URI Disambiguation in the Context of Linked Data

Afraz Jaffri
School of Electronics and Computer Science
University of Southampton
a.o.jaffri@ecs.soton.ac.uk

Hugh Glaser
School of Electronics and Computer Science
University of Southampton
hg@ecs.soton.ac.uk

Ian Millard
School of Electronics and Computer Science
University of Southampton
icm@ecs.soton.ac.uk

ABSTRACT
The Linked Data initiative has given rise to an increasing number of RDF datasets, many of which are freely accessible online. These resources often arise as a result of database exports; however sufficient consideration may not be given to the unseen implications caused when they are used in the wider context of the Semantic Web. This paper investigates two popular resources, DBLP and DBpedia, and discusses whether the issues regarding identity management and co-reference resolution have been suitably addressed. We find that a large percentage of authors in DBLP have been conflated, and that disambiguation pages have been incorrectly linked using *owl:sameAs* within DBpedia. Systems for dealing with these issues are presented, and directions are given for future research.

Categories and Subject Descriptors
H.3.5 [**Information Systems**]: Information Storage and Retrieval: Online Information Services – data sharing, web-based services.

General Terms
Management, Design, Reliability

Keywords
Linked Data, URI, Co-reference

1. INTRODUCTION
As the Linking Open Data project gathers pace, more and more repositories of knowledge are being added to the Linked Data Cloud, covering a wide range of topics. Many datasets stem from the focal point of the Linked Data Cloud, DBpedia [2]. Since DBpedia has harvested knowledge from Wikipedia, there is the potential to create links to any subject that is described in Wikipedia.

The datasets that have been interlinked so far have knowledge relating to people, places, books, songs and CYC [8] concepts as well as many others. Entities such as these are often prone to the problems of duplication and co-reference.

LDOW2008, April 22, 2008, Beijing, China.

Whilst extensive linking between datasets has been widely encouraged, there has been little analysis of the accuracy of the links or the datasets themselves.

Datasets are often converted from existing sources which can themselves be either incomplete or inaccurate. The linking process accentuates these inconsistencies and produces a snowball effect as more datasets are added. If the Semantic Web is to provide a meaningfully interconnected web of assertions and relations, there must also be some guarantee or measure of the correctness of the information.

One of the main areas in which errors occur, both in databases and in digital libraries which are the kinds of repositories that have been converted into RDF for use with linked data, is the problem of co-reference. Co-reference is the problem of ensuring that two different entities do not share the same name or identifier, and conversely identifying when two identifiers refer to the same entity. In the context of the Semantic Web we are therefore concerned with URIs.

This paper presents some analysis of datasets used to link data and raises the question of how to manage the identity and meaning of URIs in the Semantic Web. The next section describes some related work in the field of co-reference and author disambiguation, while Section 3 describes the problem of co-reference and where it occurs in DBLP and DBpedia. Section 4 goes on to describe possible solutions to the problem that are currently in deployment. Section 5 concludes and issues an invitation to help to provide an infrastructure where data can be confidently used on the Semantic Web.

2. RELATED WORK

2.1 Author Disambiguation
The issue of resolving the problem of co-reference occurs in many different disciplines. A brief overview of the problems and solutions that appear in Information Science and database design can be found in [16]. One of the main areas in which co-reference becomes a major problem is in author disambiguation. There are many authors who share the same name and distinguishing between them is a vital part of any digital library or citation system. Not only do authors share the same names but variation in the spelling of names can also lead to a single author having multiple identities. For example, the author 'Hugh Glaser' could be represented with his full name or by using 'H. Glaser', or 'Glaser, H.'

A wide variety of methods have been employed to try and solve the problem of author disambiguation. Some of these include record linkage [10] used in databases, citation matching [17],

name matching [5] and name equivalence identification [9]. These methods involve some form of string matching and word sense disambiguation.

Although these methods can help in identifying names with different spellings or written in different formats, the problem of disambiguating authors with exactly the same name remains a challenge. There have been recent attempts that use a different approach from the traditional string based systems. Using the Web as a means of author disambiguation has been highlighted as a possible solution to the problem. Since Web pages often contain information about people that are not included in citation references, automatic scripts can be made that check the results of search engine queries made on the names of authors [19]. Another web-based approach attempts to find the publication page of an author from his or her institution's website and match the publications contained in the page to citations in the repository [22]. The accuracy of such web based systems ranges from 73% to 84%. These systems also rely on there being sufficient information available on the Web about each author. This is not always the case, especially with older publications and publications not in the field of computer science.

Another method that has been put into practice is to use an unsupervised machine learning approach using k-way spectral clustering that disambiguates authors in citations [12]. This study focused on the DBLP dataset and chose the top ranked ambiguous names such as 'J. Lee', 'S. Lee', 'Y. Chen', 'C. Chen', 'J Anderson' and 'J Smith'. The unsupervised learning technique used co-author names, publication titles and publication venue titles for author disambiguation. This assumes that individuals will quite often author with the same people and publish to the same venues. The results of this experiment show that an average of around 65% of authors can be successfully disambiguated.

The purpose of mentioning the ongoing work in author disambiguation in a different domain is to highlight the importance of a problem that is only beginning to be appreciated on the Semantic Web. Section 3 will elaborate on this. The next section will look at how co-reference is being managed on the Semantic Web.

2.2 Disambiguation on the Semantic Web

There has been much discussion about identity and meaning on the Semantic Web from a theoretical point of view. Such discussions will continue as questions fundamental to the architecture of the Semantic Web are debated. Attention is now turning towards practical solutions of managing co-reference, or URI identity management. Since co-reference between datasets is essential for linked data to work properly, a perfect opportunity arises to test some of the methods and solutions that have been proposed.

The various methods that have been suggested for managing co-reference and identity on the Semantic Web range from ontology based [22, 11], object consolidation [15] to complete management systems [7, 16]. The above applications have been used with geographical data [21], wikis [11] and general Semantic Web data [7, 16].

There has been valuable work done on studying the reliability and stability of Wikipedia URIs [14] that are being used by DBpedia. This study suggests that the meaning of a URI found in Wikipedia stays stable approximately 93% of the time. While the results have been documented there has been little attempt to quantify how big a problem co-reference on the Semantic Web actually is. In the next section we will consider how well the meaning of URI's from Wikipedia have translated themselves to DBpedia. We will also present a study made on the DBLP bibliographic database which is available both as linked data and a database.

3. THE PROBLEM OF CO-REFERENCE

Co-reference on the Semantic Web can occur in two ways: Firstly, when a single URI identifies more than one resource and secondly when multiple URIs identify the same resource. Both situations occur frequently when studying linked data. For an example of the first situation, many URIs in the DBLP dataset are used for identifying a single author when, in fact, there are a number of people with the same name who are being incorrectly identified as being the same person. The second situation occurs much more frequently as different datasets use their own URIs to identify the same resource. People and places are entities which suffer from URI multiplicity. Spain, for example has at least four URIs:

http://dbpedia.org/resource/Spain
http://www4.wiwiss.fu-berlin.de/factbook/resource/Spain
http://sws.geonames.org/2510769
http://www4.wiwiss.fu-berlin.de/eurostat/resource/countries/Espa%C3%B1a

'Hugh Glaser' has at least eight URIs:

http://acm.rkbexplorer.com/rdf/resource-P112732
http://citeseer.rkbexplorer.com/rdf/resource-CSP109020
http://citeseer.rkbexplorer.com/rdf/resource-CSP109013
http://citeseer.rkbexplorer.com/rdf/resource-CSP109011
http://citeseer.rkbexplorer.com/rdf/resource-CSP109002
http://dblp.rkbexplorer.com/rdf/resource-27de9959
http://europa.eu/People/#person-0ff816fa
http://resist.ecs.soton.ac.uk/wiki/User:hugh_glaser
http://www.ecs.soton.ac.uk/info/#person-00021

This is to be expected and does not present a problem in itself. The problem occurs when these URIs are linked to other URIs via *owl:sameAs*. Since URI identity can often depend on the context in which it is used [6], there can be no guarantee that the two URIs are in fact the same entity. The next section supports this assertion by looking at the DBLP dataset and also the DBpedia dataset to reveal inconsistencies in the linking and naming of resources.

3.1 DBLP

The DBLP database reportedly contains over 900 000 articles from over 500 000 different authors in the field of computer science and related disciplines. The database can be seen as RDF by means of a D2R Server [4] and has been converted into linked data by adding *owl:sameAs* links to authors who are also in DBpedia. Whilst providing a comprehensive repository for scientific publications, there are a number of inconsistencies that appear in the data. This problem is not only found in DBLP but also in other digital repositories. Due to lack of resources there is often not enough time available to rigorously check the input for correctness or completeness. This has resulted in many authors having publications incorrectly attributed to them, with some having more titles under their name and some authors having less. This will have a major impact on the Semantic Web when such

repositories are used as data sources without any attempt to manage the inconsistencies or 'clean' the data.

To assess the quality of data stored in DBLP we looked at some of the most common names and tried to ascertain whether the name belonged to a single author. This was achieved by looking at the publications attributed to each name and performing a Web search on the publication to find out to which institution an author was affiliated. The remaining publications were then checked in the same way and authors who came from the same institutions were grouped together. Authors also frequently change institutions, to accommodate this when a name was found that belonged to a different institution, it was assumed to be different unless:

1. The co-authors of any publication were the same.

2. The publication venue was the same.

3. The area of research was similar.

The author's own publication page was also used if one could be found. This process allowed for a conservative estimate to be made of the number of different authors who appeared under the same name. Single author papers and papers where there was a difference of greater than four years between their publications were excluded as authors can change their field of research over a period of time.

Names were chosen in order to provide a worst case scenario for authors not having been disambiguated. The ten most common surnames in the UK along with a list of common forenames were used. A total of 49 names were investigated by selecting five forenames with the nine most common surnames, and four forenames with the remaining surname. The DBLP dataset that was used was from October 2006 which contains a total of 491 796 authors. Thus, the selected names were almost 0.01% of the total population.

The results showed that for 92% of names chosen there were at least two different authors whose publications had been incorrectly merged. The highest number of different authors was 15 for the name 'David Smith'. The mean number of authors for each name was 3.8 with a standard deviation of 2.6. The ten most ambiguous author names are shown in Table 1.

As well as several authors being considered as one, there are also a number of cases where an author has more than one name where initials are used instead of full names. For example, 'C.B. Jones', 'Cliff B. Jones' and 'Cliff Jones' are all the same author yet his publications appear under these three different names.

To estimate the number of names which include two separate authors in the entire population, a Laplace point estimated can be calculated using a 95% confidence interval using the Adjusted Wald Method [1]. Multiplying the total number of entries in DBLP by the Laplace point estimate (0.902) gives 443 600 names. This will not be a truly accurate estimation since common names were chosen and not random names.

Nevertheless, we can conclude that if a person has a common name. The probability of their publications being merged with other authors will be 90%. These results should provide concern to those working in the Semantic Web and especially those who deploy linked data.

When existing data sources are used for Semantic Web data integration it is important to consider the consistency and completeness of the data before assigning links to other datasets and also in the form of *owl:sameAs*.

All of the names in DBLP have their own URIs which is thought to identify one single author with that particular name. As these results show, in most situations that is not the case.

Table 1. List of names with most number of distinct authors

Name	No. Authors
David Smith	15
David Williams	10
David Jones	8
David Evans	7
Alan Williams	6
Matthew Jones	4
Andrew Taylor	4
Michael Taylor	4
Andrew Brown	4
Ben Smith	4

This identity problem is not just theoretical, but also has implications for the future when more applications will be built that reason with and use Semantic Web data. In particular, consider the attempt that is being made in the UK to allocate research funding and judge research excellence by citation impact [13]. One could naturally believe that a Semantic Web application could be made that amalgamates all bibliographic data from DBLP and other repositories and ranks people or institutions based on their publications. If the issue of co-reference is not taken into consideration then it is clear that not everyone will be fairly represented.

Now that the problem of co-reference has been highlighted in DBLP, we move on to looking at how well the problem is handled in DBpedia.

3.2 DBpedia

The huge amount of data that has been extracted from Wikipedia has led to a rapid increase in the number of URIs that can be used to identify people, places and things. At present DBpedia has identifiers for close to two million entities. This has enabled many other datasets to become linked with DBpedia entities through the use of *owl:sameAs* giving rise to the Web of Data.

Whilst providing a valuable resource for data providers and application developers, the conversion process has not taken into account the different needs that DBpedia has in comparison to Wikipedia. In particular, the issues of ambiguity and co-reference raised in this paper have not been addressed.

Wikipedia deals with the issue of co-reference by having special 'disambiguation' pages. These pages are created when there is more than one entry that has the same name but carries a different meaning. Disambiguation pages are mainly intended for humans searching on a particular topic who may need some help in locating the page that they are looking for. These same disambiguation pages have been carried over into DBpedia where

there is no real need for them. Instead of making entities unambiguous, as in Wikipedia, the DBpedia URIs actually introduce more ambiguity.

Consider a person or machine wanting to use a URI for Robert Williams, the American politician. Using the URI *http://dbpedia.org/resource/Robert_Williams* reveals that properties belonging to Sir Robert Williams of Dorset, Robbie Williams the singer and Robert Williams the actor have all been merged onto one page. This happens with a large number of pages that fall into the Wikipedia category 'Disambiguation'. DBpedia 2.0 provides a number of examples where URIs are not sufficiently disambiguated. One example is the URI *http://dbpedia.org/resource/Nancy_Wilson* if this URI refers to Nancy Wilson the singer then the *dbpedia:spouse* property is of Nancy Wilson the guitarist.

There are, of course, other URIs which have all the properties belonging to the correct person. The URI *http://dbpedia.org/resource/Nancy_Wilson_%27guitarist%28* will give the correct URI for the guitarist Nancy Wilson. This is simple for a human to work out, but machines will struggle. This is demonstrated by the fact that putting 'Robert Williams' or 'Nancy Wilson' into Sindice [20] puts the ambiguous URI at a higher rank than the 'real' URIs. Therefore the disambiguation URIs used in DBpedia only act as URI 'noise' and should probably be removed.

It is pleasing to note that DBpedia 3.0 has given much more attention to the issue of disambiguation. However, whilst a new '*disambiguates*' property has been created, rogue properties belonging to distinct URIs still appear in URIs referring to disambiguation pages. There are approximately 150 000 of these URIs which can be detected with relative ease. It is hoped that successive improvements to the method in which URIs are disambiguated will mean that the co-reference resolution of URIs can then be handled by external systems as described in Section 4.

A second problem arises due to the strong implications prescribed by the *owl:sameAs* property. By stating that one URI is *owl:sameAs* another, one is stating that the two references identify the same resource, and that each should share the properties of the other [3]. Looking at the *owl:sameAs* links in DBpedia one can see that URIs are made to be the same as several URIs with different meanings. For example, http://dbpedia.org/resource/Welsh is taken from a Wikipedia disambiguation page for the term 'Welsh', in DBpedia this URI is *owl:sameAs*:

<http://sw.cyc.com/2006/07/27/cyc/EthnicGroupOfWelsh>
<http://sw.cyc.com/2006/07/27/cyc/Welsh-TheWord>
<http://sw.cyc.com/2006/07/27/cyc/WelshLanguage>
<http://sw.cyc.com/2006/07/27/cyc/Welshing-Cheating>

None of these links are made from the pages that are actually identifying these concepts such as *http://dbpedia.org/resource/Welsh_language*. In another example the URI *http://dbpedia.org/resource/H.P._Lovecraft* is *owl:sameAs* the CYC URI identifying the author and the Zitgist URI identifying the music band. Clearly the two are not the same. When looking at the real URI for the music band in DBpedia there is no *owl:sameAs* link.

These issues demonstrate the necessity of having dedicated management systems In order to manage co-reference resolution on the Semantic Web. The next section looks at two such systems that are currently in production. Problems during the creation of these systems have shown that there is a significant problem that needs to be tackled.

4. POSSIBLE SOLUTIONS

There are two main initiatives that have been set up in order to confront the issue of co-reference on the Semantic Web. Our own ReSIST [18] project has gathered metadata from publications and institutions and exposed them as linked data. The Okkam project [7] is a relatively new project that will formally begin this year although the initial architecture has already been conceived.

4.1 Consistent Reference Services

The CRS sits in the Semantic Web as any other knowledge base or database would. Each data provider maintains one or more CRSs for their own knowledge. In the ReSIST project there are over 15 repositories each with their own CRS.

The CRS introduces the concept of a *bundle* to group together resources that have been deemed to refer to the same concept within a given context. Different bundles may be used to group together URIs of the same resource in different contexts. For example, there may be a bundle containing all of the URIs about a person in the context of institution 1; and another bundle containing all of the URIs about the same person in the context of institution 2. Each CRS can use different algorithms to identify equivalent resources. A full description of the service can be found in [16].

The system is being used on a live site at http://www.rkbexplorer.com. Extending this system for use with DBpedia and other sites would involve using the linking algorithms for each dataset and storing the links in a CRS. Each dataset would have one or more CRSs which would act as an authority for their data. An application may choose to give precedence to a CRS hosted from the same domain as the URI in question. Taking the *owl:sameAs* links out of the data ensures the knowledge is semantically correct without introducing a significant overhead. However, if *owl:sameAs* links wish to be made then the CRS can be used for this purpose.

4.2 Okkam

The Okkam project has been created to enable a 'Web of Entities' [7]. Whereas the CRS is a fully distributed system, the Okkam system is centralised. The main aims are to create a naming service for entities and a directory containing entity profiles under the single control of one authority.

The main service, OkkamCore, allows for the publishing, modifying and removing of entities and assertions of identity and a retrieval service based on a set of criteria. A prototype application has been made and will be sequentially improved and upgraded throughout the duration of the project. By holding identifiers for all types of entities the project hopes to avoid the proliferation of URIs that is currently occurring. For the purposes of linked data it is yet to be seen what the final system will

provide. The project will be monitored with interest as progression develops.

5. CONCLUSION

This paper has attempted to provide some motivation for finding solutions to the co-reference problem. With the linked data initiative in its early stages, it is important to think about the integrity of the data being provided before errors are found in the applications that attempt to use the data.

We would stress that DBLP and Wikipedia/DBpedia are valuable and hard-won facilities that deliver searchable resources very effectively to their many users. The problem that is arising is that in the context of the Semantic Web and Linked Data, different measures of quality pertain. It is the very Network Effect that the Linked Data community is seeking that causes the difference.

The issue has attracted significant theoretical debate, yet the only systems attempting to solve the problem are the two mentioned in Section 4. It would be in the interest of the whole Semantic Web community if this issue was carefully considered as a fundamental part of the architecture needed to make the Semantic Web gain widespread adoption.

6. ACKNOWLEDGEMENTS

This work is supported under the ReSIST Network of Excellence (NoE) which is sponsored by the Information Society Technology (IST) priority of the EU Sixth Framework programme (FP6) under contract number IST-4-026764-NOE.

7. REFERENCES

[1] Agresti, A. and Coull, B.A. Approximate is better than 'Exact' for Interval Estimation of Binomial Proportions. The American Statistician. 52 pp.119-126.1998

[2] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. DBpeddia: A Nucleus for a Web of Open Data. In Proceedings of the 6th International Semantic Web Conference (Busan, Korea 2007). Springer.

[3] Bechofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., Mcguiness, D.L., Schneider, P.F. and Stein, L.A.OWL Web Ontology Language Reference, Technical Report, W3C, http://www.w3.org/TR/owl-ref/

[4] Bizer, C. and Cyganiak, R. D2R Server – Publishing Relational Databases on the Web as SPARQL Endpoints. In Proceedings of the 15th International World Wide Web Conference.(Edinburgh, Scotland 2006).ACM

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

[6] Booth, D. URIs and the Myth of Resource Identity, Proceedings of the Workshop on Identity, Meaning and the Web (IMW06) at International World Wide Web Conference. (Edinburgh, Scotland. 2006) ACM

[7] Bouquet, P., Stoermer, H and Giacomuzzi, D. OKKAM: Enabling a Web of Entities. In Proceedings of the 16th International World Wide Web Conference (Banff, Canada) ACM.

[8] Cycorp Inc. http://www.cyc.com

[9] Feitelson, D.G. On Identifying Name Equivalences in Digital Libraries. Information Research, 9(4), p.192.2004

[10] Fellegi, I.P. and Sunter, A.B. A Theory for Record Linkage, Journal of the American Statistical Association, 64(328), pp.1183-1210 ,December 1969

[11] Gangemi, A., and Presutti, V. A Grounded Ontology for Identity and Reference of Web Resources. In Proceedings of the 16th International World Wide Web Conference (Banff, Canada) ACM.

[12] Han, H., Hongyuan, Z., and Giles, C.L. Name Disambiguation in Author Citations using a K-Way Spectral Clustering Method. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries.(Denver) ACM

[13] Harnad, S., Carr, L., Brody, T. and Oppenheim, C. Mandated online RAE CV's linked to univerestiy eprint archives:enhancing UK research impact and assessment. Ariadne http://www.ariadne.ac.uk/issue35/harnad/

[14] Hepp, M., Siorpaes, K. and Bachlechner, D. Harvesting Wiki Consensus Using Wikipedia Entries as Vocabulary for Knowledge Management. IEEE Internet Computing. 11(5) pp.54-65 Sep 2007.

[15] Hogan, A., Harth, A and Decker, S. A Grounded ontology for Identity and Reference of Web Resources. In Proceedings of the 16th International World Wide Web Conference (Banff, Canada) ACM.

[16] Jaffri, A., Glaser, H., and Millard, I. URI Identity Management for Semantic Web Data Integration and Linkage. In Proceedings of the Workshop on Scalable Semantic Web Systems (Vilamoura, Portugal 2007) Springer.

[17] McCallum, A., Niham, R., and Ungar, L.H. Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. (Boston, USA 2000). ACM Press.

[18] Resilience for Survivability in IST (ReSIST) Network of Excellence. http://resist-noe.eu

[19] Tan, Y.F., Kan, M.-Y. and Lee, D. Search Engine Driven Author Disambiguation, Proceedings 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp.314-315, ACM Press, New York.

[20] Tummarello, G., Delbru, R and Oren, E. Sindice.com: Weaving the Open Linked Data. In Proceedings of the 6th International Semantic Web Conference (Busan, Korea 2007) ACM

[21] Volz, R., Kleb, J., and Mueller, W. Towards Ontology-based Disambiguation of Geographical Identifiers. In Proceedings of the 16th International World Wide Web Conference (Banff, Canada) ACM.

[22] Yang, K., Jiang, J., Lee, H. and Ho, J. Extracting Citation Relationships from Web Documents for Author Disambiguation, Technical Report No.TR-IIS-06-017,Institute of Information Science, Taipei, Taiwan