Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2007-036 July 6, 2007

Using The Barton Libraries Dataset As An RDF benchmark

Daniel J. Abadi, Adam Marcus, Samuel R. Madden, and Kate Hollenbach

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Using The Barton Libraries Dataset As An RDF benchmark

Daniel J. Abadi
MIT
dna@csail.mit.edu

Adam Marcus
MIT
marcua@csail.mit.edu

Samuel R. Madden
MIT
madden@csail.mit.edu

Kate Hollenbach
MIT
kjhollen@mit.edu

ABSTRACT

This report describes the Barton Libraries RDF dataset and Longwell query benchmark that we use for our recent VLDB paper on Scalable Semantic Web Data Management Using Vertical Partitioning [4].

1. BARTON DATA

The dataset used for this benchmark is taken from the publicly available Barton Libraries dataset [1]. This data is provided by the Simile Project [3], 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, converted from raw data stored in an old library format standard called MARC (Machine Readable Catalog). Because of the multiple sources the data was derived from and the diverse nature of the data that is cataloged, the structure of the data is quite irregular.

At the time of publication of this report, there are slightly more than 50 million triples in the dataset, with a total of 221 unique properties, of which the vast majority appear infrequently. Of these properties, 82 (37%) are multi-valued, meaning that they appear more than once for a given subject; however, these properties appear more often (77% of the triples have a multi-valued property). The dataset provides a good demonstration of the relatively unstructured nature of Semantic Web data.

2. LONGWELL OVERVIEW

Longwell [2] is a tool developed by the Simile Project, which provides a graphical user interface for generic RDF data exploration in a web browser. It begins by presenting the user with a list of the values the *type* property can take (such as *Text* or *Notated Music* in the library dataset). The user can click on the types of data he desires to further explore. Longwell shows the list of currently filtered resources (RDF subjects) in the main portion of the screen, and a list of filters in panels along the side. Each panel represents a property that is defined on the current filter, with popular object values for that property and their frequency also presented in this box. If the user selects an object value, this filters the working set of resources to those that have that property-object value defined,

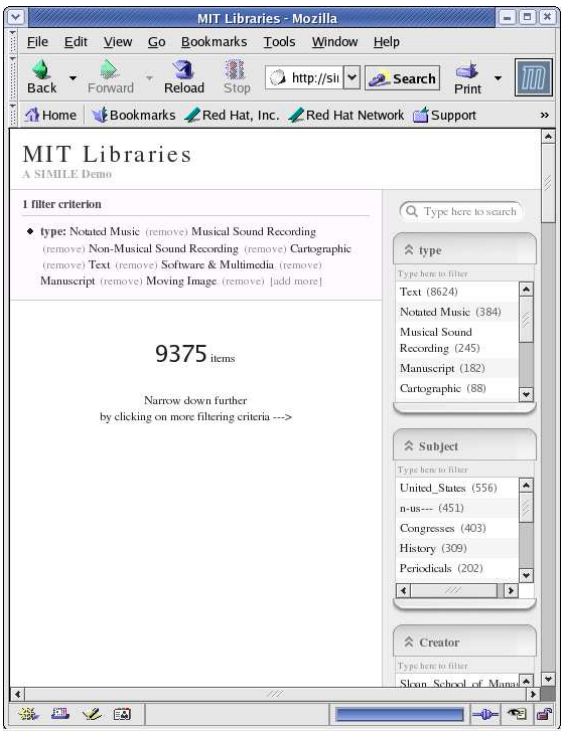

Figure 1: Longwell Opening Screen

updating the appropriate panels with the new frequency counts for this narrower set of resources.

We will now describe a sample browsing session through the Longwell interface, along with screenshots from this sample session. The opening screen with the list of different types is shown in Figure 1. Note the list of different types is shown at the top of the screen and their frequencies in the upper-right-hand side of the screen.

The path starts when the user selects *Text* from the *type* panel (at the upper-right-hand side of the screen), which filters the data into a list of text entities. This is shown in Figure 2.

On the right side of the screen, we find that popular properties on these entities. As the user scrolls down (see Figure 3), we can see that they include "subject," "creator," "genre," and "publisher." Within each property there is a list of the counts of the popular objects within this property. For example, as the user scrolls

VLDB '07, September 23-28, 2007, Vienna, Austria.

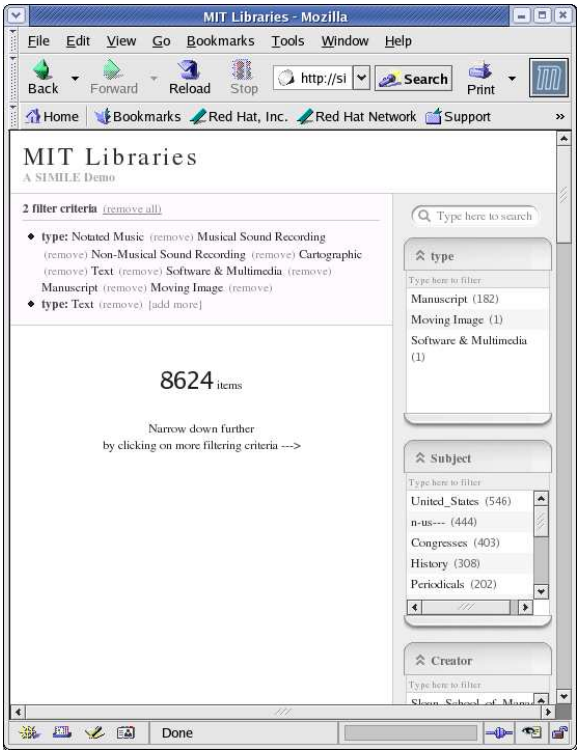

Figure 2: Longwell Screen Shot After Clicking on "Text" in the Type Property Panel

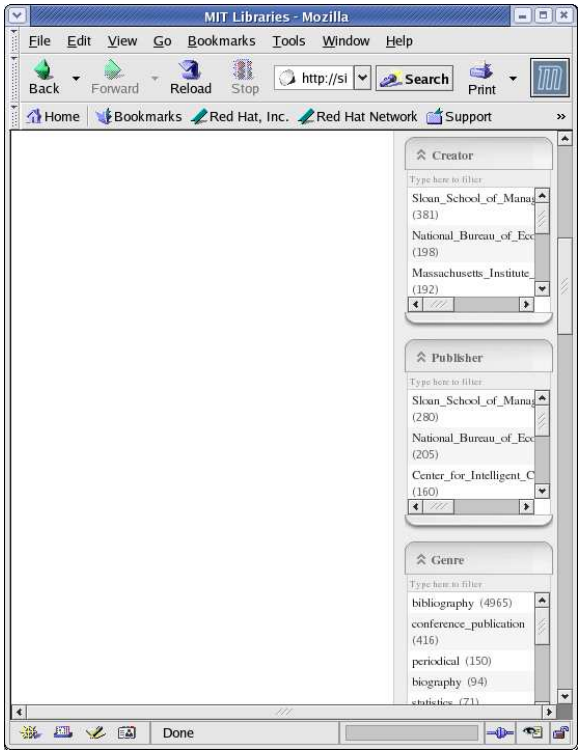

Figure 3: Longwell Screen Shot After Clicking on "Text" in the Type Property Panel and Scrolling Down

down even more (see Figure 4), we find out that the German object value appears 122 times and the French object value appears 131 times under the language property. By clicking on "fre" (French language), information about the 131 French texts in the database is presented, along with the revised set of popular properties and property values defined on these French texts. This is shown in Figure 5.

Currently, Longwell only runs on a small fraction of the Barton data – 9375 records, as its RDF triple store cannot scale to support the full 50 million triple dataset while still allowing interactive response time to queries (we show this scalability limitation in our paper [4]).

3. LONGWELL QUERIES

The benchmark features seven queries that need to be executed on a typical Longwell path through the data. These queries are based on a typical browsing session, where the user selects a few specific entities to focus on and where the aggregate results summarizing the contents of the RDF store are updated.

The full queries are described at a high level here and are provided in full in the next section as SQL queries against a triple store.

Query 1 (Q1). Calculate the opening panel displaying the counts of the different types of data in the RDF store. This requires a search for the objects and counts of those objects with property *Type*. There are 30 such objects. For example: *Type*: *Text* has a count of 1, 542, 280, and *Type*: *NotatedMusic* has a count of 36, 441.

Query 2 (Q2). The user selects *Type*: *Text* from the previous panel. Longwell must present him with a list of other defined properties for resources of *Type*: *Text*. It must also calculate the frequency of these properties. For example, the *Language* property is defined 1, 028, 826 times for resources that are of *Type*: *Text*.

Query 3 (Q3). For each property defined on items of *Type*: *Text*, populate the property panel with the counts of popular object values for that property (where popular means that an object value appears more than once). For example, the property *Edition* has 8 items with value "[1st_ed._reprinted]."

Query 4 (Q4). This query recalculates all of the property-object counts from Q3 if the user clicks on the "French" value in the "Language" property panel. Essentially this is narrowing the working set of subjects to those whose *Type* is *Text* **and** *Language* is *French*. This query has a much higher-selectivity than Q3.

Query 5 (Q5). Here we perform a type of *inference*. If there are triples of the form (*X Records Y*) and (*Y Type Z*) then we can *infer* that *X* is of type *Z*. Here *X Records Y* means that *X* records information about *Y* (for example, *X* might be a web page with information on *Y*). For this query, we want to find the inferred type of all subjects that have this *Records* property defined that also originated in the US Library of Congress (i.e. contain triples of the form (*X origin "DLC"*)). The subject and inferred type is returned for all non-*Text* entities.

Query 6 (Q6). For this query, we combine the inference first step of Q5 with the property frequency calculation of Q2 to extract information in aggregate about items that are either directly known to be of *Type*: *Text* (as in Q2) or inferred to be of *Type*: *Text* through the Q5 *Records* inference.

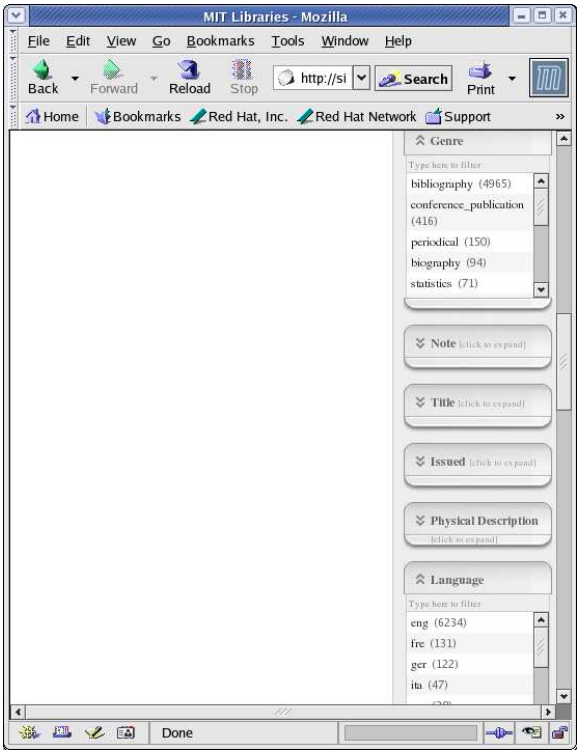

Figure 4: Longwell Screen Shot After Clicking on "Text" in the Type Property Panel and Scrolling Down to the Language Property Panel

Figure 5: Longwell Screen Shot After Clicking on "fre" in the Language Property Panel

Query 7 (Q7). Finally, we include a simple triple selection query with no aggregation or inference. The user tries to learn what a particular property (in this case *Point*) actually means by selecting other properties that are defined along with a particular value of this property. The user wishes to retrieve subject, *Encoding*, and *Type* of all resources with a *Point* value of "end." The result set indicates that all such resources are of the type *Date*. This explains why these resources can have "start" and "end" values: each of these resources represents a start or end date, depending on the value of *Point*.

We make the assumption that an administrator has selected a set of 28 interesting properties over which queries will be run. These properties are listed in Section 5. There are 26761389 triples for these properties. For queries Q2, Q3, Q4, and Q6, only these 28 properties are considered for aggregation.

4. LONGWELL QUERIES IN SQL

The queries below are the seven benchmark queries written in SQL against a triple store schema.

The properties table listed in these queries contains the list of 28 properties that are processed for queries 2, 3, 4, and 6.

Query1:

```
SELECT A.obj, count(*)
FROM triples AS A
WHERE A.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
GROUP BY A.obj
```

Query2:

```
SELECT B.prop, count(*)
FROM triples AS A, triples AS B,
  properties AS P
WHERE A.subj = B.subj
  AND A.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
  AND A.obj = "<http://simile.mit.edu/2006/01/ontologies/mods3#Text>"
  AND P.prop = B.prop
GROUP BY B.prop
```

Query3:

```
SELECT B.prop, B.obj, count(*)
FROM triples AS A, triples AS B,
  properties AS P
WHERE A.subj = B.subj
  AND A.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
  AND A.obj = "<http://simile.mit.edu/2006/01/ontologies/mods3#Text>"
  AND P.prop = B.prop
GROUP BY B.prop, B.obj
HAVING count(*) > 1
```

Query4:

```
SELECT B.prop, B.obj, count(*)
FROM triples AS A, triples AS B,
   triples AS C, properties AS P
WHERE A.subj = B.subj
  AND A.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
  AND A.obj = "<http://simile.mit.edu/2006/01/ontologies/mods3#Text>"
  AND P.prop = B.prop
  AND C.subj = B.subj
  AND C.prop = "<http://simile.mit.edu/2006/01/ontologies/mods3#language>"
  AND C.obj =
   "<http://simile.mit.edu/2006/01/language/iso639-2b/fre>"
GROUP BY B.prop, B.obj
HAVING count(*) > 1
```

Query5:

```
SELECT B.subj, C.obj
```

```
FROM triples AS A, triples AS B,
     triples AS C
WHERE A.subj = B.subj
  AND A.prop = "<http://simile.mit.edu/2006/01/ontologies/mods3#origin>"
  AND A.obj = "<info:marcorg/DLC>"
  AND B.prop = "<http://simile.mit.edu/2006/01/ontologies/mods3#records>"
  AND B.obj = C.subj
  AND C.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
  AND C.obj != "<http://simile.mit.edu/2006/01/ontologies/mods3#Text>"
```

Query6:

```
SELECT A.prop, count(*)
FROM triples AS A, properties AS P (
 (SELECT B.subj
  FROM triples AS B
  WHERE B.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
  AND B.obj = "<http://simile.mit.edu/2006/01/ontologies/mods3#Text>")
UNION
 (SELECT C.subj
  FROM triples AS C, triples AS D
  WHERE C.prop = "<http://simile.mit.edu/2006/01/ontologies/mods3#records>"
  AND C.obj = D.subject
  AND D.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>"
  AND D.obj = "<http://simile.mit.edu/2006/01/ontologies/mods3#Text>")
) AS uniontable
WHERE A.subj = uniontable.subj
  AND P.prop = A.prop
GROUP BY A.prop;
```

Query7:

```
SELECT A.subj, B.obj,
  C.obj
FROM triples AS A, triples AS B,
     triples AS C
WHERE A.prop = "<http://simile.mit.edu/2006/01/ontologies/mods3#Point>"
  AND A.obj = '"end"'
  AND A.subj = B.subject
  AND B.prop = "<http://simile.mit.edu/2006/01/ontologies/mods3#Encoding>"
  AND A.subj = C.subject
  AND C.prop = "<http://www.w3.org/1999/02/22-rdf-syntax-ns#Type>";
```

5. PROPERTIES

The 28 properties contained in the properties table are:

```
<http://simile.mit.edu/2006/01/ontologies/mods3#access>
<http://simile.mit.edu/2006/01/ontologies/mods3#address>
<http://simile.mit.edu/2006/01/ontologies/mods3#affiliation>
<http://simile.mit.edu/2006/01/ontologies/mods3#authority>
<http://simile.mit.edu/2006/01/ontologies/mods3#catalogingLanguage>
<http://simile.mit.edu/2006/01/ontologies/mods3#code>
<http://simile.mit.edu/2006/01/ontologies/mods3#contents>
<http://simile.mit.edu/2006/01/ontologies/mods3#copyrightDate>
<http://simile.mit.edu/2006/01/ontologies/mods3#dateCreated>
<http://simile.mit.edu/2006/01/ontologies/mods3#dates>
<http://simile.mit.edu/2006/01/ontologies/mods3#edition>
<http://simile.mit.edu/2006/01/ontologies/mods3#encoding>
<http://simile.mit.edu/2006/01/ontologies/mods3#extent>
<http://simile.mit.edu/2006/01/ontologies/mods3#fullName>
<http://simile.mit.edu/2006/01/ontologies/mods3#issuance>
<http://simile.mit.edu/2006/01/ontologies/mods3#language>
<http://simile.mit.edu/2006/01/ontologies/mods3#nonSort>
<http://simile.mit.edu/2006/01/ontologies/mods3#origin>
<http://simile.mit.edu/2006/01/ontologies/mods3#partName>
<http://simile.mit.edu/2006/01/ontologies/mods3#partNumber>
<http://simile.mit.edu/2006/01/ontologies/mods3#point>
<http://simile.mit.edu/2006/01/ontologies/mods3#qualifier>
<http://simile.mit.edu/2006/01/ontologies/mods3#records>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://simile.mit.edu/2006/01/ontologies/mods3#sub>
<http://simile.mit.edu/2006/01/ontologies/mods3#changed>
<http://simile.mit.edu/2006/01/ontologies/mods3#created>
<http://simile.mit.edu/2006/01/ontologies/mods3#physicalDescription>
```

6. REFERENCES

[1] Library catalog data. http://simile.mit.edu/rdf-test-data/barton/.

[2] Longwell website. http://simile.mit.edu/longwell/.

[3] Simile website. http://simile.mit.edu/.

[4] D. J. Abadi, A. Marcus, S. Madden, and K. Hollenbach. Scalable semantic web data management using vertical partitioning. In *Proc. of VLDB*, 2007.