Linked Open Corpus Models, Leveraging the Semantic Web for Adaptive Hypermedia

Ian O'Keeffe, Alexander O'Connor, Philip Cass, Séamus Lawless and Vincent Wade
Centre for Next Generation Localisation
Knowledge and Data Engineering Group
School of Computer Science and Statistics
Trinity College Dublin, Ireland
{Ian.OKeeffe,Alex.OConnor,casspm,Seamus.Lawless,Vincent.Wade}@scss.tcd.ie

ABSTRACT
Despite the recent interest in extending Adaptive Hypermedia beyond the closed corpus domain and into the open corpus world of the web, many current approaches are limited by their reliance on closed metadata model repositories. The need to produce large quantities of high quality metadata is an expensive task which results in silos of high quality metadata. These silos are often underutilized due to the proprietary nature of the content described by the metadata and the perceived value of the metadata itself. Meanwhile, the Linked Open Data movement is promoting a pragmatic approach to exposing, sharing and connecting pieces of machine-readable data and knowledge on the WWW using an agreed set of best practices. In this paper we identify the potential issues that arise from building personalization systems based on Linked Open Data.

Categories and Subject Descriptors
H.5.4 [**Hypertext/Hypermedia**]: Architectures;

H.3.5 [**Online Information Services**]: Web-based services;

Keywords
Adaptive Hypermedia, Personalization, Linked Open Data

1. INTRODUCTION
In the traditional, closed model approach to Adaptive Hypermedia (AH), proprietary models that describe aspects of the system and the environment including the user, the domain and the presentation strategy all need to be defined in advance. These systems also need the content which will be used in the generation of presentations to be in a defined format and described using a specific metadata standard. This reliance upon bespoke, proprietary content and models restricts the ease of adoption, scalability and accessibility of such technologies. This paper describes a "speed of web"' experiment carried out to assess the concern that the use of Linked Open Data for adaptivity could negatively impact upon responsiveness and reliability.

The Open Model for AH described by this paper proposes utilizing the vast volume of data available on the WWW to address the issues described above. It is proposed to not only gather content from the web, but also to gather the data which describes that content. A number of web technologies have emerged in recent years which make this Open Model achievable. Linked Open Data (LOD) is a practical approach to exposing, sharing and connecting content via the WWW. The structure and links exposed in LOD repositories can be leveraged quickly. This structure potentially has real value in AH as it describes both the content and the domain, which are both basic requirements of AH systems.

HT'12, June 25-28, 2012, Milwaukee, Wisconsin, USA.
ACM 978-1-4503-1335-3/12/06…$10.00.

2. Evaluation
The inclusion of distributed information sources in any user-facing application can reduce usability by introducing delays, unreliability and distortion in the interface. In particular, because adaptive applications seek to perform rich transformations on the web experience of the user based on their attributes, it is vital that the adaptive system be responsive and reliable. These concerns arise directly in the use of Linked Open Data for adaptivity, because such systems depend on querying remote knowledge bases with differing infrastructure and with often-complex query expressions. In order to address this concern an experiment was carried out to assess, at a qualitative level, what kind of delay remote querying introduces.

The relatively high cost of querying linked open data remotely presents an important design consideration, particularly for knowledge-intensive applications such as Adaptive Hypermedia environments. The prototype used in this experiment demonstrates a system which uses one repository, the Linked Movie Database. However, the nature of the web dictates that different linked data endpoints for different content and platforms will have different performance profiles. In order to gain a qualitative perspective on the cost of remote linked data querying, two representative queries were chosen and executed on a number of repositories, chosen from across the LOD cloud. While these results are not statistically representative, they do provide an initial view of the "speed of the LOD web".

The queries executed were designed to be representative of the kinds of operation relevant to a LOD AH environment, specifically two tasks: retrieving a long list of entities, and retrieving the detail of particular entities. These queries are represented graphically in Figure 1. One of the key challenges for creating representative queries was to define queries of approximately equivalent semantics. In order to achieve this, the three linked data sources chosen were examined to locate somewhat equivalent entities and attributes of similar cardinality and relationship to the entities.

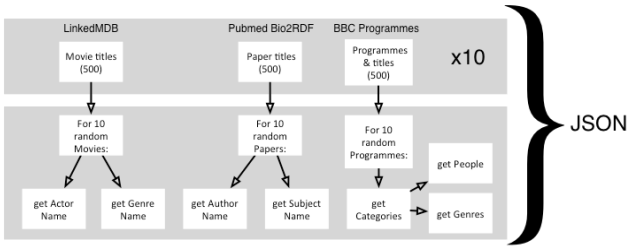

Figure 1: Schematic representation of queries used to measure response of different LOD repositories.

For the LinkedMDB data source, Movie entities were chosen as the basic object for retrieval. A query for 500 titles for entities of the Movie type were retrieved. This query was repeated ten times, and the aggregate time recorded.

From this list of retrieved titles, ten Movies were chosen at random, and the list of Actor names and Genre names for the Movie of a particular title were retrieved.
Similarly, for Pubmed Bio2RDF [1], JournalArticle entities were chosen for retrieval, with paper authors and subject keywords were selected. These are approximately similar attributes in terms of cardinality and semantics to the Movie entities.

The BBC Programme data source [2] differed slightly in the way it represented the categorizations. The basic entity under investigation was the Episode, but the graph differed from the other two data sources by having one category relationship, and the difference between Persons and Genres was decided by type inference. This also accounts for a high variance in the result timings because of a higher variance in the cardinality of the results. This is a good example of the challenges of attempting to uniformly assess different data sets: modeling differences can make equivalence only approximate. The query procedures on each Linked Data source were repeated three times, on separate days. The results were serialized as JSON. The architecture for this test harness was based on the SPARQL Endpoint Interface for Python [3] and used Python's built-in Timer library.

There are many inherent difficulties with effective performance testing, particularly with regard to http-based interfaces, where caching, proxies, network issues and other factors complicate the repetition of trials. Because of this, aggregate times for repetitions of individual times were recorded separately. This provides an informative, rather than objective guide to likely timing for queries. The results are shown in Figure 2 and Figure 3.

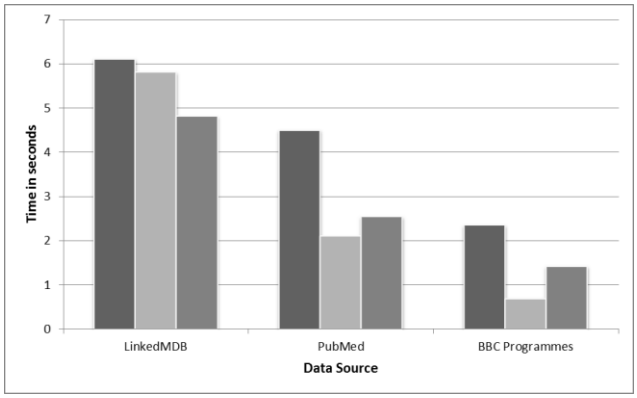

Figure 2 Time for sequences of genres for 10 random items

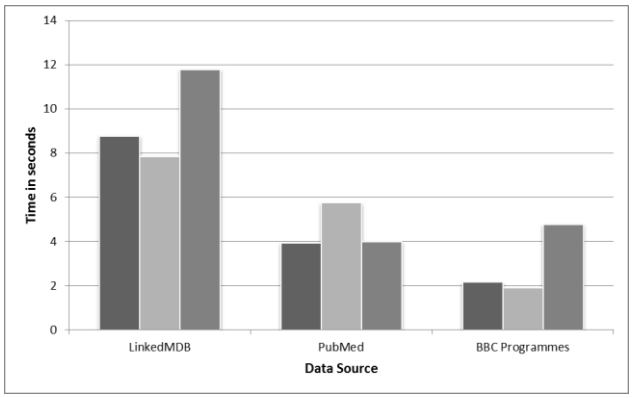

Figure 3 Time for 10 sequences of 500 title queries

These results show that the performance of remote semantic querying can vary depending on a complex series of known and unknown effects. This has implications for the design of adaptive systems, which point towards a need for caching strategies for complex or repetitive queries. Another factor observed was the limitation of connection rates by the end-points. One endpoint returned rate limits after 10 successive queries in a 12 second period. This points to another key non-functional design parameter: supporting rate limited access to end-points.

3. CONCLUSION AND FUTURE WORK

There are also some important practical considerations to be taken into account in implementing effective Open Model AH systems. The high variability of the response performance of LOD repositories motivates further research into repository and client performance, particularly through caching. It is important for future architectures for Open Model AH, and LOD in general, to be resistant to variable responses and which are able to avoid excessive load on LOD repositories. The personalized nature of Open Adaptive Hypermedia means that it is an interesting example of the general problem of deciding how to store portions of large Linked Data sets to improve client performance. The challenges of personalization are also applicable to other LOD use cases.

4. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin

5. REFERENCES

[1] Nolin, M.-A., Ansell, P., Belleau, F., Idehen, K., Rigault, P., Tourigny, N., Roe, P., Hogan, J.M. and Dumontier, M. 2008. Bio2RDF Network of Linked Data. *Semantic Web Challenge at International Semantic Web Conference (ISWC 2008)*. Karlsruhe, Germany

[2] BBC Backstage programmes API http://backstage.bbc.co.uk (accessed April 2012)

[3] SPARQL Endpoint interface to Python, http://sparql-wrapper.sourceforge.net (accessed April 2012)