Linked Open Data Framework for Serendipity in History of Art Research

Gianmaria Silvello

Dept. of Information Engineering, University of Padua, Italy
silvello@dei.unipd.it

Abstract. In this paper we outline the main lines of research for defining a framework based on *Linked Open Data (LOD)* for supporting knowledge creation in the *Cultural Heritage (CH)* field with a particular focus on History of Art research.

We delineate the main challenges we need to deal with and we explore the state-of-the-art in LOD publishing systems, LOD citation and authority management. Furthermore, we introduce the idea of computer-aided serendipity in History of Art research with the purpose of contributing to the advancement of the field and to the definition of new methodologies for entity linking and retrieval.

1 Motivation

One of the most relevant socio-economical and scientific changes in recent years has been the recognition of data as a valuable asset. The Economist magazine recently wrote that "data is the new raw material of business"; and the European Commission stated that data-related "technology and services are expected to grow from EUR 2.4 billion in 2010 to EUR 12.7 billion in 2015" [9]. The principal driver of this evolution is the Web of Data, the size of which is estimated to have exceeded 100 billion facts (i.e. semantically connected entities). The actual paradigm realizing the Web of Data is LOD [14], which by exploiting Web technologies, such as the *Resource Description Framework (RDF)* [20], allows public data in machine-readable formats to be opened up ready for consumption and re-use. LOD is becoming the de-facto standard for data publishing, accessing and sharing because it allows for flexible manipulation, enrichment and discovery of data as well as for overcoming interoperability issues. The ground breaking potential of this approach resides in the semantic connections among data enabling new knowledge creation and discovery possibilities.

The CH domain and in particular the History of Art research field, provide a fertile ground where the LOD potential can grow and bloom; indeed, in History of Art the preponderant way to produce new knowledge is to reveal connections between different items (illuminated manuscripts, pictures, frescos) that can cast a new light on an artist, an artistic movement or an art-historical period.

This potential is widely recognized by public and private agencies, which keep investing in publishing CH resources as LOD. In the CH domain, which in the EU alone “accounts for 3.3% of GDP and employs 6.7 million people (3% of total employment)” [9]; as a consequence, in the last couple of years within the Europeana digital library[1], the European Commission started a major effort for publishing as LOD millions of multilingual and multimodal CH resources (e.g. archival documents, illuminated manuscripts, pictures) gathered from more than 2300 institutions distributed across 36 countries.

Nevertheless, publishing CH resources as LOD on the Web, is just the first step to enable new knowledge creation. To realize the full potential of the Web of Data, we need to devise innovative methods for seeking and creating links between datasets and to design user-oriented services for exploiting them. To this purpose, History of Art is a fruitful domain that, not only can benefit, but also can assist the development of these new methods because it provides: rich and heterogeneous information needs, ample structured and unstructured resource datasets and a proactive research community accustomed to seek new connections between entities.

The aim of this paper is to outline the basics to design a LOD framework for supporting knowledge creation in the History of Art domain and to outline some of the scientific challenges we have to face. Overall, the framework we envision pursues two main objectives: (i) to provide LOD-ready functionalities to create, access and cite new knowledge and to represent and retain authority information; (ii) to assist domain experts in the creation of semantic connections and to empower them with (semi) automatic serendipity capabilities.

Accomplishing the first goal will overcome current LOD publishing practice where the data are produced and stored by local systems and then mapped, synchronized and published as LOD; this procedure requires a multiplication of investments and produces scarcely connected datasets given that the expert knowledge is not tackled in the publishing activity. Our idea is to conflate data creation, connection, and publishing in one phase and let computer scientists and domain experts to work back-to-back joining knowledge and efforts. In this context there is the need also to provide concrete solutions for fundamental but overlooked issues as data citation and authority management.

Achieving the second goal will push History of Art research boundaries by joining the experts’ knowledge and intuition with automatic tools able to retrieve relevant entities in the Web of Data. To this purpose, from the computer science point-of-view, we need to devise query models from the experts information needs, envisage and define methods for mapping entities to queries, use them for retrieving target entities and then provide entity rankings. Current solutions are biased toward Web search where user needs are expressed using very few terms; as a consequence, they employ flat models of user inputs where contextual information, text, images, links, categories and feedbacks, when available, are considered at the same level and equivalent one to the other.

The History of Art domain provides us with rich and heterogeneous user inputs that poses additional challenges, but that also provides the means for advancing entity retrieval state-of-the-art. The framework we envision could also

[1] `http://europeana.eu/`

push ahead the creation and distribution of CH resources offering new ways for addressing consumers' demand for data access and for greater participation in the creativity process.

The rest of the paper is organized as follows: Section 2 presents the state-of-the-art in LOD publishing in the CH context, serendipity oriented algorithm and data citation methodologies. In Section 3 we present the main design lines, the challenges we have to face for realising the envisioned framework and introduce the art research use case on which the proposed framework will rely. In Section 4 we propose an architecture that could implement the envisioned framework. Finally, in Section 5 we draw some final remarks.

2 State-of-the-art and some open questions

The efforts for disclosing the LOD potential within the CH field must go towards the design of new methodologies for creating meaningful and possibly unexpected semantic links between data and for managing the knowledge created through these connections. This endeavor is needed to make the Web of Data fully operational and so that its role can be compared to that played by search engines over the last two decades in disclosing the full potential of the Web.

The CH domain and particularly the History of Art not only can be beneficial of the full exploitation of the LOD potential, but they can also provide fertile ground where new methods and technologies can be designed, developed and applied. Indeed, in History of Art the preponderant way to produce new knowledge is to reveal connections between different items (illuminations, pictures, frescos) that can cast new light on an artist, an artistic movement or an art-historical period. The most valuable connections are the unexpected ones linking elements that may seem to have very few in common; indeed, in this domain, important discoveries have often been done thanks to associations and connections of items emerged also by chance in the research path identified by domain experts. This process of discovery can be defined as *serendipity* and it is especially encouraged by LOD where meaningful links between entities allow us to move across diverse and apparently unrelated knowledge domains. History of Art is a well suited domain where algorithms fostering serendipity within LOD can be designed, developed and evaluated because it provides: rich and heterogeneous information needs, ample structured and unstructured resource datasets and a proactive community accustomed to seek new semantic connections between entities.

Given the scientific and economic impact of LOD and the leading role of CH for triggering its potential, a lot of research is being carried out in important fields such as database systems, information retrieval and digital libraries focusing especially on efficient ways for storing and querying datasets [2,3], extending Web search to entity retrieval [5] and devising sustainable methods for publishing [6] and linking LOD [11, 13].

However, several crucial questions still remain unanswered or overlooked. What merits to be linked? How can linking towards potentially unknown data sources be eased? How can the serendipity process be enhanced and guided by semantic methodologies? How can domain experts' knowledge be exploited for creating meaningful semantic links? How the authoritative information about entities, activities and people involved in data creation be established and retained? How new created knowledge represented by a data subset can be cited?

We can point out four main research areas concerning these questions that must be taken into account:

- **LOD-based systems:** The vast majority of systems in the CH domain are not natively LOD-based and adopt a publishing paradigm involving redundancy and multiplication of costs [15]. Moreover, in most cases (e.g. Europeana), the links to external sources are established without involving domain experts and are rather syntactic than semantic. The framework we envision is aimed as developing user-oriented services for native LOD creation and exploration. Moreover, it aims to provide services for connecting data while experts are working, thus exploiting their knowledge for link creation.
- **Serendipity capabilities:** No system in the CH domain provides end-users with this function. Related methodologies are proposed in the context of entity linking and retrieval targeting Web search: they deal with sparse user inputs [1] and general datasets [16]. These solutions employ flat models of user inputs mainly focusing on query expansion techniques [4].
- **LOD citation and authority management:** Recently two EU projects (i.e. PRELIDA[2] and DIACHRON[3]) marginally considered these aspects from the permanent preservation point-of-view, but there are as yet no concrete solutions for LOD data. Other citation systems focused on relational data [17] or hierarchical data such as *eXtensible Markup Language (XML)* [7], but they are not applicable to the LOD case. [18] proposed an initial methodology based on named graphs and RDF quad semantics which enables persistent, dereferenceable, variable granularity and human- and machine-readable citations of LOD subsets, but no ready-to-use solution has been implemented yet. Several challenges are still open in data citation of LOD such as the citation of evolving datasets which involves temporal aspects of RDF, the definition of equality between two citations and the concept of citation closure [8] (something like closure of a paper references).

3 Serendipity in art research: a proposal

For what it is concerned with the serendipity algorithms the framework we envision will employ (semi) automatic methods for suggesting entities to researchers; the idea is that this methodology could help them in their daily work by triggering the exploration of new research paths leading to new knowledge creation in History of Art. Now, we present the concrete steps we are going to follow to

[2] `http://prelida.eu/`

[3] `http://www.diachron-fp7.eu/`

realize the framework that we call *Linked Open Data Enhanced Framework for Serendipity in Art Research* (LODESTAR).

We will rely on a concrete use case rooted in history of art research which aim is to identify, analyze and assess the influence of the most ancient Divine Comedy illuminated manuscript illustrations on the many representations of the Hereafter and biblical and mythological characters that can be found in late-XIV century and Renaissance art. Indeed, many inventions that can be found in the first Divine Comedy illuminated manuscripts were destined to have great success and also to be used in later manuscripts, not only with Dante's poem, but also with other texts and other works of art. In order to achieve large-scale results, beside illuminations, we will also take into account products of the major arts such as paintings, frescos and sculptures, thus bringing new knowledge on different fields of the History of Art. The object of the research is highly original and constitutes a field of investigation that has never been explored before. Given its ample scope, the large diffusion of the Divine Comedy and its influence through the centuries on many fields, this use-case offers substantial chances of success especially for applying semi-automatic serendipity methodologies; it is a valuable starting stage to design, apply and test methods that will be then re-used in other use-cases and domains.

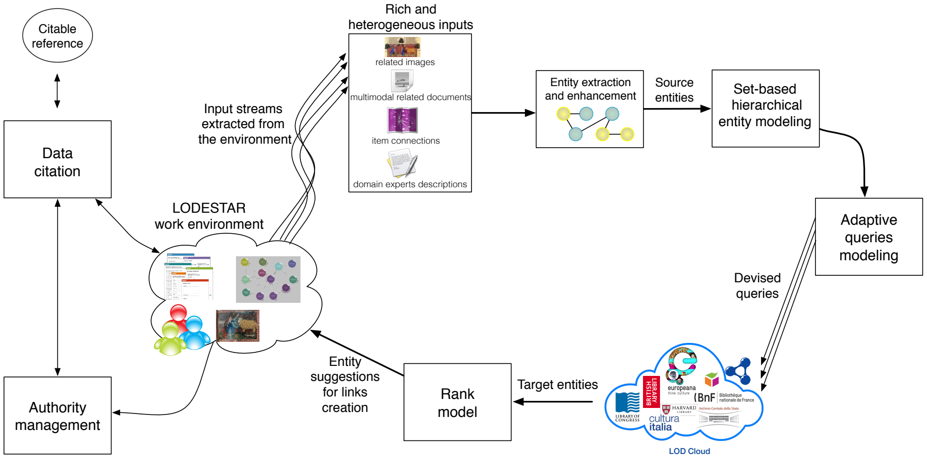
Citable reference
Data citation
Authority management
Input streams extracted from the environment
LODESTAR work environment
Rich and heterogeneous inputs
related images
multimodal related documents
item connections
domain experts descriptions
Entity extraction and enhancement
Source entities
Set-based hierarchical entity modeling
Adaptive queries modeling
Devised queries
LOD Cloud
Target entities
Rank model
Entity suggestions for links creation

Fig. 1. The workflow envisioned by the LODESTAR framework.

In Figure 1 we show the workflow realizing the objectives tackled by the LODESTAR framework. In the center there is the work environment where domain experts will carry out their daily research work. This environment provides rich and heterogeneous input streams such as free and semi-structured text, images, links, user profile and contextual information; LODESTAR is required to automatically catch these streams that are implicit expressions of user needs and to extract entities from them.

These entities will be sources for the entity retrieval algorithms realizing the LODESTAR serendipity methodology shown on the right hand side of Figure 1. This would be a major shift from state-of-the art entity retrieval solutions, which are typically employed in syntactic linking processes where input sources are constituted by few keywords. In LODESTAR users are not asked to explicitly issue queries, but the traces they leave behind while using the system will be automatically gathered and exploited for devising queries. To tackle this goal we have to employ new methods to model the extracted entities by associating different groups of entities to different levels of relevance decided on the basis of the user needs within the context under consideration; in other words we have to go beyond the current "flat models" that consider all the user inputs at the same level.

This model will allow us to devise several queries corresponding to different entities selections and compositions that will be used to retrieve "target entities" from the LOD cloud. The purpose of issuing multiple queries is to cover a wide spectrum of potential user needs as well as to ease the discovery of unknown entities that may trigger the creation of unexpected connections. The rank module gathers the target entities and orders them accordingly to the user need they tackle.

Lastly, the ranked entities are suggested to the users that will consider them for establishing semantic connections fostering new knowledge creation.

The left-hand side of Figure 1 reports data citation and authority management components needed to give credit to data creators (e.g. produce a diachronic reference to a data subset), retain context (e.g. who created a link, when a link has been created) and control data conflicts (e.g. choose the most authoritative statement among contradicting ones). LODESTAR has to provide general methodologies for tackling these issues for which, in literature, there are no concrete solutions yet. We will employ the solution proposed in [18] by extending it in order to consider also the temporal aspects of LOD datasets; indeed, a citation should be resolved by referring to the precise version of the data that was cited. To this end we plan to employ some RDF versioning methodology as described in [12, 19].

4 A possible implementing architecture

Serendipity can be realized by following the workflow shown on the right of Figure 1. Firstly, we need to trace the user information within the LODESTAR work environment such as the item being studied, annotations, items being compared to it, free text and semi-structured text (e.g. form fields), clickstream and user profiles. Then, we apply mining algorithms, extending existing entity extraction solutions, for extracting phrases, entities, categories from all these input streams; afterwards, the enhancement module, exploiting suitable ontologies, will connect the extracted entities one with the other via automatically established semantic links. At this stage we obtain an RDF graph of connected entities representing a very rich combination of user needs.

There exists no ready-to-use solution for handling this "user needs graph" and devise queries from it; LODESTAR represents the very first effort tackling this problem. The methodology we envision is to hierarchically model the user needs graph by associating different groups of entities to different levels of relevance decided on the basis of contextual information.

Machine learning methods have been proposed in literature for the task of learn classes in ontologies and constructing knowledge from user-provided examples; a possibility is to develop a similar approach to assign relevance weights or probabilities (i.e. the estimated relatedness to the user needs) to the entities in the graph. These weights will be exploited for modeling the user needs graph with a innovative set-based model, i.e. the NESTOR model [10], organizing the entities in sets with different levels relevance as shown in Figure 2.

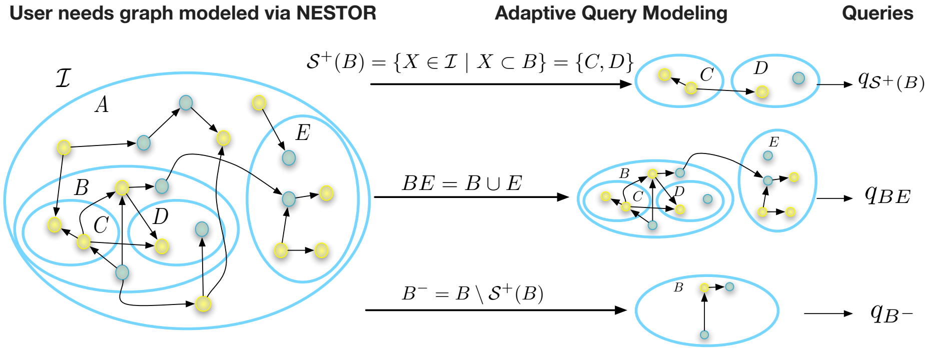
User needs graph modeled via NESTOR
Adaptive Query Modeling
Queries
$\mathcal{I}$
A
B
C
D
E
$\mathcal{S}^+(B) = \{X \in \mathcal{I} \mid X \subset B\} = \{C, D\}$
$BE = B \cup E$
$B^- = B \setminus \mathcal{S}^+(B)$
$q_{\mathcal{S}^+(B)}$
q_{BE}
q_{B^-}

Fig. 2. A methodology for handling rich and heterogeneous user inputs and devising queries via adaptive query modeling.

NESTOR comes with a whole bunch of defined set-operations that will be exploited to design an algorithm for selecting alternative entity combinations within the graph; from these selections a bunch of queries covering a wide spectrum of potential user needs will be devised.

The queries will be issued against the LOD cloud for retrieving groups of target entities; each group will be ordered by a ranking model built on solutions for structured search, e.g. statistical language models, and then aggregated by exploiting meta-search engine techniques to be presented as link suggestions to the users.

This envisioned methodology can be embedded in a more comprehensive architecture as shown in Figure 3 composed by 4-staked layers concerning important aspects of a system architecture: scalability, robustness, reactivity and suitability.

The first layer is designed by considering that LODESTAR will deal with big datasets and large user and machine requests. Therefore, in order to provide linear scalability and fault tolerance, several technologies should be taken into

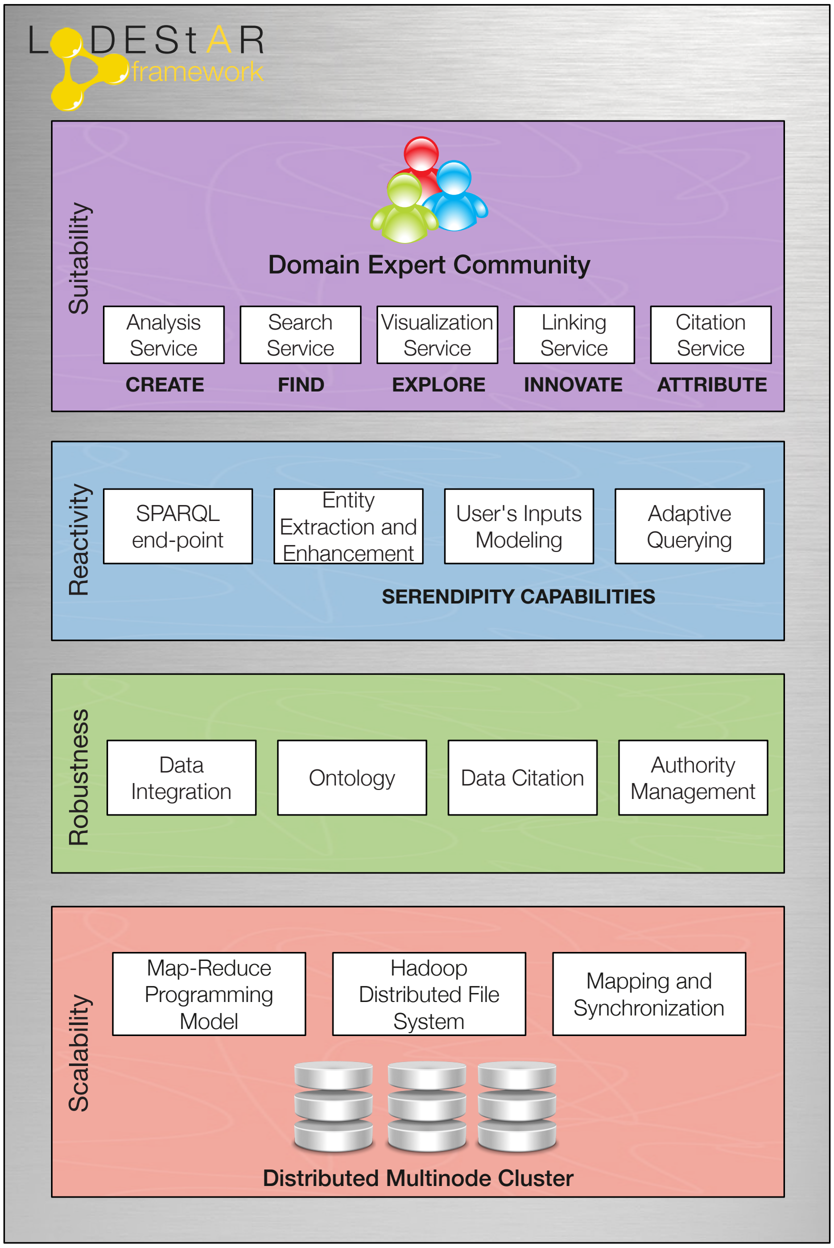

Fig. 3. A possible architecture implementing the LODESTAR framework.

account, in particular: Apache Cassandra[4] as distributed column store solution and Apache Hadoop[5] in conjunction with the Map-Reduce functionalities to provide reliable and distributed computing.

Scalability is also concerned with the selection of datasets that will serve the use-case we have in mind; in a real-world scenario we cannot assume that all relevant data will be available as LOD because even if many relevant CH institutions and systems are LOD-compliant – e.g. Europeana[6], the Library of Congress, the British Library, the French National Library, and Cultura Italia[7] – others still are not – e.g. Artstor, Dante On-line[8], and the Web gallery of Art.

To deal with all these sources, it is necessary to employ a mapping and synchronizing module that will harvest, map to LOD and store these data. An apposite data integration module will blend different resource kinds (as part of the robustness layer). All the resources will be represented accordingly to ontologies defined within the domain of interest; computer scientist and art historians will jointly carry out this activity by defining an RDF Schema and reusing existing vocabularies to establish bridges with third-party ontologies – e.g SKOS, Europeana Data Model, CIDOC-CRM – that will also be exploited for serendipity.

In the reactivity layer, we could provide a SPARQL end-point, which will work as a data provider allowing third-party services to discover, re-use and connect to the LODESTAR data. This layer also accommodates the modules implementing the serendipity algorithms.

All the presented services and solutions must be made available to end-users via pluggable user interface software components (i.e. Web portlets); the main end-user interfaces need to deal with an environment for describing items, annotating and visually comparing them, seeking and connecting entities, browsing the data and making use of the data citation and serendipity modules.

5 Final Remarks

In this paper we discussed the role of LOD in the CH field and we outlined some of the cultural and computational challenges we have to face in order to define a framework for supporting knowledge creation in CH. In particular, we selected a stimulating use case based on History of Art research which provides us with rich input data and a research community oriented to seek new connections between different items to create new knowledge.

The main challenges we outlined regard the necessity to define new methodologies for entity linking and retrieval which take into account complex user inputs, a work environment which on the one hand exploits existing LOD for knowledge creation and on the other hand stimulates the production of new and

[4] `http://cassandra.apache.org/`

[5] `https://hadoop.apache.org/`

[6] `http://www.europeana.eu/`

[7] `http://www.culturaitalia.it/`

[8] `http://www.danteonline.it/`

unexpected connections between CH resources and a data citation methodology allowing us to cite evolving subset of data with variable granularity and to produce both machine- and human-readable references.

References

1. Balog, K., Bron, M., De Rijke, M.: Query Modeling for Entity Search Based on Terms, Categories, and Examples. ACM Trans. Inf. Syst. 29(4), 22:1–22:31 (2011)
2. Blanco, R., Mika, P., Vigna, S.: Effective and Efficient Entity Search in RDF Data. In: Proc. of the 10th international conference on The semantic web - Volume Part I. pp. 83–97. Springer-Verlag, Berlin, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=2063016.2063023
3. Blanco, R., Ottaviano, G., Meij, E.: Fast and Space-Efficient Entity Linking for Queries. In: Proc. of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15. pp. 179–188. ACM Press, New York, New York, USA (Feb 2015), http://dl.acm.org/citation.cfm?id=2684822.2685317
4. Boston, C., Fang, H., Carberry, S., Wu, H., Liu, X.: Wikimantic: Toward Effective Disambiguation and Expansion of Queries. Data Knowl. Eng. 90, 22–37 (2014)
5. Bron, M., Balog, K., de Rijke, M.: Example Based Entity Search in the Web of Data. In: Proc. of the 35th European conference on Advances in Information Retrieval. pp. 392–403. Springer-Verlag, Berlin, Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-36973-5_33
6. Buccio, E., Di Nunzio, G.M.D., Silvello, G.: A Linked Open Data Approach for Geolinguistics Applications. Int. J. Metadata Semant. Ontologies 9(1), 29–41 (2014), http://dx.doi.org/10.1504/IJMSO.2014.059125
7. Buneman, P., Silvello, G.: A Rule-Based Citation System for Structured and Evolving Datasets. IEEE Data Eng. Bull. 33(3), 33–41 (2010), http://sites.computer.org/debull/A10sept/buneman.pdf
8. Buneman, P., Tannen, V., Davidson, S.B., Frew, J., Cohen-Boulakia, S.: Computational Challanges in Data Citation. Workshop report, University of Pennsylvania (2014)
9. Commission of the European Communities: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Towards a Thriving Data-Driven Economy. COMM(2014) 442 Final (2014)
10. Ferro, N., Silvello, G.: NESTOR: A Formal Model for Digital Archives. Information Processing & Management 49(6), 1206–1240 (November 2013)
11. Ferro, N., Silvello, G.: Making it Easier to Discover, Re-Use and Understand Search Engine Experimental Evaluation Data. ERCIM News 96, 26–27 (January 2014)
12. Flouris, G., Konstantinidis, G., Antoniou, G., Christophides, V.: Formal Foundations for RDF/S KB Evolution. Knowl. Inf. Syst. 35(1), 153–191 (2013)
13. Gottipati, S., Jiang, J.: Linking Entities to a Knowledge Base with Query Expansion. In: Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 804–813. Association for Computational Linguistics (Jul 2011), http://dl.acm.org/citation.cfm?id=2145432.2145523
14. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, USA (2011)

15. Marden, J., Li-Madeo, C., Whysel, N., Edelstein, J.: Linked Open Data for Cultural Heritage: Evolution of an Information Technology. In: Proc. of the 31st ACM International Conference on Design of Communication. pp. 107–112. SIGDOC '13, ACM, New York, NY, USA (2013)
16. Meij, E., Bron, M., Hollink, L., Huurnink, B., de Rijke, M.: Mapping Queries to the Linking Open Data Cloud: A Case Study Using DBpedia. Web Semant. 9(4), 418–433 (2011)
17. Pröll, S., Rauber, A.: Scalable Data Citation in Dynamic, Large Databases: Model and Reference Implementation. In: Proc. of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA. pp. 307–312 (2013)
18. Silvello, G.: A Methodology for Citing Linked Open Data Subsets. D-Lib Magazine 21(1/2) (2015), `http://dx.doi.org/10.1045/january2015-silvello`
19. Stefanidis, K., Chrysakis, I., Flouris, G.: On Designing Archiving Policies for Evolving RDF Datasets on the Web. Lecture Notes in Computer Science, vol. 8824, pp. 43–56. Springer (2014)
20. W3C: RDF 1.1 Concepts and Abstract Syntax – W3C Recommendation 25 February 2014 (February 2014)