Managing URI Synonymity to Enable Consistent Reference on the Semantic Web

Afraz Jaffri, Hugh Glaser, and Ian C. Millard

Dependable Systems and Software Engineering Group
School of Electronics and Computer Science
University of Southampton
{a.o.jaffri,hg,icm}@ecs.soton.ac.uk

Abstract. The Web of Data is growing at an ever increasing rate, with RDF datasets being produced in the order of billions of triples. The effect of this increase has meant that many entities for which knowledge is being published have developed a number of URI synonyms. Managing URI synonymity plays an important part in establishing a solid foundation for data inter-linkage. This paper sets out an architecture for managing URI equivalences on the Web of Data by using Consistent Reference Services. A Use Case is presented to highlight the importance of managing identity and several advantages and disadvantages of using the CRS over other coreference resolution methods are discussed.

Keywords: URI, Identity, Coreference, Linked Data

1 Introduction

The issue of identity has become a central area of Semantic Web research. Whilst existing in theory for a number of years, practical solutions are now required to solve the URI Identity Crisis [1]. There are two fundamental issues associated with URIs that are at the heart of Semantic Web architecture. Firstly, how can a URI be associated with the entity that it is intending to denote? Secondly, how to manage coreference and disambiguation between URIs that are deemed to denote the same entity?

The first issue has been dealt with extensively in past Identity on the Web Workshops [2]. Enabling URIs to deal with so called ónon-informationô resources and the http-range14 resolution [3] has led to the production of the first tutorial on how to produce linked open data [4]. Whilst some dispute still remains about the effectiveness of using 303 redirects to handle non-information resource URIs, data conforming to this practice has begun to appear in large quantities. This paper will not focus on theoretical debate about the worthiness of http-range14, but will instead focus on the second issue of finding a practical solution to manage the URI synonymity problems that arise when large knowledge repositories are interlinked on the Web.

The Linking Open Data initiative has led to an explosion in the number of URIs used to identify different entities, which has also provided new impetus into finding a solution for managing URI coreference. The increase in the number of information sources being exposed as RDF has also led to an increase in the number of URIs used to identify different entities. It is often the case that data in different repositories will hold information regarding the same entities. This multiplicity of URIs leads to the problem of *coreference*, where different URIs are used to describe the same entity. On an open Semantic Web this presents a problem when there is a need to link together knowledge from disparate information providers. The present approach, used by many in the Linking Open Data community, is to use various equivalence mining techniques in order to assert *owl:sameAs* relations between entities that are considered to be the same [5]. DBpedia has, for example, made an assertion that: *<http://dbpedia.org/resource/Berlin><owl:sameAs><http://sws.geonames.org/2950 159/>*.

The semantics of *owl:sameAs* mean that all the URIs linked with this predicate have the same identity [6], this means that the subject and object must be exactly the same resource with respect to all properties. The major disadvantage with this approach is that the two URIs become indistinguishable even though they may refer to different entities according to the context in which they are used. For example, consider the case where a person has a URI at one institution and then moves to another institution that provides another URI. If the person makes an *owl:sameAs* link between them then it will not be possible to differentiate between the person as they were at the first institution and the person as they are at the second institution. The knowledge about the person at each institution effectively becomes merged so, for example, the addresses would not be able to be separated.

We subscribe to the belief that the meaning of a URI may change according to the context in which it is used [7]. For example the URIs that refer to Spain given above could refer to óSpain the political entityô, or óSpain the geographic locationô, or óSpain the football teamô. Some people would be happy to use each URI interchangeably because they do not care about the precise definition, whereas others will want a URI that specifically matches their intended meaning. There is a requirement to have some form of a system that deals with URIs about the same resource that are not exactly identical. The semantics of *owl:sameAs* are too strong and other alternatives like rdfs:seeAlso do not fit the intended purpose. Such a requirement is vital if data is to be cleanly linked together between multiple sources in a consistent fashion.

This paper presents a solution for managing URI synonymity on the Semantic Web. Section 2 describes our vision of the Semantic Web within a Consistent Reference Service infrastructure. Section 3 presents two real-world scenarios to highlight the importance of identity management. Section 4 examines related work in the area and gives a critique of other solutions to the problem and Section 5 concludes with directions for future work and discussion.

2 CRS Architecture on the Semantic Web

The Consistent Reference Service has been described fully in [8]. The underlying philosophy of the CRS is to treat URIs as first-class entities and separate the equivalences of a URI into a separate knowledge base that will be aware of both intra-repository and inter-repository synonymity. Equivalent URIs are grouped into óbundlesô which are themselves given their own URI. When an application wishes to find an equivalent URI, the CRS can be queried to retrieve the corresponding bundle. In this section we will expand on the initial application of the CRS in our own Linked Data site to developing an infrastructure of multiple CRSes each attached to a different repository of Linked Data.

2.1 Coreference Bundles

A set of URI equivalences is grouped together in a bundle. An example bundle for óHugh Glaserô from the http://citeseer.rkbexplorer.com/ repository is given below in N3:

```
@prefix coref:
<http://www.resist.ecs.soton.ac.uk/ontology/coref#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://dblp.rkbexplorer.com/crs/bundle-1882749>
      a coref:Bundle     ;

coref:hasCanon
<http://dblp.rkbexplorer.com/id/people-
5c3b0c986bef5fa4e181c5830d56326b-
9118ee1bfc54e3cb07408669fc2f7c48> .

coref:duplicate
<http://dblp.rkbexplorer.com/id/people-
5c3b0c986bef5fa4e181c5830d56326b-
4c67591cb890d2a08f6ba6a9e2c03cd7>      .

<http://southampton.rkbexplorer.com/id/person-04860> .

<http://dblp.rkbexplorer.com/id/people-
5c3b0c986bef5fa4e181c5830d56326b-
9118ee1bfc54e3cb07408669fc2f7c48> .

coref:insertedOn ‘2008-02-12 14:45:39’.
```

This bundle highlights two types of equivalences: First, intra-repository equivalences of a URI, i.e. those equivalences that originate from the same dataset. Such equivalences are often ignored or overlooked which can lead to the problem of

URI disambiguation [9]. Secondly, inter-repository equivalences are shown, i.e. those equivalences that originate from a different dataset.

In the literature thus far we have refrained from describing CRSes that are attached to repositories apart from http://www.rkbexplorer.com and its sub-domains. This was because it is a requirement for the data provider themselves to construct a CRS from their own knowledge bases. However, in order to stimulate debate and demonstrate how CRSes can be used on the Semantic Web we will describe a prototype system that is using data from DBpedia and other linked data repositories.

2.2 Integrating CRSes with Linked Data

The CRS architecture recommends that each linked data repository should have at least one CRS. Multiple CRSes may be used to group together URI equivalences according to the context in which they are used. The CRS is simply another knowledge base that holds knowledge about URI synonyms contained within the repository. The data and CRS are linked through a simple predicate named *'hasCRS'*. The predicate as used on a URI for óHugh Glaserô is given below:

```
http://southampton.rkbexplorer.com/data/person-00021
resist:hasCRS
http://dblp.rkbexplorer.com/crs/person-00021 .
```

The URI that is the object of this statement is the bundle for the URI for óHugh Glaserô. In this example the CRS being used is for DBLP whilst the subject URI comes from the Southampton repository. This kind of linking makes it possible for any CRS that has a bundle for a given URI to be used. An additional benefit that arises from not having to use oneôs own CRS is that another CRS has a more complete set of URI synonyms or a CRS that is more trusted can be used for finding all equivalences of a URI.

Once a bundle for a URI has been found, the full equivalence class can be constructed by traversing the *coref:duplicate* URIs and ófollowing your noseô. To illustrate, we will go through an example of finding all the URI synonyms for http://dbpedia.org/resource/Portugal. From looking at the data we can see that there are 3 URIs that are *owl:sameAs* the DBpedia URI:

```
http://www4.wiwiss.fu-berlin.de/eurostat/resource/countries/Portugal
http://www4.wiwiss.fu-berlin.de/eurostat/resource/regions/Portugal
http://www4.wiwiss.fu-berlin.de/factbook/resource/Portugal
```

There are, in fact, more URIs for Portugal on the Web of Data. At present, no procedure exists for finding a complete set of synonyms for a given URI. We use URIs of example locations of CRSes with a possible set of URI synonyms in each. Qnames are used instead of full URIs for brevity. With a CRS mechanism the procedure would be as follows:

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geonames: <http://sws.geonames.org/>
PREFIX factbook: <http://www4.wiwiss.fu-berlin.de/factbook/resource/>
PREFIX eurostat: <http://www4.wiwiss.fu-berlin.de/eurostat/resource/>
```

1. The URI dbpedia:Portugal is dereferenced and the *coref:hasCRS* predicate is followed to http://dbpedia.org/crs/Portugal or any other external CRS.

2. The CRS gives RDF about the URI including *coref:duplicate* predicates which are:

```
geonames:2264397
factbook:Portugal.
```

3. The geonames:2264397 URI is dereferenced and the *coref:hasCRS* predicate is followed to http://sws.geonames.org/crs/2264397 or any other external CRS.

4. The CRS gives RDF about the URI including *coref:duplicate* predicates which are:

```
dbpedia:Portugal
geonames:Portugal
```

These URIs have already been found so no further following is needed.

5. The factbook:Portugal URI is derferenced and the *coref:hasCRS* predicate is followed to http://www4.wiwiss.fu-berlin.de/factbook/crs/Portugal or any other external CRS.

6. The CRS gives RDF about the URI including *coref:duplicate* predicates which are:

```
factbook:Portugal
dbpedia:Portugal
eurostat:countries/Portugal
```

The first 2 URIs have already been followed, so the third is taken.

7. The eurostat:countries/Portugal URI is dereferenced and the *coref:hasCRS* predicate is followed to
http://www4.wiwiss.fu-berlin.de/eurostat/crs/countries/Portugal

8. The CRS gives RDF about the URI including *coref:duplicate* predicates which are:

```
eurostat:regions/Portugal
dbpedia:Portugal
```

From the URIs that have been followed the full equivalence closure of http://dbpedia.org/resource/Portugal is:

```
<http://dbpedia.org/resource/Portugal>
<http://www4.wiwiss.fu-berlin.de/factbook/resource/Portugal>
<http://sws.geonames.org/2264397/>
<http://www4.wiwiss.fu-
berlin.de/eurostat/resource/countries/Portugal>
<http://www4.wiwiss.fu-
berlin.de/eurostat/resource/regions/Portugal>
```

The sequence of events is depicted in Figure 1.

There are several issues that arise when implementing the above methodology. Firstly, the difference between this approach and using *owl:sameAs* must be highlighted. As noted in the introduction the semantics of *owl:sameAs* are very strict and it is debatable whether the two Eurostat URIs should be *owl:sameAs*. The other consideration is of Semantic Web applications who must always load the data of each URI that is *owl:sameAs* the current URI. This limits performance and imposes unnecessary loading of data. The CRS architecture allows for following as many, or as few duplicate URIs as required with no significant barrier on performance. It is not our intention to remove *owl:sameAs* from linked data, rather we would definitely encourage its use in situations where the semantics of the relation are correct.

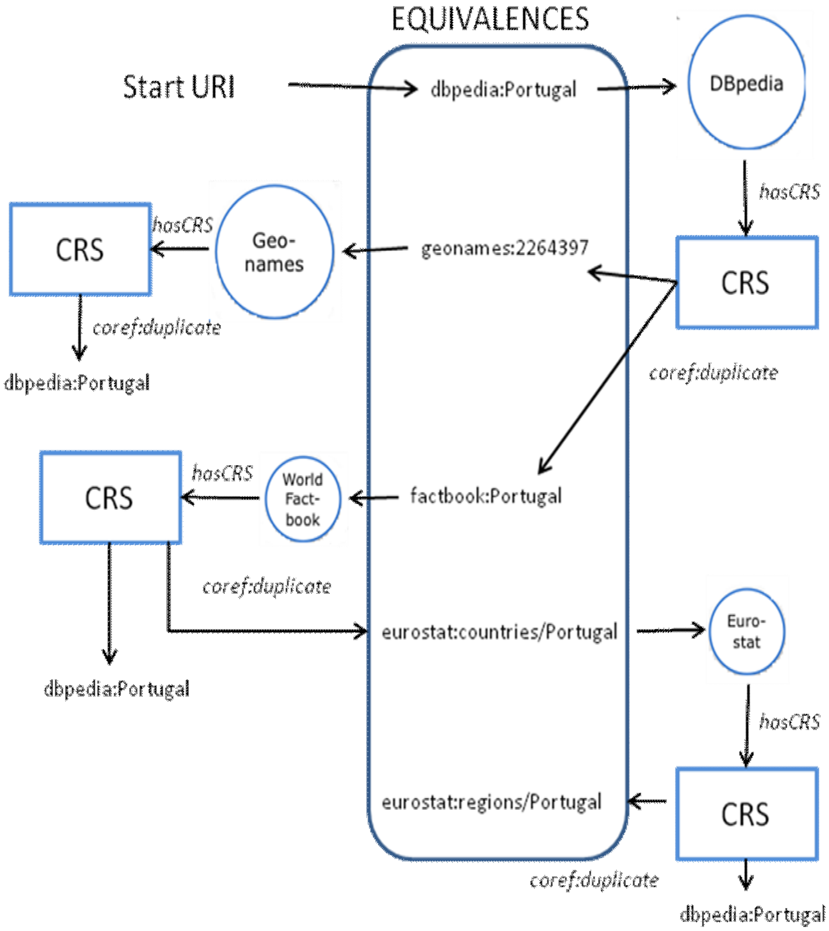

Fig.1. Finding the equivalence closure of a URI

The second issue that arises is how the URI synonyms are acquired. In our prototype application the CRSes created for each dataset were made with datasets of links that were already made available on the Web. It is simple a case of putting the same URIs that would be linked using *owl:sameAs* into a separate knowledge base. There is plenty of work needed in developing linking algorithms for detecting URI equivalence. The CRS system is envisaged to utilise these algorithms and provide links in such a way as to preserve URI equality without establishing the formal semantics of an *owl:sameAs* relation.

Another issue arises over which CRS contains which duplicate URIs. The example above uses URIs that are randomly distributed amongst the CRSes. It is entirely possible for one CRS to contain all equivalences of a URI, thus reducing the work needed to find the full equivalence set. However, the more common scenario is that data providers will not be aware of every single synonym for their URIs hence the need for multiple CRSes. As an example, we can look at the current DBpedia data for Portugal which does not contain all URI synonyms in the form of *owl:sameAs* links.

With the CRS architecture established, the next section will provide a scenario that is being used in a real life study of identity management in the UK. The example highlights the need for the Semantic Web to come up with a robust solution for managing URI coreference.

3 Identity Management Scenario

The Joint Information Systems Committee (JISC) is an organization that funds research into technological infrastructure in the UK. Recently, they have awarded a contract for a study on identity management for lifelong learning in UK higher and further education. The Invitation to Tender [10] focuses attention on how to handle the identity management lifecycle. Two scenarios are given to highlight the requirements that any identity management system must be able to handle. One of the scenarios is reproduced below:

ñGeorge is working as a recording engineer in the music industry, having achieved a Level 2 BTEC in music many years ago. He wants to improve his skills to gain more chance of promotion, so registers for a course leading to a BTEC National Award in Music Technology at a local FE college. This goes well, and he completes the course and gets a new job in another town. After another year or so, he wants to continue his studies, so he registers for a foundation degree at another college, which is validated by the local university. On successful completion of this, his employers pay for him to register with the university to complete an honours course. Ten years later, he moves to the country and fancies a change of direction, so registers for a higher education certificate in counselling.ò

One can see that each institution will give George his own, or indeed several URIs or other forms of identity designation. All of these ids will need to be tracked both inside an institution and also by other institutions that have had George as a student or employee. There are three main challenges that need to be addressed:

Provisioning of Identity ï It is highly likely that a number of different electronic identifiers will be issued to an individual that will contain data consistent with the knowledge of each issuing authority.

Maintenance of Identity ï Circumstances of individuals often change. Each identity credential that an individual has must be able to reflect changes in data over a period of time.

Deprovisioning of Identity ï When an individual leaves an institution, the knowledge about an individual still needs to be kept and made available. Other institutions may wish to examine the knowledge of a person after a long period of time.

Semantic Web applications should be at the forefront in providing solutions to problems such as these. However, the current framework for managing identity and coreference is lacking in methods for solving such issues. We only have to look at the number of identities of a person on the Web from sites such as Flickr, Facebook, and the blogosphere to realise that identity management is a core requirement for the success of the Semantic Web. Those who believe that these are minor issues that can be solved within the current climate should ask themselves if they know, or could find out every URI that denotes them on the Web? From inspecting oneûs FOAF file we can see that some people, like Hugh Glaser know of 22 URIs, where as Tim Berners-Lee only knows of (or discloses) 3 URIs.

There are currently two schools of thought when URIs and identity are talked about. The first says that there should be one canonical reference for every entity in the world [11]. The second, as is practiced within the Linking Open Data movement says that identical URIs should be linked through *owl:sameAs* and crawlers such as Sindice [12] will provide URI aggregation. With regards to the above scenario the problem with issuing a single URI that everyone must use for an individual is that there is no way of associating knowledge that an institution has about a person. The URI will be out of their domain. If knowledge is sent to a centralised repository then there is a serious risk of contradictions and inconsistencies arising in the data along with other problems with centralised repositories, such as data confidentiality.

The problem of ambiguity when using *owl:sameAs* to link all URIs was outlined in Section 1. If the identities are coming from different institutions then it will become impossible to know which knowledge has come from which institution. There is also the additional problem of keeping track of all URIs and if operations such as adding or deleting triples need to be performed, the performance cost may be excessive. The change in context of a URI will also distort knowledge in certain situations. If in the above example a person had married and changed their name, then the URIs for the person before and after marriage could not be linked with *owl:sameAs*, since any property describing ómarital statusô would have two different values.

Applying the CRS approach to the above scenario removes the restrictions with the other two kinds of approaches. Each data provider can have its own set of URIs to refer to their resources and a CRS to manage them. When a person moves from one institution to another, the old URI may be a *coref:duplicate* in the new institutionôs CRS. With this distributed approach knowledge can be created and maintained by each separate institution as is commonly the case today. The only addition is the introduction of a CRS, which involves minimal cost to the user as it is only another knowledge repository. When the identities of a person need to be amalgamated, the algorithm given in Section 2.2 can be run. In fact, in a fully CRS world additional features such as inter CRS negotiation and caching could further minimise the cost involved in finding all equivalences of a URI.

Identity management is becoming a hot topic in many different areas such as the Web, government, security and education. The Semantic Web will need to be able to address the concerns of all these different interests if it is to be taken seriously to provide the next generation of information management and integration applications. Having described the CRS architecture and provided a use case for motivation, we will now look at the related research in the area.

4 Related Research

The idea of separating links from data is not a new one. During the early stages of the Web there were competing systems that were trying to provide alternative approaches for open hypermedia systems [13]. One such project was Microcosm which featured a selection and action link following paradigm and a message passing framework that was compatible with Web architecture [14]. The feature that we wish to highlight here is the separation of content and link information into a linkbase. The linkbase was a link database that contained all information about link availability within a document. The linkbase stored specific links, contained within a source document, and generic links which could be made from any document. The purpose behind the linkbase was to counter the early navigational problems on the Web, such as only being able to access pages by following a set of specific links or knowing an address beforehand and typing it into a browser. Even though the CRS architecture is substantially different from the linkbase model, the underlying idea of separating links from data to facilitate ease of use, remains similar.

The most recent project to offer a system of URI identity management is the Okkam project [15]. The architecture used in this project aims to mimic the DNS architecture of the Web. Instead of a DNS server, an ENS (Entity Name System) server or servers are provided that aim to create an environment of unique URI provisioning and usage. The ENS acts as a global repository of URI identification which searches for entities, adds new entities and issues new identifiers. The goal of the project is to have data providers use Okkam issued URIs for entities that exist in the system.

There are several reservations that we have with such an infrastructure. Firstly the analogy with the DNS system appears incorrect. The DNS is a hierarchical system that is used for finding the *location* of a particular resource. The Semantic Web needs a system for finding the *identity* of a resource, and the two are quite difference tasks. A postal address will tell you that person A lives at the given house, but how do I find out who person A is?

Secondly the issuing of identifiers by Okkam or what is referred to as the Okkamisation of entities will only add to the proliferation of URIs on the Semantic Web. When someone mints a new URI for a resource it is because they have knowledge about the URI that they wish to disseminate. There can never be a way of accurately determining that the Okkam URI is the same entity to which a knowledge provider wishes to refer. Furthermore, if someone wishes to use a DBpedia URI because they believe it fits their purpose, then the requirement for using an Okkam URI becomes a hindrance. This also leads on to the question of how the system will determine that a URI is the same as one in their system. Equivalence determination is always prone to error and as already explained, URI similarity is subject to the context in which the URI is used.

The final and strongest criticism is that the ENS architecture is a centralised system which goes against the principles of Web architecture [16]. Furthermore, the creation and interaction between multiple ENS serves is not clear or explained in detail. Even though the ENS approach has many drawbacks, the project has given a lot of thought and consideration into the problem of URI coreference and should be applauded for giving the topic due importance in Semantic Web research.

An approach to identifying equivalent instances occurring across data sources has been used to perform object consolidation on the Semantic Web [17]. The algorithm looks for and uses inverse functional properties to detect instance equivalence and additional algorithms are used to describe how these equivalences are stored and ranked in memory. This work can be used to assist in the automated population of a CRS from crawling linked data URIs and pages. Since the major concern of any identity management application is the establishment of similarity metrics, this research provides one possible method to accomplish this task.

5 Conclusion

URI identity management needs to be at the heart of Semantic Web and Linked Data research. The enhancement that will be achieved from a consistent form of reference for all information and non-information resource will greatly increase the ease with which Semantic Web applications can be developed.

Our CRS service has been deployed on a linked data site and prototypes that use other linked data repositories have been constructed. The algorithm proposed in Section 2.2 for finding all equivalences of a URI is a simple and direct approach that does not need any new standards or protocols and conforms to current Linked Data best practice. The CRS is a fully decentralised and distributed approach to identity management that does not violate the principles of Web Architecture.

Future work will focus on developing the prototype to be used as a first point of call for finding synonyms of a URI. With increased adoption, other factors such as caching, trust rating and equivalence mining can then be investigated.

The identity management scenario presented in Section 3 provides sufficient motivation for the issue of URI coreference to be addressed and an agreed solution to be formally deployed. We hope that increased discussion and research will provide the infrastructure needed to create applications that utilise the Web of Data to its full potential.

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.

References

1. Halpin, H. Identity, Reference and Meaning on the Web, Proceedings of the Workshop on Identity, Meaning and the Web (IMW06) at WWW2006, Edinburgh, Scotland.
2. Bouquet, P., Stoermer, H., Tummarelo, G. & Halpin, H. Identity, identifiers, identifications - Entity-Centric Approaches to Information and Knowledge Management on the Web Workshop at WWW2007, Banff, Canada.
3 .Fielding, R., W3C Technical Architecture Group mailing list, June 18, 2005.[online] http://lists.w3.org/Archives/Public/www-tag/2005Jun/0039.
4. Bizer, C., Cyganiak, R. & Heath, T., How to Publish Linked Data on the Web, [online], http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/ [20 July 2007]
5. Equivalence Mining and Matching Frameworks [online] http://esw.w3.org/topic/TaskForces /CommunityProjects/LinkingOpenData/EquivalenceMining [1 July 2007]
6. Bechofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., Mcguiness, D.L., Schneider, P.F. & Stein, L.A.OWL Web Ontology Language Reference, Technical Report, W3C,[online] http://www.w3.org/TR/owl-ref/
7. Booth, D. URIs and the Myth of Resource Identity, Proceedings of the Workshop on Identity, Meaning and the Web (IMW06) at WWW2006, Edinburgh, Scotland.
8. Jaffri, A., Glaser, H., & 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.
9. Jaffri, A, Glaser, H. & Millard, I. URI Disambiguation in the Context of Linked Data. In Proceedings of the 1st Workshop on Linked Data on the Web at WWW2008, Beijing, China.
10.JISC ITT: Study on Identity Management for Lifelong Learning in UK Higher and Further Education, JISC, [online] http://www.jisc.ac.uk/media/documents/funding/2008/02/jiscittstudyonidentitymanagement.doc [10 March 2008]
11.Bouquet, P., Stoermer, H. & Cordioli D. An Enitity Name System for Linking Semantic Web Data. In Proceedings of the 1st Workshop on Linked Data on the Web at WWW2008, Beijing, China.
12.Tummarello, G., Delbru, R & Oren, E. Sindice.com: Weaving the Open Linked Data. In Proceedings of the 6th International Semantic Web Conference (Busan, Korea 2007) ACM
13.Davis, H.C., Hall, W., Heath, I., Hill, G.J. & Wilkins, R.J. Towards an Integrated Information Environment with Open Hypermedia Systems. In Proceedings of ECHTâ92, ACM Press, pp 181 - 190 (1992).
14.Carr, L., Hall, W., Davis, H. & Hollom, R. The Microcosm Link Service and its Application to the World Wide Web. In Proceedings of the 1st World Wide Web Conference, Geneva Switzerland, May 25-27, 1994, ACM Press.

15. Bouquet, P., Stoermer, H & Giacomuzzi, D. OKKAM: Enabling a Web of Entities. In Proceedings of the 16th International World Wide Web Conference (Banff, Canada) ACM.
16. Jacobs, I. & Walsh, Norman. Eds. Architecture of the World Wide Web, Volume One, W3C, [online] http://www.w3.org/TR/webarch/ [10 March 2008]
17. Hogan, A., Harth, A & Decker, S. Peforming Object Consolidation on the Semantic Web Data Graph. In Proceedings of the Workshop on Identity, Identifiers and Identification at WWW2007, Banff, Canada, 2007. ACM Press.