Towards Entity-Centric Preservation for Web Archive Enrichment

Gerhard Gossen, Elena Demidova, Thomas Risse, Giang Binh Tran

L3S Research Center and Leibniz Universität Hannover, Germany
{gossen, demidova, risse, gtran}@L3S.de

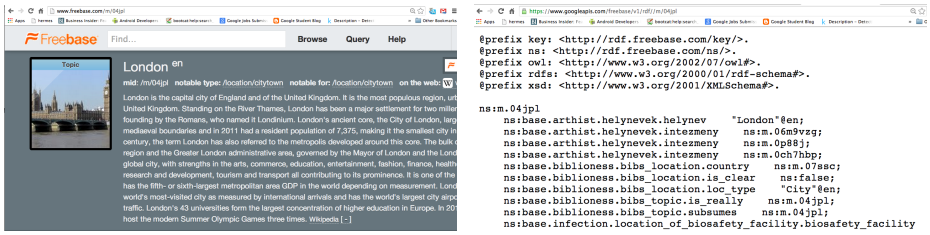
1 Introduction

Today Web content and long-term Web archives are becoming more interesting for researchers in humanities and social sciences [5]. In this context, Linked Open Data (LOD) [3] - a standardized method to publish and interlink structured semantic data - plays an increasingly important role in the area of digital libraries and archives. For example, linking entities in documents with their semantic descriptions in the LOD Cloud provides richer semantic descriptions of the document and enables better semantic-based access methods [1].

While the temporal dimension gains importance, it becomes necessary to look at LOD as a constantly evolving source of knowledge. LOD inherits the property of the Web of not having a documented history. Therefore, in the area of Web archiving it becomes more important to preserve relevant parts of the LOD Cloud along with crawled Web pages. This preservation process requires several steps, such as entity extraction from Web pages (e.g. using Named Entity Recognition (NER) techniques [2]) coupled with enrichment of extracted entities using metadata from LOD [1] as close as possible to the content collection time point and calls for Web archive enrichment approaches that collect entity context information from LOD.

To facilitate interpretation of the archived Web documents and LOD entities linked to these documents in the future, presentation of the LOD entities within Web archives should take into account content, provenance, quality, authenticity and context dimensions. In addition, prioritization methods to select the most relevant sources, as well as entities and properties for archiving are required. In the following we discuss the requirements in more detail. Then we present the resulting entity preservation process.

Content and Schema: In order to get a complete information provided by the object properties, traversal of the knowledge graph is required. However, traversal of large knowledge graphs is computationally expensive. Apart from that, while available properties are source dependent their usefulness with respect to the specific Web archive varies. Whereas some properties (e.g. entity types, or equivalence links) can be considered of crucial importance, others can be less important such that they can be excluded from the entity preservation process. Therefore, property weighting is required to guide an entity-centric crawling process.



(a) Web representation (b) RDF representation

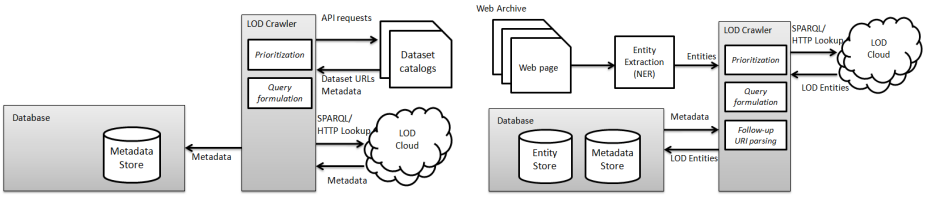
Fig. 1. "London" in Freebase

Provenance and Quality: Similar to Web document archiving, documentation of the crawling or entity extraction process as well as unique identification and verification methods for collected entities are required to ensure re-usability and citeability of the collected resources. In addition, for better interpretation of the entities stored within the Web archive, it is crucial to collect metadata describing the original data source for an entity. Optimally, these metadata should include data quality parameters such as methods used for dataset creation (e.g. automatic entity extraction, manual population, etc.), freshness (last update at the dataset and entity levels), data size, as well as completeness and consistency of data. Unfortunately, such metadata is rarely available in the LOD cloud. Therefore, archiving systems should provide functionality for statistical analysis of data sources to estimate their quality and reliability. To ensure correct access to the archived information available metadata about publisher, copyright and licenses of the sources needs to be preserved.

Authenticity and Context: Authenticity is the ability to see an entity in the same way as it was present on the Web at the crawl or extraction time. One of the main principles of LOD is the use of dereferenceable URIs that can be used to obtain a Web-based view of an entity (this view also may differ from the machine-readable representation). For example, Figure 1 presents a part of the Web and RDF representations of the entity "London" from the Freebase dataset of March, 4, 2014. To satisfy authenticity requirements, in addition to machine-readable entity representation, human-readable pages representing archived entities should be preserved. Such information can include visual resources such as photographs of people, maps snippets, snippets of Wikipedia articles, etc.

2 Entity Preservation Process

Whereas entity extraction from the archived Web documents performed by NER tools can deliver entity labels, types and eventually initial pointers (URIs) of the relevant entities in the reference LOD datasets, the collection of relevant entities should be extended beyond the LOD sources used by the extractors (such as DBpedia). Furthermore, the content of the entities needs to be collected and preserved. In these context important challenges are connected to *SPARQL endpoint*



(a) SPARQL endpoint discovery and metadata crawling

(b) Entity-centric crawling

Fig. 2. Entity preservation process

discovery and metadata crawling, *prioritization of the crawler* and *entity-centric crawling* to obtain the most relevant parts of the LOD graphs efficiently.

SPARQL endpoint discovery and metadata crawling: Existing dataset catalogs such as DataHub[1] or the LinkedUp catalogue[2] include endpoint URLs of selected datasets as well as selected statistics, mostly concerning the size of specific datasets; however, existing catalogs are highly incomplete. Metadata crawling includes several steps to obtain SPARQL endpoints URLs and generate metadata. Based on this metadata, prioritization of LOD sources and properties within these sources for preservation can be performed. Fig. 2(a) presents the overall procedure of SPARQL endpoint discovery and metadata crawling.

In total, SPARQL endpoint discovery, metadata collection and pre-processing includes the following steps:

Step 1a. Query LOD catalogues to obtain a seed list of LOD datasets, SPARQL endpoints and other available metadata.

Step 1b. For each LOD source, collect available (or generate) metadata, e.g. topics, schema, version, quality-related statistics and license.

Step 1c. For a specific Web archive, select LOD datasets w.r.t. the topical relevance and quality parameters. Establish property weighting for crawl prioritization.

Prioritization of the crawler: Similar to Web crawling, there is a trade-off between data collection efficiency and completeness in the context of LOD crawling. On the one hand, the entity preservation process should aim to create a potentially complete overview of the available entity representations and collect data from possibly many sources. On the other hand, due to the topical variety, scale and quality differences of LOD datasets, preservation should be performed in a selective manner, prioritizing data sources and properties according to their topical relevance for the Web archive, general importance as well as quality (e.g. in terms of relative completeness, mutual consistency, and freshness).

Entity-centric crawling: Entities in Linked Data Cloud can be retrieved from SPARQL endpoints, through URI lookups or from data dumps. The entity

[1] `http://datahub.io/`

[2] `http://data.linkededucation.org/linkedup/catalog/`

preservation process can be a part of a *Web archive metadata enrichment* that extracts entities from the archived Web pages and fed them into the Linked Data Crawler (e.g. [4]) for preservation. Fig. 2(b) presents the overall procedure for entity extraction and crawling. In total, entity-centric crawling includes the following steps enabling the crawler to collect comprehensive representations of entities and their sources for archiving:

Step 2a. For each entity extracted by NER, collect machine-readable entity representations from the relevant datasets (either using SPARQL or HTTP).
Step 2b. Collect human-readable entity view(s) available through the HTTP protocol (see Fig. 1 (a)) using Web crawling techniques.
Step 2c. Follow object properties (URIs) of an entity to collect related entities from the same LOD dataset. Here, property weighting together with other heuristics (e.g. path length) can be used to prioritise the crawler.
Step 2d. Follow the links (e.g. *owl:sameAs*) to external datasets to collect equivalent or related entities. In this step, prioritization of the crawler can be performed based on the estimated data quality parameters in such external datasets.

3 Conclusions and Outlook

In this paper we discussed aspects of entity-centric preservation in LOD in the context of the long-term Web archives. We described requirements and methods for entity preservation in Web archives. As the next steps towards the entity-centric LOD preservation we envision investigation of preservation-relevant quality aspects of linked datasets, as well as development of methods to automatic property weighting to achieve effective prioritization of the entity crawling.

Acknowledgments

This work was partially funded by the European Research Council under ALEXANDRIA (ERC 339233) and the COST Action IC1302 (KEYSTONE).

References

1. E. Demidova, N. Barbieri, S. Dietze, A. Funk, H. Holzmann, D. Maynard, N. Papailiou, W. Peters, T. Risse, and D. Spiliotopoulos. Analysing and enriching focused semantic web archives for parliament applications. *Future Internet*, 6(3):433–456, 2014.
2. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of the ACL*, 2005.
3. T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136.* Morgan & Claypool, 2011.
4. R. Isele, J. Umbrich, C. Bizer, and A. Harth. LDSpider: An open-source crawling framework for the web of linked data. In *Proc. of ISWC 2010*, 2010.
5. R. Rogers. *Digital Methods.* MIT Press, 2013.