Entity-Centric Preservation for Linked Open Data: Use Cases, Requirements and Models

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

L3S Research Center & Leibniz University Hannover
Appelstrasse 9, 30167 Hannover, Germany
{demidova,risse,gtran,gossen}@L3S.DE

Abstract. Linked Open Data (LOD) plays an increasingly important role in the area of digital libraries and archives. It enables better semantic-based access methods and also gives the human reader more insights into linked entities. Since semantics is evolving over time, the temporal dimension has an increasing impact on the applicability of LOD to long-term archives. LOD does not necessarily provide a history. Therefore it is not possible to go back in time and access the document related LOD content at the time of the document creation. In this paper we propose to collect LOD information along with the Web content for Web archives to also ensure a good semantic coverage of a Web crawl. We discuss use cases and requirements and derive the general approach and related data models for the implementation.

Keywords: Linked Open Data, Entity Preservation

1 Introduction and Motivation

Linked Open Data (LOD) plays an increasingly important role in the area of digital libraries and archives. LOD is a standardized method to publish and interlink structured semantic data. Major LOD hubs like DBpedia[1] or Freebase[2] extract information from Wikipedia and provide it as structured data with well-defined semantics. By linking entities in documents with their semantic descriptions in the LOD Cloud (i.e. the datasets published in LOD format), richer semantic descriptions of the document itself can be achieved. On the one hand, this enables better semantic based access method. On the other hand, since the LOD information is often based on Wikipedia and other textual entity descriptions, it also gives the human reader more insights into the entity at crawl time.

Although long-term archives can benefit from LOD, the temporal dimension has an increasing impact on the access and the interpretation of content. The human understanding of an information object at a later point in time requires, depending on the kind of information object, additional context information. While traditional materials like papers or books often bring enough context information for the understanding, this

[1] dbpedia.org

[2] www.freebase.com

is rarely the case for user generated content on the Web and Social Web. A twitter message like “The new #ipod is cool http://bit.ly/1cd7ylZ” will be hardly understandable in 50 years without additional knowledge. We do not know today if iPods will still exist or if people will know what an iPod was in 2014. Since entities are also evolving (e.g. names change), we do not even know if the iPod in 2064 is still a portable music player.

While LOD can successfully support the semantic access to content today[9], it is of limited use for long-term archives. LOD is constructed similar to the Web as a constantly evolving source of knowledge. In this way it also inherits the property of the Web of not having a documented history. For the Web this issue is partly addressed by establishing Web archives or application specific versioning support e.g. in Wikipedia. For LOD, independent steps in this direction are done by some sites like Freebase to offer snapshots of their content at different times. However, this is site specific and no general support exists.

To address these issues we propose, in the area of Web archiving, to preserve relevant parts of the LOD Cloud along with crawled Web pages. This requires a semantic analysis including entity extraction from Web pages (e.g. using NER techniques) coupled with enrichment of extracted entities using metadata from LOD as close as possible to the content collection time point and calls for integrated crawling and Web archive metadata enrichment approaches that jointly collect Web data and related LOD entities. Since the LOD graph is - like the Web - an ever growing information collection, a number of issues need to be addressed that we will discuss in this paper. These issues include integrated crawling approaches to seamlessly collect Web documents along with relevant entities from the LOD cloud, prioritization methods to select the most relevant sources, entities and properties for archiving as well as modelling of the entities and their provenance (in particular the LOD sources) within Web archives. In summary, the contributions of this paper are as follows:

- Analysis of major use cases to derive requirements for entity-centric LOD preservation in the context of Web archiving;
- An approach for the entity-centric crawling of Linked Open Data content;
- Data models for the entity preservation in Web archives.

This paper is an extended version of [13] that discussed a preliminary approach for entity-centric crawling of Linked Open Data content.

2 Use Cases

Today an increasing interest in using Web content can be observed within the scientific community. Long term Web archives are becoming more interesting for researchers in humanities and social sciences. For example, historical sciences are increasingly using Web archives for their research. However, major research activities on events and topics from historian point of view will begin in around 25 years after the event. Due to the volume of information available on the Web and partly preserved in Web archives, working with the content raises a number of challenges. Especially temporal aspects depending on the time span between crawling and usage have an increasing impact on the way how to find documents and to interpret the content.

Searching for documents on today's Web is mostly based on specifying the search intent by using a number of keywords. Web search engines, for example, take a number of assumptions to fulfill the users' informational needs. This way, Web search for the entity "Pope" will result in many pages mentioning Pope Francis as the search engine assumes that in 2014 most interest exists on the current Pope. This assumption does not hold for Web archives usage (and other long term archives) containing snapshots of parts of the Web of many years. Making assumptions about the potentially most interesting interpretation of a keyword query "Pope" in a Web archive is hard since the temporal search intentions of the users can be manifold.

Knowledge about temporal aspects of entities can help to disambiguate them and support the Web archive users. The required knowledge can be taken from LOD sources like Freebase or DBpedia. By linking entities contained in the archived documents to their descriptions in the LOD Cloud at the crawling time, the entities get a unique temporal interpretation. For example, the entity "Pope" mentioned in a newspaper article created in 2012 will most probably linked to Pope Benedict XVI, while in a document created in 2014 would be linked to Pope Francis. The interlinking helps finding other documents where the same entity has been mentioned as well as to disambiguate mentions of other entities with the same name at a different point in time.

By taking the long-term perspective the temporal dimension gains a crucial importance. One example is the evolution of entities, and in particular their name changes. These changes can often be observed for geographic names and organizations. For example, the dissolution of the USSR in 1991 resulted in renaming of entire states, cities and streets throughout the involved countries, often returning locations their historical names. A prominent example in this context is the change of the name of St. Petersburg (1703 - 1914 and from 1991) to Petrograd (1914 - 1924) and Leningrad (1924 - 1991). Also persons are changing names e.g. Pope Francis was named Jorge Mario Bergoglio before his election in 2013. This is an important information for time-aware query reformulation. In this way a query about Pope Francis would be extended with his birth name to obtain documents written before 2013.

The interlinking of the archived documents with LOD entities can also facilitate access to human readable presentation of entities on the Web to get insights into the perception of the entity at the crawling time, which can be especially beneficial for researchers in historical and social sciences. For example, the former Managing Director of the International Monetary Fund Dominique Strauss-Kahn and the former President of Germany Christian Wulff have been highly distinguished people before a number of allegations came up. For getting a better understanding of an archived document in its temporal context, the reader should know how the entities contained in this document were described at the creation time or, at latest, at the crawling time.

A time sensitive access to Linked Open Data is currently not supported. Even though Wikipedia provides a detail history for each document this is not the case for most LOD sites. This has some consequences for the access and interpretation of entities in the future. For example, the LOD description of Pope Benedict XVI in a document created 2012 will present his description of today. Even if the description is correct, it does not show the perception of the Pope Benedict in 2012. Wikipedia - the source of DBpedia, Freebase and other - is constantly updated and even if it aims at neutral descriptions, the perception of the authors has always an impact. For example, an event an entity is involved in could be highlighted for some time and later be lowered again. The differences in the descriptions of an entity at different times provide many interesting insights for researchers. Even the non-existence of a LOD entry in a specific dataset can be of interest and should be documented.

To enable future researchers to get good and comprehensive overview about topics and events today, adequate Web archiving including entity-centric LOD preservation is a necessity. This will ease the access to archived content and support its interpretation in the future.

3 Requirements

In order to facilitate interpretation of the archived Web documents and LOD entities linked to these documents in the future, presentation of the entities within Web archives should take into account content, provenance, quality, authenticity and context dimensions. In the following we present these requirements in more detail.

3.1 Content and Schema

The entities on the Web of Data are typically represented using triples, in the form of <*entity URI, property, value*>. Thereby *entity URI* is a unique identifier of the entity within the data source. The properties can be further differentiated in the datatype properties (directly providing property values) and object properties (linking the entity to other entities and facts in the knowledge graph using URIs). Object properties are also used to link entities to entity types within a classification scheme.

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 of 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.

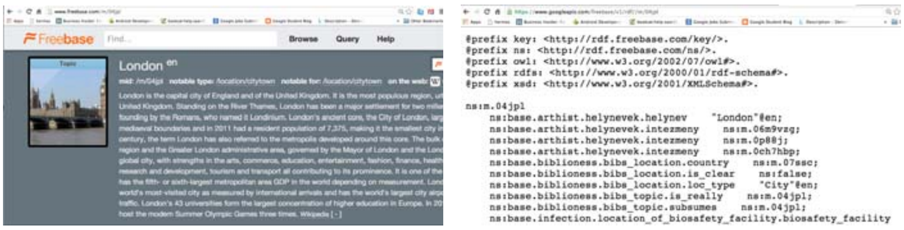
3.2 Provenance and Quality

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 data source of entity origin. 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.

3.3 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 on 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 entity “London” in Freebase dataset from March, 4, 2014.



Freebase
Find...
Browse
Query
Help
Topic
London en
mid: /m/04jpl notable type: /location/citytown notable for: /location/citytown on the web:

London is the capital city of England and of the United Kingdom. It is the most populous region, ur
United Kingdom. Standing on the River Thames, London has been a major settlement for two mille
founding by the Romans, who named it Londinium. London's ancient core, the City of London, larg
mediaeval boundaries and in 2011 had a resident population of 7,375, making it the smallest city in
century, the term London has also referred to the metropolis developed around this core. The bulk
region and the Greater London administrative area, governed by the Mayor of London and the Lond
global city, with strengths in the arts, commerce, education, entertainment, fashion, finance, health
research and development, tourism and transport all contributing to its prominence. It is one of the
has the fifth- or sixth-largest metropolitan area GDP in the world depending on measurement. Lond
world's most-visited city as measured by international arrivals and has the world's largest city airpo
traffic. London's 43 universities form the largest concentration of higher education in Europe. In 20
host the modern Summer Olympic Games three times. Wikipedia [-]

(a) Web representation

```
@prefix key: <http://rdf.freebase.com/key/>.
@prefix ns: <http://rdf.freebase.com/ns/>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

ns:m.04jpl
    ns:base.arthist.helynevek.helynev    "London"@en;
    ns:base.arthist.helynevek.intezmeny    ns:m.06m9vzg;
    ns:base.arthist.helynevek.intezmeny    ns:m.0p88j;
    ns:base.arthist.helynevek.intezmeny    ns:m.0ch7hbp;
    ns:base.biblioness.bibs_location.country    ns:m.07ssc;
    ns:base.biblioness.bibs_location.is_clear    ns:false;
    ns:base.biblioness.bibs_location.loc_type    "City"@en;
    ns:base.biblioness.bibs_topic.is_really    ns:m.04jpl;
    ns:base.biblioness.bibs_topic.subsumes    ns:m.04jpl;
    ns:base.infection.location_of_biosafety_facility.biosafety_facility
```

(b) RDF representation

Fig. 1. “London” in Freebase

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, organization logos, snippets of Wikipedia articles, etc. The archiving of human-readable entity representation can enable correct entity disambiguation going beyond the automatic processing and make archived entities valuable assets e.g. for social science researchers.

Authenticity raises a number of issues in the dynamic Web environment. One important issue for the authenticity is that the archived Web pages containing the entity, the Web pages representing the entity in LOD as well as machine-readable entity representation have to be crawled within a short time interval to collect coherent entity views in both, machine-readable and human-readable representations.

4 Entity Extraction and Preservation

In the context of Web archives, entities can be extracted from the archived Web documents using Named Entity Recognition (NER) tools as a part of *Entity Extraction* or added to the archive from external LOD sources during the *Entity Preservation* as a part of overall Web archive metadata enrichment process. Thereby, the nature of LOD, being a distributed collection of datasets, brings alone several challenges with respect to entity crawling and preservation.

4.1 Entity Extraction

Entity extraction within Web archives can typically be handled using state-of-the-art NER tools such as Stanford NER [12], Illinois Named Entity Tagger [24], SENNA [6], SpeedRead [1], Wikipedia Miner [19], TagMe [11], Illinois Wikifier [5] and others. These tools make use of different approaches to entity extraction and interlinking and indicate significant differences with respect to precision, coverage and efficiency.

The Wikipedia-based NER tools such as Wikipedia Miner and Illinois Wikifier directly support linking of extracted entities to Wikipedia. However, they have a relatively small coverage being restricted to entities mentioned in Wikipedia and are less efficient as they require access to Wikipedia dumps during the extraction phase. For example, Wikipedia Miner – one of the best tools for Wikipedia-based entity recognition – consumes a large chunk of memory to load the Wikipedia entity graph during the processing [1, 7]. Experimental results in the literature also show that text-based tools such as SENNA, Stanford NER and SpeedRead can achieve better precision and recall than Wikipedia-based tools.

4.2 Entity Preservation

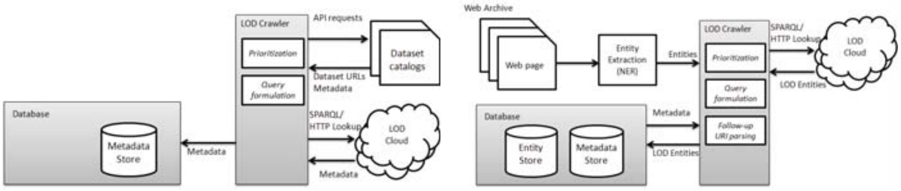
Whereas entity extraction can deliver entity labels, types and eventually initial pointers (URIs) of the relevant entities in the reference LOD datasets, the collection of relevant entities should be extended beyond the LOD sources used by the extractors. Furthermore, the content of the entities needs to be collected and preserved according to the requirements presented in Section 3. In these context important challenges are connected to *SPARQL endpoint 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[3] or the LinkedUp catalogue[4] include endpoint URLs of selected datasets as well as selected statistics, mostly concerning the size of specific datasets; however, existing catalogs are highly incomplete. Furthermore, existing approaches to distributed SPARQL querying suffer from efficiency, scalability and as a result low availability of SPARQL endpoints [27]. As an alternative, some LOD sources provide data dumps for download; however, given the dynamic nature of LOD, the dumps require frequent updates and come alone with lack of scalable tools for their processing. Approaches to alleviate these problems are being actively developed in the Semantic Web community [27].

[3] http://datahub.io/

[4] http://data.linkededucation.org/linkedup/catalog/

Metadata crawling includes several steps to obtain SPARQL endpoints URLs and generate corresponding metadata. Based on this metadata, prioritization of LOD sources and properties within these sources for preservation can be performed. Figure 2(a) presents the overall procedure of SPARQL endpoint discovery and metadata crawling. In total, SPARQL endpoint discovery, metadata collection and pre-processing include following steps:



(a) SPARQL endpoint discovery and metadata crawling

(b) Entity-centric crawling

Fig. 2. Entity preservation

Step 1a. Query LOD catalogues to obtain a seed list of LOD datasets, SPARQL endpoints and other available metadata. The crawler can also work with a pre-defined seed list of LOD sources.

Step 1b. For each available LOD source, collect and generate metadata. The metadata should include topical information to facilitate selection of relevant datasets for a specific Web archive, available schema information (especially with respect to the relevant entity types and their properties), the version of the dataset, quality-related statistics and license information.

Step 1c. For a specific Web archive, perform LOD dataset selection according to the topical relevance and quality parameters. If schema information is available, 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 possibly complete overview of the entity representations available in LOD and collect information from possibly many sources. On the other hand, due to the topical variety, scale and quality differences of LOD datasets, entity 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). As such metadata is rarely available, methods to generate this metadata need to be developed.

An important consideration when crawling LOD sources is the decision when to stop the crawl. A natural stopping criterion is when all linked resources are crawled, but because of the scale of the Linked Open Data Cloud this is neither feasible nor desirable. In Web crawling it is common to restrict the crawl to specific depth, i.e. only crawl paths of a given maximum length. This model can be generalized as a cost-based model, where each link traversal is assigned a cost and the crawler only follows paths where the sum of link costs is less than a specified budget. In this model we can assign costs to links based on their expected utility for the archive. For example, commonly used properties such as *rdfs:label* are typically useful because they are understood by more applications. By using heuristics we can dynamically prioritize which links to follow and limit the crawl to the most relevant properties.

Entity-centric crawling: Entities in Linked Data Cloud can be retrieved from SPARQL endpoints, through URI lookups or from data dumps [14]. The entity 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. [15]) for preservation. Figure 2(b) present the overall procedure for entity extraction and crawling. In total, Entity-centric crawling includes following steps:

Step 2a. For each entity extracted by NER, collect entity representations from the relevant datasets. In this step the crawler collects datatype properties (labels) and object properties (URIs) of a given entity typically using SPARQL queries. To form the queries, labels, types and (if available) initial URIs resulting from the NER process as well as available schema information can be used (see e.g. [9]).

Step 2b. Collect entity view(s) available through the HTTP protocol. Here the crawler can use Web crawling techniques to collect Web pages containing human-readable entity representations (see Figure 1 (a)).

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.

Using the steps above, the crawler can collect comprehensive representations of entities and their sources for archiving. In the next section we describe models to preserve this data.

5 Data Models for Preservation

In order to model entities within a Web archive taking into account content, schema, provenance, quality, authenticity and context dimensions, we need to preserve not only the entities, but also the metadata of the sources these entities originate from. Therefore, in this section we briefly present the **Entity Model** and **Entity Source Model** that jointly address these requirements.

In order to serialize the models presented below, an RDF representation making use of existing vocabularies (e.g. RDF Schema[5], VoID[6] for data source schemas and PROV-

[5] http://www.w3.org/TR/rdf-schema/

[6] http://www.w3.org/TR/void

O[7] for provenance properties) as well as an icrawl specific vocabulary[8] can be created and written to an RDF file or directly to a triple store within the Web archive. The data models discussed in this paper describe properties that become necessary based on the requirements described in Section 3. These models can be extended according to the needs of particular archival applications.

5.1 Entity Model

According to the requirements identified in Section 3, in the *Entity Model*, we cover the aspects of entity content, schema, provenance, authenticity and context. Our entity model is shown in Table 1.

First of all, the *Content and Schema* properties give an overview of the entity content in its original dataset. With respect to the content we differentiate between the key properties such as *label* and *type* and other properties. More detailed property weighting is provided as a part of *Entity Source Model*.

Then, *Provenance and Quality* properties include unique entity identifiers within the Web archive and its original source (e.g. a URI in the LOD source), as well as content verification properties. Furthermore, they provide information on entity relevance in its original context.

Finally, *Authenticity and Context* properties link the entity to its human-readable representation within the Web archive and specify important relationships to other entities within the archive (e.g. relationships between the entities extracted from documents and entities crawled from LOD, or equivalence relationships with other entities).

5.2 Entity Source Model

The *Entity Source Preservation Model* that represents a specific entity source is shown in Table 2. The *Content and Schema* properties of the *Entity Source Model* include domain and schema information as well as detailed property weighting to facilitate prioritization of the crawler.

The *Provenance and Quality* properties include identifier of the data sources including their URLs, versions and access times, as well as quality parameters and legal aspects.

6 Related Work

The problems discussed in this paper are particularly relevant to the fields of Web archiving and Web archive enrichment, Linked Data crawling, metadata dictionaries, as well as entity dynamics and entity interlinking. In this section we discuss related work in these areas in more detail.

Web archiving: The goal of Web archiving [18] is to preserve Web pages and sites

[7] http://www.w3.org/TR/2013/REC-prov-o-20130430/

[8] http://l3s.de/icrawl-ns#

Property	Description	Properties for RDF Serialization
	Content and Schema	
label	A human-readable entity name.	*rdfs:label*
type	An entity type (according to the original source, or a NER extractor) e.g. *Person*, *Location*, *Organization*.	*rdf:type*
properties	Properties as specified in the original source.	taken from the source
	Provenance and Quality	
URI	Persistent identifier of the entity within the Web archive	
version identifier	Unique identifier for each archived entity version.	*icrawl:hasVersion*
entity source	A reference to the entity source (see *Entity Source Model*).	*void:Dataset*
origin URI	The URI of the entity in the original source.	*prov:hadPrimarySource*
crawl timestamp	Timestamp when the entity was extracted or retrieved.	*prov:generatedAtTime*
fingerprint	Verification of the content, including algorithm and value.	*icrawl:hasFingerprint*, *prov:wasGeneratedBy*, *rdf:value*
software	Software used to extract or retrieve the entity.	*prov:SoftwareAgent*
relevance score	Relevance score of the entity including an algorithm (e.g. ObjectRank[2] or frequency) and the score.	*icrawl:hasAlgorithm*, *prov:wasGeneratedBy*, *rdf:value*
	Authenticity and Context	
visualization URI	A unique identifier of the visual representation of the entity within the Web archive.	*icrawl:visualizedBy*
snippet	A short textual representation of the entity using e.g. surrounding sentences or descriptive properties.	*dc:description*
relation	Relations to other objects in the Web archive, e.g. equivalence.	e.g. *owl:sameAs*

Table 1. Entity Model

Property	**Description**	**Properties for RDF Serialization**
Content and Schema		
domain	A topical domain of the dataset	*rdfs:domain*
class	Relevant entity types in the dataset schema	*rdfs:Class*
weighted properties	Weighted properties for archiving prioritization. The weight is in the range [0,1], where "1" stands for the highest archiving priority.	*icrawl:hasProperty*, *rdf:property*, *rdf:value*
Provenance and Quality		
URI	Persistent identifier of the source within the Web archive.	
origin URI	URL of the source on the Web at the archiving time point.	*void:uriLookupEndpoint*
version identifier	Unique and stable identifier for each version of the source.	*icrawl:hasVersion*
quality parameters	Quality metrics (e.g. data size, freshness, completeness of data).	e.g. *void:entities*, *void:properties*
legal parameters	Publisher and license of the source.	*dc:publisher*, *icrawl:license*

Table 2. Entity Source Model

for the future. This is required because Web pages disappear rapidly [25] or change their content. Therefore it is necessary to preserve their content in Web archives, of which the most prominent is the Internet Archive[9].

Web archives are typically created using Web crawlers (e.g. Heritrix [20]) which follow hyperlinks between Web pages and store the content of Web pages they encounter on the way. Depending on the policies of the archiving institution, links are followed in the order they are encountered (*breadth-first-crawling*) or are prioritized according to their relevance for a given crawl topic (*focused crawling* [4]). In this paper we pay specific attention to the aspects of entity-centric Linked Data crawling for Semantic Web resources. These aspects has not been addressed sufficiently in the related work.

Web archive enrichment: Linked Open Data has been successfully used for Web archive enrichment in the ARCOMEM project[10]. To this extent, entity representations extracted from textual documents using NER techniques have been linked with the corresponding entities in DBpedia and Freebase datasets[9]. This mapping enabled co-resolution of entity representations extracted from multiple Web documents. However, the LOD entities used for the co-resolution and Web archive enrichment have not been preserved. As a consequence, external links to the dynamic Freebase dataset became obsolescent very quickly. The entity preservation techniques discussed in this paper can alleviate this problem by providing a stable time-based representation of linked entities within the Web archive.

Linked Data crawling and versioning: Recently, there have been many studies in the areas of Linked Data crawling [15], Semantic Web search engines (e.g. Sindice[11]) as well as distributed SPARQL querying, e.g. [14]. The Memento project[12] aims at providing an HTTP-based versioning mechanism for Web pages and Linked Data and implements versioning of Linked Data as a part of content negotiation [8]. However, existing solutions do not provide entity-centric preservation facilities. In contrast, we collect Linked Data resources in a focused way by paying specific attention to their relevance in the context of Web archives.

Metadata dictionaries: Multiple models and metadata dictionaries support descriptions of entities and their specific properties at different level of details and can potentially be useful in the context of entity preservation. For example, PROV Family of Documents[13] comprises a set of W3C recommendations for representing provenance information and VoID [14] is an RDF Schema vocabulary for expressing metadata about RDF datasets, just to name some examples. While some properties from the existing models and data dictionaries can be mapped to the properties of the data models proposed in this paper, they do not cover the entire spectrum of preservation aspects of

[9] https://archive.org/

[10] www.arcomem.eu

[11] http://sindice.com

[12] http://mementoweb.org

[13] http://www.w3.org/TR/prov-overview/

[14] http://www.w3.org/TR/void/

distributed entities in LOD. To ensure interoperability, we reuse existing properties in the model serialization, where applicable.

Entity dynamics: The detection of the dynamics of entities attracted recently a number of researchers with a special focus on query reformulation. Berberich et al. [3] propose reformulating a query into terms prevalent in the past by measuring the degree of relatedness between two terms when used at different times by comparing the contexts as captured by co-occurrence statistics. Kanhabua and Nørvåg [17] incorporate Wikipedia for detecting former names of entities. They exploited the history of Wikipedia and consider anchor texts at different times pointing to the same entity as time-based synonyms. Kaluarachchi et al. [16] propose to discover semantically identical concepts (or named entities) used at different time periods using association rule mining to associate distinct entities to events. Tahmasebi et al. [26] proposed an unsupervised method for named entity evolution recognition in a high quality newspaper (i.e., New York Times). They consider all co-occurring entities in change periods as potential succeeding names and filter them afterwards using different techniques.

Entity interlinking: *Entity interlinking* identifies equal, similar or related entities [10, 23] and formally describes such relations. The state-of-the-art in link discovery is mostly focused on *owl:sameAs* relations linking entities representing the same real-world objects in distributed data collections e.g. [21, 22]. In the context of LOD preservation, we retrieve entities from distributed sources independently. To perform Web archive consolidation, we can rely on state-of-the-art entity interlinking methods and available equivalence links.

7 Conclusions and Outlook

The contributions of this paper are manyfold. First of all, we described use cases for entity preservation in Linked Open Data in the context of Web archives. These use cases include applications for entity-based access to long term archives, evolution aware entity search and time sensitive access to LOD. Then we derived requirements for entity preservation with respect to the content, schema, provenance, quality, authenticity and context dimensions. Following that, we discussed methods for entity extraction and preservation in Web archives and presented data models to represent data sources and entities within Web archives according to the identified dimensions. As the next steps towards the entity-centric LOD preservation we envision investigation of preservation-relevant quality aspects of Linked Open Datasets, as well as development of the methods to automatic property weighting to achieve effective prioritization in the entity crawling process.

Acknowledgments

This work was partially funded by the European Research Council under ALEXANDRIA (ERC 339233).

Bibliography

[1] R. Al-Rfou' and S. Skiena. Speedread: A fast named entity recognition pipeline. In *COLING*, pages 51–66, 2012.

[2] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 564–575. VLDB Endowment, 2004. ISBN 0-12-088469-0. URL http://dl.acm.org/citation.cfm?id=1316689.1316739.

[3] K. Berberich, S. J. Bedathur, M. Sozio, and G. Weikum. Bridging the terminology gap in web archive search. In *WebDB*, 2009.

[4] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-16):1623–1640, 1999. ISSN 1389-1286. doi: 10.1016/S1389-1286(99)00052-3. URL http://www.sciencedirect.com/science/article/pii/S1389128699000523.

[5] X. Cheng and D. Roth. Relational inference for wikification. In *EMNLP*, 2013. URL http://cogcomp.cs.illinois.edu/papers/ChengRo13.pdf.

[6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, Nov. 2011. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1953048.2078186.

[7] M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 249–260, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2035-1. URL http://dl.acm.org/citation.cfm?id=2488388.2488411.

[8] H. V. de Sompel, R. Sanderson, M. L. Nelson, L. Balakireva, H. Shankar, and S. Ainsworth. An http-based versioning mechanism for linked data. *CoRR*, abs/1003.3661, 2010. URL http://arxiv.org/abs/1003.3661.

[9] S. Dietze, D. Maynard, E. Demidova, T. Risse, W. Peters, K. Doka, and Y. Stavrakas. Preservation of social web content based on entity extraction and consolidation. In *2nd International Workshop on Semantic Digital Archives (SDA) in conjunction with the 16th International Conference on Theory and Practice of Digital Libraries (TPDL), Pafos, Cyprus, September 2012*, 2012.

[10] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Trans. on Knowl. and Data Eng.*, 19(1):1–16, Jan. 2007. ISSN 1041-4347. doi: 10.1109/TKDE.2007.9. URL http://dx.doi.org/10.1109/TKDE.2007.9.

[11] P. Ferragina and U. Scaiella. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1625–1628, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0099-5. doi: 10.1145/1871437.1871689. URL http://doi.acm.org/10.1145/1871437.1871689.

[12] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219885.

[13] G. Gossen, E. Demidova, T. Risse, and G. B. Tran. Towards entity-centric preservation for web archive enrichment. In *Joint Proceedings of the USEWOD '15 and the PROFILES'15 Workshops, Portorož, Slovenia, May 31 - June 1, 2015.*, volume 1362 of *CEUR*

Workshop Proceedings, pages 81–84. CEUR-WS.org, 2015. URL http://ceur-ws.org/Vol-1362/PROFILES2015_paper5.pdf.

[14] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data summaries for on-demand queries over linked data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 411–420, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772733.

[15] R. Isele, J. Umbrich, C. Bizer, and A. Harth. LDSpider: An open-source crawling framework for the web of linked data. In *Proceedings of 9th International Semantic Web Conference (ISWC 2010) Posters and Demos*, 2010.

[16] A. C. Kaluarachchi, A. S. Varde, S. J. Bedathur, G. Weikum, J. Peng, and A. Feldman. Incorporating terminology evolution for query translation in text retrieval with association rules. In *CIKM*, pages 1789–1792, 2010.

[17] N. Kanhabua and K. Nørvåg. Exploiting time-based synonyms in searching document archives. In *Proceedings of the 10th annual joint conference on Digital libraries*, JCDL '10, pages 79–88, New York, NY, USA, 2010. ACM.

[18] J. Masanés. *Web Archiving: Issues and Methods*, pages 1–53. Springer, 2006. ISBN 978-3-540-23338-1. doi: 10.1007/978-3-540-46332-0_1.

[19] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artif. Intell.*, 194: 222–239, Jan. 2013. ISSN 0004-3702. doi: 10.1016/j.artint.2012.06.007.

[20] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to Heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, 2004.

[21] A. Nikolov, M. d'Aquin, and E. Motta. Unsupervised learning of link discovery configuration. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, volume 7295 of *Lecture Notes in Computer Science*, pages 119–133. Springer, 2012. ISBN 978-3-642-30283-1.

[22] G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, and W. Nejdl. Eliminating the redundancy in blocking-based entity resolution methods. In *Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL 2011, Ottawa, ON, Canada, June 13-17, 2011*, pages 85–94, 2011.

[23] G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, and W. Nejdl. Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 53–62, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0747-5.

[24] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-29-9.

[25] H. M. SalahEldeen and M. L. Nelson. Losing my revolution: How many resources shared on social media have been lost? In *Theory and Practice of Digital Libraries*, volume 7489, pages 125–137. Springer, 2012. ISBN 978-3-642-33289-0. doi: 10.1007/978-3-642-33290-6_14.

[26] N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann, and T. Risse. Neer: An unsupervised method for named entity evolution recognition. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, Mumbai, India, 2012. URL http://www.l3s.de/neer-dataset/.

[27] R. Verborgh, M. V. Sande, P. Colpaert, S. Coppens, E. Mannens, and R. V. de Walle. Web-scale querying through linked data fragments. In *7th Workshop on Linked Data on the Web*. CEUR series of workshop proceedings, 2014. URL http://linkeddatafragments.org/publications/.