Benchmarking RDF Schemas for the Semantic Web*

Aimilia Magkanaraki, Sofia Alexaki, Vassilis Christophides,
and Dimitris Plexousakis

Institute of Computer Science, FORTH
Vassilika Vouton, P.O Box 1385, GR 711 10, Heraklion, Greece
{aimilia, alexaki, christop, dp}@ics.forth.gr

Abstract. Describing web resources using formal knowledge (i.e., creating metadata according to a formal representation of a domain of discourse) is the essence of the next evolution step of the Web, termed the Semantic Web. The W3C's RDF/S (Resource Description Framework/Schema Language) enables the creation and exchange of resource metadata as normal web data. In this paper, we investigate the use of RDFS schemas as a means of knowledge representation and exchange in diverse application domains. In order to reason about the quality of existing RDF schemas, a benchmark serves as the basis of a statistical analysis performed with the aid of VRP, the Validating RDF Parser. The statistical data extracted lead to corollaries about the size and the morphology of RDF/S schemas. Furthermore, the study of the collected schemas draws useful conclusions about the actual use of RDF modeling constructs and frequent misuses of RDF/S syntax and/or semantics.

1. Introduction

In the next evolution step of the Web, termed the Semantic Web [2], vast amounts of information resources (data, documents, programs) will be made available along with various kinds of descriptive information, i.e., metadata. Better knowledge about the meaning, usage, accessibility, validity or quality of web resources will considerably facilitate automated processing of available Web content/services. The Resource Description Framework (RDF) [31] enables the creation and exchange of resource metadata as normal Web data. To interpret these metadata within or across user communities, RDF allows the definition of appropriate schema vocabularies (RDFS) [6]. However, the fact that several communities, even with similar needs, have developed their own metadata vocabularies independently indicates the need for schema repositories facilitating knowledge sharing. In this way, already defined concepts or properties for a domain can be either reused as such or simply refined to meet the resource description needs of a particular user community, while preserving a well-defined semantic interoperability infrastructure (i.e., through the RDF/S data model primitives such as *SubClassOf* and *SubPropertyOf*).

* This work has been partially supported by the EU Projects OntoWeb (IST-2000-29243) and Question-How (IST-2000-28767).

There exist several ongoing efforts to build registries of available metadata vocabularies. Among those efforts we note the SCHEMAS Project [48], the DESIRE Registry [17], the SWAG Dictionary [50] and the Xmlns.com [61]. The SCHEMAS Project provides "a forum for metadata schema designers involved in projects under the IST Programme and national initiatives in Europe". Part of the work undertaken in this project was the construction of a registry for metadata schemas in RDF/S. The DESIRE Registry adopts the same approach as the SCHEMAS Project, while the Semantic Web Agreement Group Dictionary also highlights the need for interconnected vocabularies of terms, in order to form "a third party index, where parties can register the semantic connections between schemas". Finally, the experimental Xmlns.com intends to provide an Internet domain suitable for simple Web namespace management.

Although we share similar motivations with the above initiatives, the focus of our work is different. More precisely, we are interested in the structural analysis of the available RDF/S schemas from various applications. Our contribution is twofold: (a) we have collected (28 schemas from 9 different application contexts) and classified across two dimensions (i.e., domain of discourse, semantic depth of resource descriptions) available RDF/S schemas on the web, and (b) we provide complete statistics about the size and morphology of these schemas (e.g., number of classes/properties, breadth and depth of hierarchies). We believe that benchmarking existing RDF/S schemas, apart from being an added-value service of the above registries, is quite useful for testing the functionality and performance of existing RDF validation, storage, inference and query tools [30]. Furthermore, the conclusions presented in this paper about the actual use of RDF/S modeling constructs in real scale applications, provide a helpful feedback to the Semantic Web community regarding future versions or extensions of RDF/S. To the best of our knowledge, there has not been a previous attempt in this direction. The recent study of RDF data on the Web [21] does not comprise the harvested RDF/S schemas and mainly addresses Portal applications (e.g., Netscape Open Directory, as we have studied in [1]). The set of the schemas we collected are available on the "The ICS-FORTH RDF/S Schema Registry" Web page [44].

2. RDF/S in a Nutshell

The Resource Description Framework and Schema Language (RDF/S) ([31], [6]) aim to facilitate the encoding, exchange, processing and reuse of resource metadata while each user community is free to specify its own description semantics in a standardized, interoperable, human-readable manner via an XML-based infrastructure [59].

The RDF data model is based on the notion of "resource". Everything, concept or object, available on the web or not, can be modeled as a resource identified by a unique URI ([3]). With the constructs of the RDF data model we can describe interrelationships among resources in terms of named properties and values. Properties capture either attributes of a resource or binary relationships between resources. The definition of these attributes/relationships and their semantical attribution is accomplished through the RDF Schema Language (RDFS) [6].

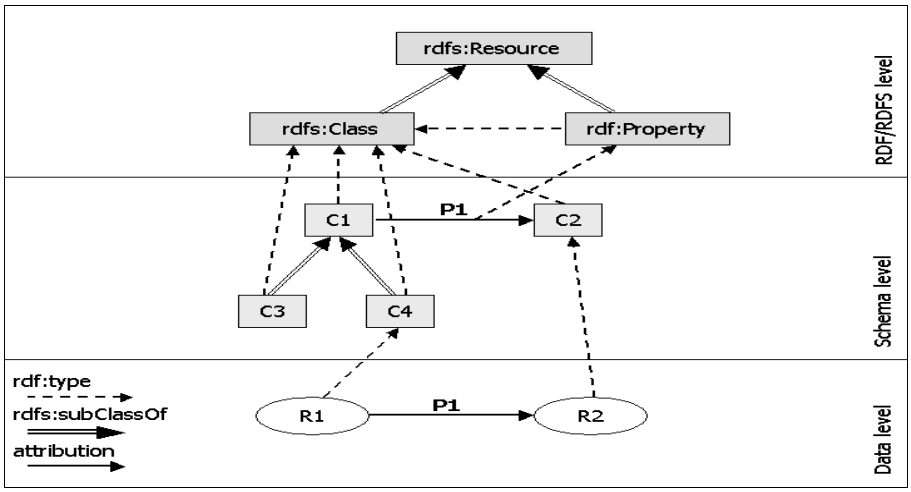

Fig. 1. Abstraction levels in a typical RDF/S schema

An RDF Schema declaration is expressed in the basic RDF Model and Syntax Specification [31] and consists of classes and properties. In other words, the RDF Schema mechanism provides a type system for RDF models i.e., a vocabulary of the valid terms that can be used to describe resources. We briefly summarize the basic RDF/S features, which are available for representing domain knowledge:

- **Core Classes**: The basic constructs of the RDF/S meta-language are *Class*, *Property* and *Container*, which correspond to entities, relations or attributes and complex or structured values, respectively.
- **Abstraction mechanisms**: RDF/S features the following abstraction mechanisms: (multiple) class or property inheritance and (multiple) classification of resources. The former is declared using the *rdfs:subClassOf* or *rdfs:subPropertyOf* core properties while the latter using the *rdf:type* core property. Typically, we identify three core abstraction levels, which are depicted in **Figure 1**.
- **Restriction mechanisms**: Although RDF/S does not provide elaborate mechanisms for defining property restrictions (as in the case of Description Logic or frame languages), we can declare simple domain and range restrictions through the *rdfs:domain* and *rdfs:range* core properties.
- **Documentation facilities**: The properties *label*, *comment*, *isDefinedBy* and *seeAlso* are used to document the development of a schema.
- **Reification mechanisms**: Although not expressible at schema level, RDF provides mechanisms for representing statements. This mechanism - formally known as *reification* - is applicable at the data level and the constructs used for this process are *statement*, *subject*, *predicate*, *object* and *type*.

The XML namespace facility [5] plays a crucial role in the development of RDF schemas, since it enables the reuse of terms from other schemas. With the use of an XML namespace, descriptive terms (i.e., class or property names) are uniquely identified by a URI (i.e., playing the role of a name prefix) as normal web resources.

To parse and validate RDF/S schemas, we have used the ICS-FORTH Validating RDF Parser (VRP) [55]. The parser analyzes syntactically RDF/XML statements according to the RDF Model and Syntax Specification [31] and the validator checks whether both the RDF schemas and related instances satisfy the semantic constraints implied by the RDF Schema Specification [6]. Additionally, VRP can extract statistics about the structure of schemas, as well as, quantitative data from related instances.

3. A Classification of Existing RDF/S Schemas

The set of the RDF/S schemas collected were classified under the following two dimensions: (a) the ***application domain*** they refer to and (b) the ***semantic depth*** in which they have been developed. **Table 1** presents the collected schemas classified according to these dimensions. The latter term refers to the degree in which the various relationships expressible in RDF/S (e.g., *subClassOf*, *subPropertyOf*, user-defined properties) are exploited in schema design.

Regarding the former dimension, we have identified the basic application domains presented below, which by no means restrict the range of possible knowledge domains that RDF/S can represent:

- **Cultural Heritage/Archives/Libraries**: The schemas of this sector serve two different functionalities: they either provide standardized definitions of concepts and processes referring to libraries, archives, museums and cultural heritage applications in general, or they provide guidelines for the encoding, structure and exchange of information.
- **Educational/Academic**: The schemas of this kind serve either as a vocabulary for facilitating the worldwide exchange of learning resources like exercises, diagrams, slides and videos or provide means to describe and formalize aspects of research activities and scientific publications.
- **Publishing/News**: The schemas of this sector provide vocabularies for the encoding and interchange of news information among individuals and mass media organizations (news agencies, newspapers etc.). They refer to any source of publication and electronic material in whichever format (CDROM, DVD, slide etc.), linkable or citable, in print or on-line and to its properties, such as its creator, period of validation, edition or subject-encoding scheme.
- **Audio-Visual**: Schemas under this category are essentially ontologies of the basic concepts used to represent information about people in the film industry, multimedia production and distribution. They also describe information about processes and events related to every aspect of film/multimedia production and selling, such as advertising, casting, acting etc. Up to now we have classified only one schema in this category but similar efforts are under development, e.g., MusicBrainz Metadata Initiative[1] [37].

[1] Although the related schema is still under development, the MusicBrainz RDF Data Dump can be downloaded from http://www.musicbrainz.org/download.html.

- **Geospatial/Environmental**: The schemas of this sector encompass a terminology of concepts and guidelines for the representation and sharing of geospatial/geographical and environmental information. Although we have classified two schemas under this sector, this domain is currently less exploited than the others from a knowledge representation perspective.
- **Biology/Medicine**: The schemas classified under this category are mainly thesauri of terms and controlled vocabularies that provide definitions and semantic relations between the terms. The main functionality of such schemas is to facilitate interoperability among systems storing, processing and querying biological or medical data and to facilitate communication between people by providing registered definitions of the terms.
- **E-Commerce**: Such schemas are mainly dictionaries or taxonomies that clarify the terms used in e-commerce applications, e.g., real estate investment management, advertisement or web-based sales. They provide a central reference of registered definitions about accounts, actors, services, economic transactions etc. that are used to facilitate the communication between clients, vendors, enterprises, providers or any other entity that participates in economic transactions.
- **Ubiquitous/Mobile/Grid Computing**: As in the case of the Audio-visual domain, more efforts of this context are under way, e.g., efforts from the WAP Forum [56]. The schemas of this sector are mainly vocabularies of concepts enabling the exchange of data (e.g., technical characteristics of the client or the network) between devices, as well as data related to resource allocation by a Grid scheduler.
- **Cross-Domain**: These schemas are usually vocabularies providing general-purpose descriptive terms from a more extensive domain and can be used in a variety of application-neutral contexts. Thus, some of the schemas presented can serve as exchange formats or as thesauri of general terms with the aim of better facilitating communication and interoperability.

The latter classification dimension refers to the structure of the RDF Schemas, and in particular the semantic depth of resource descriptions in which they have been developed, i.e., the kind of relations used for modeling a domain. In the broader sense of the term, we can characterize each schema as an ontology, since it can constitute an agreed vocabulary shared among people and organizations. For the purpose of our study, we have adopted the following semantic depth levels used in the implementation of an ontology [18]:

- **Dictionaries and Vocabularies**: the schemas developed at this level define simple lists of concepts and their definitions. Most of the times, they consist only of class definitions and their structure is almost flat.
- **Taxonomies**: the characteristic of taxonomies is that the main relation they define between concepts is that of specialization. The hierarchy depth of taxonomies depends on the detail in which a schema implementer decides to refine domain concepts.
- **Thesauri**: besides defining relations among broader/narrower terms through the definition of hierarchies, a thesaurus also declares relations of equivalence, association and synonymy. The nature of these semantic relations is what distinguishes thesauri from taxonomies.

Table 1. Classification of RDF Schemas according to application domain and semantic depth

Application Domain	Dictionary/ Vocabulary	Taxonomy	Thesaurus	Reference Model
Cultural Heritage/ Archives/Libraries	•**Euler** [22] •**RSLP-CLD** [46]			•**CIDOC** [13]
Educational/ Academic	•**IMS** [29] •**Universal** [52]	•**Mathem. International** [34]		•**CERIF** [12]
Publishing/News	•**BibLink** [4] •**DOI** [19] •**SlinkS** [49] •**RSS** [47]			
Audio-Visual				•**IMDB** [28]
Geospatial/ Environmental	•**CZM** [14]			•**GML** [23]
Biology/Medicine			•**Gene Ontology** [24]	
E-Commerce		•**BSR** [7] •**UNSPSC** [53]		•**RED** [16]
Ubiquitous/ Mobile/Grid Computing	•**CC/PP** [10]			•**P3P** [42] •**RDF Calendar** [43] •**Scheduler's Allocation Schema** [26]
Cross-Domain	•**CERES/NBII** [11] •**Dublin Core** [20] •**Lexical WordNet** [58]		•**MetaNet** [35]	•**Limber Thesaurus** [32] •**Top Level Ontology** [51]

- **Reference Models**: a conceptual reference model combines all the previously stated relations to capture the semantics of a domain. This body of knowledge, describing a domain or subject matter, comprises a representation vocabulary for referring to the concepts in the subject area and the logical statements that describe the nature of the terms, the relations among the terms and the way the terms can or cannot be related to each other.

The set of schemas presented in **Table 1** indicates that RDF/S, due to its domain-neutral nature, is gaining acceptance for simple ontology construction (i.e., no logical axioms) in various sectors. Hence, we can argue that useful lessons can be learned from performing a detailed analysis of the defined schemas. Such an analysis reveals the degree to which RDF/S has been understood and adopted, as well as, common misunderstandings or mistakes. Its results can be used as feedback to schema designers. They also substantiate the need for tools for schema validation. The proliferation of schemas defined in RDF/S also calls for scalable tools for their storage and querying, such as the tools provided by the ICS-FORTH RDFSuite [45]. The analysis of the schemas is the topic of the next section.

4. Analyzing the Structure of RDF/S Schemas

Before presenting the statistics we extracted for the RDF/S schemas of our testbed, we consider useful to give some general comments regarding our overall experiment. First of all, harvesting schemas on the web was a time-consuming task due to inexistence of complete (RDF/S) schema repositories. This fact stresses the need for rich RDF/S schema registries. A second observation made was that a considerable number of schemas were developed with errors ranging from missing or wrong declarations of classes and properties, to misuse of the RDF/S modeling constructs and to confusion between the rdf and rdfs namespaces. This fact indicates the need for generally accepted RDF authoring, parsing and validation tools. We furthermore observed that schema designers utilize mostly the core RDF/S constructs (i.e., simple definitions of classes or properties). A last observation is related to the use of the Dublin Core Element Set [20] as a widely accepted top-level ontology that is either reused as such or refined by the schemas of our testbed (i.e., direct relationships between the schemas was not encountered). As suggested in [18] and [25], richer, cross-domain (top-level) ontologies (schemas) are needed to provide more elaborate forms of semantic interoperability between various application domains.

Table 2 illustrates the statistics extracted by our testbed. The columns of this table correspond to various structural characteristics of a schema. In particular, the sub-columns "*Total*" under the "*Classes*" and "*Properties*" columns refer respectively to the total number of classes and properties either locally defined in a specific schema or reused from an external namespace. The sub-columns "*Hierarchies*" are used to present respectively the number of class and property hierarchies declared and refer to hierarchies whose depth is greater than 0. Note that class hierarchies' roots are the direct subclasses of *rdfs:Resource,* while as a property hierarchy root we consider any property without superproperties. We can consider each hierarchy as a different "facet" of the schema implemented, that is orthogonal information assets under which resources can be classified. These statistics can be used to measure the "size" of a schema. Column "*subClassOf*" refers to detailed statistics about the class hierarchies defined and column "*subPropertyOf*" refers to statistics about the property hierarchies. "*Depth"* records the depth of class and property hierarchies, while "*Subnodes"* and "*Supernodes"* refer respectively to the in- and out-degrees of theses hierarchies (or schema DAGs in case of multiple inheritance). These statistics can be used to measure the "morphology" of a schema. In the case of class hierarchies, "*Depth*" is the length of the *subClassOf*-path from a node to the root. Depth is defined similarly for the case of property hierarchies, as the length of the *subPropertyOf*-path from a given property to the hierarchy root. "*Subnodes*" and "*Supernodes*" characterize, respectively, the number of subnodes and supernodes attached to a node when multiple *rdfs:SubClassOf* and *subPropertyOf* RDF/S properties are used. For each of the above 3 cases, we provide the maximum and average occurrence. The gathered statistics are reported for all schema hierarchies.

One general observation we can make from the data of **Table 2** is that most of the schemas define few classes and properties, with the exception of Real Estate Data Consortium [16], Basic Semantic Registry [7], UNSPSC [53] and Gene Ontology [24]. We can consider these schemas as rich domain models of the application to which they refer. Via the extensibility mechanisms of RDF/S, a designer can extend

Table 2. Statistical data about the structure of schemas

Schema	Classes		Properties		SubClassOf						SubPropertyOf					
	Total	Hierar-chies	Total	Hierar-chies	Depth		Sub Nodes		Super Nodes		Depth		Sub Nodes		Super Nodes	
					Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg
CIDOC	77	3	205	20	8	4.4	7	1.1	2	1.1	2	1.2	10	0.4	2	0.4
Euler	20	2	22	4	1	1	14	0.8	1	0.8	1	1	1	0.2	1	0.2
RSLP-CLD	11	2	43	7	1	1	3	0.4	1	0.4	1	1	7	0.5	1	0.5
CERIF	42	2	142	3	1	1	13	0.3	1	0.3	1	1	18	0.3	1	0.3
IMS	17	1	8	1	2	2	5	0.7	1	0.7	1	1	2	0.3	1	0.3
Math. Internat.	211	1	0	0	11	7.9	43	1.6	9	1.6	-	-	-	-	-	-
Univers.	5	0	13	0	-	-	-	-	-	-	-	-	-	-	-	-
BibLink	14	2	20	2	1	1	5	0.6	1	0.6	1	1	1	0.1	1	0.1
DOI	13	1	13	0	1	1	7	0.6	1	0.6	-	-	-	-	-	-
SLinkS	20	1	56	4	2	1.6	2	0.2	1	0.2	1	1	1	0.1	1	0.1
RSS	6	0	9	3	-	-	-	-	-	-	1	1	2	0.4	1	0.4
IMDB	65	2	182	0	2	1.8	37	0.9	1	0.9	-	-	-	-	-	-
GML	20	3	33	1	3	1.9	5	0.8	2	0.8	2	2	6	0.6	1	0.6
CZM	77	1	66	0	6	4.3	4	0.9	1	0.9	-	-	-	-	-	-
Gene Ontology	6993	175	9	0	12	5	106	1.2	6	1.2	-	-	-	-	-	-
BSR	2714	230	1754	0	4	1.7	62	0.6	1	1	-	-	-	-	-	-
RED	5073	5	285	1	5	2.4	763	1.9	5	1.9	3	1.5	233	1.6	2	1.6
UNSPSC	16506	57	2	0	3	3	63	1	1	1	-	-	-	-	-	-
CC/PP	18	3	3	1	2	1.2	4	0.7	1	0.8	1	1	1	0.3	1	0.3
P3P	414	14	365	4	3	2.1	245	1.6	4	1.7	1	1	312	0.9	1	0.9
RDF Calendar	57	17	92	2	3	1.3	4	0.6	3	0.6	1	1	3	0.1	1	0.1
Schedul. Allocat.	16	1	23	2	1	1	3	0.2	1	0.3	1	1	2	0.2	1	0.2
Dublin Core	2	0	22	0	-	-	-	-	-	-	-	-	-	-	-	-
CERES/ NBII	8	1	14	0	1	1	3	0.4	1	0.6	-	-	-	-	-	-
Lexical WordNet	9	1	5	0	2	1.3	4	0.6	1	0.6	-	-	-	-	-	-
Limber Thesaur.	11	1	17	3	2	1.3	3	0.4	1	0.5	1	1	4	0.6	1	0.6
MetaNet	66	3	11	2	2	1.6	17	0.9	2	1	1	1	2	0.3	1	0.3
TopLevel Ontology	189	1	141	1	11	6.3	11	1	3	1.1	6	3	18	1	2	1

them by defining application-specific concepts. We can additionally observe that, when many classes are defined, the number of properties declared is relatively low and vice versa. It could be claimed that schema implementation is *property-centric* or

Table 3. Percentage of Multiple Inheritance

Schema	% Multiple Inheritance of classes	% Multiple Inheritance of Properties
Real Estate Data Consort.	**0.840**	**0.757**
P3P	**0.644**	**-**
RDF Calendar	**0.100**	**-**
CIDOC	**0.168**	**0.068**
Mathematics International	**0.333**	**-**
GML	**0.050**	**-**
Gene Ontology	**0.184**	**-**
MetaNet	**0.030**	**-**
Top Level Ontology	**0.068**	**0.035**

class-centric, depending on whether the designer decides to model concepts as classes or properties. This choice is a design decision that has to be made in order to better capture the semantics of the modeled domain.

In general, the schemas examined are shallow and they tend to be developed in breadth rather than depth. The maximum depth observed was 12 (Gene Ontology), while the maximum breadth (i.e., number of subnodes) was 763 (Real Estate Data Consortium). The fact that an ontology exhibits a sizable number of subnodes for a given node might indicate that there is a modeling deficiency and that the schema implementer should consider the addition of intermediate nodes [40]. Similarly, the sizable number of supernodes might signify that there are repeated declarations of subsumption relationships or that the modeling of the domain knowledge is not clear. The average number of subnodes, however, tends to be less than 1.0, a fact that indicates the existence of nodes not attached to a hierarchy. The number of hierarchies (whose depth is greater than 0) defined is also low, regardless of the number of classes or properties declared. This fact indicates the centralization of concepts around some top-level terms and the formulation of few large hierarchies instead of many small hierarchies of terms.

In particular, the number of schemas using the *subPropertyOf* construct is relatively small. The majority of schemas do not use this construct or they use it to a limited extent. Our study has shown that, when this construct is used, the top-level property is most of the times unconstrained (i.e., there are no imposed domain and range restrictions). Furthermore, *subPropertyOf* is used mainly for relationships between classes rather than attributes of a class. The phenomenon of properties with undefined range or domain was also encountered frequently for a set of non top-level properties. However, when domain restrictions were defined, it was noticed that several properties were declared with multiple domain definitions.

Additionally, from **Table 3**, we can see that *multiple inheritance* for classes, although not widely used (only in 9 out of 28 schemas), was more frequent than *multiple inheritance* for properties (only in 3 out of 28 schemas). The percentage of classes with multiple inheritance ranges from 33.3% (in Mathematics International) to 3% (in MetaNet) while for properties ranges from 6.8% (in CIDOC) to 3.5% (in Top-level Ontology), with the exception of the Real Estate Ontology, which is 84% for classes and 75.7% for properties and P3P, which is 64.4% for classes. Unfortunately, in the Real Estate Ontology the large number of multiple inheritance occurrences, is

due to the repeated declaration of *SubClassOf*/*SubPropertyOf* of a class/property to all its ancestors in the corresponding hierarchy (no cycles in class/property hierarchies were detected). The same phenomenon is partially observed in P3P.

The examination of instance files reveals that *multiple classification* of resources was rarely used, apart from the case of the CIDOC ontology instance files. However, we must state that we have not found a substantial number of instance files for the examined RDF schemas (with the exception of the RSS schema widely used by Portals like CNET.com and xmlTree[2]). At schema level, multiple classification was observed only in the P3P ontology. Furthermore, we have not encountered at all the *reification* mechanism. Reification is not expressible at schema level in RDF/S and it is also highly likely that the mechanism is not widely understood. Furthermore, we can credit its absence to the fact that a schema/ontology designer wishes to represent domain knowledge and not statements about information resources. Finally, one construct that was not used was that of containers (Sequences, Bags, Alternatives). On the contrary, the domain and range restriction mechanisms for properties as well as the documentation facilities (*comment*, *label* etc) were extensively used.

One last corollary refers to the correlation between the richness of modeling techniques used and the semantic depth. As we can observe from **Table 2**, the majority of schemas classified as "Reference Models" in **Table 1** exhibit a rather complete use of RDF/S modeling constructs (e.g., Real Estate Data Consortium [16], CIDOC [13], CERIF [12], GML [23], P3P [42] and Top Level Ontology [51]). In contrast to other schemas, they define deep and/or broad hierarchies of both classes and properties. Furthermore, they utilize multiple inheritance for classes and/or properties to a greater extent than other schemas. Although it is rather premature to draw general conclusions about the morphological construction of RDF/S schemas, the evidence collected by our experiments points to a tight correlation of the notion of semantic depth to the variety of modeling constructs used by schema designers.

5. Towards Richer RDF/S Modeling Constructs

Besides commenting on the morphology of the examined schemas, the whole process of this survey gave us the stimulus to also study the modeling techniques actually used by schema designers. In this section, we will present the most common semantic errors made and will discuss the involved RDF/S modeling constructs. These errors are mainly due to modeling deficiencies that future RDF/S versions should cover. These deficiencies are partially addressed by current RDF/S extensions, such as DAML+OIL ([15]), and real-scale Semantic Web applications seems to demand the incorporation of credible solutions in the core RDF/S standard.

5.1 Meta-schemas

An important number of the RDF schemas of our testbed extend the core RDF/S meta-model. This is mainly performed by refining the classes *rdfs:Class* and *rdf:Property* using the *rdfs:subClassOf* relation (see **Figure 2**). We should note that

[2] http://home.cnet.com, http://www.xmltree.com

the separation of meta-schemas, schemas and resource data is not clear in either RDF/S [6] or in the recent RDF M &T [27]. In fact, as eloquently commented in [39], RDF/S does not distinguish between the data and schema levels and all information is represented uniformly in the form of a graph. As a consequence, a number of redundancies or semantic inconsistencies in class or property declarations arise, as explained in the sequel. We believe that a clear separation is useful for application designers as previous experience in semantic-networks suggests (e.g., Telos [38]).

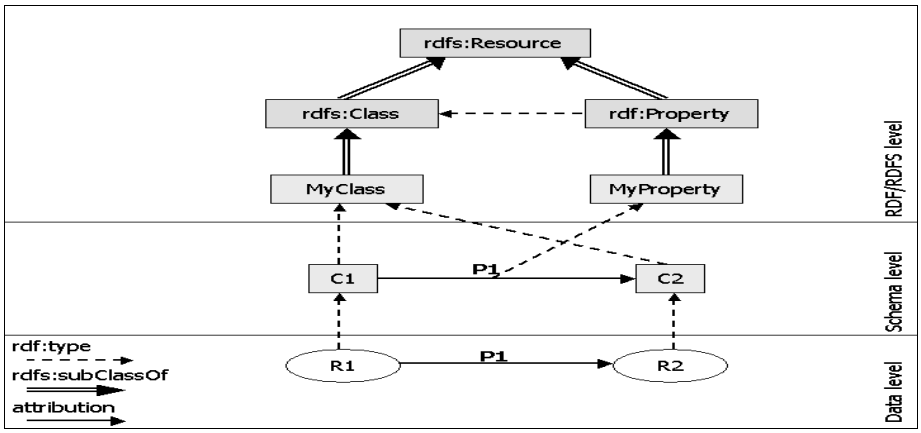

Fig. 2. Modeling Meta-schemas

The resources which extend the classes *rdfs:Class* and *rdf:Property* (e.g., *MyClass* and *MyProperty*) are of type *rdfs:Class* (see RDF M&T rule 9b). Although this kind of information should be inferred by the RDF processors, it has been explicitly stated in a number of RDF/S schemas of our testbed. Furthermore, since *MyClass* is subclass of *rdfs:Class* and *MyProperty* is subclass of *rdf:Property*, the resources that are declared instances of the class *MyClass* (e.g., C1) are classes, whereas the resources that are declared instances of the class *MyProperty* (e.g., P1) are properties (see RDF M&T rule 11). Hence, it is redundant to declare that a class (property) is both instance of a subclass of the *rdfs:Class* (*rdf:Property*) and instance of the *rdfs:Class* (*rdf:Property*).

We should also point out that in RDF/S a class can have as instances other classes without being declared as a subclass of *rdfs:Class* (i.e., as meta-class). However, it is advisable to declare explicitly such a class as a metaclass (as shown above) so that this knowledge can be exploited at the application level. In this manner, the separation between the different levels becomes clear.

Finally, although the RDF/S specification claims that properties are first-class citizens, properties are not treated as equally as classes. In RDF/S both a meta-class of classes and a meta-class of properties is a class, in contrast to the knowledge representation language Telos [38] where a meta-class of individuals is a class but a meta-class of properties (meta-property) is a property. Hence, while in Telos a meta-property can have domain and range, in the RDF/S model it cannot. Furthermore,

notice that, at the data level, *P1* cannot be "of type" *P1*, as is the case for classes, where we say that a resource *R1* is of "type" *C1*. This is attributed to the fact that the *rdf:type* property is applicable only for classes and RDF/S does not provide us with an instantiation mechanism for properties at the data level.

5.2 Non-Binary Relations

The RDF data model is based on binary relations, i.e., relations between two classes. However, there are modeling circumstances where the use of ternary or higher arity relations is needed. At the data level, we can implicitly represent ternary relations by using the *rdf:value* property and an intermediate resource [31]. The *rdf:value* property is used to denote the principal value of the main relation. To illustrate the representation of ternary relations, we use the following example in RDF/XML serialization.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:a="http://www.ics.forth.gr/schemas/testschema#">
  <rdf:Description
         rdf:about="http://www.monitors.com/Trademark1">
         <a:size rdf:value="17"
          a:measure="inches"/>
  </rdf:Description>
</rdf:RDF>
```

This example illustrates the case where we need to represent, apart from the size of a monitor, the measuring system used (e.g., centimeters or inches). We would like to model such a relation in the schema as presented in the left part of **Figure 3**. The RDF/XML serialization of this ternary relation is given at the right part of **Figure 3**. The inability to model ternary or higher arity relations at the schema level stems from the fact that the domain of a property should always be a class. Thus, the syntax in the above format is not valid. In our testbed, the definition of a property as the domain of another property was encountered in several schemas.

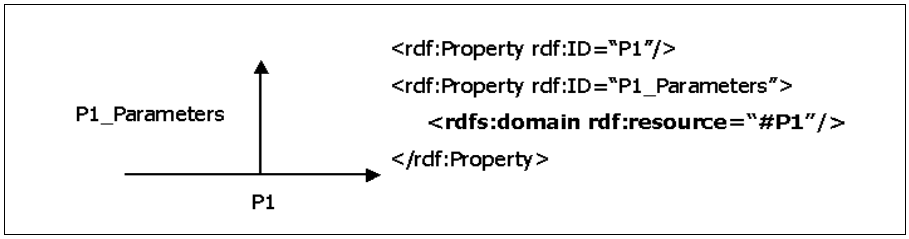

Fig. 3. Ternary relations at schema level

5.3 Enumerated Types and Specialization of rdfs:Literal

In our study, we observed the need for enumerated types, e.g., to define the possible values that a property can have (e.g., *Value1*, *Value2*, *Value3*). Although not

explicitly supported by RDF/S, schema designers treated enumerated types by representing them as shown in **Figure 4**. The possible values the property can have are defined to be instances of its range class. Unfortunately, the same mechanism is not applicable in the case of the *rdfs:Literal*, i.e., we cannot define *Value1* or *Value2* as instances of *rdfs:Literal*.

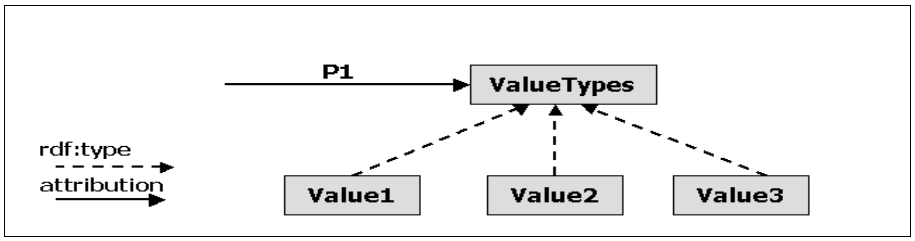

Fig. 4. Representing enumerated types

Additionally, a frequently encountered situation was the specialization of the *rdfs:Literal* class in order to support a richer set of data types. XML data types [60] can be used for this purpose while ensuring interoperability with other XML-based applications.

6. Related Work and Summary

During the last decade several studies have been conducted on the formal aspects of knowledge representation languages and ontologies ([33], [36], [54], [57]). In [9] a structured bibliography of studies related to ontologies is provided, while [8] and [25] list a set of B2B standards and content standardization efforts respectively, as well as classification criteria for them. In [41], a framework for comparing ontologies is developed and 10 representative ontologies are examined (e.g., CYC, Generalized Upper Model, UMLS, WordNet). From the set of qualitative comparison criteria proposed for ontologies, we can distinguish criteria referring to their general characteristics (e.g., purpose, coverage, size, formalism used, accessibility, design process, evaluation methods) and criteria about the content of the ontology (e.g., taxonomic organization, top-level divisions, internal structure of concepts, granularity). Our work is complementary, in the sense that it proposes *quantitative* criteria about the structure of RDF/S schemas representing various kinds of ontologies. The relationship between qualitative and quantitative ontology comparison is a subject, which deserves further study and experimentation. Intuitively, a qualitative criterion can be interpreted into a set of quantitative criteria. For example, the taxonomic organization of an ontology could possibly be determined by making a reduction to the number of property and class hierarchies and the in- and out-degrees of theses hierarchies. These statistics indicate whether an ontology is organized in large or a number of smaller taxonomies, as well as, the degree to which an ontology is well-structured and complete. We believe that the set of the schemas presented in this paper form a suitable testbed for the application of such qualitative/quantitative comparison of domain ontologies.

References

[1] S. Alexaki, V. Christophides, G. Karvounarakis, D. Plexousakis, K. Tolle. "*The ICS-FORTH RDFSuite: Managing Voluminous RDF Description Bases*". In Proceedings of the 2nd International Workshop on the Semantic Web (SemWeb'01), in conjunction with WWW10, pp. 1-13, Hong Kong. May 1, 2001.

[2] Tim Berners-Lee, James Hendler, Ora Lassila. "*The Semantic Web*". Scientific American. May 2001. <http://www.sciam.com/2001/0501issue/0501berners-lee.html>

[3] T. Berners-Lee, R. Fielding, L. Masinter. "*Uniform Resource Identifiers (URI): Generic Syntax*". RFC 2396. August 1998. <http://www.ietf.org/rfc/rfc2396.txt>

[4] BIBLINK <http://hosted.ukoln.ac.uk/biblink>

[5] Tim Bray, Dave Hollander, Andrew Layman. "*Namespaces in XML*". W3C Recommendation. January 14, 1999.

[6] D. Brickley, R.V. Guha. "*Resource Description Framework Schema (RDF/S) Specification 1.0*". W3C Candidate Recommendation. March 27, 2000.

[7] *B*asic *S*emantic *R*egistry < http://www.ubsr.org/ >

[8] Christoph Bussler. "*B2B Protocol Standards and their Role in Semantic B2B Integration Engines*". In Bulletin of the Technical Committee on Data Engineering. Vol. 24, No. 1, pp. 3-11. IEEE Computer Society. March 2001.

[9] Massimiliano Carrara, Nicola Guarino. "*Formal Ontology and Conceptual Analysis: A Structured Bibliography*". Version 2.5. March 22, 1999. <http://www.ladseb.pd.cnr.it/infor/ontology/Papers/Ontobiblio/TOC.html>

[10] *C*omposite *C*apability/*P*reference *P*rofiles <http://www.w3.org/TR/CCPP-struct-vocab/>

[11] CERES/NBII Thesaurus Partnership Project <http://ceres.ca.gov/thesaurus/RDF.html>

[12] *C*ommon *E*uropean *R*esearch *I*nformation *F*ormat <http://www.cordis.lu/cerif>

[13] CIDOC Reference Model <http://www.cidoc.icom.org/guide/guideint.htm>

[14] *C*ostal *Z*one *M*anagement Ontology <http://dlforum.external.forth.gr:8080/>

[15] DAML+ OIL (March 2001) <http://www.daml.org/2001/03/daml+oil-index>

[16] *R*eal *E*state *D*ata Consortium <http://www.dataconsortium.org/>

[17] DESIRE Metadata Registry <http://desire.ukoln.ac.uk/registry/>

[18] Martin Doerr, Nicola Guarino, Mariano Fernandez Lopez, Ellen Schulten, Milena Stefanova, Austin Tate. "*State of the Art in Content Standards*". OntoWeb. Deliverable 3.1. Version 1.0. November 2001.

[19] *D*igital *O*bject *I*nfrastructure <http://www.doi.org/index.html>

[20] DUBLIN CORE Metadata Initiative <http://dublincore.org/>

[21] Andreas Eberhart. "*Survey of RDF data on the Web*". Technical Report. International University in Germany. 2001 <http://www.i-u.de/schools/eberhart/rdf/rdf-survey.htm>

[22] *Eu*ropean *L*ibraries and *E*lectronic *R*esources in Mathematical Sciences <http://www.emis.de/projects/EULER/>

[23] *G*eography *M*arkup *L*anguage <http://www.opengis.net/gml/00-029/GML.html>

[24] GENE ONTOLOGY <http://www.geneontology.org/GO.doc.html>

[25] Guarino, N., Welty, C., and Partridge, C. "*Towards Ontology-based harmonization of Web content standards*". In S. Liddle, H. Mayr and B. Thalheim (eds.), Conceptual Modeling for E-Business and the Web: Proceedings of the ER-2000 Workshops. Springer Verlag, pp. 1-6. 2000.

[26] Dan Gunter, Keith Jackson. "*The Applicability of RDF-Schema as a Syntax for Describing Grid Resource Metadata*". Document: GWD-GIS-020-1. June 2001.

[27] Patrick Hayes. "*RDF Model Theory*", Working Draft, W3C. September 25, 2001

[28] *I*nternet *M*ovie *D*atabase <http://www.csee.umbc.edu/~skallu1/>

[29] IMS Global Learning Consortium <http://www.imsproject.org/rdf/index.html>

[30] Ora Lassila. "*Taking the RDF Model Theory Out For a Spin*". To appear in Proceedings of the 1st International Semantic Web Conference, Sardinia, 2002.

[31] O. Lassila, R. Swick. "*Resource Description Framework (RDF) Model and Syntax Specification*". W3C Recommendation. February 1999.
[32] *L*anguage *I*ndependent *M*etadata *B*rowsing of *E*uropean *R*esources <http://www.limber.rl.ac.uk/>
[33] Alexander Maedhe, Steffen Staab. "*Comparing Ontologies- Similarity Measures and a Comparison Study"*. Internal Report No. 408, Institute AIFB, University of Karlsruhe, Germany. March 2001.
[34] MATHEMATICS INTERNATIONAL <http://www.mathematik.uni-kl.de/~ontology/>
[35] METANET <http://archive.dstc.edu.au/RDU/staff/jane-hunter/harmony/jodi_article.html>
[36] Rajatish Mukherjee, Partha Sarathi Dutta, Sandip Sen. "*Analysis of domain specific ontologies for agent-oriented information retrieval*". In the Working Notes of the AAAI-2000 Workshop on Agent-Oriented Information Systems (AOIS). 2000.
[37] MUSIC BRAINZ <http://www.musicbrainz.org>
[38] J. Mylopoulos, A. Borgida, M. Jarke, M. Koubarakis. "*Telos - a language for representing knowledge about information systems*". ACM Transactions on Information Systems, 8(4):325-362. 1990.
[39] W. Nejdl, H. Dhraief, and M. Wolpers, "*O-telos-rdf: a Resource Description Format with Enhanced Meta-Modeling Functionalities Based on O-telos"*. In Workshop on Knowledge Markup and Semantic Annotation at the 1st Inrternational Conference on Knowledge Capture (K-CAP 2001), Victoria, BC., Canada. 2001.
[40] Natalya Fridman Noy, Deborah L. McGuinness. "*Ontology Development 101: A Guide to Creating Your First Ontology*". Stanford Knowledge Systems Laboratory Technical Report KSL-01-05. March 2001.
[41] Noy, N. F. and Hafner, C. D. "*The state of art in ontology design*". IA Magazine, 18 (3), pp. 53-74. Fall 1997.
[42] The *P*latform for *P*rivacy *P*references Project <http://www.w3.org/P3P/>
[43] RDF CALENDAR <http://ilrt.org/discovery/2001/04/calendar/>
[44] RDF SUITE REGISTRY <http://139.91.183.30:9090/RDF/Examples.html>
[45] RDF SUITE <http://139.91.183.30:9090/RDF/>
[46] RSLP *C*ollection *L*evel *D*escription <http://www.ukoln.ac.uk/metadata/rslp/>
[47] *R*DF *S*ite *S*ummary 1.0 <http://groups.yahoo.com/group/rss-dev/files/specification.html>
[48] SCHEMAS PROJECT <http://www.schemas-forum.org/>
[49] *S*cholarly *Link S*pecification <http://www.openly.com/slinks/>
[50] SWAGD: WebNS.net - The SWAG Dictionary <http://webns.net/>
[51] TOP LEVEL ONTOLOGY <http://www-sop.inria.fr/acacia/personnel/phmartin/RDF/phOntology.html>
[52] UNIVERSAL <http://www.ist-universal.org/>
[53] *Un*iversal *S*tandard *P*roducts and *S*ervices *C*lassification <http://eccma.org/unspsc/>
[54] Pepijn R.S Visser, Dean M. Jones, T.J.M Bench-Capon, M.J.R Shave. "*An Analysis of Ontology Mismatches; Heterogenity versus Interoperability*". AAAI 1997 Spring Symposium on Ontological Engineering, Stanford University, Canada.
[55] The Validating RDF Parser <http://139.91.183.30:9090/RDF/VRP/index.html>
[56] WAP: The WAP Forum Specifications <http://www.wapforum.org/what/review.htm>
[57] Peter C. Weinstein, William P. Birmingham. "*Comparing Concepts in Differentiated Ontologies*". In Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management (KAW' 99). Banff, Alberta, Canada. October 1999.
[58] Princeton WordNet <http://www.cogsci.princeton.edu/~wn>
[59] XML <http://www.w3.org/XML/>
[60] The XML Schema Data types <http://www.w3.org/1999/XMLSchema-datatypes>
[61] XMLNS.com <http://www.xmlns.com>