A New Framework For Querying Semantic Networks

Katerina Tzompanaki, Martin Doerr

Institute of Computer Science, F.O.R.T.H. Crete–Greece

http://www.ics.forth.gr/

Abstract

The upcoming large-scale metadata repositories, semantic networks of RDF triples integrating large amounts of cultural–historical data, are not easily accessible to global query paradigms, such as “query by example” or keyword search. ISO21127 (CIDOC Conceptual Reference Model) is an adequate global schema for such systems, but querying individually hundreds of different kinds of properties leaves a huge recall gap compared to text retrieval, whereas a global restriction to “core metadata,” such as Dublin Core, deprives the systems of any more advanced integration and reasoning capability. We therefore propose and have implemented a new query paradigm: Intuitive “fundamental” categories and relationships, as we are used to from core metadata, are presented to the user as complex deductions from a rich underlying network of more specialized actual metadata, rather than being primary documentation elements. In addition to efficiency, we also provide simplicity, as the user does not need a deep understanding of a complex schema in order to obtain the desired result. Application of the framework can easily be adjusted to many domains and user preferences.

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

Introduction

Content management systems can successfully be accessed by text search engines that use keywords without prior knowledge of the structure of the content. The great success of search engines and their providers such as Google is based on a long history of optimizing *recall*—i.e., getting back *all* documents in a collection relevant to a particular question (set of keywords) together with *other*, unrelated documents—versus *precision*—i.e., getting back *only* relevant documents, but possibly *losing* relevant ones. This technology has reached its limits now much below the desired recall and precision goals of cultural–historical research. Only the great redundancy of resources on the web with respect to popular topics simulates perfection. Any further improvement of precision and recall needs elaborate metadata to be added to the content to be searched.

These metadata must follow a particular schema ("vocabulary"), such that the user can know which fields to query. The traditional query paradigm is based on a stable and simple structure of the information-organized tables, as in relational databases. Only if all fields are filled in with all applicable data about the objects or subject of the information system will the query paradigm yield 100 percent recall and precision. We call such systems to form a "Closed World," in which all applicable data of a domain and restricted context are known. This holds typically for office automation systems.

Information about cultural heritage kept and collected by museums or systems like digital libraries is complex and by nature incomplete. It may be organized by different people, using a schema in different ways, or even using different schemata and languages, making the correlation among different sources or even the same source a great challenge. It forms an "Open World." Currently, the most popular search method in Open World systems is the keyword*based search in text documents, image captions, and database fields. It usually yields high recall rate and a medium to low precision rate.

The Semantic Web promises to overcome the recall and precision problem for information not backed up by high redundancy by resorting to rich, formally structured metadata for documents and objects of interest and links between them, and combining them with general, formal background knowledge. Large-scale metadata repositories, semantic networks of Resource Description Framework (RDF) triples integrating large amounts of cultural–historical data have been developed that are globally accessible via the Internet, such as the Europeana (http://www.europeana.eu/portal/), cultureSampo (http://www.kulttuurisampo.fi/index.shtml), and many other projects world-wide. In these systems, the CIDOC Conceptual Reference Model (CRM) is becoming more and more popular as a rich RDF schema adequate to integrate complex museum data, for instance submitted in the form of Lightweight Information Describing Objects (LIDO) Extensible Markup Language (XML) records. Among those projects are the German Digital Library (http://www.deutsche-digitale-bibliothek.de/), the ResearchSpace Project (www.researchspace.org), the WISSKI Project (http://wiss-ki.eu), and the CLAROS project (http://explore.clarosnet.org). "Linked Open Data" are advocated for cultural institutions, in which RDF data reside on local servers, but accessible under published RDF schemata from the Internet.

However, the Semantic Web is an Open World system. Users cannot know precisely what's out there, nor in which terms exactly things have been documented. Querying individually hundreds of different kinds of properties creates a huge recall gap compared to text retrieval. Particular queries become so "precise" that real data rarely fit to the question, and querying a combination of even a few properties tends to frustrate the users with empty answers (Fernandez et al., 2008). A global restriction of the semantic network to "core metadata" on the other side, as propagated by defenders of Dublin Core or VRA, deprives the systems of the reasoning capability and information integration that the Semantic Web promises and researchers need, and does not solve the inherent problem of incomplete knowledge.

In order to overcome the problem of effective searching as described above, we propose a querying system for semantic networks based on a few fundamental categories (FCs) and (binary) relationships (FRs). These categories are “base classes” covering the domain, and the relationships are deductions from complex path expressions of all sorts of deep relationships and documentation alternatives in a much richer and more specialized semantic network, rather than a reduction of the primary documentation to core metadata. The deductions are formulated to ensure the high recall of each FR by comprising the respective formulation variants of the facts documented directly or indirectly in the network. The FRs simulate a much simpler network of a small number of intuitive relationships, which can be combined in an advanced search to cover a wide range of frequent and relevant queries. However—we cannot produce wonders—recall is no longer 100 percent, as people may have been used to in their local collection system. Many of the deductions only entail the probability that the relationship holds; e.g., we may conclude that a painting from a particular workshop was most probably created at the location of the workshop.

The idea of the Fundamental Categories and Fundamental Relationships has already been presented to the public, in the “Museums and the Web 2012” conference (Tzompanaki-Doerr, 2012a). This paper is a first recommendation to CIDOC for adapting to this new querying mechanism which can effectively replace flat approaches like the Dublin Core Elements and allows for structuring the metadata using rich schemata like the CIDOC-CRM. The specific set of the paths that formulate the Fundamental Relationships (http://www.ics.forth.gr/tech-reports/2012/2012.TR429_Intuitive_querying_CIDOC-CRM.pdf) are published for discussion with users of the CIDOC-CRM in order to conclude with a standardized set.

In the framework of the European Integrating Project 3D-COFORM (http://www.3d-coform.eu/), funded by the European Community’s Seventh Framework Program (FP7/2007-2013, no 231809), together with project partners, we are currently implementing such a system based on the CIDOC CRM (ISO21127) (Doerr, 2003) and extensions describing the digital provenance for empirical three-dimensional (3D) modeling processes. Digital provenance data (Theodoridou et al., 2010; Moreau et al., 2007) form deep chains of events connected by output–input, 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 a high recall rate, as not only explicitly documented properties but also derived properties across independently created metadata records can be combined for calculating the desired results. In parallel the Research Space project (http://www.researchspace.org/) is also implementing this approach.

Problem statement

As described above, the only way to provide better recall and precision than keyword search is to use suitable metadata and associative queries. We maintain that the poor performance of associative queries in Open World systems, in particular in information aggregation services such as cultural metadata repositories, has the following causes.

The relationship the user is looking for was not documented, or was represented in a different way. For instance, someone may look for things “*made from*: steel,” but some objects were registered as “*has part,”* which is “*made from:* steel,” or as “*has type*: steel object.” Our experience developing ISO21127 showed that 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.

Most frequent associative queries include “AND” conditions (conjunctions). A typical query may ask for a particular type of thing from a particular period of time, a particular geographic area,

used/made by a particular group of people. The problem is that the lack of recall in each field, due to missing alternative associations or incompleteness of knowledge, roughly multiplies with the number of fields put in the conjunction. If each element of the conjunction has a recall of 90 percent, the conjunction of four has about 66 percent if there is no other correlation between the fields. Therefore, a successful advanced search facility must strive to increase recall of each individual parameter (as the "4W," or "who-when-where-what" queries). Even better would be to indicate to the user which parameter may have been most "catastrophic" to the result.

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" of properties along those property paths. Propagation may be very complex to formulate as query, but is also very powerful when it comes to recall improvement. For instance, one should assume that the actors, place and time that are reported for the building of a house (the "super-event") also apply for or include the building of its walls (a "sub-event"); or that materials a part is made of are considered to be among the materials the whole is made of; or that the subjects a thing represents also apply to its copy or derivative, etc.

Such reasoning allows for querying facts that are not stated within a single metadata record; for instance, we have the British Museum website *(*http://www.britishmuseum.org/) saying that the "Horsemen from the west frieze of the Parthenon" is part of the Parthenon, and from the other side there is the Acropolis Museum (http://www.theacropolismuseum.gr/) stating that Parthenon was created by Pheidias. Using the CIDOC-CRM the metadata would be:

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

The respective query on the integrated metadata could assume that Pheidias was involved in the making of the Horsemen as well.

Last but not least, another problem of formulating queries is SPARQL (**S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage). Most favored by information technology (IT) experts, it has transferred the old relational paradigm onto the graph structure of the Semantic Web, creating an incredibly complex system, even for specialists. In our applications, no IT expert was able to verify that a SPARQL query of the kind we present in this paper will yield the results intended by a domain expert simply by reading it.

Related work

One category of already-existing systems aims at helping users to formulate path queries with terms from an ontology (RDF schema). This can be done with menu-guided user interfaces to specify subject-property-object triples, combined with a look-ahead enhanced search, such as the one realized in DBpedia (Auer et al., 2007). Other systems, like NightLite (Ding et al., 2004), employ a query formulation facility with graph representations of the ontology, but still require SPARQL knowledge. Another approach is the use of natural language queries, which are automatically mapped to associations of triples of the implemented ontology by a built-in dictionary and some inference mechanism, such as the Power Aqua system (Lopez et al., 2006). This approach relieves the user from learning the ontology terms, but it inherits all the well-known polysemy of natural language, which deteriorates precision, and often provides even worse recall than the explicit use of ontology terms. Other natural language search systems, such as Swoogle (Ding et al., 2004) and SemSearch (Lei et al., 2006) do not interface to a triple store.

The most common approach to reduce the complexity of querying is to reduce the complexity of the Semantic Network itself. The Dublin Core Metadata Elements (Hilman, 2006) , VRA Core (VRA Core 4.0 Element Description, 2007), and other metadata standards reduce the network to flat relationships. Similarly, the Consortium of Interchange of Museum Information (http://www.cni.org/projects/cimi/) has proposed the metadata elements: who, what, when, where ("4W") as a domain of independent relations to four kinds of entities (person, thing, time, place), a kind of "faceted search." In other terms, whatever relation a thing may have to a person is an answer to the question "who," etc. This works relatively well for metadata describing only the history of objects. Otherwise, the ambiguity, for instance between history and subject, becomes overwhelming. Systems like Artifacts Canada (http://www.pro.rcip-chin.gc.ca/artefact/index-eng.jsp) provide an advanced search facility based on this paradigm (plus a "how").

However, these metadata elements are too poor to allow for reasoning with integrated metadata or query refinement. The lack of precision in the primary documentation, in particular the missing concept of events, results in the inability to *integrate* related data from different sources, as has been shown (Doerr and Iorizzo, 2008). If such "simple" metadata are to be created individually for all elements of the complex correlation graphs characteristic for history, interesting works of arts and e-science data, the same facts may have to be repeated manually hundreds to tens of thousands of times, which is ineffective and error prone, and in no way "simple." Consider, for instance, the over twenty implementations of the "Thinker" by Rodin and all related artifacts. All these systems cannot be scaled up to higher precision, because the precise knowledge is lost in the documentation in favor of recall.

An interesting intermediate between a full-fledged semantic network and faceted search is the very successful Finnish CultureSampo (Hyvoenen et al., 2006). It uses nine instead of four facets, including material, events, and object types. It uses for each facet rich term hierarchies of inclusion or subsumption, and provides multiple explicit, direct relationships among facets, such as fifty kinds of social agent-agent relationships. It even provides a natural language search. It comes closest to our approach. It avoids the error-prone search for suitable properties to query, provides deductions from term hierarchies, class and property subsumption, selection of valid query parameters, and faceted search. It still misses, however, other deductions, such as property propagation along part-whole relations and derivation chains.

In this paper, we propose a method that tries to combine the best from the aforementioned approaches and to go beyond them.

Realization

The 3D-COFORM project aims at providing integrated technologies to make the large-scale production of 3D models feasible for the systematic documentation and study of material cultural heritage. For that purpose, it combines leading-edge technologies for 3D model generation from acquired data (photographic or laser), generation of synthetic models and presentations. Underlying is a scalable repository infrastructure (RI) to manage integrated, distributed data and metadata about cultural-heritage objects themselves, digital representations of them, and scholarly and scientific annotations. The RI contains a metadata repository (Doerr et al., 2010b) we have implemented on SESAME (http://www.openrdf.org/) and an OWLIM (http://www.ontotext.com/owlim/) reasoner that provides the platform and semantics to manage objects, 3D modeling, models, and presentations alike; and supports the scholarly discourse on recent and past object features in archaeology, sites and monuments management, museum disciplines, and conservation. The reasoner is used to optimize the performance of the queries described below by pre-calculating frequent deductions and storing them physically in the network. The reasoner ensures that these deductions are updated following changes of the primary data. The query system we have designed and describe in this paper is part of the "Integrated Viewer and Browser"(IVB) component that is running on the RI,

and which we are implementing together with partners of the project, in particular ISTI – CNR (http://www.isti.cnr.it/) in Pisa.

3.1 Designing fundamental categories

Whereas our current implementation is based on the CIDOC CRM and extensions for the domain of digitization, our approach can be applied to other ontologies in an analogous way. Much of its reasoning capability depends on explicit representation of events, which is also the case in the ABC Harmony model (Lagoze and Hunter, 2001), DOLCE (Gangemi et al., 2002), BFO (http://www.ifomis.org/bfo), the Europeana EDM (Doerr et al., 2010a), various digital provenance models, and other ontologies. Our primary target is the generic search for things, ideas, people, and facts from the past characteristic for digital libraries, cultural–historical research, science, business intelligence and political inquiries. We draw on rich previous experience in the cultural domain such as the Polemon Project (Bekiari et al., 1998) and explicit queries collected from archaeologists and museum curators and analyzed by us in 3D-COFORM.

In a typical web search engine, searches would homogeneously return just web pages, or, in a digital library, only documents. In a semantic network, however, users can retrieve any instance of any class known to the system by any kinds of direct or indirect relations to other things in the network. Therefore, we first divide the entities of our universe of discourse into a set of relevant FCs that appear to be founded deeply in our intuitive understanding of the world in this or a similar form. These FCs serve as domains and ranges of FRs described below. Similar to core metadata, we try to *cover* the domain with as few FRs as possible, which a user can easily learn, but still to be able to make some powerful distinctions keyword search cannot do, such as discerning places from people with the same name. The idea is to try to satisfy as many different kinds of questions as possible by asking a few more general ones, and not only the most frequently asked questions. For the selection of the FCs, we follow the tradition of Ranganathan (1965), CIMI's 4Ws, and others. Still, our method is mostly intuitive and by insight; the future may give us the chance for wider explicit user studies.

In our implementation, we have selected (the *crm* namespace refers to the CIDOC-CRM schema:http://www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.2_english_label.rdfs):

1. *Thing* = *crm:E70.Thing*, comprising material and immaterial things, a special case of "what" and Ranganathan's "Matter."
2. *Actor* = *crm:E39.Actor,* comprising persons, organizations, offices, and informal groups, equal to "who" and Ranganathan's "Personality."
3. *Event* = *crm:E2.Temporal_Entity*, comprising states, historical and other periods in the sense of the CRM (crm:E4.Period), and events (crm:E5.Event) and activities (crm:E7.Activity) in the narrower sense. It is equal to Ranganathan's "Energy." In some cases, periods can be regarded as a "when."
   Time = *crm:E52.Time-Span*, a date-time interval, a special case of "when" and equal to Ranganathan's "Time."
4. *Place* = *crm:E53.Place*, geometric extents in space, on earth and on objects, often related to or even identified by some stable and prominent configuration of matter, such as a settlement. It is equal to "where" and Ranganathan's "Space."
5. *Concept* = *crm:E55.Type*, comprising all kinds of universals, such as types of things, people, events, places, species, etc. This is a special case of "what." Ranganathan and many library subject catalogues do not distinguish between particular things and types of things; however, FRBR(O'Neill, 2002) introduces the notion of "Concept."

These categories should cover the domain of interest as a “base level” distinction similar to Lakoff (Lakoff, 1987), but they are neither completely disjointed nor absolute. Disjointedness is actually not helpful for recall. For instance, a settlement can be at least a “Thing” and a “Place.” A person (Actor) undergoing surgery or as a body in an excavated tomb, may be described, besides others, in terms of properties of a “Thing.” Such implications may appear odd in other contexts. A modern biologist may regard species as “Things”—i.e., human inventions with creators and other historical attributes—whereas other domains may see species only as “Concepts.” Therefore, the FCs should be adjustable/adjusted to the audience by adding or subtracting “less prototypical” subclasses (Lakoff, 1987), or even by extending them.

In the cultural–historical context, which we initially anticipated, queries with numerical values as parameters are rather rare (except for dates and geo-coordinates). However, in the 3D processing domain, such queries do occur. Therefore we will add in the future “crm:E54_Dimension” to the FCs, but a generic treatment of different metrics in intuitive user queries we have not (yet) explored.

3.2 Designing the Relationships

In addition to the URIs (Uniform Resource Identifiers), we assign to all RDF nodes in the RI textual (non-unique) labels with names or titles. This is becoming a good practice in RDF databases. Some also have descriptions in form of an rdf:literal. A user formulating a query in our system may first type in a keyword. A full-text search into all literals returns the associated nodes in the browser, together with minimal metadata and icons. Each node is marked by the FC it is an instance of.

For a more precise query, a user must first “select” (in the sense of the Structured Query Language (SQL) “Select” statement) the FC from which the question should return instances. In a normal digital library, this may be fixed to “document.” Then the user must compose a sort of “Where Clause.” The most simple one consists of a flat list of properties with the selected FC as domain and with range values combined by AND or OR. The design challenge is to find a minimal set of FRs intuitive to the user and easy to learn, which widely cover the respective discourse with high recall and a precision great enough not to be flooded by unrelated answers.

Pustejovsky (1995) observed how language disambiguates words by the relations to other words in a phrase. For instance, “He spoke to the museum” versus, “He walked around in the museum” seems contradictory in an ontology, but does not surprise people in whatever language we translate it to. This “complementary polysemy,” as Pustejovsky calls it, can be explained by classifying contextual expressions into relatively few, language-neutral categories (“quales”). When a user selects a relationship term and a value, we use a similar mechanism to disambiguate the relationship term as a further help to the user: The term is interpreted according to the selected FC and the FC the range value is instance of, rather than forbidding “illegal values.” Of course, the user may also filter values by an FC he or she chooses.

A good example is the term *from*, a very natural relationship term describing any sort of origin or provenance. For instance, in good museum practice and intuition “Things *from* New Guinea” may mean things *found*, *produced*, or *used* in New Guinea or things with *parts from* there. It may also mean things produced *by* people coming *from* New Guinea. This interpretation is common for all Place values. Museum metadata frequently contain the term “provenance” in this sense. However, “Things from J.W. Goethe” (an Actor) has a different interpretation: It could mean things *created*, *produced*, *modified*, *said*, *acquired*, *owned*, *kept,* or *used* by him or his household; gifts he gave or *received*; or awards he *received*. “Things from the Parthenon” (a Thing) may mean *parts* or pieces of the Parthenon, but it may also comprise inscriptions found on it. Quite differently, we would interpret “Actors (people) from New Guinea” as a sort of nationality concept, whereas “Actors (people) from Siemens Company” (Actor) would pertain to

membership. “Places from Time” make no sense. All interpretations correspond to composite path expressions in the CIDOC CRM. Constrained to a particular combination of FCs as domain and range, it is feasible to find all relevant expressions in the ontology for this interpretation.

Our empirical sources for the FR are “simple” metadata schemata, such as Dublin Core and VRA, but also the Europeana Data Model (EDM), experiences from structuring museum information (Bekiari et al., 2008), generalizations of the CRM itself, and intuition. We divide the relationships into those describing (1) how and what something is (classification, part-whole structure), (2) what an item has undergone in its history, and (3) what it may “show,” say, or refer to. We have not looked at relationships of intention, motivation, or cause, because they are rarely documented. This may be the subject of future extensions. In our current implementation, we have selected:

1. ***has type****:* denotes relations of an item to a classification, category, type, essential role or other unary property, such as a format, material, color. It generalizes over dc:type, dc:classification, dc:format, dc:language. The relationship is applicable to all FCs and has always range Concept.
2. ***is type of***: the inverse of “*has type”.* The relationship is applicable to all FCs and has always domain Concept.
3. ***has part :*** the inverse of *is part of*. Denotes structural relations of an item to a narrower unit it contains. The relationship is applicable to all FCs, except for Concept. In case of Actors, one would rather speak of “*has member*”, and persons are the minimal elements. Domain and range must be identical.
4. ***is part of***: denotes structural relations of an item to a wider unit it is contained in. The relationship is applicable to all FCs, except for Concept. In case of Actors, one would rather speak of “***is member of***”, and persons are the minimal elements. Domain and range must be identical.
5. ***from, has generator****:* denotes the relations of an item to constituents of a context in its history which is either significant for the item, or the item is significant for the context, “provenance” in the widest sense, including time intervals and places. In case of genealogy or group formation, natural language prefers the terms parent and founder respectively in order to refer to Actors. The relationship is a special case of has met.
6. ***is origin of, generator of***: the inverse of from, has founder or parent. In case of Actor as domain, one would rather speak of ***“is owner or creator of “.***
7. ***is similar or the same with***: denotes the symmetric relation between items that share features or are possibly identical. It is only usual for Things to document similarity manually. There exist enough comparison algorithms that deduce degrees of similarity automatically. We do not deal with these in this work.
8. ***has met****:* denotes the symmetric relation between items that were present in the same event, including time intervals and places. Applicable to any combination of FCs, except for Concepts.
9. ***refers to or is about****:* denotes the relation of an item that is information, contains information or has produced information to the item this information refers to or is about. The relation can even be extended to a Place from where such information originated.
10. ***is referred by/ is referred to at****:* the inverse of *refers to.*

11. ***borders or overlaps with***: this symmetric Relationship denotes the relationship between instances of the category place that limit with one another or overlap.
12. ***by***: denotes the active participation of an actor upon a Thing or Event

Table 1 describes which of the above relationships are applicable to respective combinations of FCs as domain and range. Each relationship has a different interpretation for each applicable combination of domain and range, which adapts the general meaning described above to the concrete case. All in all, we define ninety-one FR combinations. Nevertheless, the user is displayed a maximum of six FR per each FC combination.

(select)	Range(query parameter)					
	Thing	Actor	Place	Event	Time	Concept
Thing	8.has met	8.has met	9.refers to	9.refers to	5.from	1.has type
	9.refers to or is about	5.from	10.is referred to at	10.is referred to by	Destroyed on	
	10.is referred to by	9.refers to or is about	5.from	5.from	Created on	
	3.has part	10.is referred to by	Used at	Destroyed in	Modified on	
	7.is similar or same with	12.by	Created at	Created in	Used on	
	5. from	Used by	Found or	Modified in		
	4.is part of	Created by	acquired at	Used in		
		Modified by	Was created /produced by person from			
		Found or acquired by	Is/was located at			
Actor	8.has met	4.is member of	8.has met	9.refers to	9.refers to	1.has type
	6.is owner or creator of	3.has member	5.from	10.is referred to by	5.from	
	9. refers to	8. has met	9.refers to	5.from	8.has met	
	10.is referred to by	5.has generator	10.is referred to at	8.has met	Was brought into existence at	
		6.is generator of		Was brought into existence at	Was taken out of existence at	
		9.refers to		Was taken out of existence at	Performed action at	
		10.is referred to by		Performed action at	Influenced	
				Influenced		
Place	8.has met	8.has met	4.is part of	9.refers to	5.from	1.has type
	6.Is origin of	6.Is origin of	3.has part	10.is referred to by	10.refers to	
	9.refers to or is about	9.refers to or is about	11.borders or overlaps with	8.has met	8.has met	
	10.is referred to by	10.is referred to by				
		8.has met				
Event	6.is origin of	12.by	9.refers to or is about	9.refers to or is about	9.refers to or is	1.has type
	10.is referred to by	10.is referred to by	10. is referred to at	10.is referred to by	5.from	
	9.refers to or is about	9.refers to or is about	5.from	3.has part	starts	
	8.has met	8.has met		5.from	ends	
	created	brought into existence			has duration	
	destroyed	took out of existence				
	modified					
	used					
Time	6.is origin of	6.is origin of	6.is origin of	6.is origin of	4.is part of 3.has part	1.has type
Concept	2.is type of	2.is type of	2.is type of	2.is type of	2.is type of	1.has type 2.is type of

Table 1. Fundamental Categories and Fundamental Relationships.

The category Concept plays a special role. Concepts can be subdivided into subtypes of the FCs themselves, such as “Thing-Concepts,” “Place-Concepts,” etc. The FR 1.*has type* has domain all FCs, but the range is restricted to subtypes of the domain, such as “Thing. *has type*: Thing-Concepts,” “Place. *has type* Place-Concept, etc. Further, all relationships in table 1 can

be extended into “categorical” questions: for instance, the relation “Things *from* Place” can be extended into “Things *from type of* Place” via a join with “Place has type Place-Concept.” This is implemented as generic mechanism. The relationships in table 1 are not all disjointed. There are some subsumption relations between them: for instance, *has met* is in many cases a generalization of *from.*

Our framework foresees open-ended specializations of each of the FRs to dynamically meet demands for increased precision, down to the source ontology level. For instance, “Thing *was created at* Place” is an obvious specialization of “Thing *from* Place.” The user will be able to browse from the basic FRs to their specializations. This mechanism is absolutely impossible in an implementation based on core metadata. The user interface will further allow the user to combine the FRs even to simple path expressions, as if they were properties of the network; for instance, “all Things *from* Events *of type* ‘Excavation’ *at* time ‘1890-1910’ AND *at* place ‘Crete’.” In the 1990s, we implemented a similar system of customizable predefined “where” clauses that are presented as simple properties and can be combined to other queries for the Greek Archive of Monuments (Bekiari et al., 1998). It is, however, based on relational technology and a knowledge base managing the query mediation system. It is still in use. All together, these mechanisms allow for orders of magnitudes more queries than any core metadata system, and yet preserve their simplicity of access to the user.

Experimentation

In order to prove the practical evaluation and the usefulness of the proposed framework we have tested some of the most profound FRs against a semantic network of 272447 explicitly declared triples from legacy and digitization provenance metadata generated by partners from the 3D-COFORM project.

The most complex case we could demonstrate so far is the one referring to the “Ivory Panel A.15-1955,” an object from the Victoria and Albert Museum that shows the ”Ascension.” The object was digitized twice, which comprised dozens of processing steps and intermediate files. Based on it, we queried for the following FRs getting the number of results shown in Figure 1:

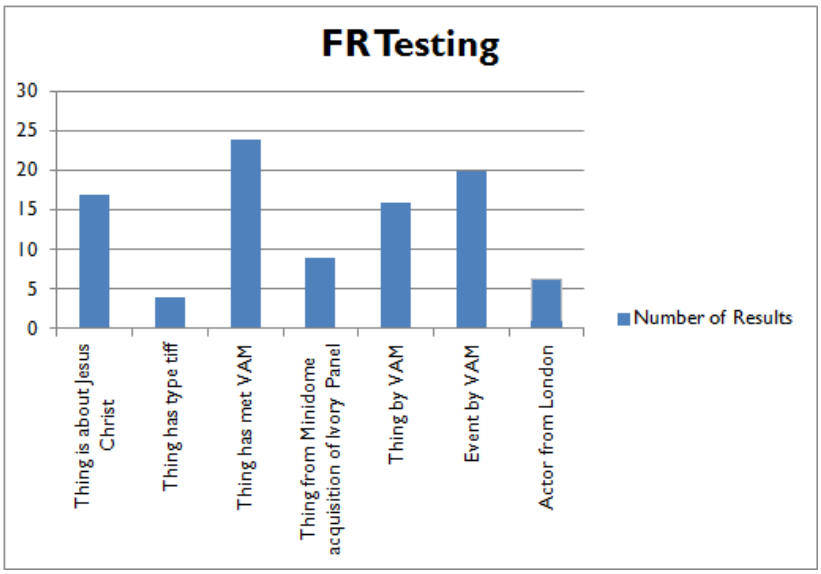

Figure 1: FR testing results

The queries collected all the expected results. The most profound one, the “Thing is about Jesus Christ,” collected all things referring to Jesus, including things that are digitization

products of things that refer to Jesus or derivatives of them. So, we are able to retrieve things that result from about eight transitive closures of part-whole and derivation properties.

Conclusions and further work

We propose and are implementing a new framework for querying semantic networks: For formulating queries, the user is presented a small list of configurable fundamental relationships and relevant specializations, easy to comprehend, that abstract by rich deductions from an underlying semantic network of much more specialized metadata comprising explicit event descriptions. These FRs simulate to the user a much simpler semantic network, which covers as many generic questions as possible with a high recall. Following the selection of a value for a query parameter, the user is presented with at most a dozen relationships, fewer than Dublin Core. The specializations of the FRs allow for systematically increasing the precision of queries on demand, down to the level of detail of the underlying network and ontology.

With this method, we can overcome the recall–precision gap between keyword and semantic search, the problems of formulating powerful queries in complex semantic networks, and the problems of simplifying the metadata themselves; but, of course, we rely on an efficient database technology. Still, our method of selecting FCs and FRs is mostly intuitive, but the method we present here is independent from specific configurations. Much work has been already done on testing, consolidating, and refining the FRs with respect to real user questions, including practical 3D data management and scholarly queries. Nevertheless, there is space for further improvement as well as for testing and verification in different environments except for the 3D-COFORM project..The complete analysis of the FRs is published in a technical paper (Tzompanaki-Doerr, 2012b) to which members of the CIDOC-CRM group can refer in order to start a discussion on the interpretation of each FR and the conformity with the users' needs.

Acknowledgements

We gratefully acknowledge the generous support from the European Commission for the Integrated Project 3D-COFORM (grant no. 231809). Also, we would like to thank Federico Ponchio (ISTI - CNR) for our flawless cooperation and Christos Asaridis (ICS-FORTH) for his active participation in part of the implementation of this framework.

References

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives. (2007). "DBpedia: a nucleus for a Web of open data." In *6th international The semantic Web and 2nd Asian conference on Asian semantic Web conference*. Heidelberg: Springer. 722–735.

Bekiari, Ch., Ch.D. Gritzapi, and D. Kalomoirakis. (1998) "POLEMON: A Federated Database Management System for the Documentation, Management and Promotion of Cultural Heritage." In *26th Conference on Computer Applications in Archaeology*. Barcelona.

Bekiari, C., L. Charami, M. Doerr, C. Georgis, and A. Kritsotaki. (2008). "Documenting Cultural Heritage in Small Museums." In *Annual Conference of CIDOC 2008*.

Ding, L., T. Finin, A. Joshi, R. Pan, R.S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs(2004). "Swoogle: a search and metadata engine for the semantic Web." In *the thirteenth ACM international conference on Information and knowledge management*. New York: ACM. 652–659.

Doerr, M. (2003). "The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine*, 24(3), 75–92. Available at: http://www.cidoc-crm.org/index.html

Doerr, M., and D. Iorizzo. (2008). “The dream of a global knowledge network-A new approach.” ACM, *Journal on Computing and Cultural Heritage*, 1(1). 1–23.

Doerr, M., S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, and H. van de Sompel. (2010a). The Europeana Data Model (EDM). In World Library and Information Congress: 76th IFLA General Conference and Assembly.. Available at: http://www.europeana.eu/portal/

Doerr, M., K. Tzompanaki, M. Theodoridou, Ch. Georgis, A. Axaridou, and S. Havemann. (2010b). “A Repository for 3D Model Production and Interpretation in Culture and Beyond.” In VAST 2010: 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage. Paris: Palais du Louvre. 97–104.

Fernandez, M., V. Lopez, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells. (2008). “Semantic Search meets the Web.” In The IEEE International Conference on Semantic Computing. 253–260.

Gangemi A., N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. (2002). “Sweetening Ontologies with DOLCE.” In EKAW 2002, Spain.166–178.

Hilman, D. (2005) Using Dublin Core - The Elements. 2005, last updated 28-Aug-2006. Consulted October 13, 2011. Available http://dublincore.org/documents/usageguide/elements.shtml

Hyvoenen, E., T. Ruotsalo, T. Haeggstroem, M. Salminen, M. Junnila, M. Virkkilae, M. Haaramo, E. Maekelae, T. Kauppinen, and K. Viljanen. (2006). “CultureSampo—Finnish Culture on the SemanticWeb: The Vision and First Results.” In 12th Finnish Artificial Intelligence Conference STeP. Available at: http://www.kulttuurisampo.fi/index.shtml

Lagoze, C., and J. Hunter. (2001). “The ABC Ontology and Model.” *Journal of Digital Information.*

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Lei, Y., V. Uren, and E. Motta. (2006). “SemSearch: A Search Engine for the Semantic Web.” In Staab, S., and V. Svátek (eds.). *Managing Knowledge in a World of Networks*. Heidelberg: Springer LNCS. Volume 4248. 238–245.

Lopez, V., E. Motta, and V. Uren. (2006). “PowerAqua: Fishing the Semantic Web.” In Sure, Y., and J. Domingue (eds.). *The Semantic Web: research and applications*. Volume 4011. 393–410. Available at: http://technologies.kmi.open.ac.uk/poweraqua/

Moreau, L., J. Freire, J. Myers, J. Futrelle, and P. Paulson. (2007). *The Open Provenance Model*. University of Southampton. Available at: http://openprovenance.org/

O’Neill, E.T. (2002) “FRBR: Functional Requirements for Bibliographic Records, Application of the Entity-Relationship Model to Humphry.” In *Library Resources & Technical Services* (46)4.

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Ranganathan, S.R. (1965) *A Descriptive Account of Colon Classification*, Bangalore: Sarada Ranganathan Endowment for Library Science.

Theodoridou M., Y. Tzitzikas, M. Doerr, Y. Marketakis, and V. Melessanakis. (2010) “Modeling and Querying Provenance by Extending CIDOC CR.M. Distributed and Parallel Databases.” In *Distributed and Parallel Databases*, (27)2.

Tzompanaki, K., and M. Doerr (2012a). “A New Framework For Querying Semantic Networks”.In the *Museums amd the Web 2012: the international conference for culture and heritage on-line*. April 11-14, San Diego, CA, USA. Also available at: http://www.museumsandtheweb.com/mw2012/papers/a_new_framework_for_querying_semantic_networks

Tzompanaki, K. and M.Doerr (2012b). Fundamental Categories and Relationships for Intuitive querying CIDOC-CRM based repositories. Technical Report 2012. Available at:

http://www.ics.forth.gr/tech-reports/2012/2012.TR429_Intuitive_querying_CIDOC-CRM.pdf

VRA Core 4.0 Element Description. (2007). Last updated May 4, 2007. Consulted October 14, 2011. Available at: http://www.loc.gov/standards/vracore/VRA_Core4_Element_Description.pdf