Binary RDF Representation for Publication and Exchange (HDT)

Javier D. Fernández, Miguel A. Martínez-Prieto, Claudio Gutierrez
Axel Polleres
Interchange of large RDF data sets…

- Existing serialization syntaxes oriented mainly to represent documents
  - RDF/XML, N3, Turtle, JSON, etc.
  - Document-centric → data-centric view
- Redundancy
- No structure (chunks)
  - Lack of metadata
  - Sequentiality of the information
- Need for more efficient exchange formats!
- Examples:
  - Billion Triple 2010 (~3200M triples, 318 gzipped chunks, ~27GB)
  - Uniprot (~845M, 12 gzipped chunks, ~23GB)
Goals for a more suitable interchange format for large RDF datasets:

- Clean publication
  - Metadata
  - Compactness (binary encoding)
- Efficient exchange
  - RDF compression
- Basic data operations
- compare: W3C EXI WG for XML!

Our approach:
HDT in short:

- **RDF**: 
  - Logical and physical metadata describing the RDF data set. It serves as an entrance point to the information.

- **Header**
  - Mapping between elements in the data set and unique IDs, thus contributing to compactness.

- **Dictionary**
  - Structure of the data after the ID replacement, in a compressed form.
Main ideas:

- **Header**
  - Allows to store metadata (statistics, point to dataset description, e.g. VoID)
  - Point to chunks of data

- **Dictionary**
  - Map prefixes/URIs to numeric identifiers
  - Represented compactly by grouping
    - (1) Common **subject**-**objects**
    - (2) The non common **subjects**
    - (3) The non common **objects**
    - (4) **Predicates**

- **Triples**
  - Compactly represented in a tailored binary format
  - Specific compression:
    - More efficient compression than just bzip on N3!
    - Allows direct patterns \((s,p,o)\), \((s,?p,?o)\) and \((s,p,?o)\). lookups on the compressed file!
## HDT Bitmap Triples Compression Results

<table>
<thead>
<tr>
<th>Data set</th>
<th>Triples (millions)</th>
<th>Size (MB)</th>
<th>HDT Plain Compress</th>
<th>Universal Compressors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Billion Triples</td>
<td>106.9</td>
<td>15081.74</td>
<td>31.87%</td>
<td>9.54% 6.83% 5.32%</td>
</tr>
<tr>
<td>Uniprot RDF</td>
<td>79.2</td>
<td>7083.22</td>
<td>14.33%</td>
<td>8.71% 5.04% 3.99%</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>47</td>
<td>6882.20</td>
<td>6.62%</td>
<td>6.97% 5.11% 4.10%</td>
</tr>
</tbody>
</table>
Now acknowledged as W3C member submission:

- [http://www.w3.org/Submission/2011/03/](http://www.w3.org/Submission/2011/03/)

---

**Binary RDF Representation for Publication and Exchange (HDT)**

W3C Member Submission 30 March 2011

This version: [http://www.w3.org/Submission/2011/SUBM-HDT-20110330/](http://www.w3.org/Submission/2011/SUBM-HDT-20110330/)

Latest version: [http://www.w3.org/Submission/HDT/](http://www.w3.org/Submission/HDT/)

Editor: Javier D. Fernández

**Authors:**

- Javier D. Fernández
- Miguel A. Martínez-Prieto
- Claudio Gutiérrez
- Axel Polleres

Copyright © 2011 DERI Galway at the National University of Ireland, Galway, Ireland, Free University of Bozen-Bolzano, The Open University, Universidad Politécnica de Madrid, Alcatel-Lucent, Cisco, OpenLink Software and Profium Ltd. All rights reserved.

This document is available under the W3C Document Licence. See the W3C Intellectual Rights Notice and Legal Disclaimers.

---

**Supported by:**

- Alcatel-Lucent
- Cisco
- OpenLink Software
- Profium
- The Open University
- DERI