HDT-it: Storing, Sharing and Visualizing Huge RDF Datasets.

Mario Arias Gallego[1], Javier D. Fernández[1], Miguel A. Martínez-Prieto[1,2], and Claudio Gutierrez[2]

[1] Computer Science Department, University of Valladolid (Spain)
mario.arias@gmail.com, {jfergar,migumar2}@infor.uva.es
[2] Department of Computer Science, University of Chile (Chile)
cgutierr@dcc.uchile.cl

Abstract. Huge RDF Datasets are currently being published in the Linked-Open-Data cloud. Appropriate data structures are required to address scalability and performance issues when storing, sharing and querying these datasets. HDT (Header, Dictionary, Triples) is a binary format that represents RDF data in a compressed manner, therefore saving space whilst providing fast search operations directly on the compressed representation. These facts make it an excellent format when storing or sharing millions of triples. HDT-it is a tool that can generate and consume RDF files in HDT format. It demonstrates the capabilities of HDT by allowing to search basic triple patterns against HDT files, and also visualizes the 3D adjacency matrix of the underlying RDF graph to provide an overview of the dataset distribution.

Keywords: rdf, hdt, linked-open-data, rdf visualization

1 Introduction

The success of RDF[3] has favored the availability of many public big-scale datasets in this format. Some studies about the Linked-Open-Data Cloud[4] estimate that 25 billion triples are currently being published and interconnected. Some examples of big-scale datasets[5] include DBPedia, Freebase, Bio2RDF or Geonames, which contain more than 100 million triples each.

The users typically query these datasets using SPARQL Endpoints[6], but sometimes it is more convenient to download the whole dataset to use it locally without a network connection, or when they plan to perform costly operations without overloading the server. For example biologists might need their own copy of Bio2RDF to data mine millions of proteins related to a disease.

[3] http://www.w3.org/RDF/

[4] http://www4.wiwiss.fu-berlin.de/lodcloud/

[5] http://linkeddata.org/data-sets

[6] http://www.w3.org/wiki/SparqlEndpoints

In that scenario, the user is required to download a dump-file, typically in NTRIPLES[7] or RDF/XML[8] format. Since these syntaxes are very verbose, they are compressed using generic techniques such as GZIP or BZIP2 to save bandwidth. Once the dump is downloaded, it must be preprocessed in order to perform any kind of browsing or query: The file needs to be uncompressed, parsed, loaded into a RDF-Store, and complemented with the appropriate indices.

It is easy to reckon that this whole process is arduous and pointless. The HDT[3] format aims at simplifying it by generating a compressed and searchable version of the dataset that can be directly consumed. Due to its advantages, HDT has been accepted as a member submission by the W3C[9], as a seed to standardize the format and make it compatible throughout the Web.

The HDT-it tool demonstrates the advantages of HDT to store, share, and consume big RDF datasets. It can generate HDT files from traditional RDF serializations and back, and also allows to browse and perform basic searches against the compressed representation. In addition, it shows the adjacency matrix of the RDF graph to help understand the overal structure of the dataset.

2 Proposal

HDT achieves very compact RDF representations by proposing a natural division of three components to facilitate RDF management in real applications:

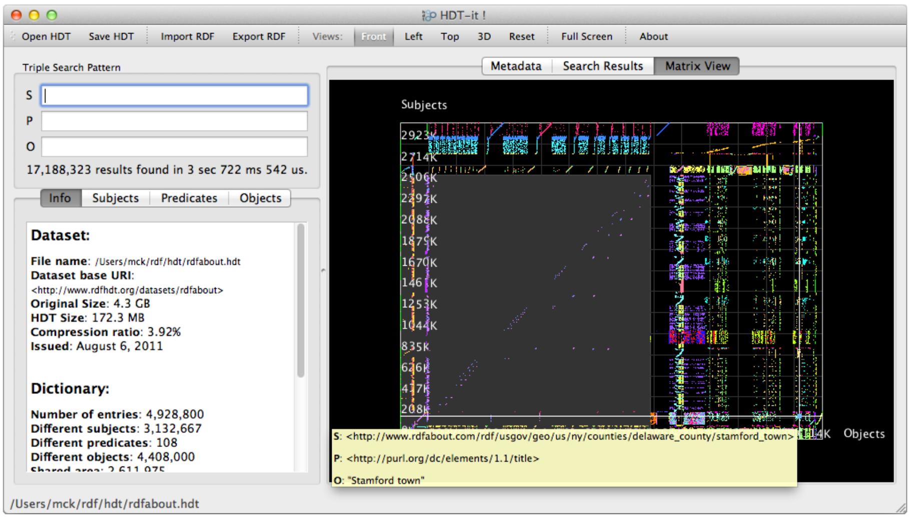
1. **Header**: Expresses metadata about the dataset in plain RDF format. For example who was the creator, when was it issued, the namespaces used, a textual description, the number of triples, statistics and other internal details about the dataset organization.
2. **Dictionary**: It provides a mapping between the dataset strings (URIs, blanks and literals) and numeric identifiers. Thereafter each triple can be represented as a tuple composed of 3 identifiers, instead of three strings. By applying dictionary compression algorithms [2], the compression ratio can be reduced while maintaining *extract* and *locate* operations. Some techniques can even perform prefix, substring or approximate searches.
3. **Triples**: Once that the triples have been converted to identifiers by means of the dictionary, the structure of the RDF Graph is saved in the Triples section. A very simple approach is using sorted adjacency lists to reduce component repetitions. If adjacency list boundaries are represented with compact bitmap structures [3], the exact location of a specific item can be located in constant order, consequently providing very fast searches.

Thanks to HDT, the previous dataset downloading scenario is vastly simplified, and the waiting time reduced. The user downloads the data already in HDT format, and she is ready to use HDT-it to start browsing and performing queries against it without the need of pre-processing.

[7] http://www.w3.org/2001/sw/RDFCore/ntriples/

[8] http://www.w3.org/TR/REC-rdf-syntax/

[9] http://www.w3.org/Submission/2011/03/



17,188,323 results found in 3 sec 722 ms 542 us.

Dataset:

File name: /Users/mck/rdf/hdt/rdfabout.hdt
Dataset base URI:
<http://www.rdfhdt.org/datasets/rdfabout>
Original Size: 4.3 GB
HDT Size: 172.3 MB
Compression ratio: 3.92%
Issued: August 6, 2011

Dictionary:

Number of entries: 4,928,800
Different subjects: 3,132,667
Different predicates: 108
Different objects: 4,408,000
Shared area: 2,611,975

S: <http://www.rdfabout.com/rdf/usgov/geo/us/ny/counties/delaware_county/stamford_town>
P: <http://purl.org/dc/elements/1.1/title>
O: "Stamford town"

/Users/mck/rdf/hdt/rdfabout.hdt

Fig. 1. HDT-it! browsing RDFAbout in matrix mode.

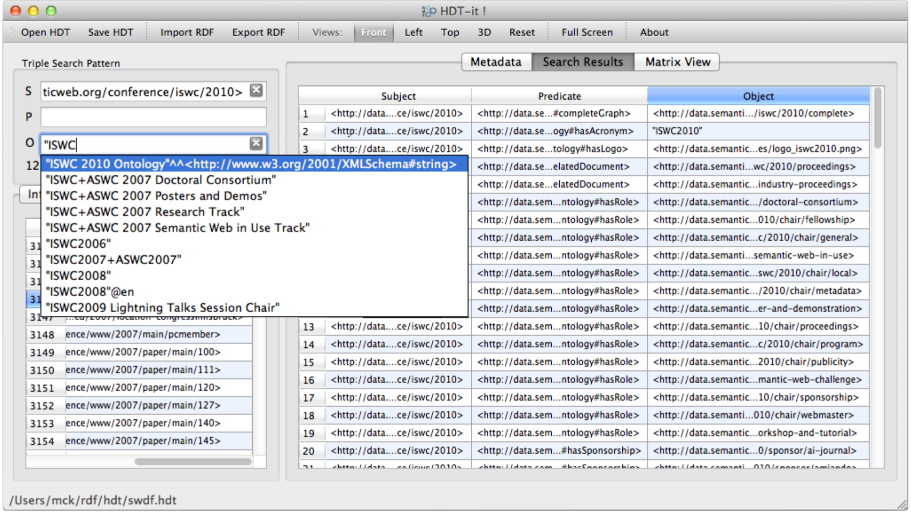
Once the data is loaded on HDT-it, the Info part on the left side (Figure 1) provides all the information about the dataset available in the Header. For example the original and HDT sizes are shown, demonstrating the size reduction. The metadata can be browsed in plain RDF format in the *Metadata* tab.

The user can search for any triple pattern by filling the text boxes on the top left part. As the user types, she will be proposed suggestions thanks to the dictionary prefix search capabilities. The resulting triples can be browsed both as a plain list in the *Search Results* tab (Figure 2), or visually in the *Matrix View*. They can also be exported in RDF or HDT format, making the generation of subsets of a dataset straightforward.

On the bottom left part, the user can browse a list of every different subject, predicate and object. When the user double-clicks on any of them, they are automatically added to the triple search pattern box, launching a new query that reveals the details about that specific item.

RDF Visualization. HDT-it also includes a new visualization technique based on the 3D adjacency matrix of the triples [1]. It basically interprets each triple of identifiers as a 3D coordinate in space, and renders them in a scatter plot. This representation is valuable to have an overall view of the distribution and complexity of RDF data within a data set.

HDT-it can render the matrix directly from an HDT file (right part of figure 1), and allows to rotate, zoom and pan it interactively. Each different predicate is assigned a different color so they are easily distinguishable. When the user hovers the mouse over any point, it is highlighted and a tooltip message reveals the triple that is under the cursor. This helps understand what kind of



S ticweb.org/conference/iswc/2010>

P

O "ISWC

- "ISWC 2010 Ontology"^^<http://www.w3.org/2001/XMLSchema#string>
- "ISWC+ASWC 2007 Doctoral Consortium"
- "ISWC+ASWC 2007 Posters and Demos"
- "ISWC+ASWC 2007 Research Track"
- "ISWC+ASWC 2007 Semantic Web in Use Track"
- "ISWC2006"
- "ISWC2007+ASWC2007"
- "ISWC2008"
- "ISWC2008"@en
- "ISWC2009 Lightning Talks Session Chair"

3148	ence/www/2007/main/pcmember>
3149	ence/www/2007/paper/main/100>
3150	ence/www/2007/paper/main/111>
3151	ence/www/2007/paper/main/120>
3152	ence/www/2007/paper/main/127>
3153	ence/www/2007/paper/main/140>
3154	ence/www/2007/paper/main/145>

	Subject	Predicate	Object
1	<http://data....ce/iswc/2010>	<http://data.se...#completeGraph>	<http://data.semanti.../iswc/2010/complete>
2	<http://data....ce/iswc/2010>	<http://data.se...ogy#hasAcronym>	"ISWC2010"
3	<http://data....ce/iswc/2010>	<http://data.se...tology#hasLogo>	<http://data.semantic...es/logo_iswc2010.png>
		<http://data.se...elatedDocument>	<http://data.semanti...wc/2010/proceedings>
		<http://data.se...elatedDocument>	<http://data.semantic...industry-proceedings>
		<http://data.sem...ntology#hasRole>	<http://data.semantic.../doctoral-consortium>
		<http://data.sem...ntology#hasRole>	<http://data.semantic...010/chair/fellowship>
		<http://data.sem...ntology#hasRole>	<http://data.semantic...c/2010/chair/general>
		<http://data.sem...ntology#hasRole>	<http://data.semanti...semantic-web-in-use>
		<http://data.sem...ntology#hasRole>	<http://data.semantic...swc/2010/chair/local>
		<http://data.sem...ntology#hasRole>	<http://data.semantic.../2010/chair/metadata>
		<http://data.sem...ntology#hasRole>	<http://data.semantic...er-and-demonstration>
13	<http://data....ce/iswc/2010>	<http://data.sem...ntology#hasRole>	<http://data.semantic...10/chair/proceedings>
14	<http://data....ce/iswc/2010>	<http://data.sem...ntology#hasRole>	<http://data.semantic...c/2010/chair/program>
15	<http://data....ce/iswc/2010>	<http://data.sem...ntology#hasRole>	<http://data.semantic...2010/chair/publicity>
16	<http://data....ce/iswc/2010>	<http://data.sem...ntology#hasRole>	<http://data.semantic...mantic-web-challenge>
17	<http://data....ce/iswc/2010>	<http://data.sem...ntology#hasRole>	<http://data.semantic...10/chair/sponsorship>
18	<http://data....ce/iswc/2010>	<http://data.sem...ntology#hasRole>	<http://data.semanti...010/chair/webmaster>
19	<http://data....ce/iswc/2010>	<http://data.sem...ntology#hasRole>	<http://data.semantic...orkshop-and-tutorial>
20	<http://data....ce/iswc/2010>	<http://data.sem...#hasSponsorship>	<http://data.semantic...0/sponsor/ai-journal>

/Users/mck/rdf/hdt/swdf.hdt

Fig. 2. Semantic Web Dog Food dataset in HDT, showing subjects and results.

data exists in each area of the dataset. If the user clicks, the current predicate is selected and the rest grayed out, so the user can browse them separately.

A full screencast of HDT-it can be watched online at RDF-HDT webpage [10].

3 Conclusions and Future work

HDT is a convenient format to store and share RDF datasets due to its small output size and data management capabilities. HDT-it shows its advantages by allowing to load HDT files and browse, search and visualize them interactively.

The future work of HDT-it is focused on providing additional search features, namely executing SPARQL queries directly on the compressed representation.

References

1. M. Arias, J.D. Fernández, and M.A. Martínez-Prieto. RDF Visualization using a Three-Dimensional Adjacency Matrix. In *Proc. of SemSearch*, 2011. Available at `http://km.aifb.kit.edu/ws/semsearch11/8.pdf`.
2. N. Brisaboa, R. Cánovas, F. Claude, M. A. Martínez-Prieto, and G. Navarro. Compressed String Dictionaries. In *Proc. of SEA*, pages 136–147, 2011.
3. J.D. Fernández, M.A. Martínez-Prieto, and C. Gutierrez. Compact Representation of Large RDF Data Sets for Publishing and Exchange. In *Proc. of ISWC*, pages 193–208, 2010.

[10] `http://www.rdfhdt.org/screencast`