Improving Bibliographic Search through Dataset Enrichment Using Linked Data

Fattane Zarrinkalam
Dept. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
zarrinkalam.fattane@stu-mail.um.ac.ir

Mohsen Kahani
Dept. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
kahani@um.ac.ir

***Abstract*—Bibliographic search is an important activity every researcher is involved in. Currently, there are different digital libraries that provide a web-based system for bibliographic search. All these systems have the limitation that they are based on using a single local private dataset. This closed-world view over data reduces quality of search results since no single dataset can be regarded as being complete. In this paper an approach is introduced for enriching a local bibliographic dataset with Linked Data sources. The experimental evaluation shows that this Linked Data driven enrichment is effective in terms of improving bibliographic search results.**

Keywords- bibliographic search; Linked Data; enrichment;

I. INTRODUCTION

An important activity every researcher is involved in is searching web-based digital libraries for finding researchers and publications related to his field of interest. With the fast growing number of publications in scientific domains, this requirement has become more essential than before. There are different bibliographic search systems which can be used for this purpose. The richer is the bibliographic dataset of a system, the better results it provides for its users. Currently all such systems are based on using a single private dataset. From a practical point of view, none of these datasets is complete. Utilizing multiple datasets in a bibliographic search system provides a richer data layer, and it clearly improves quality of search service.

Linked Data provides a powerful method for publishing and accessing structured data on the web. It uses URI as naming mechanism, semantic data model for representation of data, and HTTP as the access and retrieval mechanism. An important task for Linked Data publishers is to create semantic links between their own data and related data from other sources. Such semantic links increase the possibility of gathering and integrating data from different Linked Data sources. Therefore, by reducing cost of retrieving and integrating data from different datasets, Linked Data enables data consumer applications to more easily benefit from multiple datasets.

In this paper an approach is presented for automatically enriching a local bibliographic dataset with related Linked Data sources, in order to improve bibliographic search. The proposed approach is evaluated through experiments and results prove that it is effective. The main contributions of this paper are:

1) It presents an algorithm for enriching a local bibliographic dataset with external Linked Data sources.
2) By quantitatively evaluating the proposed enrichment algorithm, it measures effect of this enrichment process on improving bibliographic search. Therefore it provides a practical groundwork for understanding how bibliographic search can benefit from Linked Data.

This paper is structured as follows. Section II briefly discusses related work. In section III the proposed enrichment approach is described. Evaluation of the proposed approach is presented in section IV, and finally section V concludes the paper.

II. RELATED WORK

There are some works that instead of relying on a local private dataset, access multiple public datasets on the web for gathering their required data. For example this approach is used by Shani et al. [1] to improve quality of a recommender system. Their results demonstrate that recommender systems can benefit by this move from a local private dataset to multiple public and open datasets. Although, from theoretical point of view, it seems promising to apply this idea on every data-centric application, but from practical perspective its cost and complexity cannot be underestimated.

Traditional web based systems are developed with high level of heterogeneity in terms of representation model, access and retrieval mechanisms, and this reduces chances of effectively applying the above idea. Linked Data, due to its benefits has been interesting for tackling with the problem of retrieving data from multiple datasets. For instance Heitmann and Hayes [3] discuss how open recommender systems can utilize Linked Data to gather data from different sources. Specifically, they focus on augmenting a closed collaborative music recommender system with Linked Data. Results show significant improvement of system performance in terms of precision and recall. Waitelonis and Sack [2] apply similar idea in the video search engine yovisto and enhance user experience by providing a semantically supported explorative search.

There are also some works that use Linked Data to add semantic annotations to their local dataset, in order to improve search and retrieval of its data. For example, Haslhofer et al. [4] use Linked Data for publishing

annotations of the Europeana dataset, which contains information about millions of digital resources in European institutions. They also enrich annotations by creating links to related resources from Linked Data cloud. In [5] YUMA Map Annotation Tool is used to augment map annotations with relevant information from Linked Data sources. Rusu et al. [6] discusses using three Linked Data sources WordNet, OpenCyc, and DBpedia for annotating textual data.

As literature study shows, the idea of using multiple Linked Data sources for improving data layer of an application is applied mostly for annotating a dataset with further information. Few works like [3] consider eliminating missing values. The point is that when semantic annotations are added to a dataset, original search procedures developed for that dataset must be modified to consider these annotations, in addition to the original dataset. But when a dataset is enriched by eliminating its missing values and adding new related data, the original search procedures do not require any modification. In this paper the second approach is applied, which has a lower cost.

To the best of our knowledge currently there is no work specifically focusing on enriching bibliographic datasets with Linked Data sources with the goal of improving bibliographic search. Further, although theoretically it is obvious that enriching a dataset improves search over that dataset, but in this paper the effect of using Linked Data for enrichment is quantitatively evaluated to provide a practical view over how much improvement is gained.

III. PROPOSED APPROACH

In this section, after providing required background, the proposed approach is introduced in detail.

A. *Background*

Currently all systems that enable bibliographic search are based on a closed-world view. It means they use a single local private dataset as the data layer. This reliance on a single dataset reduces quality in terms of completeness of the search results, since no single dataset can be claimed to be complete.

In some domains there is a single dataset that can be considered as standard or de facto standard source, and almost complete (e.g. IMDB for movie). But in some domains there is no such a single dataset. For example, in case of digital libraries, there are a number of datasets, each containing data about some publications. For instance some publications are indexed in IEEE, but not in DBLP, or ACM, and vice versa. Even if a publication exists in two datasets, they may contain different details about the same publication. Abstract of a paper might exist in a digital library, but not in another one. Therefore, it is not a reasonable assumption that a single dataset contains all the data that is required for a bibliographic search system.

Therefore it is interesting to have a bibliographic search facility that relies on multiple datasets. In order to have such a facility, it is possible to develop a system that uses a crawler to gather required data from different related online sources. Although this is possible but from practical point of view it is not interesting, because each online source might represent its data in a different format, and require special mechanisms for accessing, searching, and retrieving data. This heterogeneity and diversity is known in the literature as a real challenge in data integration.

By introduction of Linked Data, this integration problem is considerably relaxed. Tim Berners-Lee, the inventor of the web, presented a set of four rules for publishing data on the web [7] in such a way that it provides a global, web-scale, machin-processable data space, i.e. the web of data. These rules are known as Linked Data principles. The idea of Linked Data is to assign a HTTP URI to every resource which its data is to be published on the web, and use RDF representation for describing resources. Further, HTTP is used for dereferencing URIs and accessing information of resources. When URI of a resource is retrieved by a client, its description is sent to him. Using RDF links inside description, it is possible to move from one resource to another related one, and browse objects, similar to browsing traditional web pages [8].

The most important example of adoption of the Linked Data principles is the Linking Open Data (LOD) project. It is a community project which its goal is to publish existing open datasets using Linked Data principles, and moving toward creation of the web of data. Datasets published in LOD have formed what is known as the LOD cloud, an illustration of the current web of data. As shown in Fig. 1 this cloud contains a large number of datasets, containing data about different types of entities. Despite their diversity, these datasets can be viewed as a number of clusters, each cluster focusing on a special domain. For instance, there is a cluster for the bibliographic domain. It includes datasets which contain data about academic publications (e.g. title, abstract, keywords, publish date, venue).

Most Linked Data publishers provide a SPARQL endpoint for their dataset which allows the clients to query over data. The standard publishing and accessing mechanisms provided by Linked Data, has the great benefit that it eliminates the heterogeneity and integration problems of previous data publishing methods, and it enables easy utilization of different datasets published based on Linked Data principles.

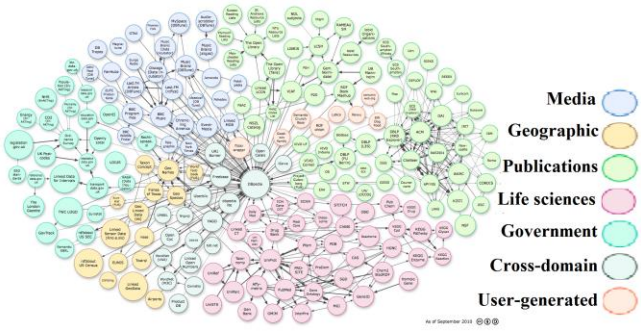

Figure 1. LOD cloud as in September 2010[1]

1 http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.html

There are two approaches for using multiple Linked Data sources in bibliographic search. Both approaches use a local dataset as the primary source, and also multiple linked data sources as the secondary sources. In the first approach, whenever a search is initiated by the user, in addition to the primary source, the secondary sources are also simultaneously searched to find additional data. In the second approach, before any search is performed by users, first the local dataset is enriched by related data from secondary sources, and after this offline process is finished, the system is available for users.

The former approach has the advantage that since secondary sources are searched on-the-fly, results are always up-to-date. Clearly its disadvantage is increased cost in terms of bandwidth and time consumed for searching secondary sources. It is not efficient to initiate search on external source, whenever local dataset is searched.

The later approach has a better utilization of bandwidth since the enrichment is performed offline and only once. Further, when the user is involved, no time is spent for searching external sources. Therefore it is more efficient from the end user's point of view. This approach has the disadvantage that the results might be out of date. Also it is possible to miss some results due to lack of access to new data which is added to the secondary sources after the enrichment process. Periodic execution of the enrichment process will reduce this possibility of missing results.

The motivation of this paper is to assess possibility and effectiveness of using Linked Data sources for enriching a local bibliographic dataset, in order to improve search results. It is possible that the local dataset has missing values, i.e. there is no data available for some attributes of some publications. The goal is to first reduce such missing values in the local dataset, and then to add related data to it.

B. *Dataset Enrichment*

The enrichment process uses K predefined external Linked Data sources S_i ($1 \le i \le K$) that publishes bibliographic data. For each external source S_i, an attribute $trust_i$ is defined with initial value of 0. This attribute is used as an indicator of the importance of the external source from the point of view of its consistency with the local dataset.

The local dataset contains data about N publications P_i ($1 \le i \le N$), where:

$P_i=(source_i, T_i, abs_i, keyList_i, authList_i, V_i, year_i, equList_i)$

$source_i$: the dataset in which the publication exists
T_i: title of P_i
abs_i: abstract of P_i
$keyList_i$: list of the keywords of P_i
$authList_i$: list of the authors of P_i
V_i: venue of P_i
$year_i$: publication year of P_i
$equList_i$: $\{P_j \mid P_j$ is a publication equivalent to P_i and $source_i \neq source_j\}$

In addition to publications, there is a list of M authors named *AuthList* where:

$$\text{AuthList}=\bigcup_{i=1}^{N} authList_i$$

$$M= |\text{AuthList}|$$

The enrichment process includes two steps. The algorithms of these steps are illustrated as pseudo-code in Fig. 2 and Fig. 3.

Here, these algorithms are briefly described.

In the first step, for each author in the local dataset, the external Linked Data sources are searched to find authors which are equivalent to that author, i.e. have the same name. This is done by method *findEquiAuthors*, which executes SPARQL queries on the endpoint of external sources to find equivalent authors. This process is known as "Co-reference resolution" in the web of data which is still a research issue [9].

```
foreach author A_i in AuthList {
  localPubs = findPublications(A_i, local)
  foreach external source S_j, (1≤j≤K) {
    temp = findEquiAuthors(A_i, S_j)
    foreach author A_m in temp {
      externalPubs = findPublications(A_m, S_j)
      foreach publication P_x in externalPubs{
        P_match = findMatch(P_x, localPubs)
        if P_match != -1
           add P_x to equList_match
        else
           add P_x to local dataset
      }
    }
  }
}
```

Figure 2. Algorithm of the first step of enrichment process

```
foreach publication P_i in the local dataset {
   foreach attribute a_i,m of P_i {
      init candidate list
      foreach P_j in equList_i {
      if value of a_i,m is missing
         if value of a_j,m is not missing
             add pair P_i to candidate list
      else
         if value of a_j,m == a_i,m
             increase trust of source_j by 1
      }
      If value of a_i,m is missing
         a_i,m = selectValue(candidate, m)
   }
}
```

Figure 3. Algorithm of the second step of enrichment process

Then for each of these equivalent authors, his publications are retrieved from the corresponding external source. The publications of the original author, from the local dataset, are compared to the publications of the equivalent external author. If for an external publication, a publication is matched in the local dataset, then the former is added to the equivalent list of the later. Otherwise, the unmatched external publication is added to the local dataset. This is performed in order to reach the second goal, i.e. adding new data to the local dataset. Simply stated, this means other publications of an author that do not exist in the local dataset are added to it.

In the second step, for each publication in the local dataset, each of its attributes is checked. If the value of that attribute is missing, then it is required to find a value for it. This is done by creating a list of candidate values from the equivalent list of the publication. Then the best candidate is selected by *selectValue* method as the value of the corresponding attribute. Otherwise, i.e. if the value of the attribute is not missing, its value is used to update trust of external sources. For each publication P_j from the equivalent list of the publication P_i, if the value of that attribute in P_j is the same as in P_i, then trust of the $source_j$ is increased by one. The idea is that the more similar is the equivalent publication to the local publication, the greater is trustworthiness of the corresponding external source. The *selectValue* method checks the candidate values. If they all are the same, then one of them is returned as the output. Otherwise, if different values exist in the candidate list, then a conflict resolution mechanism is used to produce the result. This mechanism is based on the trust value of external sources. The value from the candidate list which corresponds to the most trusted source is selected as the return value of the *selectValue* method.

IV. EVALUATION

To evaluate the proposed idea of enriching a local bibliographic dataset with Linked Data an experiment is performed which is discussed in this section.

To generate the local dataset, a specific crawler is developed to collect bibliographic data from CiteSeerX. This crawler communicates with the OAI service[2] of CiteSeerX which provides a harvesting mechanism for retrieving data from this dataset.

After collecting required data, a filtering process is performed to remove publications with high level of missing data, e.g. publications that their title, abstract and publication year are all missing. The idea is that such publications are non-promising in the sense that their missing values cannot be eliminated by enriching process. After this filtering step, the local dataset contains about 3700 publications.

Three Linked Data resources IEEE[3], DBLP[4], and ACM[5] are selected for enriching the local dataset. These are major sources of Linked Data which publish bibliographic information about publications in computer science and engineering. Further, they all use the same ontology, i.e. AKT ontology[6] for representation of their data. Therefore it is possible to create a SPARQL query and execute it on the endpoint of all these Linked Data sources, hence reducing complexity of accessing these sources for retrieving data. This is one of the benefits of Linked Data which, in comparison to the traditional web solutions, greatly reduces the burden of data integration from multiple sources.

The dataset enrichment algorithms proposed in Fig. 2 and Fig. 3 are implemented in Java programming language. The *findmatch* method uses Levenshtein distance algorithm for determining similarity of two titles. If two publications have titles which their Levenshtein similarity is greater than a specified threshold, then they are considered to be equal publications. In the experiments, this threshold is set to value 0.90.

It must be noted that based on manual evaluation of local dataset it was observed that the meaning of 'missing' needs to be specifically defined for each attribute of the publications. For instance, a string shorter than 30 characters is not a valid value for the *abstract* attribute, while it can be a valid value for the *keyword* attribute.

Since the main goal of the dataset enrichment process is to improve the bibliographic search, it is required to measure the effect of this process on search results. Therefore a set of scenarios was devised, each for searching publications based on different bibliographic attributes. In the first scenario, the goal is to find publications of a specific author. Next scenario is focused on finding publications that are published in a specific year. In the third scenario, publications with a specific subject are found. Finally, last scenario finds publications published in a specific venue and in a specific time interval. For each scenario, a set of queries is created and these queries are performed first on the initial dataset, and then on the enriched dataset. By comparing results of these queries on the two datasets, effect of the enrichment process can be evaluated.

To perform the first scenario, 100 authors are randomly selected from the initial dataset. For the second scenario, each year from 1990 to 2010 is selected as the specified year. In the third scenario, a set of 49 concepts was extracted from the subject areas of the 14th international CSI[7] computer conference (CSICC'09)[8]. These concepts were considered as important keywords searched by researchers in computer engineering field. For each concept, related publications are those that the concept is included in their title, abstract, keywords, or venue. In the last scenario, 50 venues are randomly chosen from the initial dataset and for each of these venues, publications that are published in that venue in the time interval of 1990-2010 are found.

Figures 4-7 illustrate number of search results for each scenario for both initial and enriched version of the local

2 http://citeseerx.ist.psu.edu/oai.html

3 http://ieee.rkbexplorer.com

4 http://dblp.rkbexplorer.com

5 http://acm.rkbexplorer.com

6 http://www.aktors.org/ontology

7 Computer society of Iran

8 http://csicc2009.aut.ac.ir/en/callforpaper.htm

dataset. For all four scenarios, number of results is increased in most cases.

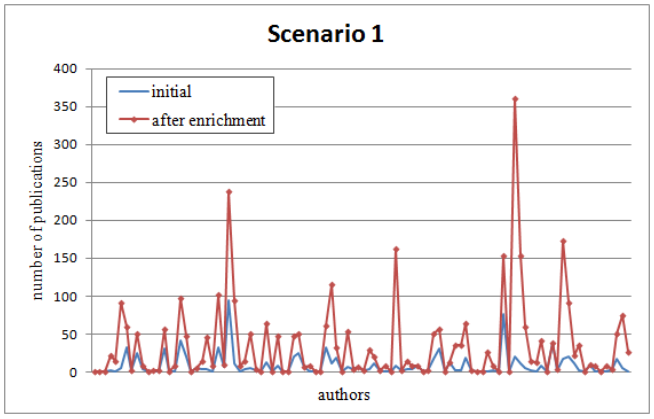

Figure 4. Number of search results for scenario 1

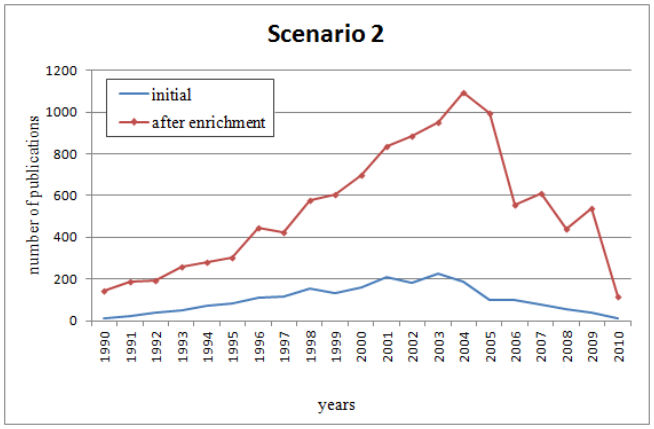

Figure 5. Number of search results for scenario 2

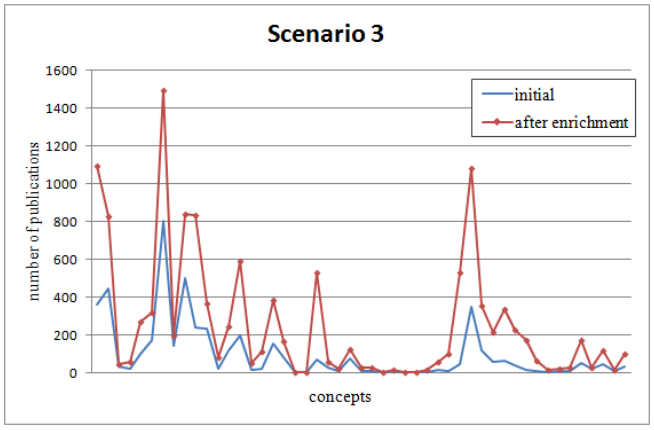

Figure 6. Number of search results for scenario 3

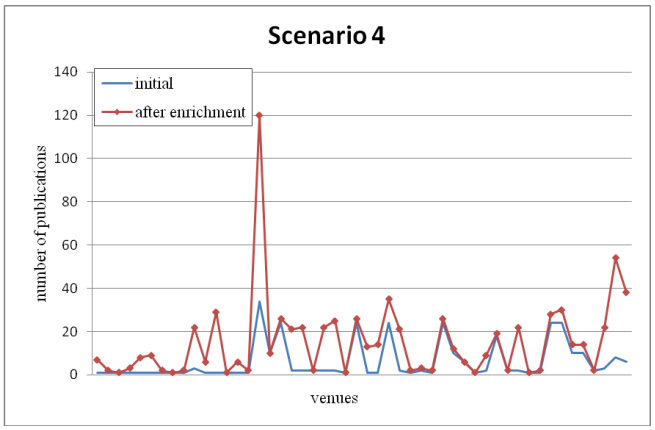

Figure 7. Number of search results for scenario 4

TABLE I. represents average, standard deviation, and coefficient of variation of the results for each scenario before and after enrichment. Coefficient of variation is a measure of dispersion and variability, and it is calculated as the ratio of standard deviation to the average. It has the benefits that unlike standard deviation, it is not context dependent (it enables comparing two sets of values independent of their different contexts), and it is not easily influenced by odd extreme values, e.g. outliers [10].

As this table shows, in all the scenarios, the average number of results is significantly increased. It must be noted that this increase, has not increased fluctuations in number of results. The reason is that coefficient of variation is decreased after enrichment. It means that the enrichment has increased number of results for different scenarios in an almost uniform way.

It is interesting to evaluate the enrichment process in terms of precision and recall, two common metrics in information retrieval literature. It is assumed that, except for the values that are considered as missing, the initial dataset and also external Linked Data sources do not contain incorrect data. Therefore it can be concluded that all the four scenarios have a precision of 100% both before and after enrichment, hence the enrichment has no effect on precision of search results.

There are two reasons for this claim. The first reason is that results of each scenario are retrieved by executing SQL queries on the local dataset, which deterministically determine the results. There is no fuzziness or probability involved. A tuple is included in the result set of a query if and only if it is exactly matched with the query pattern. Simply stated, each result returned by queries of a scenario is a correct result (of course based on the assumption that the queries are correct).

The second reason is that the enrichment process is based on an accurate algorithm. The local dataset is enriched by selecting some tuples from external sources and adding them to the local dataset. If the tuples of the external sources are correct, then the tuples added by the enrichment process are also correct.

TABLE I. ANALYSIS RESULTS OF THE EXPERIMENT

scenario	Average		Standard Deviation		Coefficient of variation	
	initial	*enriched*	*initial*	*enriched*	*initial*	*enriched*
# 1	9.06	34.80	14.70	55.02	1.62	1.58
# 2	97.10	253.86	157.22	339.12	1.62	1.34
# 3	102.43	530.33	65.31	294.60	0.64	0.56
# 4	6.12	15.36	8.67	19.42	1.42	1.26

The enrichment process generally improves recall of bibliographic search. The reason is that, as shown in figures 4-7, it usually increases number of search results which all of them are correct results.

It is not possible to exactly calculate the value of recall for each scenario because the number of all correct results is not known. The reason is that, it is possible that some missing values in the local dataset are not eliminated by the enrichment process. In other words, the local dataset still has some missing values after enrichment process.

Although the value of recall cannot be exactly measured, but it is possible to calculate its increase rate. If before and after enrichment there are respectively n and m results for a scenario ($m \geq n$), then recall is increased by a rate of $(m-n)/n$. Since for each scenario a number of cases are tested, (e.g. for the first scenario 100 different authors are used), here n and m are the average number of results among all cases of a scenario, respectively before and after enrichment.

Results of this calculation for the four scenarios is presented in TABLE II. which proves the effectiveness of the proposed approach.

V. CONCLUSION

In this paper a Linked Data driven enrichment approach is proposed for enriching a local bibliographic dataset, with the goal of improving bibliographic search. The proposed approach is implemented and experimentally evaluated by a scenario-based method. A local dataset is created which includes data collected from CiteSeerX. and then it is enriched by three Linked Data sources ACM, DBLP, and IEEE. Analysis of results demonstrates that the proposed enrichment algorithm is successful in improving search, in terms of increasing number of results. Therefore the proposed approach is sound and effective in improving bibliographic search.

TABLE II. RATE OF RECALL INCREASE

scenario	Rate of Recall Increase
# 1	2.84
# 2	4.18
# 3	1.61
# 4	1.51

REFERENCES

[1] G. Shani, M. Chickering, and C. Meek, “Mining recommendations from the web,” In ACM Conference on Recommender Systems, 35–42. ACM New York, NY, USA, 2008.

[2] J. Waitelonis and H. Sack. “Augmenting video search with linked open data,” In Proc. of Int. Conf. on Semantic Systems 2009 (i-Semantics 2009), 2009.

[3] B. Heitmann and C. Hayes, “Using Linked Data to build open, collaborative recommender systems,” In Proceedings of the AAAI Spring Symposium "Linked Data Meets Artificial Intelligence", 2010.

[4] B. Haslhofer, E.M. Roochi, M. Gay, and R. Simon, "Augmenting Europeana content with linked data resources", in Proc. I-SEMANTICS, 2010.

[5] R. Simon, B. Haslhofer, W. Robitza, and E. Momeni, “Semantically Augmented Annotations in Digitized Map Collections,” Proceedings of the Joint Conference on Digital Libraries (JCDL), Ottawa, Canada, June 13 - 17, 2011.

[6] D. Rusu, B. Fortuna, and D. Mladenic. “Automatically Annotating Text with Linked Open Data,” Linked Data on the Web Workshop, the World Wide Web Conference, Hyderabad, India, 2011.

[7] T. Berners-Lee, “Linked Data. Design Issues for the World Wide Web,” http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[8] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - The Story So Far,” Internatioal Journal on Semantic Web and Information Systems, 5 (3), pp. 1-22, 2009.

[9] H. Glaser, I. Millard, A. Jari, T. Lewy, and B. Dowling, “On coreference and the Semantic web,” In 7’h International Semantic Web Conference (ISWC 2008), Karlsruhe, Germany, 2008.

[10] A.W. Hendricks and K.W. Robey, "The Sampling Distribution of the Coefficient of Variation," In Annals of Mathematical Statistics, 7 (3), pp. 129–132, 1936.