Reasoning based on property propagation on CIDOC-CRM and CRMdig based repositories

Katerina Tzompanaki[1,2], Martin Doerr[1], Maria Theodoridou[1], Irini Fundulaki[1]

FORTH – Institute of Computer Science
N. Plastira 100 Vassilika Vouton
GR-700 13 Heraklion, Crete, Greece

[1]{katetzob, martin, maria, fundul}@ics.forth.gr
[2]{tzompana}@lri.fr

Abstract. Reasoning on provenance information and property propagation is of significant importance in e-science since it helps scientists manage derived metadata in order to understand the source of an object, reproduce results of processes and facilitate quality control of results and processes. In this paper we introduce a simple, yet powerful reasoning mechanism based on property propagation along the transitive part-of and derivation chains, in order to trace the provenance of an object and to carry useful inferences. We apply our reasoning in semantic repositories using the CIDOC-CRM conceptual schema and its extension CRMdig, which has been develop for representing the digital and empirical provenance of digital objects.

Keywords: Semantic networks, information access, semantic search, metadata, reasoning, provenance

1 Introduction

Over the last decade, semantic repositories that integrate heterogeneous data sources under semantic schemata such as *ontologies* have become an important component of the Semantic Web. These repositories usually support limited forms of *reasoning* that are used to infer *implicit knowledge* along subsumption relationships. Large-scale metadata repositories, i.e., semantic networks of RDF [1] triples integrating large amounts of data, have been developed and are globally accessible via the Internet. The list of such projects about cultural-historical data is long, including the Europeana[2], cultureSampo[3], German Digital Library[4], ResearchSpace[5], WISSKI[6], and

[1] http://www.w3.org/RDF/
[2] http://www.europeana.eu/
[3] http://www.kulttuurisampo.fi/?lang=en
[4] http://www.deutsche-digitale-bibliothek.de/
[5] http://www.researchspace.org

CLAROS[7] Projects. Linked Open Data[8] are advocated for cultural institutions, in which RDF data reside on local servers, and are accessible under published RDF schemata. In these systems, the CIDOC-CRM[9] [1] is becoming more and more popular as a rich RDF schema adequate to integrate complex cultural data.
These semantic repositories naturally follow the "Open World Assumption", where knowledge is regarded as *incomplete* since metadata may be created by different people who state different facts about the same artifact and even may use the schema in different albeit correct ways. For instance, someone may say that an artifact is from Athens and someone else that the same artifact is part of the Parthenon Frieze in London. Thus establishing the correlation among information coming from multiple sources or even the same source becomes a necessity but simultaneously a great challenge. As in any Open World system, also in cultural heritage semantic repositories, users cannot know precisely what has been documented and how. So, while searching in the metadata they may ask for *implicit* knowledge like:

- characteristics (properties) of artifacts that have been recorded somewhere in the semantic network but are not directly associated to the object of interest [2]. For instance, the material from which an object is made of is recorded for the object parts and not for the object itself.
- characteristics that have multiple modeling alternatives. For instance, the "place of origin" of an object may be perceived as anything like its (a) place of creation, (b) place of discovery, (c) place of use and/or (d) creator's birthplace
- characteristics that are generalizations of sets of more specific properties. For instance, the *"has met"* property [3-4] denotes the symmetric relation among items and people that were present in the same event, including time intervals and places. More specifically the *"has met"* property can be considered as the super-property of many properties, such as *"carried out by"* or *"used"* and their inverse ones.

In this paper we introduce a simple yet powerful reasoning mechanism based on inference and completion of metadata, as a means to help scientists query a semantic repository in order to trace and understand the source of their results, to reproduce results and to ease quality control of results and processes. Generalization and inferring of metadata from related objects is achieved by using the *propagation* of some object properties along the transitive part-of and derivation chains of information. We base our reasoning on a semantic repository which uses CIDOC-CRM[10] and its extension CRMdig[11] [5-6] appropriate for representing provenance. The implementation of this mechanism is feasible and indeed simplifies the querying process of scientists upon complex semantic repositories in the cultural heritage field and beyond [4]. The described framework has been applied in the framework of the European IP 3D-

---

6 http://wiss-ki.eu
7 http://explore.clarosnet.org
8 http://linkeddata.org
9 http://www.cidoc-crm.org/
10 CIDOC CRM v5.0.4 Encoded in RDFS. http://www.cidoc-crm.org/rdfs/cidoc-crm
11 CRMdig 3.0 Encoded in RDFS. http://www.ics.forth.gr/isl/CRMext/CRMdig.rdfs

COFORM[12], funded by the European Community (FP7/2007-2013, no 231809). In this project metadata describing the digital provenance for empirical 3D modeling and digitization processes are recorded along with metadata about the physical objects. Digital provenance data form deep chains of events connected by input-output, with up to tens of thousands of intermediary products that "inherit" many properties along the processing chains up to data about the digitized objects themselves. Using reasoning rules, we result in high recall rates, as not only explicitly documented properties but also derived properties across independently created metadata records can be combined for calculating the desired results, as long as referential integrity along these chains is preserved. In parallel, the Research Space project has also implemented this approach following our model.
This paper is organized as follows: we first review related work in Sec. 2 before introducing the reader to the problem in Sec. 3; Sec. 4 describes our approach; conclusions are provided in Sec. 5.

2 Related work

Data provenance is one kind of metadata that can be used to answer basic questions such as "*who created this artifact?*", "*where and when was this artifact created?*", "*when was this artifact modified and by whom?*" [7]. Provenance can support a large number of applications [8]: (a) *data quality & reliability*, (b) *audit trail* (c) *replication recipes* and (d) *attribution*. Provenance information can be used to determine the use of resources, to detect errors in data generation, that is to provide an *audit trail* for the data. Repeatability of experiments is an essential problem in scientific data management. Having fine grained provenance information about the processes used to create a data product, allows one to *replicate* the results of experiments in order to verify or debate scientific results. Knowing the author/creator of an artifact allows one to determine the ownership of data and hence liability in the case of errors (*attribution*) [9]. The problem of storing, accessing, and querying provenance has received a lot of attention in the last years. Research has focused in the areas of workflow and database systems which deal with different levels of provenance granularity regarding the type of data collected about a specific product (a data product or the result of a process).

1. ***Workflow systems***: A workflow can be a *process* (a series of steps that leads to the creation of a real world artifact) or a program (e.g., a series of computations that produce a data item). The provenance of a workflow *(coarse-grained provenance)* can be thought of as the entire history of the derivation of the result of the process [7], [10]. The information stored for the specific process can include the *different versions* of the *software* and the *hardware* used, the *agents* that were involved in the workflow chain (processes, human agents) and the *"things"* (e.g. data) employed by the processes. The ability to query the provenance of workflows allows users to explore and better understand results and enables knowledge re-use [7]. A large number of work-

[12] http://www.3d-coform.eu/

flow provenance models have been developed to represent provenance such as OPM [11], Provenir Ontology [12] and latest the W3C Recommendation Provenance Ontology (PROV-O) [13]. OPM and Provenir represent information of computational processes only, whereas PROV-O models provenance information that is generated by different systems and exchanged under different contexts.

2. ***Database systems***: At the other end of the spectrum, *data provenance (fine-grained provenance)* provides a detailed trace of how a piece of data has been obtained from a transformation process (i.e. query) [10]. Data provenance may indicate (a) the tuples involved in the computation of a result tuple (*why-provenance*) (b) where these tuples reside (*where-provenance*) (c) the *query operators* used to obtain the result tuple (*how provenance*) [14]. The above types of provenance have been extensively studied for relational databases and only recently for Linked Data [15].
Despite the research that has been conducted in the above topics there has been no explicit approach developed for representing and reasoning about provenance along the transitive part-of and derivation chains. The above approaches deal only with computational processes on digital artifacts whereas in our approach we are able to reason combining metadata of real world objects with metadata of digital objects and to deduct useful inferences with multiple applications such as maintenance of repositories of digitization products and completion of metadata by implicit knowledge, in applications where production chains comprise thousands of intermediates and dozens of final products without need to manage this redundancy in the repository explicitly.

3 The problem

It is quite common that a user might be interested in a property that is not explicitly documented for the object, but can only be implicitly inferred from related data. For instance, someone may search for things "*made from: steel*", when objects may have been registered as *having parts* (using the "*is composed of*" property) that are "*made from: steel*". From this part-of property chain, we can deduct that the "whole" object is also made from steel, because it has parts made from steel. Moreover, the information may be represented in a different way than the one the user expects, for example instead of "*made from: steel*", objects may be defined with "*has type: steel object*". As the making of CIDOC-CRM demonstrated, it is impossible to normalize a global model for information integration to one unique representation for each property. Rather, in aggregation systems and the Semantic Web, one has to accept that properties are represented by sets of reasonable alternatives that can be related to each other by deductions.
The more analytical and precise a global model is, the less obvious it is for the user how a simple, intuitive question relates to the ontology. Transitive properties (such as parts of parts or derivatives of derivatives) cause "propagation" [16] of properties along those property paths. Propagation may be very complex to formulate as query, but is also very powerful when it comes to query recall improvement. For instance, one could assume that the actors, place and time that are reported for the building of

Parthenon (the "super-event") also apply for or include the building of its friezes (a "sub-event"); or that materials a frieze is made of, are considered to be among the materials the whole Parthenon is made of; or that the subjects a frieze represents also apply to its copies or derivatives, etc. Such reasoning allows for inferring facts that are not stated within a single metadata record. Take for example the following information taken from two different sites. On one hand, we have the British Museum[13] website saying that the object with the description "Horsemen from the west frieze of the Parthenon" is part of the Parthenon, and on the other hand, there is the Acropolis Museum[14] stating that Parthenon was created by Pheidias. Using the CIDOC-CRM schema (prefixed with "crm") the metadata describing these pieces of information are:

- "Horsemen from the west frieze of the Parthenon" *crm: forms part of* "Parthenon"
- "Parthenon" *crm: was produced by* "Construction of Parthenon" *crm: carried out by* "Pheidias"

Using reasoning on the integrated metadata we could infer that Pheidias was involved in the making of the Horsemen as well. In other words, in a query about the maker of the Horsemen, Pheidias would be deducted as a plausible answer. Thus, flat queries that do not take into account such inferences are more likely to have poor or even empty results. In another perspective, metadata built without including such inference rules, provide poorer knowledge. Such inference takes advantage of the transitivity property of *crm: forms part of* and *crm: carried out by* [2] and combined with application dependent relevance criteria can improve significantly the query results in specific application domains.
In provenance data, property propagation along part-of hierarchies can be observed between complex processes and their individual actions, between measurement devices and their components, between digital products and their parts. It must clearly be understood that virtually **none** of these inferences holds in a strictly logical sense. There is a **likelihood** for instance that the same lense of my camera was used throughout an image capture if not stated otherwise. Therefore all inferences we describe increase **recall** with respect to the documented reality**,** even though the mechanism is not an information retrieval technique. Assessing the respective probabilities is not the target of this paper and may be due to future work.
In the next section we propose a framework that utilizes rules to derive useful deductions about transitive properties, based on property propagation in cultural heritage semantic networks.

4 Reasoning using provenance information

Up to this point, we have discussed the necessity of a mechanism to reason upon complex structured metadata. In this section we propose such a mechanism that takes

[13] http://www.britishmuseum.org/
[14] http://www.theacropolismuseum.gr/en

advantage of the property propagation along transitive derivation and part-of chains, in order to derive useful inferences. Our priority is to improve query recall and resolve relevance issues with additional application specific constraints. To help the user understand the meaning and practical usefulness of the framework, we present it in the context of exploiting semantic networks and completing metadata. For this reason, we also include a set of real research questions from the Cultural Heritage domain that have been analyzed in terms of queries in the 3D-COFORM project metadata repository that consists of a semantic network containing rich cultural information [17] and supports the study of such research topics. Here we show that they can be answered easily with semantic, associative queries that make use of the proposed rules. The 3D-COFORM metadata repository consists of 1M RDF triples and is the result of over one year of intensive work, testing and validating the semantic reliability regarding the inference results of our conceptual modeling. We used the BigOWLIM reasoner and query optimization was achieved by implementing shortcuts for certain paths and defining specific reasoning rules. The proposed approach is also studied and validated in fields such as geology and biology.

Assuming that the reader is familiar with the basic semantic web notions, we attach to each query its graphical representation using terms from the CIDOC-CRM that adopts the following notation: Boxes represent classes, the upper part of which is the name of the CIDOC-CRM class (orange) or CRMdig class (blue) and the lower part is the value of an instance of that class, either fixed or represented by a variable. Arrows connecting two boxes denote properties between the two respective classes, and the name of the property is printed over the arrow. Variables are represented with the letters X, Y, Z, U, V, W and denote any node of the metadata graph fitting the respective path. Query parameters include terms, numbers, dates, and strings. The variables that are returned by the query are denoted with variables prefixed with '$', e.g. *$Material*, *$Monument*, *$Height*. We are now ready to introduce the first rule, which is based on the transitivity of properties in *part-of* chains.

Rule 1: *The property of an object is the aggregation of the explicitly defined property in the object itself and the respective properties of all its subparts.*

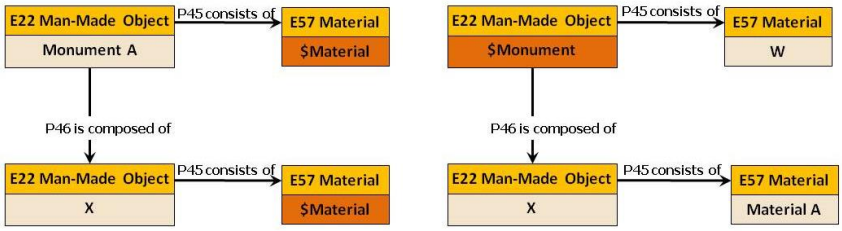

Fig. 1. Forward and backward traversal of the *part-of* chain

According to Rule 1, we can do reasoning by traversing the *part-of* chain either forward or backward (Fig. 1) and we can answer queries such as:

1. Find the material of Monument A: The material of Monument A includes its explicitly stated materials but also the materials of its parts. The query will forward traverse the *part-of* chain and collect all the Materials that have been registered both to Monument A and its parts.

2. Find Monuments constructed from Material A: The information regarding the material of an object might be registered in its parts and not directly in the object itself. So the query should search both the explicitly stated materials of the object and the materials of its parts too.

Fig. 2 presents an example of a monument which is composed of four subparts made of different materials. The object (statue of Queen Victoria[15]) does not have material information in its immediate, explicitly defined metadata but its subparts do have. Our reasoning approach will include this object in the answer set of the query "*Find all statues made of Bronze*" whereas queries relying only on explicitly defined metadata, would fail to retrieve it. Similarly, with our approach, the answer set of the query "*Find the material of the Queen Victoria Monument*" is {Grey granite, Grey marble, Bronze} while the traditional query would get an empty answer set. Using property propagation results in high recall rates however a statistical factor that may deteriorate precision is introduced, since a property is not necessarily propagated along a path or it's significance is not important. For example consider the case of The Kissing Bridge[16] sculpture, which is composed of, (i) two bases made of concrete, and (ii) two statues made of bronze. The significant information in this case is that the statues are made of bronze. Our reasoning approach will influence recall since we will infer that the Kissing Bridge sculpture is made of concrete and bronze. Precision can be improved by adding constraints on the queries.

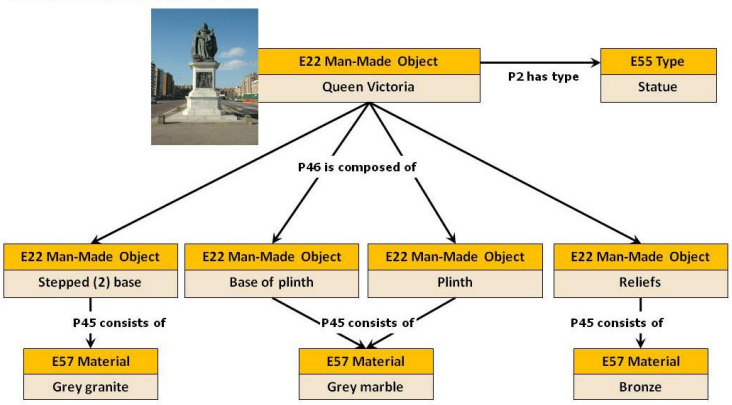

Fig. 2. An example of Queen Victoria statue metadata

[15] Public sculptures of Sussex http://www.publicsculpturesofsussex.co.uk/object?id=71
[16] Public sculptures of Sussex http://www.publicsculpturesofsussex.co.uk/object?id=127

Except from the part-of chains, the derivation chains can also be used for transfer of properties among material and immaterial objects. More specifically, CRMdig Digitization Process class marks property transfers from physical to digital objects while CRMdig Formal Derivation class marks property transfers from digital to digital objects [6]. We make the assumption, that the transformation of a physical object to its digital representation is achieved through "subject preserving" events, which means that the physical object depicted in the derivatives remains the same as the one in the derivation source. Based on this principle, we proceed to our second rule below.

Rule 2: *Physical objects may share properties with their digital representations and their derivatives.*

According to Rule 2, we can do reasoning by traversing the derivation chain (Fig. 3) either forward or backward and we can answer queries such as:

1. Find objects that depict Actor A: Physical Object A has an explicit declaration of the depicted Actor A. This property is propagated to the digital representations of Object A and thus we can infer that all Data Objects (X, Y, … Z) depict Actor A.

2. Find the size of Object A: An object's 3D model may have the size of the object automatically calculated and stored in its metadata. This property is backwards propagated through the derivation chain and thus we can infer the size of the physical object through the size registered in the metadata of its 3D representation.

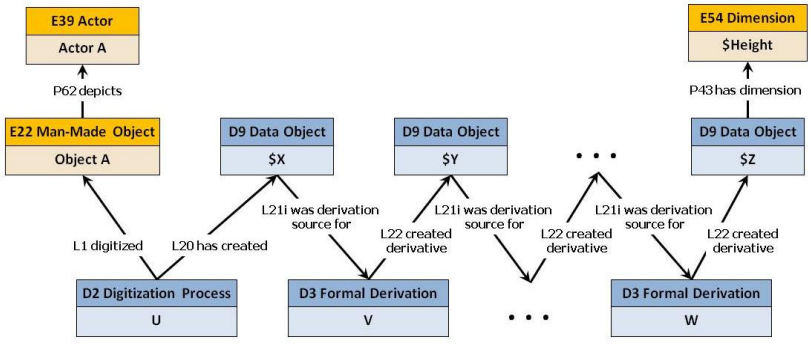

Fig. 3. Property propagation along the *derivation* chain

Fig. **4.** presents an example of a statue that depicts Ramesses II. The statue has been laser scanned and processed by MeshLab to produce its 3D model. The object "Ramesses Statue 1" does not have any size information in its immediate, explicitly defined metadata. However, our reasoning approach can answer the query "*Find the size of the Ramesses Statue 1 Object*" by retrieving the size calculated in the "3D model of Ramesses Statue 1" object and inferring that it also applies to the original physical object. Similarly, with our approach, the answer set of the query "*Find all the objects that depict Ramesses II*" is {"Ramesses Statue 1", "Scanned Ramesses

Statue 1", "3D model of Ramesses Statue 1"} while a query without inference capabilities would retrieve only {"Ramesses Statue 1"}.

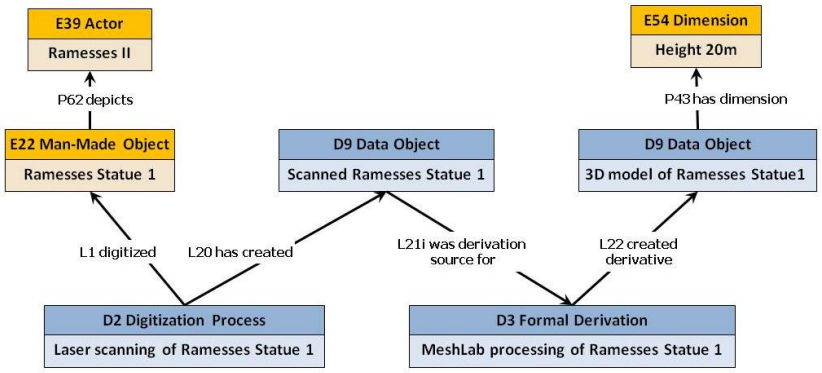

Fig. 4. An example of property propagation along the *derivation* chain

The combination of the property propagation along the two chains described above can help solve research questions that cannot be answered without reasoning. Consider the following research question: "Find Temples where Ramesses II and his wife Nefertari have the same size". If we apply both our rules on the metadata graph displayed in Fig. 5, we will get the set {"Abu Simbel Temples", "The Small Temple"} as an answer to our research question.

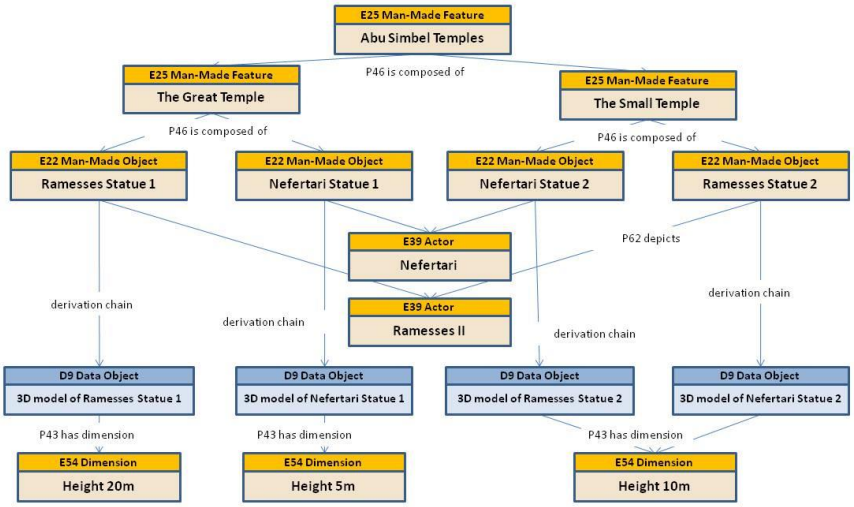

Fig. 5. Property propagation along the derivation and part-of chains

Here we have displayed two basic rules that can be used in a variety of applications, like quality control or querying. We encourage readers especially interested in the application of reasoning rules for querying purposes, to refer to the technical report in

[18] for a thorough study of the matter. As reported in [18], an exhaustive set of such rules has been implemented and tested by our team. The number of necessary rules is considerably reduced by property subsumption, but nevertheless we had to produce over a hundred counting all combinations.

5 Conclusion

In this paper, we have demonstrated a simple yet powerful mechanism of reasoning on provenance information by propagation properties along derivation and part-of chains. Moreover, we report an implementation on metadata built on the CIDOC-CRM and CRMdig schemas in the cultural heritage domain. In this implementation, it can be verified that the combination of structuring the metadata with rich schemas and applying reasoning upon them leads to the deduction of useful inferences with multiple usages. A number of such example use cases can be listed: (1) maintenance of repositories of digitization products, (2) garbage collection on reproducible intermediate files, (3) trace dependencies of products on tools and algorithms that should not become obsolete for long time preservation, (4) (re)production of valid, complete metadata at a loss of intermediate files, (5) completion of metadata by implicit knowledge, when production chains comprise thousands of intermediates and dozens of final products without need to manage this redundancy in the repository explicitly.

6 References

1. Doerr M.: The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata, AI Magazine, Volume 24 (3), pp.75-92 (2003)
2. Strubulis, Ch., et al.: Evolution of Workflow Provenance Information in the Presence of Custom Inference Rules. *SWPM2012-Proceedings of the 3rd International Workshop on Semantic Web in Provenance Management.* May 28, Heraklion, Greece (2012)
3. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini C., van de Sompel H.: The Europeana Data Model (EDM). *World Library and Information Congress: 76th IFLA General Conference and Assembly.* August 10-15, Gothenburg, Sweden (2010)
4. Tzompanaki, K., & Doerr, M.: A New Framework For Querying Semantic Networks. *Museums and the Web 2012,* San Diego, CA, USA (2012)
5. Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y., Melessanakis, V.: Modeling & Querying Provenance by Extending CIDOC CRM. Distributed and Parallel Databases, Vol. 27 (2), (2010)
6. Doerr, M., & Theodoridou, M.: CRMdig: A generic digital provenance model for scientific observation. *TaPP'11, 3rd USENIX Workshop*. June 20-21, Heraklion, Crete (2011)
7. Davidson, B. and Freire, J.: Provenance and Scientific Workflows: Challenges and Opportunities. *SIGMOD*, (2008) (Tutorial Track).
8. Goble, C.: "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics", Workshop on Data Derivation and Provenance, (2002)
9. Simmhan, Y. L. and Plale, B. and Gannon, D.: A Survey of Data Provenance in e-Science. *SIGMOD Record* Vol. 34 (3), (2005)
10. Tan, W-C.: Provenance in Databases: Past, Current and Future. IEEE Data Eng. Bulletin 30(4), (2007)

11. Moreau, L. et al.: The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, 27(6), (2011)
12. Sahoo, S., Thomas, C., Sheth, A., York, W.S., and Tartir., S.: Knowledge modeling and its application in life sciences: a tale of two ontologies. In Proceedings of WWW, (2006)
13. Lebo, T., Sahoo, S., McGuiness, D.: PROV-O: The PROV Ontology. W3C Recommendation, (2013). Available at http://www.w3.org/TR/prov-o/.
14. Cheney, J., Chiticariu, L., and Tan, W.-C.: *Provenance in Databases: Why, How, and Where.* Foundations and Trends in Databases. Vol 1(4), (2007)
15. Theoharis, Y., Fundulaki, I., Karvounarakis, G., and Christophides, V.: *On Provenance of Queries on Semantic Web Data.* IEEE Internet Computing 15(1), 31-39 (2011)
16. Wickett, K.M., Renear, A.H., Urban, R.: "Rule Categories for Collection/Item Metadata Relationships" In *Proceedings of the 73rd Annual Meeting of the American Society for Information Science and Technology.* Pittsburgh (2010)
17. Doerr, M., Tzompanaki, K., Theodoridou, M., Georgis, Ch., Axaridou, A., & Havemann, S.: A Repository for 3D Model Production and Interpretation in Culture and Beyond. *Proceedings of VAST 2010: The 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*. 21-24 September, Paris, France (2010)
18. Tzompanaki, K., & Doerr, M.: FORTH-ICS Technical Report TR429, Fundamental Categories and Relationships for Intuitive querying CIDOC-CRM based repositories (2012)