Are Your Citations Clean? New Scenarios and Challenges in Maintaining Digital Libraries

Dongwon Lee, Jaewoo Kang*, Prasenjit Mitra, C. Lee Giles, and Byung-Won On

The Pennsylvania State University and North Carolina State University*

Introduction

In many scientific-publication digital libraries (DLs) such as CiteSeer, arXiv e-Print, DBLP, or Google Scholar, "*citations"* play an important role. (The term "citation" refers to the collection of bibliographic information such as author name, title, publication venue, or year that are pertinent to a particular article.) Users often use citations to find information of interest in DLs, and researchers depend on citations to determine the impact of an article in DLs. In addition, when DLs are integrated, citations act as unique identifiers of associated documents. Therefore, it is important for DLs to keep citations of stored documents consistent and up-to-date. However, in general, keeping citations clean and consistent is a non-trivial task. Some of the challenges include: (1) data entry errors, (2) various citation formats, (3) lack of (the enforcement of) a standard, (4) imperfect citation gathering software, (5) common author names or abbreviations of publication venues, and (6) large-scale citation data.

People have noticed that many of these problems can be solved by using "global IDs" – no matter how different two citations appear, if both carry the same global ID, then they are considered the same citation. Some of the popular global IDs are ISBNs or Digital Object Identifiers (DOI) [10]. Despite their many benefits, however, such global IDs have been only partially adopted among publishers, while largely ignored by end users (especially, on the Web). That is, a scholar who posts her "publication list" to her home page usually does not put a DOI in front of each citation. Similarly, she usually does not use DOIs in the reference when she writes scientific documents (although people in some scientific disciplines such as Physics often use DOIs). Even if all such users adopt global IDs, inter-operation among different global IDs (e.g., ISBN vs. DOI) is still a remaining issue. Moreover, marking existing documents with global IDs involves substantial costs. For DLs whose data are manually curated by human experts such as ISI's SCI or DBLP, the issue of erroneous and duplicate citations is less obvious, although it still exists. However, for DLs whose data are automatically gathered and generated by software agents such as CiteSeer or Google Scholar, the problem is exacerbated [8]. Since automated indexing methods [4] are not as accurate as human experts, and because human users use diverse citation formats to refer to what is really the same article, many citation errors are included in such DLs. For large-scale DLs where human indexing methods are not sustainable, good performing automated methods are essential.

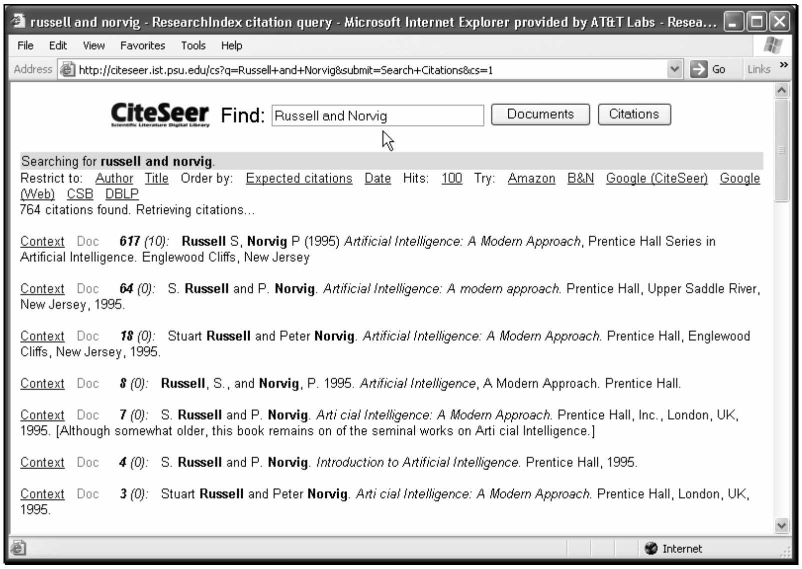
As a result, to maintain clean citations, DLs have to routinely search their collections and fix incorrect citations or remove duplicates. This so-called ***Citation Matching (CM)*** **problem** is a specialized version of the more general problem known as the "*Record Linkage (RL)*" problem [3,12], which has been extensively researched in various disciplines under various names (e.g., [2][11][6][9]). Formally, the CM problem can be stated as follows:

> **Given two lists of citations, *A* and *B*, for each citation *a* in *A*, find a set of citations *b* in *B* such that both *a* and *b* refer to the same article.**

In practice, to determine whether two citations refer to the same real-world document or not (without using global IDs), people use some distance metrics (e.g., Levenstein, Jaro, or Cosine) and a pre-defined similarity threshold. That is, according to some distance function, if the distance between two citations, *a* and *b,* is within the threshold, then two citations are marked to be "duplicates."

Motivation

To demonstrate the need for a solution to the CM problem, let us present three problems drawn from real applications. The first is the example introduced by [8]. Figure 1 is the screen-shot of CiteSeer when a user searches for a book by S. Russell and P. Norvig ("*Artificial Intelligence: A Modern Approach*"). Note that CiteSeer currently keeps 23 citations (with different formats) of the same book, mistakenly thinking they are all different. However, all 23 citations in fact refer to the identical book published by the same authors, and thus should have been consolidated in the digital library.



russell and norvig - ResearchIndex citation query - Microsoft Internet Explorer provided by AT&T Labs - Resea...
File Edit View Favorites Tools Help
Address
Go
Links
CiteSeer
Find:
Documents
Citations
Internet

http://citeseer.ist.psu.edu/cs?q=Russell+and+Norvig&submit=Search+Citations&cs=1

Find: Russell and Norvig

Searching for **russell and norvig**.
Restrict to: Author Title Order by: Expected citations Date Hits: 100 Try: Amazon B&N Google (CiteSeer) Google (Web) CSB DBLP
764 citations found. Retrieving citations...

Context Doc ***617*** *(10):* **Russell** S, **Norvig** P (1995) *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence. Englewood Cliffs, New Jersey

Context Doc ***64*** *(0):* S. **Russell** and P. **Norvig**. *Artificial Intelligence: A modern approach.* Prentice Hall, Upper Saddle River, New Jersey, 1995.

Context Doc ***18*** *(0):* Stuart **Russell** and Peter **Norvig**. *Artificial Intelligence: A Modern Approach.* Prentice Hall, Englewood Cliffs, New Jersey, 1995.

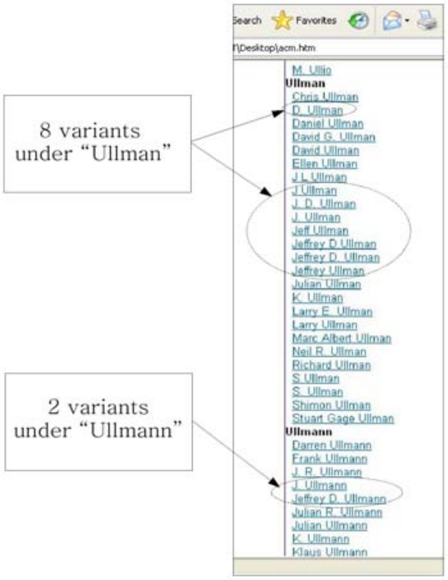
Context Doc ***8*** *(0):* **Russell**, S., and **Norvig**, P. 1995. *Artificial Intelligence*, A Modern Approach. Prentice Hall.

Context Doc ***7*** *(0):* S. **Russell** and P. **Norvig**. *Arti cial Intelligence: A Modern Approach.* Prentice Hall, Inc., London, UK, 1995. [Although somewhat older, this book remains on of the seminal works on Arti cial Intelligence.]

Context Doc ***4*** *(0):* S. **Russell** and P. **Norvig**. *Introduction to Artificial Intelligence.* Prentice Hall, 1995.

Context Doc ***3*** *(0):* Stuart **Russell** and Peter **Norvig**. *Arti cial Intelligence: A Modern Approach.* Prentice Hall, London, UK, 1995.

Figure 1. Screen-shot of citation search for "Russell and Norvig" in CiteSeer. Note the result includes 23 redundant citations all referring to the same book.



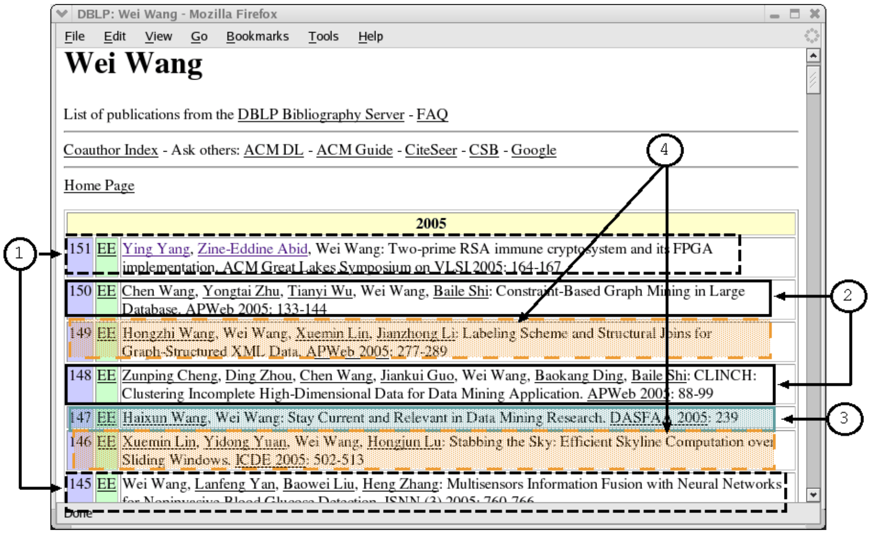
M. Ullio
Ullman
Chris Ullman
D. Ullman
Daniel Ullman
David G. Ullman
David Ullman
Ellen Ullman
J L Ullman
J Ullman
J. D. Ullman
J. Ullman
Jeff Ullman
Jeffrey D Ullman
Jeffrey D. Ullman
Jeffrey Ullman
Julian Ullman
K. Ullman
Larry E. Ullman
Larry Ullman
Marc Albert Ullman
Neil R. Ullman
Richard Ullman
S Ullman
S. Ullman
Shimon Ullman
Stuart Gage Ullman
Ullmann
Darren Ullmann
Frank Ullmann
J. R. Ullmann
J. Ullmann
Jeffrey D. Ullmann
Julian R. Ullmann
Julian Ullmann
K. Ullmann
Klaus Ullmann

Figure 2. Screen-shot of author index for "Ull*" in the ACM Portal. Note that the citations of "Jeffrey D. Ullman" appear as eight variants under "Ullman" and two variants under "Ullmann."

The second problematic example is drawn from the ACM Portal[1], which contains the name list of authors who have ever published an article in the ACM DL. As shown in Figure 2, however, the name of the same author, "Jeffrey D. Ullman", appears as a variety of spellings (eight variants under "Ullman" and two variants under "Ullmann"). As a consequence, Ullman's citations are divided and mislabeled into 10 different duplicate author entries. Such errors often indirectly contribute to the CM problem. The third example is an inverse case of the second example. It is drawn from DBLP, a popular computer science DL, where users can browse a collection of articles grouped by author's full name (i.e., author's full name acts as a primary key). In particular, Figure 3 is a screen-shot of a collection of articles by "Wei Wang". However, there are at least four (possibly up to eleven) active computer scientists with the same name

[1] Since our first report around January of 2005, authors were told that the ACM portal team has undertaken a massive Author Name Normalization project to resolve the CM problem.

spelling of “Wei Wang.” Not surprisingly, their citations are all mixed here. Any bibliometric analysis using this data would be, needless to say, faulty.


DBLP: Wei Wang - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
1
2
3
4
Done

Wei Wang

List of publications from the DBLP Bibliography Server - FAQ

Coauthor Index - Ask others: ACM DL - ACM Guide - CiteSeer - CSB - Google

Home Page

		2005
151	EE	Ying Yang, Zine-Eddine Abid, Wei Wang: Two-prime RSA immune cryptosystem and its FPGA implementation. ACM Great Lakes Symposium on VLSI 2005: 164-167
150	EE	Chen Wang, Yongtai Zhu, Tianyi Wu, Wei Wang, Baile Shi: Constraint-Based Graph Mining in Large Database. APWeb 2005: 133-144
149	EE	Hongzhi Wang, Wei Wang, Xuemin Lin, Jianzhong Li: Labeling Scheme and Structural Joins for Graph-Structured XML Data. APWeb 2005: 277-289
148	EE	Zunping Cheng, Ding Zhou, Chen Wang, Jiankui Guo, Wei Wang, Baokang Ding, Baile Shi: CLINCH: Clustering Incomplete High-Dimensional Data for Data Mining Application. APWeb 2005: 88-99
147	EE	Haixun Wang, Wei Wang: Stay Current and Relevant in Data Mining Research. DASFAA 2005: 239
146	EE	Xuemin Lin, Yidong Yuan, Wei Wang, Hongjun Lu: Stabbing the Sky: Efficient Skyline Computation over Sliding Windows. ICDE 2005: 502-513
145	EE	Wei Wang, Lanfeng Yan, Baowei Liu, Heng Zhang: Multisensors Information Fusion with Neural Networks for Noninvasive Blood Glucose Detection. ISNN (3) 2005: 760-766

Figure 3. Screen-shot of a collection of citations under author “Wei Wang” in DBLP. Note that there are at least four distinct computer scientists with the same name “Wei Wang.”

In general, since different users use different citation formats, DLs may contain a variety of citations referring to the same document. Automatically determining (and eliminating) duplicates in such DLs is not a trivial task, if not an impossible one. Nonetheless, the CM problem in DLs is an important problem to tackle. If we can precisely identify and match citations, then, we can enable precise bibliometric analyses. This would result in attributing credit to the correct authors, identifying all citations to a given article, and analyzing the impact of scholarly articles more accurately. Moreover, the CM problem arises not only in the context of DLs but also in many other related contexts. For example, online product catalog services (e.g., Google’s Froogle) face similar problems. They extract product descriptions, such as product name, price, manufacturer etc., from different Web pages, and consolidate the extracted information into lists such that all information related to the same product goes into the same list. This problem in a broad sense is a CM problem. Different Web pages use different conventions to represent the same information. Solutions addressing a CM problem should also be applicable to this problem.

Scenarios

We refer to a set of citations (or a DL) where the CM problem has not been solved as “*Dirty,*” otherwise, we refer to the DL as “*Clean*”. That is, in a clean DL, there is at most one citation that refers to a distinct article in the real world, while in a dirty DL, more than one citations referring to the same real-world document may exist --- for instance, the CiteSeer is currently a “dirty” DL as demonstrated in Figure 1. We contend that the CM problem so far has been considered in a rather narrow sense, and argue DLs of the new generation face new scenarios as follows:

1. **Creation**. When a new DL is created from a collection of digital literature, typically, the citation entries are extracted first from the literature, and then the extracted citation entries are cleaned and matched. Citation matching in this scenario is generally done in two steps in order to handle a large number of citations: (1) in the first step (known as *Blocking*), the citation entries are grouped into blocks based on some inexpensive distance metrics or by sorting on some key values (e.g., the title or the first author's last name), and then (2) in the second step, the algorithms visit each block separately and perform more elaborate matching within a block. Most of the previous work on the CM problem attempted to address this scenario. Formally,

   > Given a set of *dirty* citation entries, *S*, find all clusters *C* ($\subset$ *S*), such that all entries in *C* are close to each other with respect to some distance function.

2. **Insertion.** Once DLs are created, they need to be maintained up-to-date by adding new articles and their citations into the DLs over time. Unlike the *Creation* scenario, *Insertion* occurs almost on a daily basis throughout the lifetime of a DL. For instance, CiteSeer crawls the Web, searching for new literature, and indexes them as new documents are found. In this scenario, the set of newly found citations are inserted into an already established “clean” DL in operation (where all duplicates have already been consolidated). Although the citation matching problem in the *Insertion* scenario occurs frequently, this problem has largely

been ignored by both the citation matching and record linkage communities. We argue that efficient handling of the *Insertion* is important to maintain a large-scale DL efficiently. Formally,

> Given a set of *dirty* citation entries S_a (that are newly found) and a set of *clean* citation entries S_b (i.e., existing DL), for each entry a ($\in S_a$), find a closest entry b ($\in S_b$), such that $dist(a,b) \leq \theta$, where *dist* is some distance function, and θ is a threshold.

3. **Integration**. This scenario occurs in merging multiple DLs (e.g., merging CiteSeer and arXiv). The basic assumption here is that in each DL (established and in operation), citation entries are already cleaned and in most cases duplicates are eliminated (by possibly going through the previous *Creation* and *Insertion* scenarios). Therefore, in this scenario, citation matching mainly concerns the problem of linking citation entries across the DLs that are referring to the same object. Like the *Insertion* scenario, to the best of our knowledge, there has been little citation matching work done in this context. Formally,

   > Given two sets of *clean* citation entries, S_a and S_b, find a one-to-one mapping between entries, a ($\in S_a$) and b ($\in S_b$), such that $dist(a,b) \leq \theta$.

4. **Interoperation**. In response to a query over a federated system of DLs, citation matching must be performed on the intermediate results obtained from the individual DLs before they are returned to the end-user. Like the integration scenario listed above, the citations, referring to the same real-world article but presumably obtained from the different DLs and potentially having different formats, must be matched. Again, like the integration scenario, we assume that the DLs themselves are clean and the duplicates have been eliminated, and yet, duplicates in the intermediate results from different DLs need to be removed. Formally,

   > Let S_a and S_b be the sets of clean citation entries in the results returned from two different DLs in response to a federated search. Find a one-to-one mapping between entries, a ($\in S_a$) and b ($\in S_b$), such that $dist(a,b) \leq \theta$.

   As seen by the similarities in the definitions of the interoperation and integration cases, the same citation matching algorithms can be used for both.

The characteristics of the three scenarios are summarized in Table 1.

Scenario	S_a	S_b	Characteristics
Creation	Dirty	-	-
Insertion	Dirty	Clean	$S_a \neq S_b$ and $\mid S_a \mid << \mid S_b \mid$
Integration & Interoperation	Clean	Clean	$S_a \neq S_b$

Table 1. Three scenarios of creating and maintaining DLs.

Challenges

Although the CM problem (and its general version, the RL problem) has been extensively studied in many disciplines including databases, statistics, digital libraries, and artificial intelligence, to name a few, we argue that existing techniques are insufficient to cope with the new challenges that DLs currently face. The challenges include:

- Existing CM solutions have mainly focused on the *Creation* scenario. However, as DLs proliferate rapidly, their usage patterns and working scenarios change as well. For instance, the federation of multiple DLs using Open Archive Initiative (OAI) is no longer a dream. Also the characteristics of each scenario are slightly different, and thus an efficient solution for one scenario does not necessarily work well for the other scenarios. Therefore, the ability to handle the *Insertion* and *Merge* scenarios is crucial in the new generation DLs.

  We witness a dramatic increase of both the number of DLs available on the Web and the volume of data maintained in DLs. For instance, there are about 356 known DLs developed through the NSF NSDL program as of 2004. Furthermore, some of the existing DLs have a large number of citations in it (in the order of tens of millions), as summarized in Table 2. However, most of the developed CM solutions so far have focused on a rather static collection of small to medium-sized DLs (in the range of 1,000-10,000 citations) [2][7][8][11]. According to current estimates, CiteSeer indexes ten million citation records [4]. Detecting and reconciling variants among ten million citations efficiently without compromising the accuracy (recall the problem illustrated in Figure 1), is not a trivial task at all. The accuracy of existing CM solutions leaves much room for improvement. Although several previous work has reported an impressive 80-95% accuracy in their experiments (e.g., [2][8][11]), we predict that their applicability is limited when they are applied to truly large-scale DLs. Note that a plain nested-loop based CM algorithm requires all pair-wise comparisons of citations – a quadratic time complexity. Since it is computationally expensive for a large data set, typical CM algorithms has a pre-processing stage called "blocking" to select smaller candidate set for further examination. Although it varies by the adopted blocking scheme, it is not uncommon to have thousands of citations in the candidate set to do further examination after blocking. Therefore, when such CM methods need to be applied to "very large" citation data "repeatedly," the performance issue is still important. In this age of supercomputers with over ten teraflops of processing power, this computation may seem achievable. However, note that these citations typically reside on disk. Though disk speeds have increased, quadratic computations over very large data sets are still not feasible. Besides, oftentimes, the DLs may not even be able to employ large supercomputers to perform these computations for financial reasons. Furthermore, because of the quality of service implications, the hosts of the DLs that are being merged may not want these computations run over the DLs for a long time. Therefore, developing novel solutions that can achieve the goals of scalability and accuracy remains a challenge.

- Despite recent efforts for standardizing citation formats (e.g., Open Citation Project), authors have used (and will continue to use) various non-standard formats. Due to the lack of enforcement mechanisms, these formats vary by personal tastes, journal policy, discipline, etc. For instance, citations in some engineering fields require at least the author names and the paper title, while ones in physics may not even require a paper title. Citations in the engineering and the physical sciences may use unique identifiers for citations, while those in the social sciences may not have identifiers. Similarly, a recommended citation format in one journal tends to be quite different from the citation format in another. To make matters worse, citation formats that are posted to the Web are even more diverse. Therefore, DLs whose citations are collected from the Web tend to suffer from more serious ambiguity. For instance, consider the following 6 "real" citations taken from the example shown in Figure 1. Although they all refer to the same book, and some minor problems like the variations due to different spacing or line breaks or hyphenation can be resolved using simple rules, the problems due to the different format of each citation is much difficult to resolve. These differences occur in the many aspects: (1) number of fields used, (2) order of fields, (3) field values, (4) typos or personal comments, (5) use of special characters like space or hyphen, or (6) use of XML, etc.

```
#1: Russell S, Norvig P (1995) Artificial Intelligence: A Modern Approach,
Prentice Hall Series in Artificial Intelligence. Englewood Cliffs, New
Jersey

#2: S. Russell and P. Norvig. Artificial Intelligence: A modern approach.
Prentice Hall,  Upper Saddle River, New Jersey, 1995.

#3: Stuart Russell and Peter Norvig. Artificial Intelligence: A Modern
Approach. Prentice Hall, Englewood Cliffs, New Jersey, 1995.

#4: S. Russell and P. Norvig. Arti cial Intelligence: A Modern Approach.
Prentice Hall, Inc., London, UK, 1995. [Although somewhat older, this book
remains on of the seminal works on AI]

#5: [RN95] Artificial Intelligence-aModern Approach by S. Russell and P.
Norvig. Prentice Hall International, Englewood Cliffs, NJ,USA,1995.

#6: <reference><author>S. Russell and P. Norvig</author><title>Artificial
Intelligence: A modern approach</title><publisher>Prentice
Hall</publisher><year>1995</year></reference>
```

Therefore, developing solutions that can handle a variety of formats using appropriate domain knowledge is a challenging task.

Digital Library	Domain	# of Citations (in Millions)	Automatically Constructed?
ISI/SCI	General Science	25	No
CAS	Chemistry	23	No
MEDLINE/PubMed	Life Science	12	No
CiteSeer	General Science, Engineering	10	Yes
arXiv e-Print	Physics, Mathematics	0.3	No
SPIRES HEP	High-energy Physics	0.5	No
DBLP	Computer Science	0.6	No
CSB	Computer Science	1.4	Yes
NetBib	Network	0.05	No

Table 2. Characteristics of a few well-known scientific publication DLs.

- While the record linkage industry continues to grow (estimated to be more than 300 companies in the sector as of 2004), there are few known citation matching systems (or even record linkage systems) available to the research community (e.g., CMU's SecondString , GNU EPrints, ParaTools). It is important to have a system developed and made available for easy access of the public.

Conclusion

Despite their importance and potential impact to the digital library community, we believe the CM problem to be seriously under-researched. Due to the unique properties that exist in the CM problem such as the large number of available fields, generic solutions developed for the RL problem do not necessarily work that well. Furthermore, the novel challenges that current DLs face cannot be easily handled by existing solutions. To advocate the importance of the problem, in this article, we presented a preliminary "re-thinking" on a myriad of new challenges that we felt important for contemporary DLs.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. "Modern Information Retrieval", Addison-Wesley, 1999, 020139829X.

[2] M. Bilenko, R. Mooney, W. W. Cohen, P. Ravikumar and S. Fienberg, "Adaptive Name-Matching in Information Integration", IEEE Intelligent Systems 18(5): 16-23, 2003.

[3] I. P. Fellegi and A. B. Sunter. "A Theory for Record Linkage", J. of the American Statistical Society, 64:1183-1210, 1969.

[4] C.L. Giles, K. Bollacker, and S. Lawrence, "CiteSeer: An Automatic Citation Indexing System", ACM Conf. on Digital Libraries (DL), pp 89-98, 1998.

[5] Y. Hong, B.-W. On, and D. Lee, "System Support for Name Authority Control Problem in Digital Libraries: OpenDBLP Approach", European Conf. on Digital Libraries (ECDL), pp. 134-144, Bath, UK, 2004.

[6] M. A. Jaro. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", J. of the American Statistical Association, 84(406), pp 414-420, Jun. 1989.

[7] L. Jin, C. Li, and S. Mehrotra, "Efficient Record Linkage in Large Data Sets", Int'l Conf. on Database Systems for Advanced Applications (DASFAA), Kyoto, Japan, pp 137-148, Mar. 2003.

[8] S. Lawrence, C. L. Giles and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing", IEEE Computer, 32(6):67-71, 1999.

[9] B.-W. On, D. Lee, J. Kang, and P. Mitra, "Comparative Study of Name Disambiguation Problem using a Scalable Blocking-based Framework," ACM/IEEE Joint Conf. on Digital Libraries (JCDL), Denver, USA, pp 344-353, 2005.

[10] N. Paskin. "DOI: a 2003 Progress Report". D-Lib Magazine, 9(6), Jun. 2003.

[11] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. "Identity Uncertainty and Citation Matching", In Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA 2003.

[12] W. E. Winkler. "The State of Record Linkage and Current Research Problems", Technical report, US Bureau of the Census, Apr. 1999.