This is a preprint of the short paper accepted for publication in:
Proceedings of the 5th International Euro-Mediterranean Conference (EuroMed 2014) on Cultural Heritage: Documentation, Preservation and Protection, Limassol, Cyprus, 3-8 November 2014.
Multi-Science Publishing (ISBN 978 1 907132 47 6)

Co-reference Resolution via Annotation

A. Axaridou, M. Theodoridou, M. Doerr
Institution of Computer Science - Foundation for Research and Technology – Hellas (FORTH), 100 Nikolaou Plastira str., Vassilika Vouton, Heraklion, Crete GR, 70013, Greece - (axaridou, maria, martin)@ics.forth.gr

[Received date; Accepted date] – to be inserted later

Abstract
Co-reference is the implicit or explicit reference to an entity, in multiple information sources, possibly with different names. Being primary knowledge and a prerequisite of scientific research for all disciplines, co-reference resolution constitutes the integration of real world knowledge by human anticipation. Co-reference resolution appears as a topic of interest in many different fields such as databases and Digital Libraries, where it is often treated as "data cleaning", or by Natural Language Processing in text analysis. In the Semantic Web world, where data integration plays a crucial role, co-reference resolution appears primarily as the problem of determining whether two identifiers (two URI's) from the same or different sources, refer to the same object. In this paper we address the problem of co-reference resolution among semantic resources and we propose a simple, yet efficient, solution for managing identical resources related to Cultural Heritage artefacts. The proposed solution comprises two phases. During the first phase redundant identification is detected, based on consistency criteria between the co-referring resources. The second phase addresses the persistency and management of co-references. It is based on an ontological approach via annotation of the detected identical resources and proposes merging of redundancy via "data cleaning". The entire methodology works in the context of a semantic network being compliant to the CIDOC-CRM standard.

1. INTRODUCTION

The term co-reference often refers to the understanding of multiple explicit or implicit references to a notion appearing in free-text. Humans can resolve easily the meaning of a well written natural language text and realize the way the referred notions are related to each other. But it turns into a complicated issue, very hard to describe and interpret, for the machine intelligence. The term has been introduced by Levesque [1] as a problem also in knowledge representation data.

Successful co-reference resolution becomes of high importance and interest as the challenge is not just to deliver documents, but to leverage on the latent knowledge [2]. Discovering and managing the knowledge derived from co-reference among parts of documentation becomes the subject matter for several scientific and technology sectors. The Natural Language Processing domain aims to produce the maximum of understanding of the notions existing in a natural language text providing co-reference solutions based on natural language modelling algorithms and statistics [3]. The Digital Libraries domain is still suffering from the lack of effective interconnection of the collected knowledge. Even though a lot of work has been done to the latter, mainly by keyword classification, it is far from having overcome the knowledge "fragmentation" due to unresolved co-reference.

A partial solution to the co-reference problem is the well-known data cleaning or data scrubbing, aiming at the automated detection and elimination of the duplicate identifiers. Up to year 2000, data cleaning has not preoccupied researchers, as much as it should. Only the flourishing of data warehouse applications together with the needs of the industry exposed the problems in syntactic and semantic differences among data sources and the need to define ways to normalize data for information exchange [4]. Significant research has been done in this area but the world of Internet with the vast amounts of shared data resources imposes new challenges now

described as “instance matching”. However, “false hits” are inevitable, and spoil the reliability of integrated resources.

The Semantic Web community is faced with the problem of matching similar or identical resources spread all over the Web where each resource ignores the existence of the others. The redundant use of names referring to the same entity has become a widespread problem causing limited knowledge and inconvenient semantic communication thus obstructing data integration. The various documentation authorities of digital content experience the lack of a unified and systematic way for naming the documented entities. Centralized “authority” catalogues cannot scale up with the resources. Moreover, the inadequacy of searching the already existing knowledge about entities may also result in redundant identification of new resources increasing thus the necessity for co-reference detection and management.

At the moment, the human mediation to interconnection of the knowledge seems to be inevitable, if high quality information integration is the goal. The understanding of co-reference regarding the identification of specific citations and/or references of persons, places, objects and events, by a reader of a document, is fully built on associations found in digital objects or in metadata, on (human) background knowledge or is derived from the communication with the authors themselves. The ultimate knowledge would be what the author meant by “her/him/it/etc”. Does he refer to a part-of-speech, or a database record/element or an occurrence of a name or identifier? In principle, the situation is not different from URIs appearing in an RDF record. The URI may be locally defined, ignoring other URIs for the same thing, even if it is a Linked Open Data record about this thing, or a URI may be referred in the record misinterpreting its source, or the source of the referred URI cannot be found. The actual fact of a co-reference is primary knowledge about the referred world, prerequisite of scientific research and therefore should be persistent and curated. The ultimate challenge are not just the matching algorithms – they do what they do, but managing the manual verification.

In this paper we present a simple and efficient approach for co-reference resolution on a semantic network based on the CIDOC-CRM (ISO 21127:2006) [5, 6] standard and its extension CRMdig [7, 8]. Our approach copes with the inconvenience of the redundant use of different URIs for identifying the semantic resources. A phase for detection of the URIs co-reference via consistency estimation on probably matching resources is applied first. Finally, a phase of materializing and making this new knowledge persistent by applying annotation and data cleaning options is proposed.

2. RELATED WORK

Work on co-reference resolution and management has been developed by several research teams. There are parallel approaches providing algorithms for automatic detection of co-reference, focusing on centralized solutions where the definition of a reference basis for the co-referred entities is necessary.

A centralized approach presented by Doerr & Iorizzo [9] suggests setting up a Web 2.0 Co-reference Service supported by a grid service oriented architecture for digital libraries, so that anyone anywhere can publish a co-reference under an international scope. The main aim is to logically separate documentation units and primary sources from the network level and instead to derive data for the network level from the documentation units and primary sources.

Glaser, et al. introduced the Co-reference Resolution Service (CRS) [10] considering the notion of equivalence within a given context. It asserts that in many of the cases the use of owl:sameAs may be applied inappropriately affecting this way the definition of the involved resources. CRS comes as an external manager (not affecting the semantic repository) delivering a separate co-reference knowledge base that can be used by semantic applications as a source of co-reference resolution. Although this approach seems to facilitate co-reference management it does not provide any quality control of the accumulated knowledge, which ultimately may lead anything to become related to everything.

The self-training approach of Hu, et al. [11] for object co-reference resolution on the Semantic Web, leverages the semantic equivalency inferred by OWL between the resources, in parallel with the similarity computation between property-value pairs. It suggests for every object URI to establish a kernel that consists of semantically co-referring URIs based on owl:sameAs, (inverse)

functional properties and (max-)cardinalities, and then extend this kernel iteratively in terms of discriminative property-value pairs.

An extended work on decentralized co-reference management introduced by Meghini, et al. [12] assumes that there are ways of varying reliability to identify co-reference or non-co-reference relations; it deals with the fundamental problem what to do with this knowledge in a distributed environment, in order to increase local and global consistency and the degree of knowledge sharing. A model of co-reference knowledge is presented, which is based on the epistemological distinction of (i) a model of a common reality, (ii) a model of an agent's opinion about reality, and (iii) a model of agents' opinions whether they talk about the same object or not. This work relies on a different aspect of information integration. Independently if data are brought under a common schema or the sources operate in their own independent ways, they simply try to connect them into a coherent network of knowledge. The two basic requirements for such a network to be created and managed are the following: (i) decide whether two data elements residing in different sources refer to the same thing, and (ii) maintain persistent co-reference relation and make it accessible to the individual sources (agents) participating in the network. The same work describes how the current approaches can be deployed more effectively to support co-reference detection. These approaches can roughly be divided into pro-active and reactive methods. In pro-active methods, sometimes also found under data cleaning, identifiers of things are normalized before integration takes place, in order to increase the chance that other sources will normalize their identifiers in precisely the same way. In reactive methods, typically, data are first integrated under a common schema and then possible duplicate data entries are detected and merged.

Most of the approaches consider the automatic detection of co-reference, but the performance is frustrating. More than 75% of resolution accuracy has been mentioned [13] in some cases and are considered as high. According to Bennett et al. [14] for any reliable integration there is always the need of human intervention for the control and completion of co-reference resolution. Our implementation is considered as a typical application of reactive approach taking into account the property matching of resources as a result of human consideration. It provides a flexible way for controlled recording of the co-referred URIs enabling the estimated equivalency of the resources, via annotation, to be directly applied onto the semantic network. The Open Annotation Data Model [15] proposes a generic annotation schema where an Annotation is considered to be a set of connected resources, typically including a body and target, where the body is somehow about the target. This approach seems to be complete and detailed for common annotations but not satisfying the co-reference requirement where two or more resources have to be correlated as equivalent members of a relation (no body and target is participating). Conversely, the applied CRMdig Annotation Model compliant to CIDOC-CRM can be efficiently used in co-reference cases as presented below. Finally, after determining co-reference a step of data cleaning is suggested for eliminating the redundancy of resource referencing.

3. METHODOLOGY FOR THE CO-REFERENCE RESOLUTION

The co-reference resolution on resources in a semantic network is the topic of our interest. We assume a semantic network that is based on CIDOC-CRM and its extension CRMdig that allows representation of not only human material history and cultural objects, but also of provenance metadata, annotations and co-reference information [16]. The resources are instances of classes that belong to the CIDOC-CRM and CRMdig ontologies. All the resources of the semantic network are identified with URIs, as it is recommended by the RDF/OWL framework. The duplicates of referred resources have to be detected and resolved. Our approach to this direction, with regard to the redundancy of identifiers in the semantic network comprises two essential phases. The first attempts the detection of possibly co-referred resources based on consistency criteria complying with our data model. Co-reference management requires the positive or the negative knowledge of co-reference as the result of knowledge negotiation [12]. On positive agreement, new knowledge is inferred and a second phase attempts the propagation of this new knowledge directly into the semantic network. This is a reactive operation triggered on an already built environment, causing the direct update of the existing data, by either adding appropriate semantic relations or by removing redundant identification. This direct knowledge update accomplishes the best of query performance. The two phases are presented next.

3.1 Co-reference detection with consistency criteria

Quite often, the common characteristics between the semantic resources, especially in their appellations, their "about" descriptions and the relations to other resources, "alerts" the probability of resource co-reference. In order to safely infer that two resources with different identifiers refer to the same entity, a set of criteria compliant to the data model need to be satisfied. Four principal criteria for resource matching have to be considered, presented in the next sections.

3.1.1 Resource casting

As a starting point, a general criterion that applies to all resources and should always be checked is the 'type' of the resources. This kind of check involves the classification of the compared resources. If the resources that are assumed to be the same entity are of different types, which means they belong to either one or multiple different ontology classes, then if there is at least one case of disjoint between any of their classes the resources cannot be identical. If no classification incompatibility is detected, this is the first step for co-reference verification but more matching consistency evidence is required.

The rest criteria cannot be commonly applied to all the resources since they depend on the type of the examined resources.

3.1.2 Event based

The 'where' and 'when' properties of the compared resources imply an event basis criterion. It should be applied to resources being related to a place and time (i.e. a date or a time interval) with a very strong and distinguishing relationship. Hence an inconsistency of this level between two resources makes it unlikely for the resources to be identical. This criterion concerns the resources of the class Event. If two Events are inconsistent about 'where' or 'when' they have taken place, then apparently, they are not identical. It also concerns classes that are tightly related to the Event class such as the class Person. For example, Persons have a distinguishing relationship with a birth Event. If two Persons having the same first/last name were born on different times and/or locations then they are undeniable two different Persons.

3.1.3 Result based

The criterion of 'resulting' of the compared resources can be applied to resources that produce a result. The products and the procedures are unambiguously related. The products are identical when the producing resources are identical and the opposite. This criterion concerns resources that either where produced by Events or other entities related to instances of the Event class. An example for this criterion may be the creation of a Thing (e.g. an artefact) by an Actor. The Thing as a product of the Actor both are related to the same creation Event. It is subject to investigation that a Thing was created by different Actor creators. Especially if it is well known that the creator is one then any multiple creators found in such cases they most probably may be identical. Another example is the creation of a thing by a Production Event. A procedure can result in multiple products but each of them cannot be produced by multiple procedures. Thus the procedure resources producing the same result might be identical.

3.1.4 Cardinality based

At this point, we can extend the co-reference matching criteria to all entity relations of type one-to-one and one-to-many, as the constraint for only "one" related resource may help to the discovery of the redundant instances of an entity. Example of one-to-one relations is the Place where a specific type of Event has happened. If multiple places are referred to as the places of the Event then there is a probability for redundant identification of the same place. Example of the one-to-many relation is the Actor with a specific role of an Event, such as the photographer of a single image capture. Although a photographer can take multiple photos, each photo must have only one photographer. Thus if multiple photographers are found related to a single photo, then the redundancy of the Actor resource is quite possible and easily detected.

3.2 Reactive resolution of co-reference

When two or more resources of the semantic network, having different identifiers and satisfying

the co-reference matching criteria described above, are found to be identical, then two options are proposed for handling the co-reference ambiguity: (i) co-reference via annotation (ontological approach); and (ii) co-reference via "trusted opinion" (data cleaning).

3.2.1 Ontological approach: co-reference via annotation

Co-reference via annotation can be performed by all users, since annotations are handled as "personal opinion" information. It is a simple and low risk solution as it can be updated without affecting the identification information in the semantic network. In particular, the annotation option enables the users to co-reference identified entities by publishing extracted knowledge about similar, dissimilar, conflicting or any other detected status of co-reference between resources. It is important for the cultural heritage documentation to know the complete "story" of the resources' correlation, such as since when and who is thinking that two resources are e.g. the same or conflicting. CIDOC-CRM can support this requirement providing the class of Event which is specialized with the Annotation Event in the CRMdig ontology extension. The Event class offers details about "who-when-where" is doing something or is affected by something and accordingly the Annotation Event records this kind of information. The proposed CRMdig Annotation Model for annotating co-reference is shown in Figure 1.

In short, an Annotation Event is a subclass of Creation and is always defined in order to create an Annotation Object. A Knowledge Object is a special type of Annotation Object, containing the co-reference information which is primary knowledge. The Knowledge Object is packaged as a Named Graph to form an entire independent unit of knowledge so to be easily updatable or removable when necessary without particular consideration. When we need to be more specific referring also to the sources (named as Information Objects) of the co-reference a specific Knowledge Object is defined named as Knowledge Extraction. So, the real annotation content resides in the Knowledge Object and comprises specific CRMdig relations between the co-referred resources. This architecture enhances the extensibility of our data model with more co-reference relations that might be used in the context of a Knowledge Object. The relation that we are focusing in the current work to manage the inconvenience of multiple identical resources is the same-as relation.

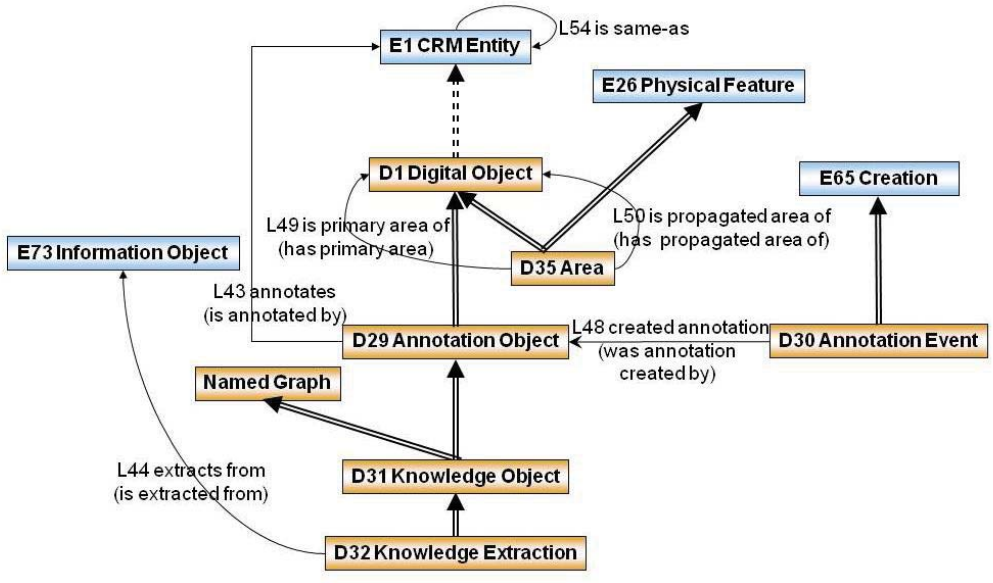

Figure 1. The CRMdig 3.0 Annotation Model

3.2.1.1 The same-as relation

The most common relation for co-referencing identical resources is to declare them 'as the same'. This is a capability already supported in RDF/OWL by the well-known property owl:sameAs.

However, for better control, we would prefer to model a dedicated property, because owl:sameAs will not allow to distinguish an identity assumption from approved identity and never questioned identity. So, we define the symmetric and transitive property named as crmdig:L54_is_same-as, with rdfs:domain/range the crm:E1_CRM_Entity of CIDOC-CRM, which practically means it can be applied to all CIDOC-CRM entities.

Declaring that A crmdig:L54_is_same-as B, B crmdig:L54_is_same-as C, etc., a rdf reasoning engine can make the correlated resources equivalent, inferring further all the rest of equality relationships e.g. that A crmdig:L54_is_same-as C, etc. Thus the equivalency of the co-referred resources is propagated in the semantic network, allowing for sharing their properties. The same-as co-reference announcement enables more accurate query results as more resources match the querying criteria. An advantage of this solution is that even though the 'same-as' co-referred resources became one and are enabled to share their properties, yet they keep their independence. Nevertheless the penalty of this approach may be a lower query performance as many same-as inferred relations have to be managed by the triple-store. An expected question to the same-as approach may be "is there any reason for keeping independent instances of the same resource?". The easy answer is "Rather no. There is no reason for a resource to be expressed with multiple identities each having their own properties. Otherwise, most probably, we are talking about different resources". This is the reason of proceeding with a second approach to simplify the co-reference issue into a closer to reality context.

3.2.2 Co-reference via "trusted opinion"

Co-reference via "trusted opinion" can be performed only by authorized users. This reactive data cleaning option aims to merge the redundant identifiers by replacing them with a single one, causing a permanent change in the semantic network. An example of the case is shown in Figure 2 and Figure 3. A simple way to proceed with this option is to follow the next steps.

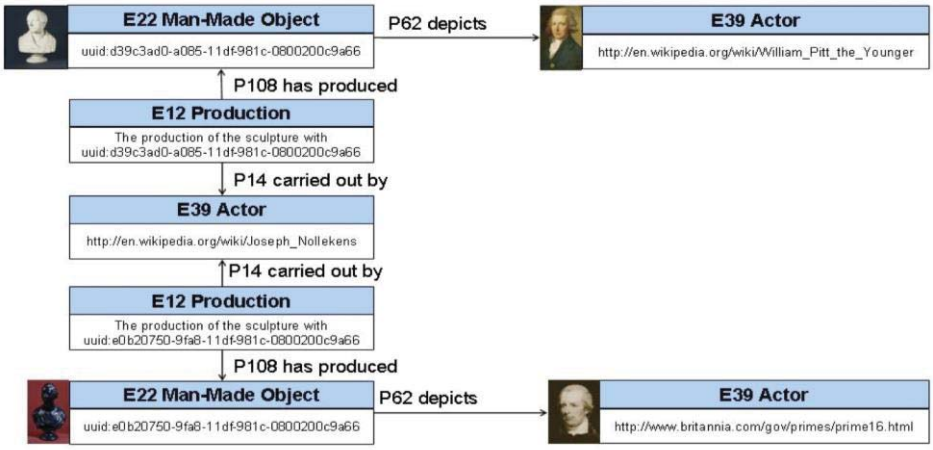

Figure 2. Reactive data cleaning - Initial data

The first step is the retrieval and collection of the multiple instances of a resource which requires querying the semantic network. These instances may either simply fulfil the co-reference criteria or already be co-referred with the same-as annotation as described above.

Next step is the selection of a surrogate identifier. The collection of instances should be ordered by descending frequency of appearance in the semantic network. The criterion for the latter is simply the count of the explicit triples they participate. It's a good practice to keep the most used identifier as the preferred one for all the instances of the resource.

The final step is merging the redundant identifiers, which is realized as the replacement of all the identifiers of the resource with the preferred one updating the appropriate triples. It is necessary, for the history completeness of the semantic network and the provenance data preservation to keep the previously used identifiers. The last requirement can be accomplished by

linking the prevailed identifier via the crm:P1_is_identified_by property to a crm:E41.Appelation object form CIDOC-CRM ontology pointing to the replaced identifier, as shown in Figure 3.

Merging of identifiers should be a serious decision. Although the older identifiers are kept in the semantic network as related appellations, it is recommended to avoid reversing after a merge. The main goal of this option is to offer the highest query performance capability regarding the URI redundancy and thus it should be permanently applied.

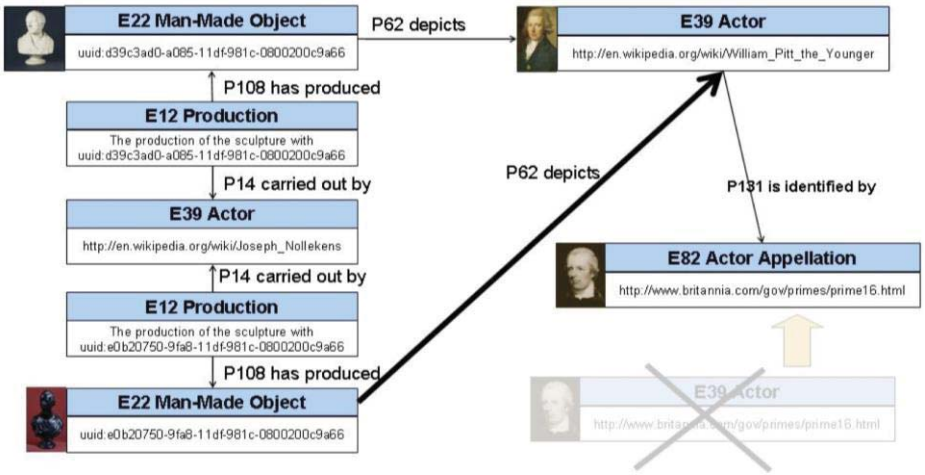

Figure 3. Reactive data cleaning - Result of merging

4. CONCLUSIONS

Co-reference is primary knowledge and prerequisite of scientific research that must be persistent and curated. In semantic networks co-referring resources using different URIs for the same entity, is a common issue. In this paper, coping with the inconvenience of URIs redundancy, we proposed an efficient approach for co-reference resolution on resources adhering to the CIDOC-CRM standard. Two main phases are important for this purpose. A phase for detection of URI co-reference via consistency estimation on probably matching resources is applied first. Consistency is considered as the positive compliance to the applied data model. When the examined resources are found consistent and a co-reference extraction is concluded, a phase of materializing and persisting this new knowledge by using annotation and data cleaning options is proposed. Annotating identical resources with the same-as relation is a personal opinion of the Annotator and offers the ability of easy update of this new knowledge. Finally, the data cleaning option is proposed for permanently merging the redundant identifiers; leading hence to a better repository performance.

REFERENCES

[1] Levesque, H.J., *Foundations of a functional approach to knowledge representation*. Artificial Intelligence, 1984, 23(2):155-212.

[2] Doerr, M. & Meghini, C., *Leveraging on Associations - a New Challenge for Digital Libraries*. In Proc. of the First International Workshop on Digital Libraries Foundations In conjunction with ACM IEEE Joint Conference on Digital Libraries (JCDL), Vancouver, Canada, 23 June, 2007.

[3] Mamakis, G., Malamos, A.G., Axaridou, A., Kaliakatsos, Y. & Ware, A., *An Algorithm for Automatic Content Summarization in Modern Greek Language*. In Proc. of ITI 3rd International Conference on Information & Communication Technology (ICICT): Enabling Technologies for the New Knowledge Society, Cairo, 5-6 December, 2005, pp. 579-591.

[4] Bulletin of the Technical Committee on Data Engineering, *Special Issue on Data Cleaning*, December 2000, 23(4).

[5] Doerr, M., *The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata*, AI Magazine, 2003, Volume 24, Number 3.

[6] CIDOC-CRM SIG, *Definition of the CIDOC Conceptual Reference Model, version 5.0.4*, Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (Eds.). http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf, 2011, (accessed

25 June 2014).

[7] Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y. & Melessanakis, V., *Modeling and querying provenance by extending CIDOC CRM*. Distributed and Parallel Databases, Springer Netherlands, 2010, 27(2), 169-210.

[8] Doerr, M. & Theodoridou, M., *CRMdig: A generic digital provenance model for scientific observation*. TaPP'11, 3rd USENIX Workshop on the Theory and Practice of Provenance, Heraklion, Crete, Greece, 20-21 June, 2011.

[9] Doerr, M. & Iorizzo, D., *Epistemic Networks in Grid + Web 2.0 Digital Libraries*. In Proc. of the First International Workshop on Digital Libraries Foundations In conjunction with ACM IEEE Joint Conference on Digital Libraries (JCDL), Vancouver, Canada, 23 June, 2007.

[10] Glaser, H., Jaffri, A. & Millard, I., *Managing co-reference on the semantic web*. In Proc. of WWW Workshop on Linked Data on the Web (LDOW), Madrid, Spain, 2009.

[11] Hu, W., Chen J. & Qu, Y., *A Self-Training Approach for Resolving Object Coreference on the Semantic Web*. In Proc. of the WWW '11, 20th international conference on World Wide Web, ACM NY, Hyderabad, India, March 28-April 01, 2011, pp. 87-96.

[12] Meghini C., Doerr M. & Spyratos N., Managing Co-reference Knowledge for Data Integration. In Proc. of EJC2008, the 18th European-Japanese Conference on Information Modelling and Knowledge Bases, Tsukuba, Japan, June 2008.

[13] Lee, H., (et al.), *Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task*. In Proc. of the 15th Conference on Computational Natural Language Learning: Shared Task, Portland, Oregon, 23-24 June, 2011, pp. 28-34.

[14] Bennett, R., (et al.), *VIAF (Virtual International Authority File): Linking Die Deutsche Bibliothek and Library of Congress Name Authority Files*. World Library and Information Congress: 72nd IFLA General Conference and Council, 2006.

[15] Open Annotation Data Model: http://www.openannotation.org/spec/core/ (accessed 25 June 2014)

[16] Rodriguez-Echavarria, K., Theodoridou, M., Georgis, Ch., Arnold, D., Doerr, M., Stork, A. & Peña Serna, S., *Semantically Rich 3D Documentation for the Preservation of Tangible Heritage*. In Proc. of Eurographics Association, VAST12: The 13th International Symposium on Virtual Reality, Archaeology and Intelligent Cultural Heritage, Brighton, UK, 2012, pp. 41-48.