PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

January/February 2015
Volume 21, Number 1/2

A Methodology for Citing Linked Open Data Subsets

Gianmaria Silvello
University of Padua, Italy
silvello@dei.unipd.it

DOI: 10.1045/january2015-silvello

Abstract

In this paper we discuss the problem of data citation with a specific focus on Linked Open Data. We outline the main requirements a data citation methodology must fulfill: (i) uniquely identify the cited objects; (ii) provide descriptive metadata; (iii) enable variable granularity citations; and (iv) produce both human- and machine-readable references. We propose a methodology based on named graphs and RDF quad semantics that allows us to create citation meta-graphs respecting the outlined requirements. We also present a compelling use case based on search engines experimental evaluation data and possible applications of the citation methodology.

1 Introduction

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" [H2020 WP, 2014—2015]. 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 the Linked Open Data (LOD), which by exploiting Web technologies, such as the Resource Framework Description (RDF), 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 in addition to overcoming interoperability issues.

Nevertheless, LOD publishing is just the first step for revealing the ground-breaking potential of this approach residing in the semantic connections between data enabling new knowledge creation and discovery possibilities. Current efforts for disclosing this potential are being concentrated on 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 shifting LOD from a publishing paradigm to a knowledge creation and sharing one.

Borgman in [Borgman, 2012b] outlined four rationales for sharing data that we think are gaining even more traction as the LOD paradigm extends its reach; Borgman pointed out that sharing and citing data is important for: (i) reproducing or verifying research, (ii) making results of publicly funded research available to the public; (iii) enabling others to ask new questions of extant data; and (iv) advancing the state of research and innovation. These rationales are to a varying extent rooted in the LOD paradigm, which makes data sharing a priority; we believe that along with data sharing, also data citation should be considered a prime concern of the research community. Indeed, together with data sharing, data citation is fundamental for giving credit to data creators and curators (*attribution*), to reference data in order to identify, discover and retrieve them [Borgman, 2012a] and for building and propagating knowledge [Buneman, 2006; Buneman and Silvello, 2010; Lawrence, *et al.*, 2011].

In the context of LOD a dedicated methodology for citing a dataset or a data subset has not yet been defined or proposed. Recently, two EU projects — i.e. PRELIDA and DIACHRON — considered these aspects from the permanent preservation point-of-view [Auer, *et al.*, 2012], but there are as yet no concrete solutions we can employ for data citation of LOD subsets.

In this paper we build on the newly defined "RDF Quad Semantics" [Klyne, *et al.*, 2014] to pinpoint a methodology for automatically generating citations of LOD subsets, which are machine-readable, but at the same time are understandable to a human. This methodology allows for citing LOD subsets with *variable granularity* (i.e. we can cite a single entity, a single statement, a subset of statements and the whole dataset) and produce citations composed by a unique identifier (i.e. a *reference*) used to retrieve the cited data subset in a human- and machine-readable format and some human- and machine-readable descriptive metadata assessing the citation (i.e. its quality and currency) and enabling data attribution [Borgman, 2012a]. A further property of the methodology being proposed here is that it is defined within the boundaries of the LOD paradigm and related — widely accepted and used — technologies; this means that if a given organization already has in place an infrastructure for creating and exposing LOD on the Web, the very same infrastructure can be exploited as is for data citation purposes.

The rest of the paper is organized as follows: in Section 2 we report on the LOD paradigm and RDF model highlighting the role of named graphs and quad semantics; furthermore, we outline the main requirements that a data citation methodology must fulfill and discuss some existing data citation systems. In Section 3 we present a use case based on search engine experimental data discussing why a data citation methodology for LOD is required. In Section 4 we describe the data citation methodology for LOD and in Section 5 we relate it to the presented use case reporting some possible applications. In Section 6 we draw some final remarks.

2 Background

2.1 Linked Open Data and RDF

The LOD paradigm [Heath and Bizer, 2011] refers to a set of best practices for publishing data on the Web [1] and it is based on a standardized data model, the Resource Description Framework (RDF). RDF is designed to represent information in a minimally constraining way and it is based on the following building blocks: graph data model, IRI-based vocabulary[2], data types, literals, and several serialization syntaxes.

The basic structural construct of RDF is a triple (subject, property, and object), which can be represented in a graph; the nodes of this graph are subjects and objects and the arcs are properties. IRIs identify nodes and arcs. RDF adopts a property-centric approach allowing anyone to extend the description of existing resources; properties represent relationships between resources, but they may also be thought of as attributes of resources, like traditional attribute-value pairs. RDF graphs are defined as mathematical sets; adding or removing triples from an RDF graph yields a different RDF graph.

RDF 1.1 [Klyne, *et al.*, 2014] specifications introduced the concept of *RDF dataset*, which is a collection of RDF graphs composed of: (i) a default RDF graph which may be empty and (ii) a set of *named graphs* which is a pair consisting of an IRI (i.e. the name of the graph) and an RDF graph.

The semantics as well as the formal definition of such named graphs are still debated by the research community, but a consensus about a limited number of options has been reached as described by Zimmermann in [Zimmermann, 2014]; in the following we consider two definitions: named graph and quad semantics.

Named graph: The graph name denotes an RDF graph or a particular occurrence of that graph. An example of named graphs is:

```
ex:g1  { ex:a ex:b ex:c
         ex:d ex:e ex:f  }
```

In the example above there is an RDF graph named "ex:g1" composed of two triples.

Quad semantics: The named graph is considered as a set of quadruples where the first three elements are subject, property and object as usual and the fourth is the name of the graph as shown in the example below where "ex:x" is the name of the graph:

```
ex:a rdf:type         ex:c ex:x .
ex:c rdfs:subClassOf  ex:d ex:x .
```

In general the fourth element can also be used as a statement identifier, a model identifier, or to refer to the "context" of a statement. In the literature, the fourth element has been used to denote a time frame in [Gutiérrez, *et al.*, 2007], to deal with uncertainty in [Straccia, 2009] and to handle provenance in [Carroll, *et al.*, 2005]. In all these cases the fourth element is used with a semantics tailored to the specific need of the application under exam.

In the following we use the fourth element as a triple identifier in order to label statements and use them for building citation graphs.

2.2 Requirements and Existing Systems for Data Citation

The "Joint Declaration of Data Citation Principles" produced by the Data Citation Synthesis Group outlined the main principles of data citation. Leveraging on the insights and considerations outlined by [Altman and Crosas, 2013] and [Ball and Duke, 2012] we point-out four main requirements a *data citation methodology* must fulfill:

1. provide a description of the data in order to give scholarly credit and normative and legal attribution to data creators and curators. Description metadata that would make up a complete citation are open to debate, but there is a certain agreement about the minimum required set, which must contain: *author*, *title*, *date* and *location* (i.e. a persistent reference to the cited data);
2. uniquely identify the cited object and the associated metadata;
3. enable variable granularity data citation (i.e. to cite a dataset as a whole, a single unit or a subset of data);
4. produce references that are both human- and machine-readable.

Many of the existing approaches to data citation allow us to reference datasets as a single unit having textual data serving as

metadata source. As pointed out by [Proll and Rauber, 2013] most data citations "can often not be generated automatically and they are often not machine interpretable."

The rule-based citation system proposed by [Buneman and Silvello, 2010] meets the desired features for data citation because it allows for citing data with variable granularity, creates both human- and machine-readable citations and associates description metadata with the cited data. On the other hand, this system works under the assumption that data are hierarchically structured (e.g. XML files) and thus it cannot be straightforwardly adopted in the context of LOD where we deal with RDF graphs.

[Proll and Rauber, 2013] proposed an approach based on assigning persistent identifiers to time-stamped queries, which are executed against time-stamped and versioned relational databases. While this system also meets the data citation requirements, it is defined for working with relational databases and there is no extension to RDF graphs.

[Groth, *et al.*, 2010] proposed the nano-publication model where a single statement (expressed as an RDF triple) is made citable in its own right; the idea is to enrich a statement via annotations adding context information such as time, authority and provenance. The statement becomes a publication itself carrying all the information to be understood, validated and re-used. The model proposed by Groth et alii is close to the RDF reification process [Klyne, *et al.*, 2014] where we can make claims about an RDF statement; in the nano-publication model a URI is assigned to the statement in order to make it a dereferenceable entity to be used in the RDF graph enriching it. Then, a name is associated to the RDF graph making it citable.

This model is not specifically defined for citing RDF sub-graphs with variable granularity, but it is centered around a single statement and the possibility of enriching it. Nevertheless, in the following we extend and improve this very idea in order to cite RDF graphs satisfying the four requirements outlined above.

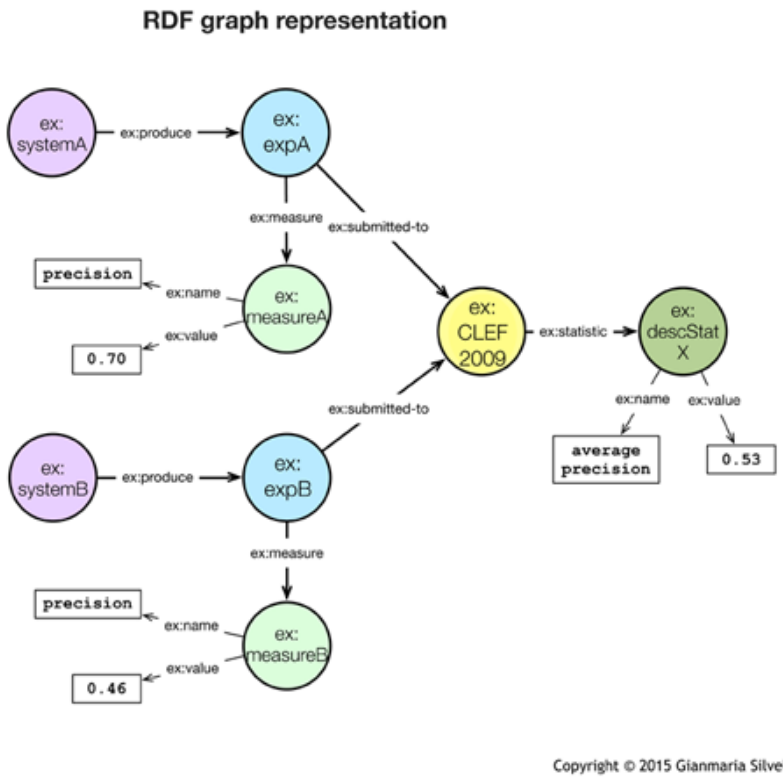
In this context, it is interesting to mention the "Research Objects" initiative which has the aim of bringing together several international activities with the common goal of defining a new approach to publications in order to improve reuse and reproducibility of research. LOD plays a central role in this context and several activities comprised by the Research Objects initiative are based on LOD-related methodologies and technologies — e.g. the Open Archive Initiative for Object Re-Use and Exchange (OAI-ORE), which exploits the RDF Framework for sharing compound objects on the Web. This initiative does not propose a methodology for citing LOD subsets, but it could exploit it within the research objects it defines; from this perspective the methodology proposed here is a companion of research objects rather than an alternative to them.

3 Use case: Search Engines Experimental Evaluation

We present a use case based on experimental evaluation of search engines, which produces scientific data that are highly valuable from both a research and financial point of view [Rowe, *et al.*, 2010]. Experimental evaluation of search engines is a demanding activity that benefits from shared infrastructures and datasets that favor the adoption of common resources, allow for replication of the experiments, and foster comparison among state-of-the-art approaches. Therefore, experimental evaluation is carried out in large-scale *evaluation campaigns* at an international level, such as the Conference and Labs of the Evaluation Forum (CLEF) in Europe and the Text Retrieval Evaluation Conference (TREC) in the USA. The evaluation activities produce huge amounts of scientific and experimental data, which are the foundation for all the subsequent scientific production and development of new systems. For this reason, these data need to be *discoverable*, *understandable* and *citable* [Harman, 2011].

As a consequence, the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) system [Agosti, *et al.*, 2012] has been defined with the aim of modeling the experimental data and developing a software infrastructure able to manage and curate them. The data made available by means of DIRECT have been mapped in RDF with the purpose of exposing them as LOD on the Web in the near future. This will increase the discoverability and the re-use of the experimental data; furthermore, it will enable a seamless integration of datasets produced by different international campaigns as well as the standardization of terms and concepts used to label data across research groups [Ferro and Silvello, 2014].

In Figure 1 we report a portion of the RDF graph and its triple representation showing a sample of experimental data[3] shareable by means of DIRECT. In this case we show two sample systems (system A and system B) which produce two experiments (exp A and exp B) submitted to an evaluation campaign (CLEF 2009); by considering a certain evaluation measure (precision) each experiment achieved a certain value (0.70 for system A and 0.46 for system B). Furthermore, the evaluation campaign is associated with a descriptive statistic indicating that the average precision of all the considered systems is 0.53.



RDF triple representation

Subject	Property	Object
ex:systemA	ex:produce	ex:expA
ex:expA	ex:measure	ex:measureA
ex:measureA	ex:name	`precision`
ex:measureA	ex:value	`0.7`
ex:expA	ex:submitted-to	ex:CLEF2009
ex:systemB	ex:produce	ex:expB
ex:expB	ex:measure	ex:measureB
ex:measureB	ex:name	`precision`
ex:measureB	ex:value	`0.46`
ex:expB	ex:submitted-to	ex:CLEF2009
ex:CLEF2009	ex:statistic	ex:descStatX
ex:descStatX	ex:name	`average precision`
ex:descStatX	ex:value	`0.53`

Figure 1: Sample RDF graph and triple representation of experimental evaluation data.

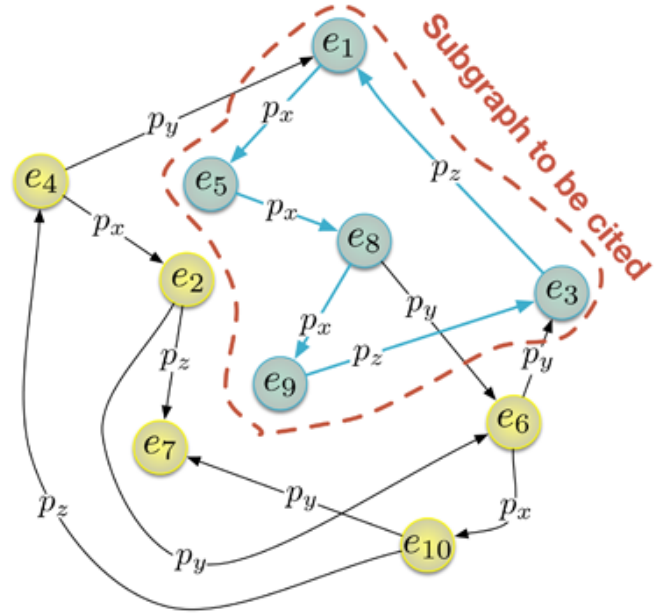
In the information retrieval field it is very common to report the data in Figure 1 in scientific papers with the purpose of describing the experiments, discussing the innovative methods employed and comparing the outcomes of new systems with previously achieved ones.

In this context data citation is fundamental for supporting claims and new knowledge built on these data as well as for giving credit to the researchers and practitioners that developed systems, produced experiments and carried out the evaluation. These data need to be cited with different granularity; indeed, we may need to cite: (a) all the available data about "CLEF 2009" if we are writing a report about the evaluation campaign; (b) a subset of statements such as "System A" participated in "CLEF 2009" achieving a "precision" of 0.7; or (c) a single statement such as "System B" produced "Experiment B".

In addition, we may need to highlight some evidence drawn from the data such as "System A" performs better than "System B" in terms of precision or "System A" performed 24% better than the average system for "CLEF 2009". In these cases, we need to enrich the cited RDF statements with additional data that make new claims clear and easily verifiable by humans as well as machines.

4 LOD Citation Methodology

In Figure 2 we consider a generic RDF graph (and its triple representation) representing a dataset presented as LOD on the Web. We use this generic instance of RDF graph as a guide to describe our citation methodology for LOD subsets.



	Subject	Property	Object
*	e1	px	e5
	e4	py	e1
	e4	px	e2
*	e5	px	e8
*	e3	pz	e1
*	e9	px	e8
*	e9	pz	e3
	e6	py	e3
	e6	px	e10
	e10	py	e7
	e10	pz	e4
	e8	py	e6
	e2	pz	e7
	e2	py	e6

RDF graph representation of the dataset

RDF triple representation

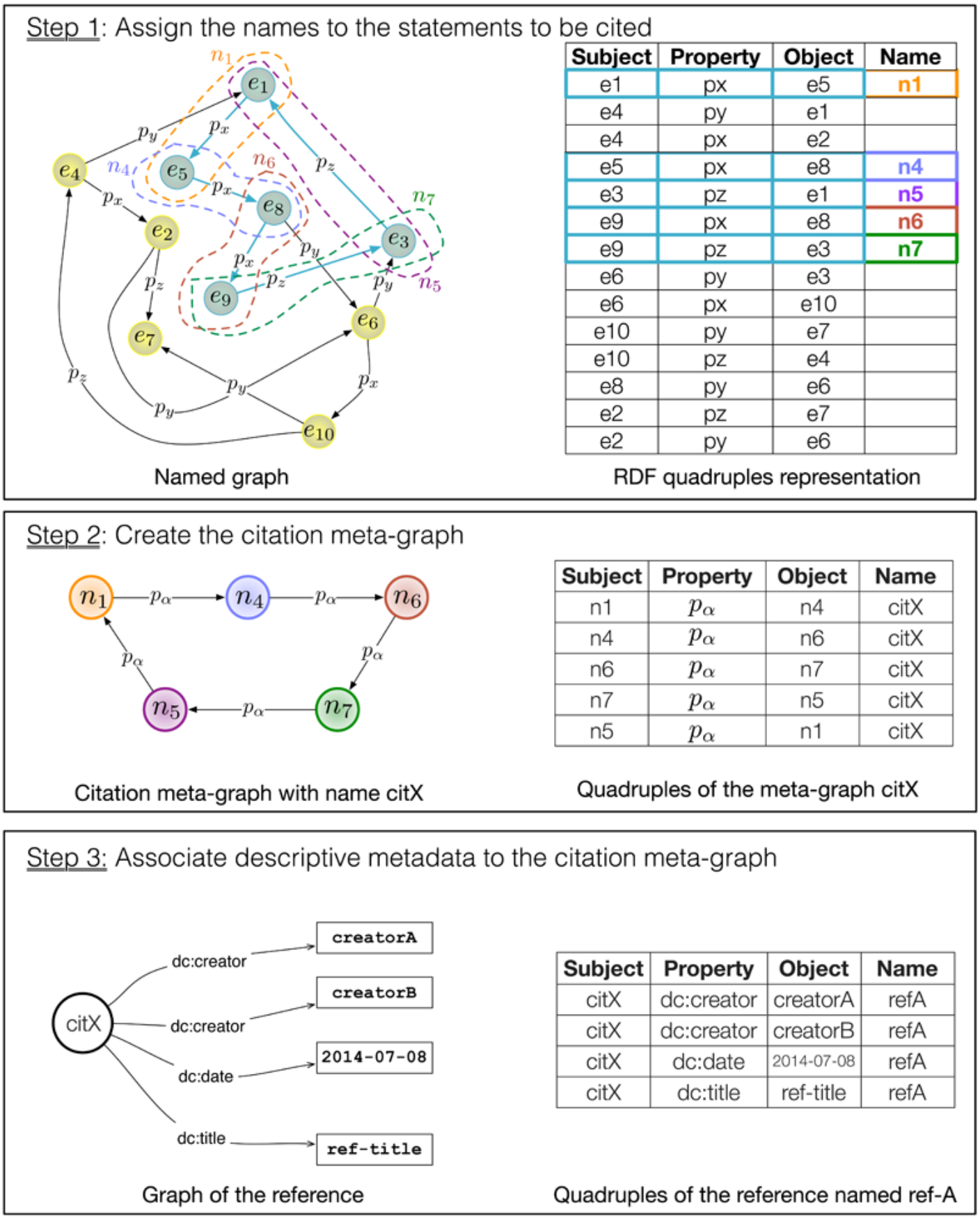
Figure 2: A representation of a generic RDF graph and the corresponding set of triples.

The citation methodology we present satisfies the requirements discussed in Section 2 and can be outlined as a three-step procedure (see Figure 3 for a graphical representation based on the generic RDF graph shown above); so, given a LOD dataset composed by some statements (i.e. RDF triples) to cite a subset of statements, we:

1. assign a *name* to each statement to be cited;
2. build a *citation meta-graph* relating the cited statements to one another;
3. build a *reference named graph* by enriching the citation meta-graph with description metadata.

The first step exploits RDF quad semantics that allows us to assign a name to every statement in an RDF graph by transforming it into a set of quadruples; as shown in Figure 3 the names can be associated only to the statements to be cited.

Once that the statements are named, they are dereferenceable entities that can be used as subjects and objects in a newly defined RDF graph; in this context, the new graph created from the named statements is called "*citation meta-graph*" because it is an RDF dataset describing the original graph — i.e. a graph describing another graph. In Figure 3 we can see that a generic property p_α is used to relate the named statements to one another; the methodology proposed is agnostic to the choice of this property that can be set manually by the creator of the citation or automatically by a system assigning the same property to each statement. For instance, a property that can be automatically set to relate one named statement with another is the property "is-related-to" of the *schema.org* vocabulary as we show below.



Subject	Property	Object	Name
e1	px	e5	n1
e4	py	e1	
e4	px	e2	
e5	px	e8	n4
e3	pz	e1	n5
e9	px	e8	n6
e9	pz	e3	n7
e6	py	e3	
e6	px	e10	
e10	py	e7	
e10	pz	e4	
e8	py	e6	
e2	pz	e7	
e2	py	e6	

Subject	Property	Object	Name
n1	p_α	n4	citX
n4	p_α	n6	citX
n6	p_α	n7	citX
n7	p_α	n5	citX
n5	p_α	n1	citX

Subject	Property	Object	Name
citX	dc:creator	creatorA	refA
citX	dc:creator	creatorB	refA
citX	dc:date	2014-07-08	refA
citX	dc:title	ref-title	refA

Figure 3: The three-step methodology for citing LOD subsets.

The citation meta-graph allows us to establish explicit connections between the statements to be cited and to set a name (i.e. an IRI) that uniquely identifies the cited set of statements as a whole. The properties connecting the statements in the meta-graph define a path through the RDF triples that can be used both by humans and machines to interpret and understand the cited data.

The citation meta-graph is associated with a name — for this reason we may want to talk about citation named meta-graph. A

possible serialization of the citation *named* meta-graph, which follows the W3C recommendation described above and an alternative to the representation given in Figure 2, is the following:

citX { n1 p_a n4
n4 p_a n6
n6 p_a n7
n7 p_a n5
n5 p_a n1 }

In the final step of the methodology, which relates the name of the citation meta-graph with descriptive metadata, creates a reference named graph; the minimum set of metadata is composed of the creator of the dataset, a date, and a title[4]. In Figure 3 we can see that with the last step of the methodology we obtain a graph with name "ref-A" which is an IRI that uniquely identifies and locates the reference and the cited data. The serialization of the reference named graph is shown in Figure 3 as follows:

refA { citX dc:creator creatorA
citX dc:creator creatorB
citX dc:date 2014-07-08
citX dc:title ref-title }

This citation methodology respects all the requirements outlined above; indeed, it creates a reference bearing the minimum required set of description metadata, but can be easily customized, uniquely identifies (by means of an IRI) the cited object and the associated metadata, allows us to create citations with variable granularity since we can cite a single resource, a single statement or a set of statements as big as the whole dataset, and creates a reference which is both human- (i.e. a serialization of the reference graph shown at step 3 of Figure 3) and machine-readable.

5 Applications of the citation methodology

Let us see how the outlined methodology can be applied to the use case presented above. In Figure 4 we show how a reference to a LOD subset in a scientific paper may look; we can see that the reference is human-readable, but activating the IRI — i.e. the IRI corresponding to the name of the reference graph — makes it possible to access the machine-readable RDF data. In this case we report the RDF graph and its N-Quads serialization [Carothers, 2014].

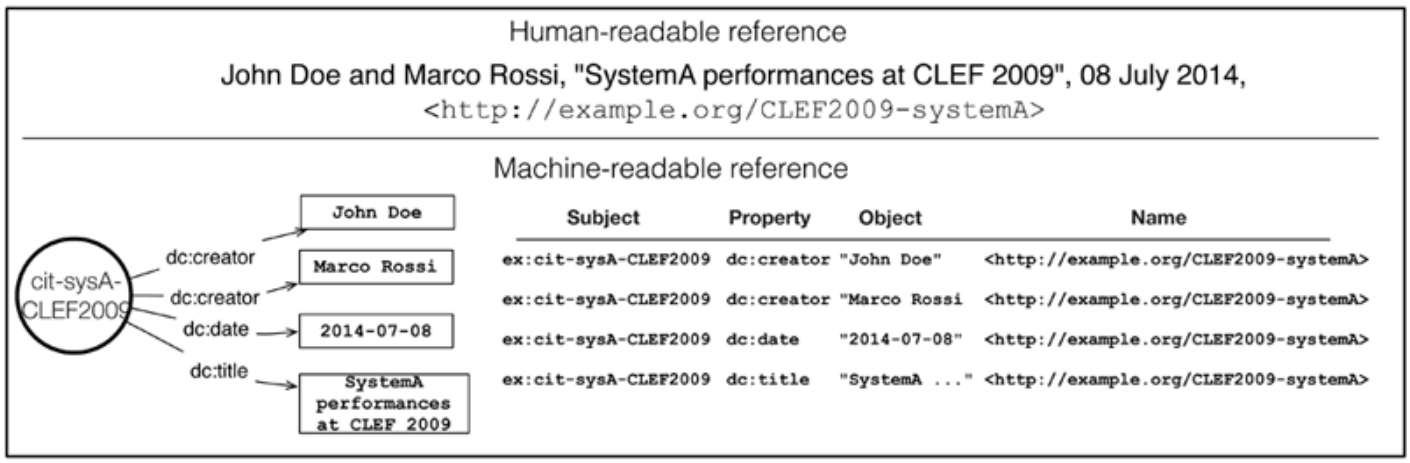


Figure 4: A human-readable reference to a LOD subset and its machine-readable translation.

In Figure 5 we see how the reference graph with name "http://example.org/CLEF2009-systemA" leads us to the citation meta-graph named "ex:cit-sysA-CLEF2009" and then to the original cited LOD subset.

As we can see the reference IRI reported in Figure 4 is a unique and persistent (according to the LOD paradigm) pointer that allows us to retrieve all the necessary information about the cited data.

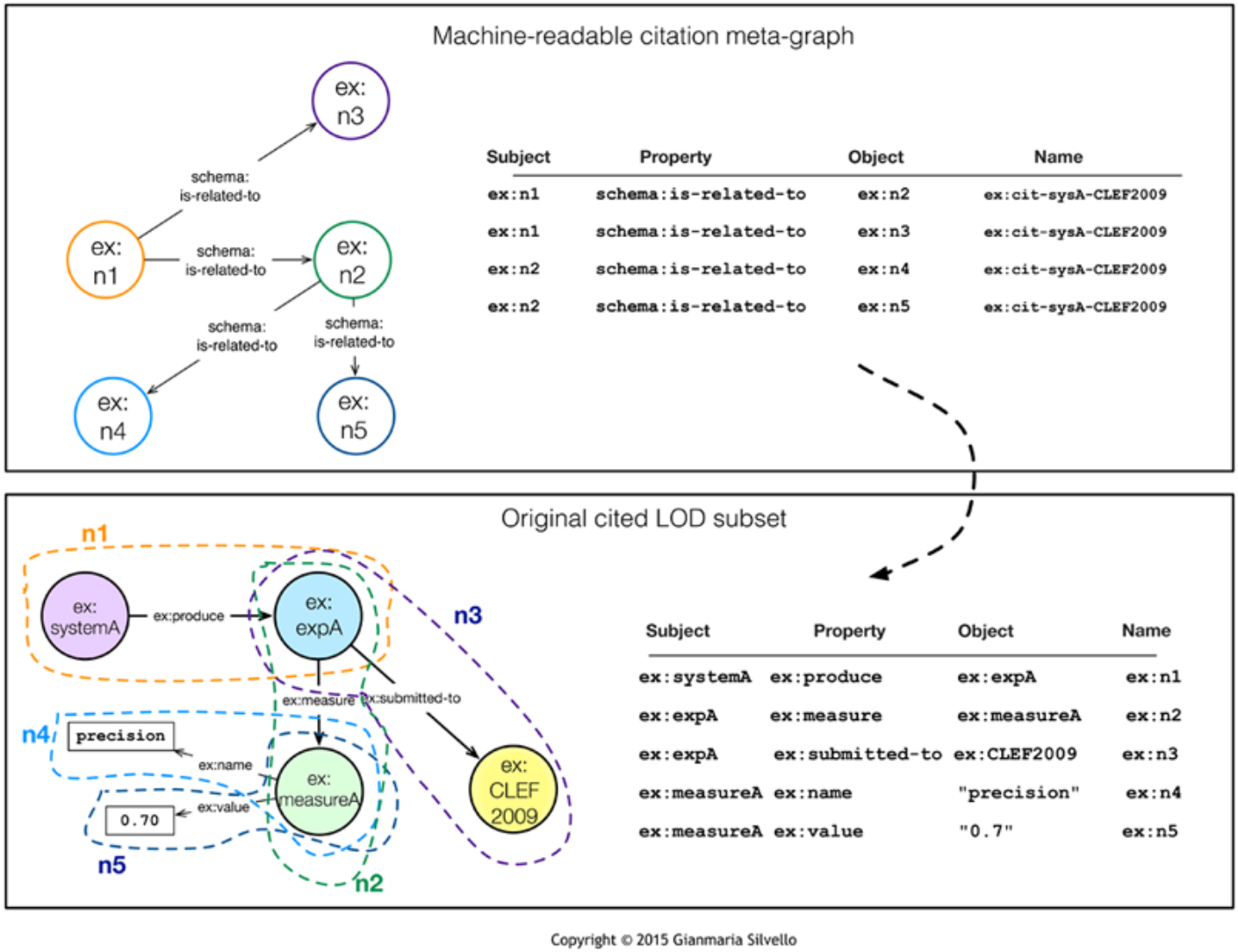


Figure 5: A citation meta-graph drawn from the use case and the correspondent cited LOD subset.

Citation meta-graphs can also be used to report inferred information from the cited LOD subset; for instance, we can build a meta-graph for supporting the following claim:

> ***Claim 1:*** "precision of system A is 24% higher than the average precision of systems which participated in CLEF 2009"

Claims like this are frequent in the information retrieval literature because experimental evaluation is a cornerstone of the field and the reported experiments about a new system or method must be compared with previous experiments, which are often carried out in international evaluation campaigns such as CLEF. Furthermore, evaluation campaigns produce papers with high scientific impacts (e.g. papers receiving many citations) called "overview papers" which report about the experiments conducted in a campaign; these papers are built around claims based on experimental data like: "The best performance for the Indian sub-task is 76.12% of the best bilingual English system and 67.06% of the monolingual baseline" [Di Nunzio, *et al.*, 2007].

Citation meta-graphs such as the one reported in Figure 5 can support such claims because every single statement composing the claim can be linked with a reference to the raw data presented as LOD on the Web, thus providing evidence for verifying the claim itself. On the other hand, these claims are the results of elaborations (e.g. statistical analyses) of the raw data and thus are not easily verifiable for a human or a machine interpreting the RDF data.

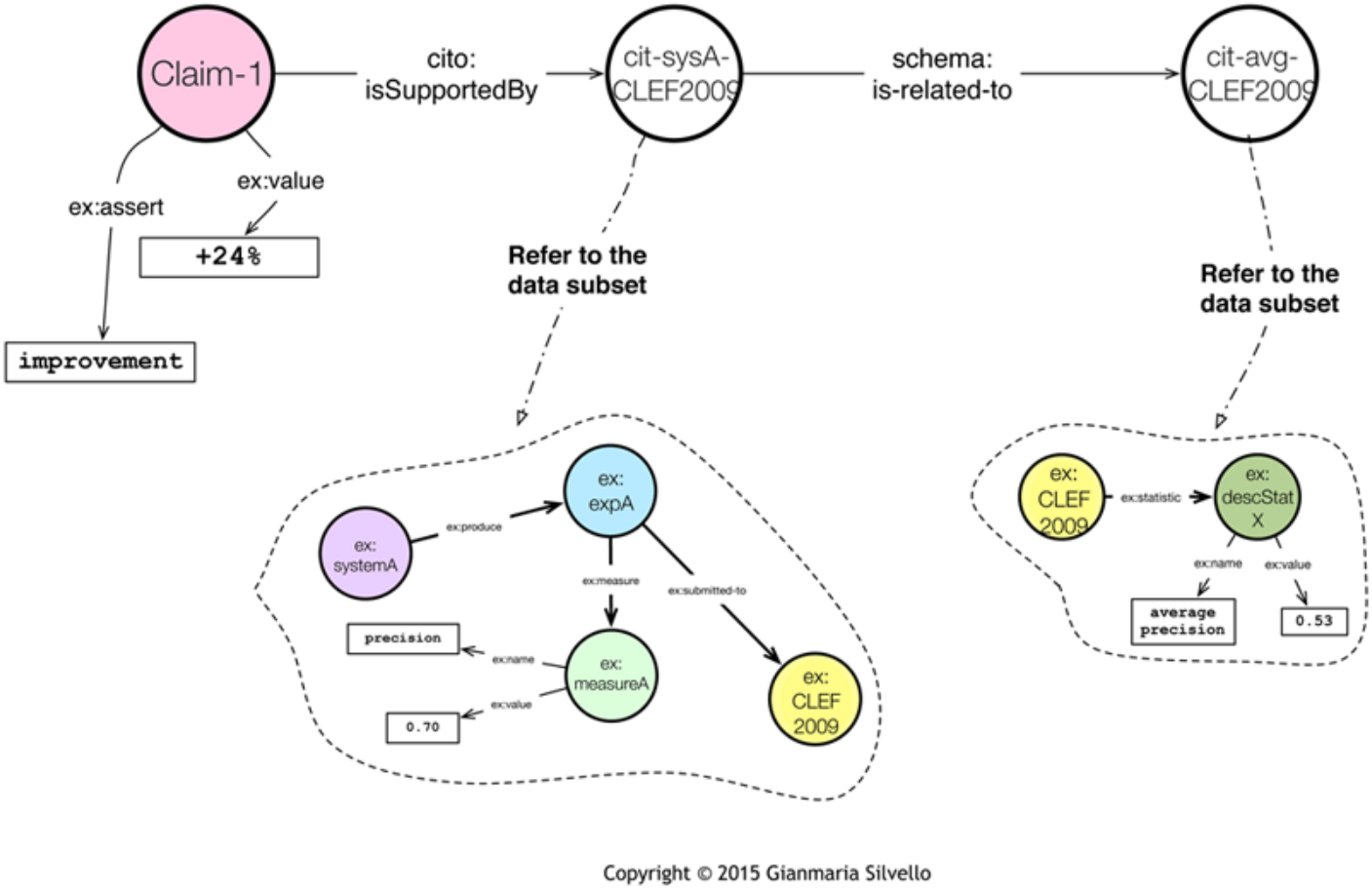


Figure 6: Machine-readable combination of citation meta-graphs supporting a claim.

A combination of citation meta-graphs can be exploited for creating human- and machine-readable references supporting complex claims built on raw data as shown in Figure 6.

We can see that "Claim 1" reported above can be associated to the RDF graph shown in Figure 6. This graph is composed of a dereferenceable entity called "claim-1" representing the claim itself and in this case it is related to two literals providing data corroborating it. The claim entity is related via the "isSupportedBy" property of the CiTO ontology to the citation meta-graph referring to the raw data about system A, which is required for verifying the first part of the claim. This meta-graph is related to a second citation meta-graph, which refers to the descriptive statistics of CLEF 2009 (i.e. another data subset) necessary to verify the second part of the claim. Basically, the graph connecting these two meta-graphs creates an explicit link between two data subsets and adds information to support "Claim 1"; both a human and a machine can easily interpret this graph and verify the claim by analyzing how the subsets are related to each other.

6 Final Remarks

In this paper we presented a simple methodology, which by exploiting the LOD paradigm enables persistent, dereferenceable, variable granularity and human- and machine-readable citations of LOD subsets. We reported a compelling use case based on search engine experimental evaluation data showing why it is necessary to cite these data and how this could be done with the presented methodology. Furthermore, we outlined a possible extension of the methodology, which can be used to sustain complex claims built on experimental data.

The presented methodology along with the preponderant role of LOD for data publishing, sharing and knowledge creation could set the ground for a factual integration between scientific papers and the data on which they are based. Furthermore, the proposed methodology provides a concrete means to (automatically) verify scientific claims based on raw data. We believe that this may represent a further step towards the steady employment of "enhanced publications" allowing us to process publications and related objects together [Vernooy-Gerritsen, 2009].

Future works will concern the implementation of the citation methodology in the DIRECT system in order to provide a concrete means for citing experimental evaluation data and verifying scientific claims. Furthermore, we want to investigate how data citation can be exploited for estimating the scientific impact of evaluation campaigns, an activity that currently is mostly based on bibliometrical indicators based on scientific paper citations [Tsikrika, *et al.*, 2014].

Notes

[1] See Tim Berners-Lee, Design Issues, Linked Data.

[2] IRIs (Internationalized Resource Identifiers) are a generalization of URIs that permits a wider range of Unicode characters [Klyne, *et al.*, 2014].

[3] For sake of understandability we present a reduced and over-semplified RDF graph reporting sample experimental evaluation data.

[4] The properties are derived from the DataCite Ontology which defines a list of core metadata properties chosen for "the accurate and consistent identification of a resource for citation and retrieval purposes".

References

[1] M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo and G. Silvello. "DIRECTions: Design and Specification of an IR Evaluation Infrastructure", in *Proc. of the Third International Conference of the Cross-Language Evaluation Forum, CLEF 2012*, In Lecture Notes in Computer Science 7488, pp. 88—99, Springer, 2012.

[2] M. Altman and M. Crosas. "The Evolution of Data Citation: From Principles to Implementation". *IASSIST Quarterly*, Spring, pp. 62—70, 2013.

[3] S. Auer, T. Dalamagas, H. Parkinson, F. Bancilhon, G. Flouris, D. Sacharidis, P. Buneman, D. Kotzinos, Y. Stavrakas, V. Christophides, G. Papastefanatos, and K. Thiveos. "Diachronic linked data: towards long-term preservation of structured interrelated information". In *Proc. of the First International Workshop on Open Data (WOD '12)*. ACM, New York, NY, USA, 31—39, 2012.

[4] A. Ball and M. Duke. "How to Cite Datasets and Link to Publications". DCC How-to Guides. Edinburgh: Digital Curation Centre, 2012.

[5] C. L. Borgman. "Why are the attribution and citation of scientific data important? In: Uhlir, Paul and Cohen, Daniel (eds.). Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop." National Academy of Sciences' Board on Research Data and Information. National Academies Press: Washington DC, 2012.

[6] C. L. Borgman. "The Conundrum of Sharing Research", *Journal of the American Society for Information Science and Technology*, 63(6):1059—1078, 2012.

[7] P. Buneman. "How to cite curated databases and how to make them citable". In *Proc. of the 18th International Conference on Scientific and Statistical Database Management (SSDBM)*. Vienna, Austria: IEEE Computer Society, 2006.

[8] P. Buneman and G. Silvello. "A Rule-Based Citation System for Structured and Evolving Datasets". *IEEE Bulletin of the Technical Committee on Data Engineering*, 3(3): 33—41, 2010.

[9] G. Carothers. "RDF 1.1 N-Quads: A line-based syntax for RDF datasets", W3C Recommendation, 25-Feb-2014.

[10] J. J. Carroll, C. Bizer, P. J. Hayes and P. Stickler. "Named Graphs, Provenance and Trust". In *Proc. of the 14th Int. Conf. on World Wide Web, WWW 2005*, ACM Press, pp. 613—622, 2005.

[11] G. M. Di Nunzio, N. Ferro, T. Mandl and C. Peters. "CLEF 2007: Ad Hoc Track Overview". In *Advances in Multilingual and Multimodal Information Retrieval: Eighth Workshop of the Cross-Language Evaluation Forum (CLEF 2007)*, C. Peters et al. eds. Revised Selected Papers, pages 13—32. *Lecture Notes in Computer Science (LNCS)* 5152, Springer, Heidelberg, Germany, 2007. Springer version here; Free working copy version here.

[12] N. Ferro and G. Silvello. "Making it Easier to Discover, Re-Use and Understand Search Engine Experimental Evaluation Data." *ERCIM News*: 96, 2014.

[13] P. Groth, A. Gibson, and J. Velterop. "The anatomy of a nanopublication". *Information Services and Use* 30, 1—2 (January 2010), 51—56, 2010.

[14] C. Gutiérrez, C. A. Hurtado, A. A. Vaisman, "Introducing Time into RDF", *IEEE Transactions on Knowledge and Data Engineering* 19 (2) pp. 207—218, 2007.

[15] D. K. Harman. "Information Retrieval Evaluation", Morgan & Claypool Publishers, USA, 2011.

[16] T. Heath and C. Bizer. "Linked Data: Evolving the Web into a Global Data Space". *Synthesis Lectures on the Semantic Web*. Morgan & Claypool Publishers, 2011.

[17] HORIZON 2020, Work Programme 2014-2015, Information and Communication Technologies, Leadership in enabling and industrial technologies, European Commission Decision C (2013)8631 of 10 December 2013).

[18] G. Klyne, J. J. Carroll and B. McBride. "RDF 1.1 Concepts and Abstract Syntax". W3C Recommendation, 25-Feb-2014.

[19] B. Lawrence, C. Jones, B. Matthews, S. Pepler, S. Callaghan. "Citation and Peer Review of Data: Moving Towards Formal Data Publication". *The International Journal of Digital Curation*, 6(2): 4—37, 2011. http://doi.org/10.2218/ijdc.v6i2.205

[20] S. Proll and A. Rauber. "Scalable data citation in dynamic, large databases: Model and reference implementation", *IEEE International Conference on Big Data*, pp. 307-312, 2013. http://dx.doi.org/10.1109/BigData.2013.6691588

[21] B. R. Rowe, D. W. Wood, A. L. Link, D. A. Simoni. "Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program2, RTI International, USA, 2010.

[22] U. Straccia. "A Minimal Deductive System for General Fuzzy RDF". In *Proc. of Web Reasoning and Rule Systems, Third International Conference, RR 2009*, Springer, LNCS 5837, pp. 166—181, 2009.

[23] T. Tsikrika, B. Larsen, H. Müller, S. Endrullis and E. Rahm. "The Scholarly Impact of CLEF (2000—2009)", in Proc. of the 4th International Conference of the CLEF Initiative, CLEF 2013, Forner et al. eds., *Lecture Notes in Computer Science* 8138, pp. 1—12, Springer, Berlin Heidelberg, 2013.

[24] M. Vernooy-Gerritsen. "Enhanced Publications: Linking Publications and Research Data in Digital Repositories", Amsterdam University Press, 2009.

[25] A. Zimmermann. "RDF 1.1: On Semantics of RDF Datasets". W3C Working Group Note, 25-Feb-2014.

About the Author

Gianmaria Silvello took the master degree in Computer Engineering, University of Padua in 2006, a post-graduate master on Design, Management and Conservation of Public and Private Digital Archives in 2007 and the Ph.D. in Information Engineering from the Doctorate School in Information Engineering of University of Padua in 2011. Since 2011, Dr. Silvello is post-doc researcher at University of Padua. Since 2006 he has been working on the design and development of a digital archive system called SIAR (Regional Archival Information System) in cooperation with the Italian Veneto Region and the Archival Supervising Office for the Italian Veneto Region of the Italian Ministry of Cultural Heritage. Since 2010 he has been working on the field of Information Retrieval Evaluation with a specific focus on effectiveness measures. His main interests are Digital Libraries and Archives, Information Retrieval Evaluation and Models and Technologies for the Web of Data.

PRINTER-FRIENDLY FORMAT Return to Article