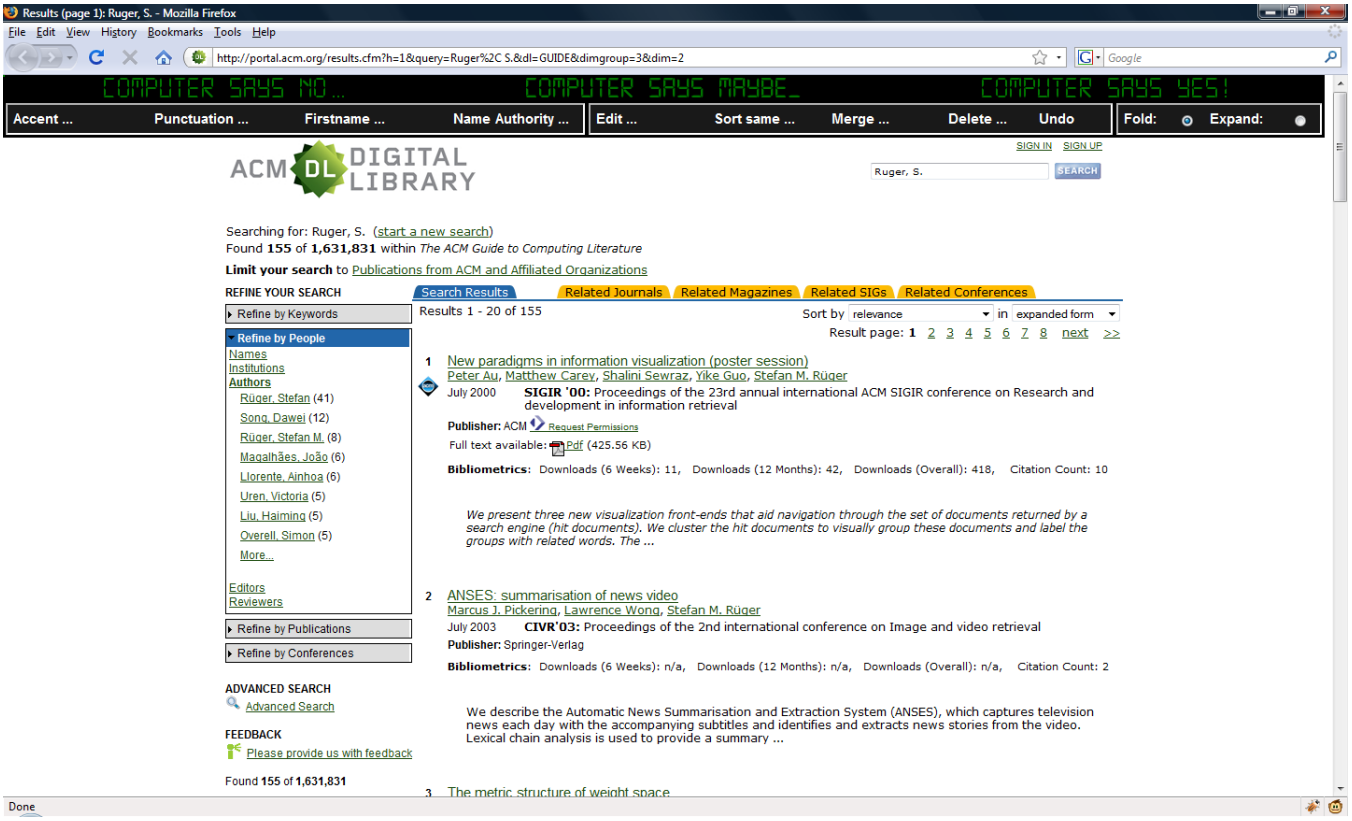


Figure 1: Searching for *Rüger, Stefan* in the ACM Digital Library: initial screen.

Figure 2: The options available for the Firstname action.

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. Figure 6 shows the sequence of transforms the user makes, with Figure 6a showing the initial names produced by I-Share. In this particular circumstance our author of interest is incorrectly represented three times. While in all three 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 6b, 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 6c. Performing the merge action (again on the previous region) has the desired result of collecting the 15 matching articles by Schön gathered together (Figure 6d).

As a point of technical interest, in this 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 instinctive to wait until all the necessary information has come up on the screen, and then start to apply the capabilities of CNS 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 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 items for a second or two brings up a popup window like

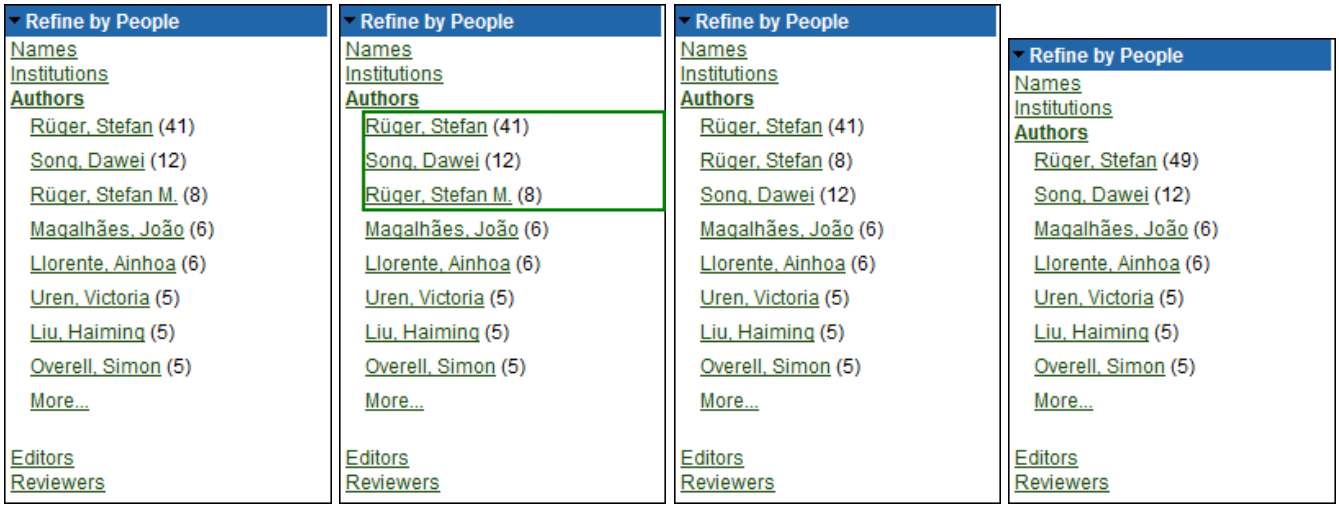


Figure 3: The different stages of transforming *Rüger, Stefan M.* (a) the initial list of names presented (b) dragging a selected region (c) folding the selected region by firstname (d) merging identical results

Figure 4: Options for merging

the ones shown in Figure 8, 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 VIAF service (http://viaf.org/) via SRW/U 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 example; for each person, the canonical version of their name is displayed along with any known alternative variations the person has used when publishing. Figure 9 shows the raw response (all possible matches) to searching for *Schon* as the surname, and *Donald* as the firstname. The returned result is in MARC-XML format, and it is the task of CSN to parse this data to glean the necessary information for display in the popup window.

Continuing with the two authors from our initial two examples, having hovered over the top most author name in the Ruger example, in Figure 8a 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. 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 saved centrally (in a Greenstone digital library). This can subsequently be search by author name to yield the desired result.

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 8a 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

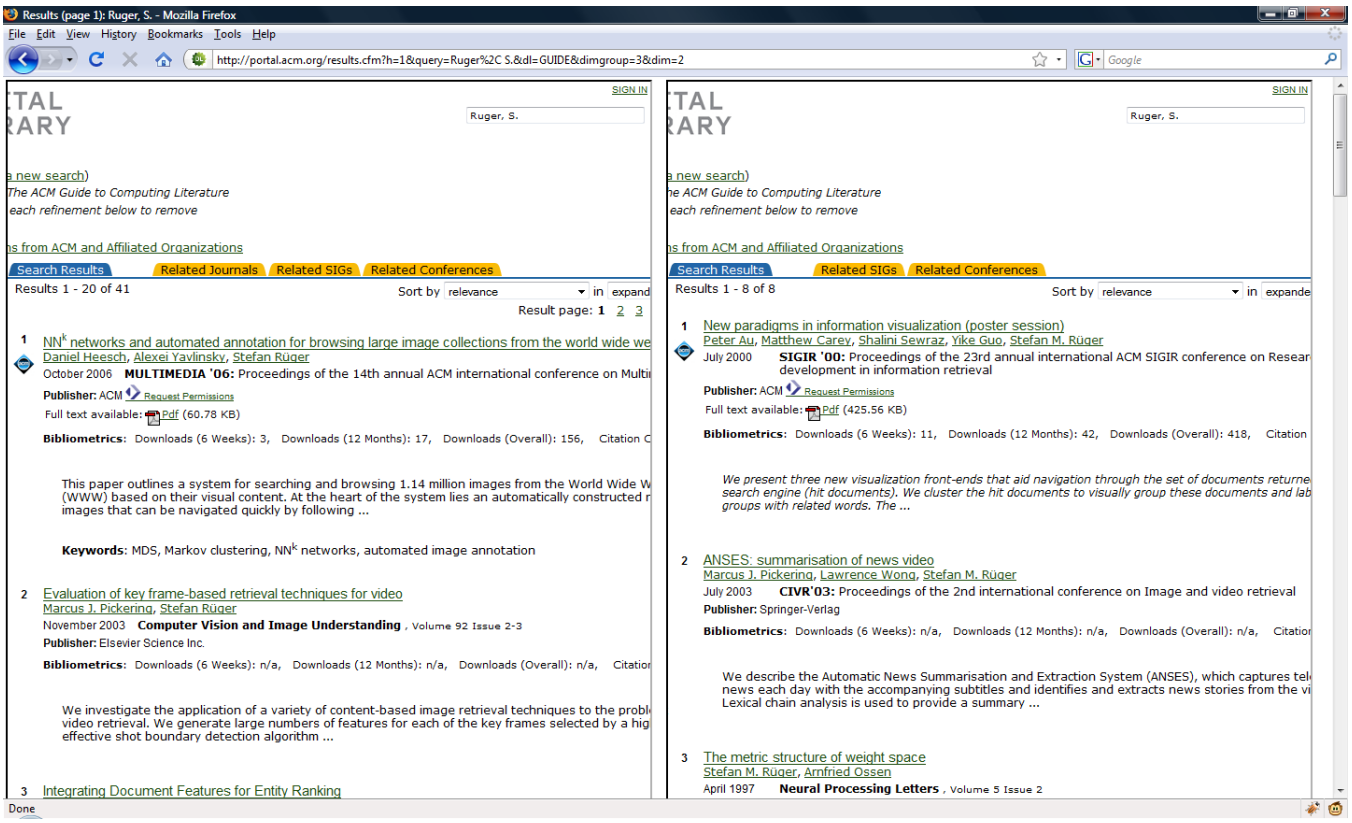


Figure 5: Searching for the merged result of *Rüger, Stefan* and *Rüger, Stefan M.* in the ACM Digital Library.

the selected region, signifying that it can be edited directly. Using the technique of Seamless Web Editing [1], 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 6e editing has been switched on for the pre-merged list of elements. A subtle point to convey through a figure, in the first line of the authors there is a blinking cursor (captured as a solid vertical line). The line represents the position of the current edit point for typing; other editing functions include 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.

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). 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 revering to the original version is desired.

3.5 Expanding terms

So far our examples have demonstrated the folding capability of CSN. The radio button choice in the top right-hand side of the tool bar has been set to Fold (see for example Figure 2). For our final example we illustrate one example of its expansion capability.

With the radio button changed to the Expand setting, now when the user selects an action, it replaces the highlighted text with an expanded version of the text. Here 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. For example, in Figure 10 we use IEEE Xplore, and enter the query *Witten, Ian.* Using the Name Authority action in expand mode, followed by selecting the text in this query box results in Figure 11, where the query term has been expanded to include the variants *Witten, I. H.* and *Witten, Ian H.*

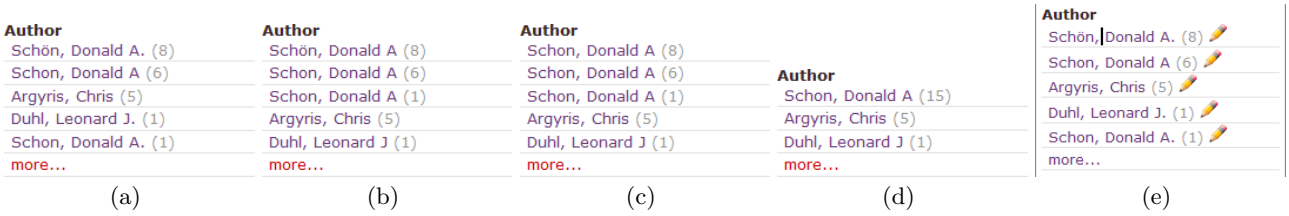


Figure 6: The different stages of transforming *Schön, Donald A.* (a) the initial list of names presented (b) folding the selected region by punctuation (c) folding by accent (d) merging identical results (e) editing content manually with Seaweed

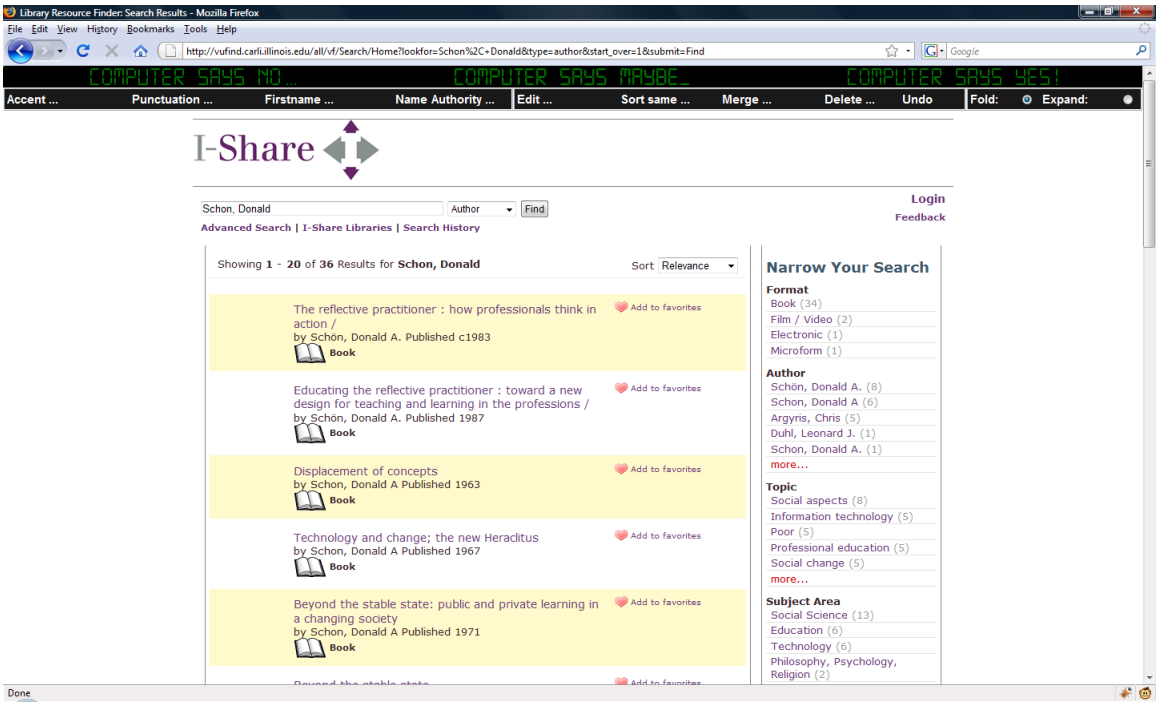


Figure 7: Searching for *Schön, Donald A.* in the I-Share Digital Library: initial screen

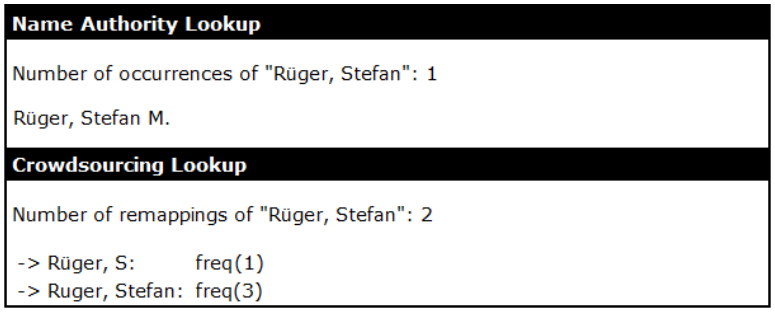
4. DISCUSSION

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.

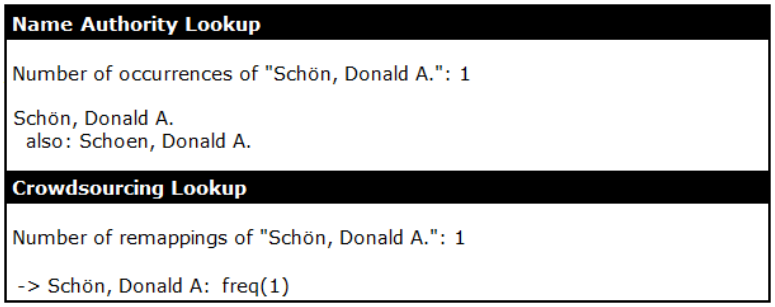
CSN uses different heuristics to address particular common problems in name disambiguation, including omissions of diacriticals, first and middle name truncation and variable punctuation. Further heuristics can be added, but we have to be careful that these additions do not have undesirable interactions, 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 additional disambiguation information online including 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



(a)



(b)

Figure 8: Exploiting name authority metadata and crowdsourcing (a) *Rüger, Stefan* (b) *Schön, Donald A.*

in various automated approaches (see [15] 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.

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 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 terms. 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 searches occur on 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.

```
<?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>
```

Figure 9: XML response (abridged) from OCLC's Virtual International Authority File SRW/U service for the query, FamilyName=Schon, FirstName=Donald

- 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.

5. CONCLUSION

In this paper we have moved authority data from server to client, and placed it under interactive control of the user. The system developed is intended to be a proof of concept. It is 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 that sociotechnical 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 best practices), and also those used to clean up large sets of data acquired *en masse* through merges or automated process (where machine learning processes predominate).

Our method uses a set of relatively simple heuristics directly derived from observed patterns in the data that can make suggestions for ways to combine resources likely to have been created by the same person. However the method relies on the end user to make the final judgement call. This clearly requires the end user to do more work, and will

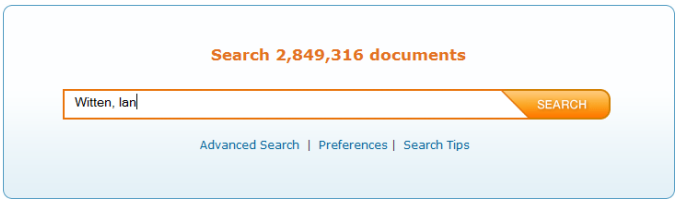


Figure 10: Entering *Witten, Ian* as a query term into IEEE Xplore

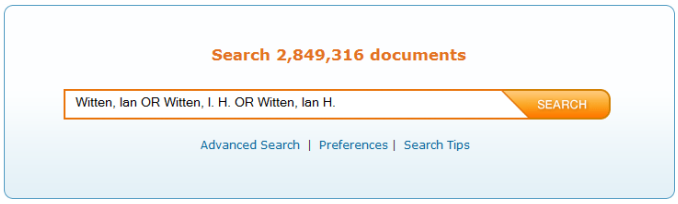


Figure 11: The result of expanding *Witten, Ian* in the IEEE query box using the Name Authority action in CSN

only be successful with users having 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 messy real-life data, across numerous web-based databases without needing access to the raw data.

The 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. We do not propose a 'wikification' of the digital library data, allowing any users to arbitrarily change any entry they may consider erroneous to the value they may consider correct. Rather our 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.

In an ideal world, systems such as these would be unnecessary. Digital libraries would be so well funded that they could afford teams of data quality managers to constantly scour the bibliographic database for errors and fix them. But of course we do not live in an ideal world, and so we want to promote systems such as these that try to ameliorate inevitable errors in low-cost, low-effort ways that offer the potential of improving overall data quality.

6. REFERENCES

[1] 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.

[2] J. Beall. Metadata for name disambiguation and collocation. *Future Internet*, 2(1):1–15, 2010.

[3] R. Bennett, C. Hengel-Dittrich, E. OŠNeill, 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.

[4] T. Burrows. Identity parade: building web portals about people. *OCLC Systems & Services: International digital library perspectives*, 23(4):329–331, 2003.

[5] D. Feitelson. On identifying name equivalences in digital libraries. *Information Research*, 9(4):paper 192, 2004. http://InformationR.net/ir/9-4/paper192.html.

[6] A. Hill. WhatŠs 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.

[7] 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.

[8] 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.

[9] 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.

[10] D. McKay, R. Parker, and S. Sanchez. 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.

[11] 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.

[12] M. Pilgrim. *Greasemonkey Hacks: Tips & Tools for Remixing the Web with Firefox*. O'Reilly Media, Inc., 2005.

[13] J. Qiu. Scientific publishing: Identity crisis. *Nature*, 451(7180):766–767, 2008.

[14] D. Salo. Name authority control in institutional repositories. *Cataloging & Classification Quarterly*, 47(3):249–261, 2009.

[15] N. Smalheiser and V. Torvik. Author name disambiguation. *Annual Review of Information Science and Technology*, 43:287–313, 2009. Medford, New Jersey: Information Today.

[16] B. Tillett. Authority control: State of the art and new perspectives. *Cataloging & Classification Quarterly*, 38(3):23–41, 2004.

[17] V. I. Torvik and N. R. Smalheiser. Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, 3(3):Article 11, 2009.