Evaluation Dataset

The objective of this research is performance evaluation of scalable semantic stores. A real dataset allows realistic and accurate quantification of semantic stores. Therefore, we are interested in using a real and public data set to evaluate the performance of semantic stores. Different RDF datasets are available online for testing purpose, such as, Barton libraries [1], DBpedia [3] and DBLP [2]. They are amongst the most commonly used RDF datasets for semantic stores evaluation.

The Barton Libraries dataset [1] is used for the performance evaluation. This data is provided by the Simile Project [4], which develops tools for library data management and interoperability. The data contains records that compose an RDF-formatted dump of the MIT Libraries Barton catalog. The data was derived from multiple sources and their diverse nature was preserved. The structure of the data is quite irregular so it provides a good demonstration of the relatively unstructured nature of semantic web data.

Barton Libraries Dataset

The dataset is analyzed using longwell [5], a faceted browser that combines the flexibility of RDF data model with user interface paradigm. Data is categorized on the basis of the types of resources present in the dataset. The index page of longwell is shown in Figure 1. The right panel shows Barton dataset types arranged in alphabetical order. The integer number with each type gives instance count of that particular type.

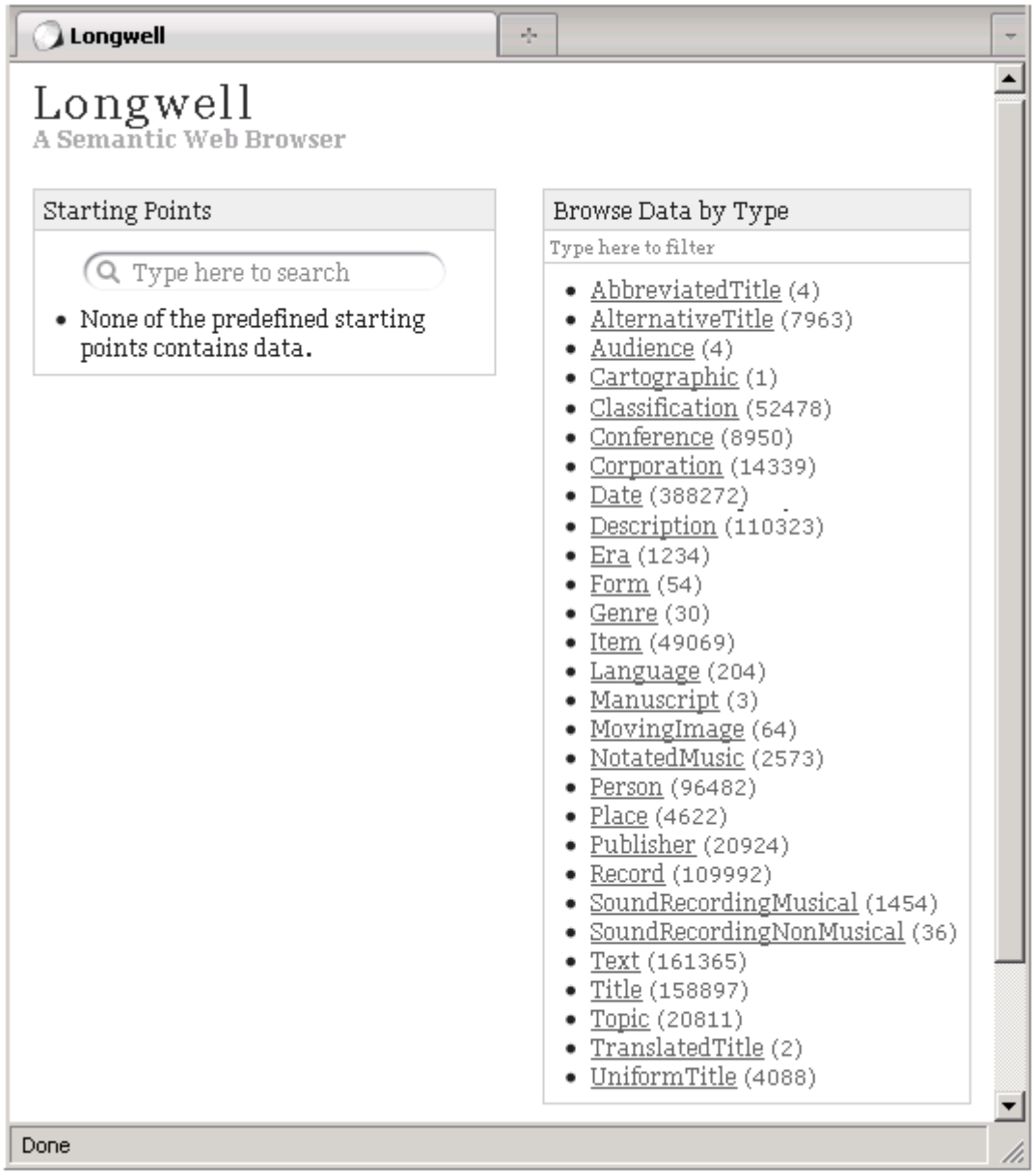

Figure 1: Longwell Screenshot

By clicking on any type, facets are shown to further drill down the data. The facets of each class represent the predicates that are associated with resources of class type. Figure 2 shows the facets for *Date* type resources.

Figure 2: Longwell screen shot of facets for *Date* type resources

An overview of the Barton data model, analyzed through longwell, is presented for the better understanding of dataset in Figure 3. In this diagram classes represent the types of resources in dataset, attributes of each class presents all those predicates whose domain is this

class. The relationship of a class with other classes shows the range of the predicates of that class.

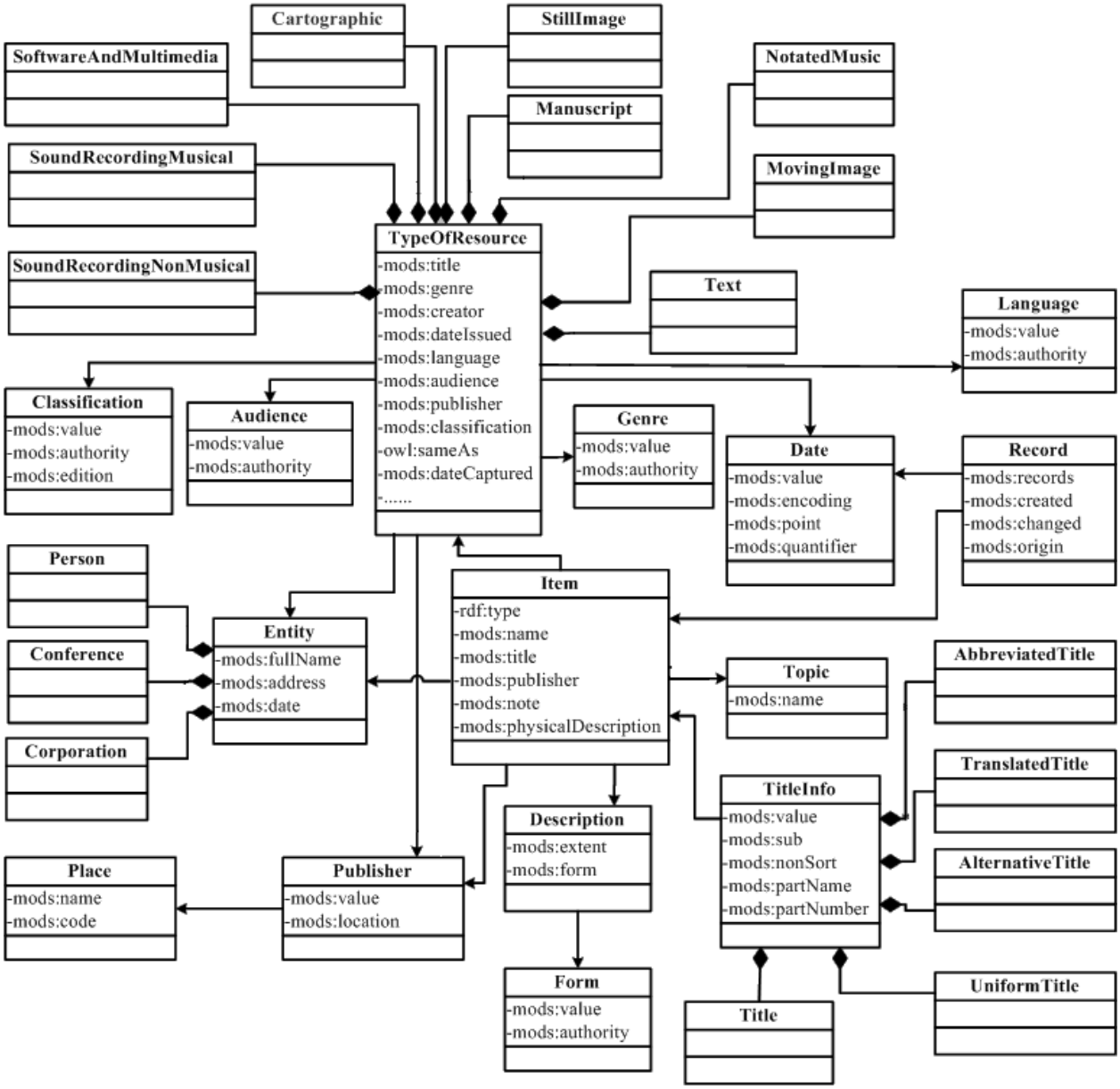

Figure 3: Overview of Barton Data Model

A brief description of classes is given in Table 1. A detailed description of attributes and their associated object type are given in [6].

Table 1: Barton Dataset Classes Description

CLASSES DESCRIPTION	
Class Name	**Definition**
Record	Record present information about the metadata record
Item:	Item is a resource that is being described
TitleInfo	TitleInfo is an abstract class represents all those words, phrases, characters, or group of characters, that constitutes the chief title, abbreviated title, translated title, alternative title and uniform title of a resource.
Topic	Topic class models all those subjects of resources that are not appropriate under title class
Date	Date class represents the information about date on which a record is created, changed, issued, and copyrighted or any other date that needs to be specified
Description	Description may be used to give a textual description for a resource when necessary
Form	Forms provides the information about the designation of physical presentation of the resource
Publisher	Publisher is the entity that published, printed, distributed, released, issued, or produced the resource
Place	Place describes the all those places that are associated with the issuing, publication, release, distribution, manufacture, production, or origin of a resource
Entity	Entity class represents all those persons, corporations and events (e.g. conference) who can be related to a resource in some way
Language	language class provides all those languages in which contents of the resources of dataset is expressed
Audience	Audience class provides a description of the intellectual level of the

	audience for which the resource is intended
TypeOfResource	Type of resource defines the term that specifies the characteristics and general type of content of the resource. Type of resource may be from one of text, cartographic, notated music, sound recording musical, sound recording nonmusical, still images, moving images, software and multimedia, and manuscripts
Classification	Classification class indicates all those categories in which resources can be organized according to subject area

Statistics of Barton Dataset

A detail analysis of Barton dataset provides us the characteristics of evaluated data that are presented in Table 2.

Table 2: Summary Statistics of Barton Dataset

Dataset Characteristics	
Total Number of Triples	25176626
Total Nodes	9716253
Total types of Instances	30
Total Unique Properties	199
Multi-valued Properties[1]	72
Single-valued Properties[2]	127

There are slightly more than 25 million triples in dataset, previously it was claimed that this dataset contains 50 million triples [7].

[1] Multi-valued properties means these properties appear more than once for a given subject
[2] Single-valued properties means these appear only once for a given subject

Namespaces

Similar string at the start of properties of dataset belongs to some predefined schemas, such as RDF, OWL or some other schemas. These are declared as namespaces in the document. URIs or namespaces for the Barton dataset are given in Table 3.

Table 3: Prefix to URI mapping

Prefix	URI
modsrdf:	http://simile.mit.edu/2006/01/ontologies/mods
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
role:	http://simile.mit.edu/2006/01/role/
owl:	http://www.w3.org/2002/07/owl#

Dataset Scaling and population

We will perform our evaluation on four different sizes of Barton dataset. Dataset 1, dataset 2, dataset 3 and dataset 4 contains 200K, 1 Million, 5 Million and 25 Million triples respectively. As we sampled our datasets from a large dataset, we tried to make our sampling fair and realistic. For fair sampling, we selected ten different samples of each dataset 1, dataset 2 and dataset 3. Average of ten samples for each dataset represents its population. Detail of each dataset is given in Table 4.

Table 4: Dataset Scaling and Population

Class Name	Scaling Factor			
	Dataset 1 (200K)	Dataset 2 (1Million)	Dataset 3 (5Million)	Dataset 4 (25Million)
Number of Date	15697	78382	391951	1959758
Number of Title	6453	32706	163502	817508
Number of Text	5499	30400	152230	760564
Number of Description	4421	21760	108802	544011
Number of Record	4430	20569	102813	514067
Number of Classification	3219	15538	77588	387942
Number of Person	2856	14198	90561	353635
Number of Item	2516	12511	62671	312270
Number of Alternative Title	780	3810	19098	95489
Number of Publisher	746	3852	23698	89589
Number of Corporation	485	2345	15838	58639
Number of Topic	459	2100	22804	52681
Number of Uniform Title	286	1162	5736	28679
Number of Conference	237	1125	5702	28139
Number of Place	189	853	5103	15542
Number of Notated Music	126	716	3008	15016
Number of Sound Recording Musical	91	449	2208	11022
Number of Abbreviated Title	87	498	2167	10835
Number of Cartographic	28	103	477	2353
Number of Manuscript	15	79	351	1753
Number of Moving Image	26	44	222	1109
Number of Language	35	58	178	399
Number of Genre	9	28	111	352
Number of Sound Recording Nonmusical	5	17	57	286

Number of Software and Multimedia	10	15	52	260
Number of Form	6	17	65	242
Number of Translated Title	5	7	16	82
Number of Audience	0	5	17	66
Number of Still Image	0	2	4	18

Dataset Cleaner

Publically available dump of Barton libraries dataset contains illegal URIs. Some semantic stores such as AllegroGraph and Mulgara do not allow loading datasets that contains illegal URIs. To load Barton dataset, illegal URIs are identified and transform into legal URIs before testing different stores using this dataset.

REFERENCES

[1] Barton Library Dataset: http://simile.mit.edu/wiki/Dataset:_Barton

[2] DBLP dataset: http://kdl.cs.umass.edu/data/dblp/dblp-info.html

[3] DBpedia dataset: http://wiki.dbpedia.org/Datasets

[4] SIMILE Project: http://simile.mit.edu/

[5] Longwell: http://simile.mit.edu/wiki/Longwell

[6] MODS User guide: http://www.loc.gov/standards/mods/v3/mods-userguide.html

[7] D. Abadi, A. Marcus, S. Madden, and K. Hollenbach. "Using the Barton libraries dataset as an RDF benchmark", Technical Report MIT-CSAIL-TR-2007-036, MIT.