Using Semantic Indexing to Improve Searching Performance in Web Archives

Arshad Khan*, David Martin[Ψ], and Thanassis Tiropanis[Ψ]
*National Centre for Research Methods (NCRM), University of Southampton
Southampton, UK
Email: a.khan@soton.ac.uk
[Ψ] Geography and Environment, University of Southampton, Southampton, UK
Email: D.J.Martin@soton.ac.uk
[Ψ] School of Electronics & Computer Sciences (ECS), University of Southampton, Southampton UK
Email: tt2@ecs.soton.ac.uk

***Abstract*—The sheer volume of electronic documents being published on the Web can be overwhelming for users if the searching aspect is not properly addressed. This problem is particularly acute inside archives and repositories containing large collections of web resources or, more precisely, web pages and other web objects. Using the existing search capabilities in web archives, results can be compromised because of the size of data, content heterogeneity and changes in scientific terminologies and meanings. During the course of this research, we will explore whether semantic web technologies, particularly ontology-based annotation and retrieval, could improve precision in search results in multi-disciplinary web archives.**

***Keywords*—Semantic indexing; archive searching; multi-disciplinary web archives; semantic searching; linked data**

I. INTRODUCTION

Information scientists have long been struggling to find a system that can help them organize disparate collections of web archives (or, in a general sense, web resource archives) so that users can have access to complete and coherent collections [1] in a much more meaningful way. Although both web archives and web repositories are sometimes used to refer to archives of web resources, the term web archive will be used through the remainder of this paper.

The prevalent access in web archives is based on the search over automatically extracted metadata from web documents [2] which have to be indexed for keyword searching. Providing broader access (unlike the current keyword search) to the collection of those web archives via an ontological framework structure could not only increase the utilization of these hard-earned resources but also make them more research-oriented, structured and flexible to cope with the changing needs of users [1], especially the research community.

The problem with conventional text based searching in web archives is that it only categorises the content in the archive on the basis of instance occurrence and query weighting with no attention paid to context, relevance, terminological coherence or relationship between web pages, nor does it relate the web pages of the repository to external sources on the web of data.

Web resources archives contain complex collections of research materials in online domains that can serve distinct communities, for example social scientists or historians who desire to search information based on contextual and provenance information [3]. We understand that semantic web technologies will be of tremendous help in identifying and integrating such heterogeneous documents inside web archives and enabling context and meaning-based search in them by exploiting existing vocabularies and domain-specific ontologies.

To further investigate the above issues, a thorough review of the most relevant research was carried out under the umbrella of web archiving, searching strategies in web archives and the application of next generation semantic web technologies. Particular attention has been given to linked data to see if users' searching experience could be improved by locating more precise and relevant information in web archives and repositories. In almost all cases, past research has focused on individual components of a typical web page in a particular domain such as semantically annotating and linking research datasets in biology, e.g., [4] multimedia objects in a newspaper archive, or [5] research publications (usually in proprietary formats) in scientific publication archives, to generate semantic metadata and generate improved search results.

This paper will describe the application and extension of keyword searching in web archives of multi-disciplinary research data by annotating web documents using more specialized and archival domain-specific ontologies and subsequently searching over the annotated metadata and instances of data i.e. Knowledge Base (KB). We will also take into account the existing and widely used classification systems e.g., thesaurus, typologies and taxonomies of social science to see if they can be synthesized into building an annotation ontology for creating annotation metadata.

 ISBN: 978-1-61208-248-6

II. CORE AREA OF RESEARCH

The development of a single on-line resource built by subject experts (e.g., a site presenting the methods and results from a research project in a social science discipline) can be time-consuming and expensive (in part, because they are not IT experts) and the full value of the resource only comes into play close to the point at which project funding ends. Those online resources not only provide a valuable personal development resource for researchers wishing to benefit from the training or data provided, but also provide an important repository of social, economic, historical and human knowledge. They are frequently created as the end-product of particular projects by academics and researchers associated with particular disciplines.

Following the completion of a project, the materials contained in some of these resources are considered for archival purposes so that they remain available despite the end of funding and dispersal of teams. This is an increasingly familiar situation as funders and institutions seek to develop repositories to increase the impact and availability of their work. Such archives or repositories of web resources enable users to search for information in the collection using basic keyword searching which proves to be ineffective simply because they retrieve web pages, datasets, and research papers merely on the basis of incidental mentions of a term in users' queries. Such a searching strategy misses the context, relationship, historical details and other attributes of a particular resource which would have enabled users to extend their exploration to other related records in the archive. One such web resources repository is the ReStore repository [6] which will be used as a test-bed for the experimentation and search performance evaluation described here.

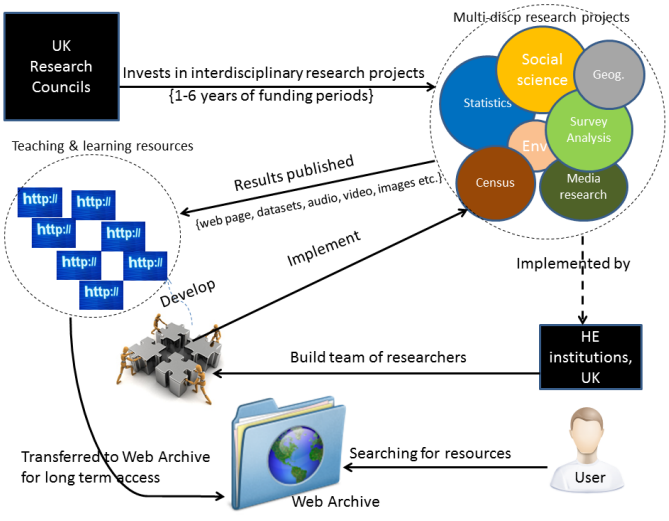

Figure 1: An overview of web resource, creation, development and its archival into a web resource archive, e.g., ReStore repository

Figure 1 shows a typical web resource development and archival process involving funding of a research project, development of a web resource by the project team based in higher education institutions, and the archival of the resource in a web archive such as ReStore repository.

Figure 1 also presents an opportunity to think about the problem which may arise at the time of searching the respective archive having different online materials created as part of various social science research projects.

III. THE PROBLEM

Some of the frequently observed problems stemming from the current searching methodologies include, but are not limited to:

- The typical keyword-search, usually based on keyword-matching or phrase-matching algorithms, hardly takes into account context, relevance and relationship between web documents, data, people, organizations, projects and other artefacts.
- Too many web pages make it almost impossible for users to go through and look for relevant information as a result of searching.
- The outputs of the same research project may be spread out over several repository/archive services, e.g., published research papers uploaded to an institutional repository, core research data uploaded to another data management service such as Economic & Social Data Service (ESDS) and videos/audios uploaded to video sharing services such as YouTube, Google videos etc.

Essentially, the content of valuable online resource collections becomes disintegrated and spread across different online archives. These are in most cases findable by standard search engines but with no classification hierarchy or associated relationships that links to their siblings (parts of the same web resource) or external, related objects that could be made available to users. An example would be research on a similar topic undertaken either in the past or currently. Similarly, a research funding organization may be funding another project in the same field or related workshops may have been offered as part of different projects. These types of relationship are well understood by the academic researcher but not readily amenable to discovery via existing searchable metadata.

IV. EXPERIMENTAL SETUP

The overall purpose of this research is to establish whether accuracy and relevance in search results are improved both at system level and in terms of user experience when searching for information in a repository of web resources having a meaningful semantic index.

We have outlined the current state of searching in website archives which largely employ only keyword-based searching techniques to retrieve information. We

 ISBN: 978-1-61208-248-6

are now able to formulate the following research questions which will act as a driving force for this particular research.

1. Can keyword-based searching be applied to semantically indexed metadata created by the semantic annotation process to enhance information retrieval in web resource archives?
2. Can domain-specific terms in an Ontology and their expressions in KB improve precision and recall in search results when added to the keywords search?
3. Does semantic indexing of ontology and KB along with inferencing cater to the heterogeneous types of data contained in web archives in terms of relationships, relevance?
4. Does the scale of searching over large multi-disciplinary web resources in web archives effect system performance in terms of the number of queries submitted at a particular point of time and the time it takes for the system to retrieve relevant information from multiple data sources?

It is important to remember that in all of the above, users of such archives have to be kept in the forefront of the research as they will be the ultimate beneficiaries of any new systems based on such research in the future.

V. PROPOSED ONTOLOGY-BASED SEARCHING SYSTEM

Figure 2 shows the building blocks of the experimentation environment which will enable us to conduct various experiments and evaluate search results performance in different scenarios.

The results have to be evaluated while keeping in view traditional RDBMS keyword-based searching techniques which are usually applied in many web resources archives including the ReStore repository.

Figure 2 shows basic components of the experimental setup including KIM (Knowledge & Information Management) platform with inherent support of GATE (General Architecture for Text Engineering), Open source full-text search engine Lucene, ReStore repository domain ontology fully mapped to KIM ontology and the resulting semantic metadata or RDF triples store with searching interface on top of it. It also follows that by incorporating the existing semantic web technologies and linked data concepts, we can enhance searching performance in web archives. A quantitative approach towards measuring performance has been adopted to evaluate accuracy of search results.

Sesame RDF repository will be an OWLIM semantic repository for managing RDF triples in addition to inferencing and SPARQL query evaluation during the course of various experimentation.

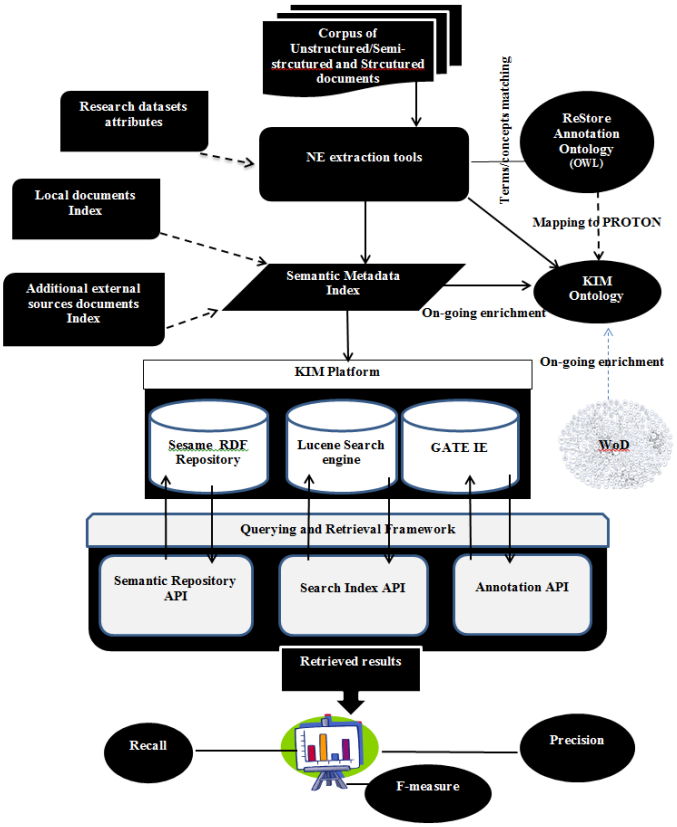

Figure 2: Proposed Semantic indexing based searching architecture built on KIM platform incorporating RDF(s) Sesame semantic repository and distributed (shared) ontological framework.

VI. THE CHALLENGE OF SCALE

Annotation of web documents containing different types of data on a large scale is certainly an issue, but the presence of a well-modelled ontology representing the domain of interest could address that issue to a large extent. Also the trend of submitting lengthy search queries [7] is making it increasingly difficult for keyword-based search engines to perform well as users are becoming more and more interested in the context precision, related information and the source and provenance of what is being searched.

The challenge is therefore of improving performance of the keyword-based search using semantics without losing search scalability [8] and users' interaction experience, to which they have become accustomed as the web has become an essential component of their lives. However it still remains to be seen whether the different environments in which web resources have originally been created (e.g., web served, harvested or manually collected), archived and indexed could influence the accuracy of results and performance of systems.

During our experimental setup we will analyse the above from different perspectives in order to support our research and solidify the basis of our findings. We have

 ISBN: 978-1-61208-248-6

already started the actual annotation work along with indexing and retrieval of information. We have already made progress in formulating the research framework and will proceed to undertake it in order to be able to evaluate performance in different scenarios, e.g., scale, heterogeneity of documents and retrieval of related information.

VII. ONTOLOGY ARCHITECTURE OF RESTORE REPOSITORY

In Figure 3, we have put together very basic terms and concepts (classes, properties) in a container with a view to mapping and extending them to other ontologies and storing the resulting annotation metadata (instances of classes or KB together in the form of integrated ontologies. The KB and mapped network of ontologies are assumed to be searched at a time to produce relevant and accurate results with a view to evolving it iteratively as new terms are added. GATE will process structured/semi-structured content in documents and OWLIM represents RDF triple store containing inference, querying engine for information retrieval.

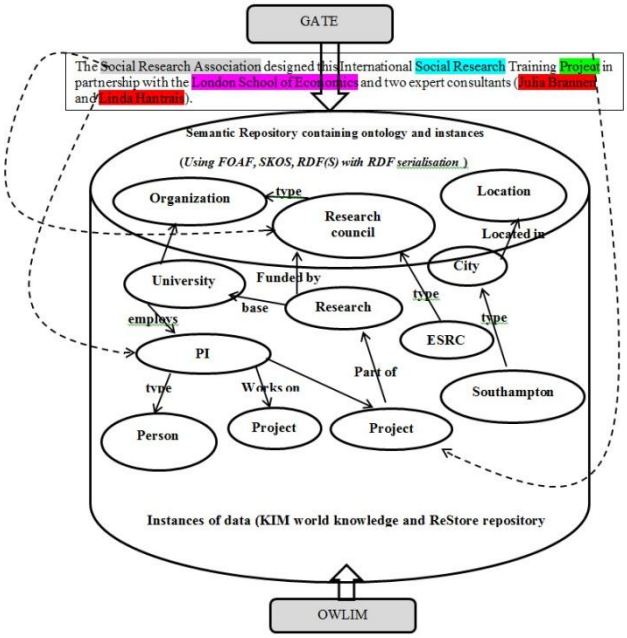

Figure 3: ReStore synthetic ontology architecture to be integrated into the KIM's bootstrap ontology KIMO to aid in (a) annotating domain-specific (Social Science Research Methods) resources in ReStore repository (b) searching relevant information in integrated ontologies.

VIII. CONCLUSION AND FUTURE WORK

The core purpose of conducting this research is to find out if searching over the content of archived web resources could produce more accurate, meaningful and trust-worthy results by using semantic web technologies and linked data techniques. In today's world of blogs, wikis, CMSs and social networking, the speed of publishing information on the web has greatly increased thereby affecting information consumption due to information overload. The problem is particularly acute in relation to web resources developed as part of specific research projects. Despite their value, in most cases these disappear from the effective searching spectrum due to non-availability of meaningful metadata (except basic bibliographic page-specific metadata), lack of linkage to contemporary research and lack of full exposure to the mainstream web. Such highly specialised web resources become almost un-searchable, or in some cases misrepresented in search results in web archives where searching is performed using keywords matching algorithms.

This paper describes a research agenda which is focused on the addition of extra semantic meaning to the existing content of a web resources archive (in our case the Restore repository so as to permit more effective searching which takes account of similar content within and beyond their parental domain in order to make them part of the structured and meaningful web of data.

ACKNOWLEDGMENT

The authors acknowledge the support of ESRC Award No. RES-576-25-0023.

REFERENCES

[1] P. Wu, A. Heok, and I. Tamsir, "Annotating the Web Archives–An Exploration of Web Archives Cataloging and Semantic Web," *Digital Libraries: Achievements, Challenges and Opportunities*, pp. 12-21, 2006.

[2] M. Costa, and M. Silva, "Towards information retrieval evaluation over web archives." pp. 37-38.

[3] P. H. J. Wu, A. K. H. Heok, and I. P. Tamsir, "Annotating Web archives—structure, provenance, and context through archival cataloguing," *New Review of Hypermedia and Multimedia*, vol. 13, no. 1, pp. 55-75, 2007/07/01, 2007.

[4] C. Berkley, S. Bowers, M. B. Jones, J. S. Madin, and M. Schildhauer, "Improving data discovery in metadata repositories through semantic search," *Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2009*. pp. 1152-1159.

[5] P. Castells, F. Perdrix, E. Pulido, M. Rico, R. Benjamins, J. Contreras, and J. Lorés, "Neptuno: Semantic Web Technologies for a Digital Newspaper Archive," Lecture Notes in Computer Science C. Bussler, J. Davies, D. Fensel *et al.*, eds., pp. 445-458: Springer Berlin / Heidelberg, 2004.

[6] "ReStore repository: A sustainable web resources repository," 14/01/2013, 2013; http://www.restore.ac.uk.

[7] M. A. Hearst, "'Natural' Search User Interfaces," *Communications of the ACM*, vol. 54, no. 11, pp. 60-6767, Nov., 2011.

[8] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan, "An ontology-based retrieval system using semantic indexing," *Information Systems*, 2011.

 ISBN: 978-1-61208-248-6