Scalable RDF Management in the Web of Data

Toward Efficient Interchange of RDF Data Streams

Javier D. Fernández

Ontology Engineering Group
Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid

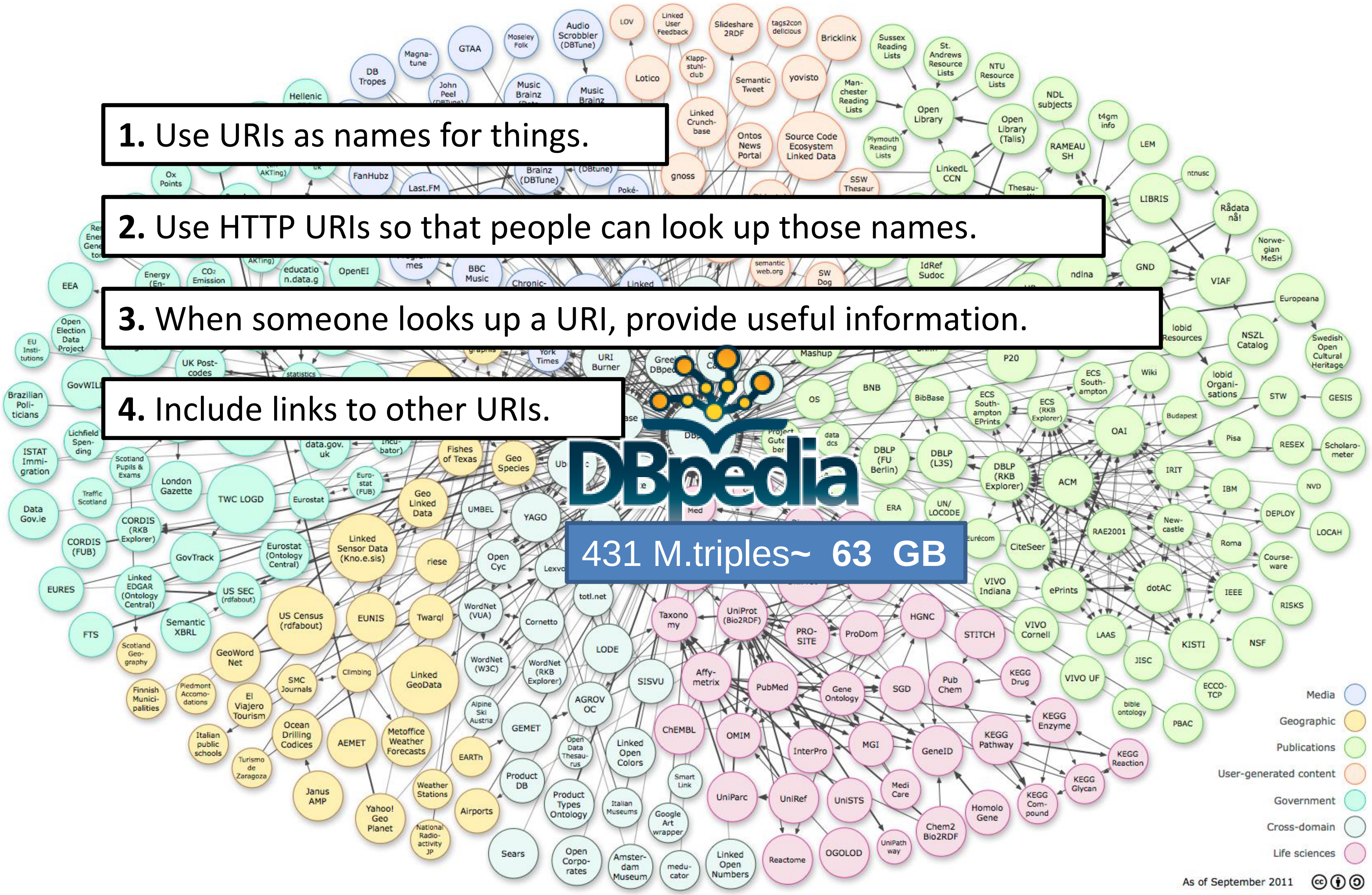
Date: 24/04/2014

Outline

Index

1. Background
2. Motivation
3. What have I done so far?
4. What am I currently doing?
5. Next steps

What is Linked Data?



1. Use URIs as names for things.

2. Use HTTP URIs so that people can look up those names.

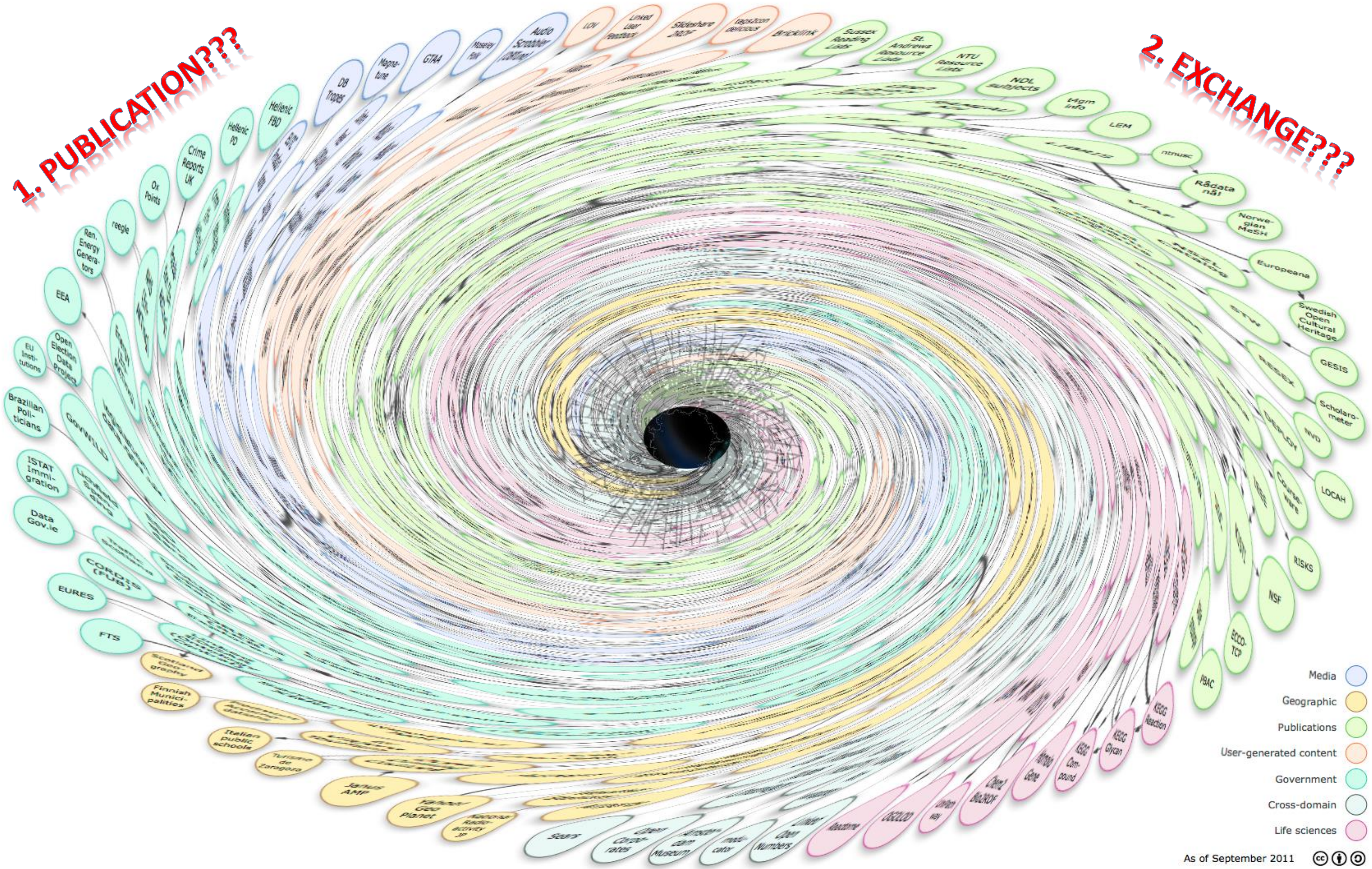
3. When someone looks up a URI, provide useful information.

4. Include links to other URIs.

431 M.triples~ **63 GB**

Linked Data cloud: > 62 billion triples

What is Linked Data?

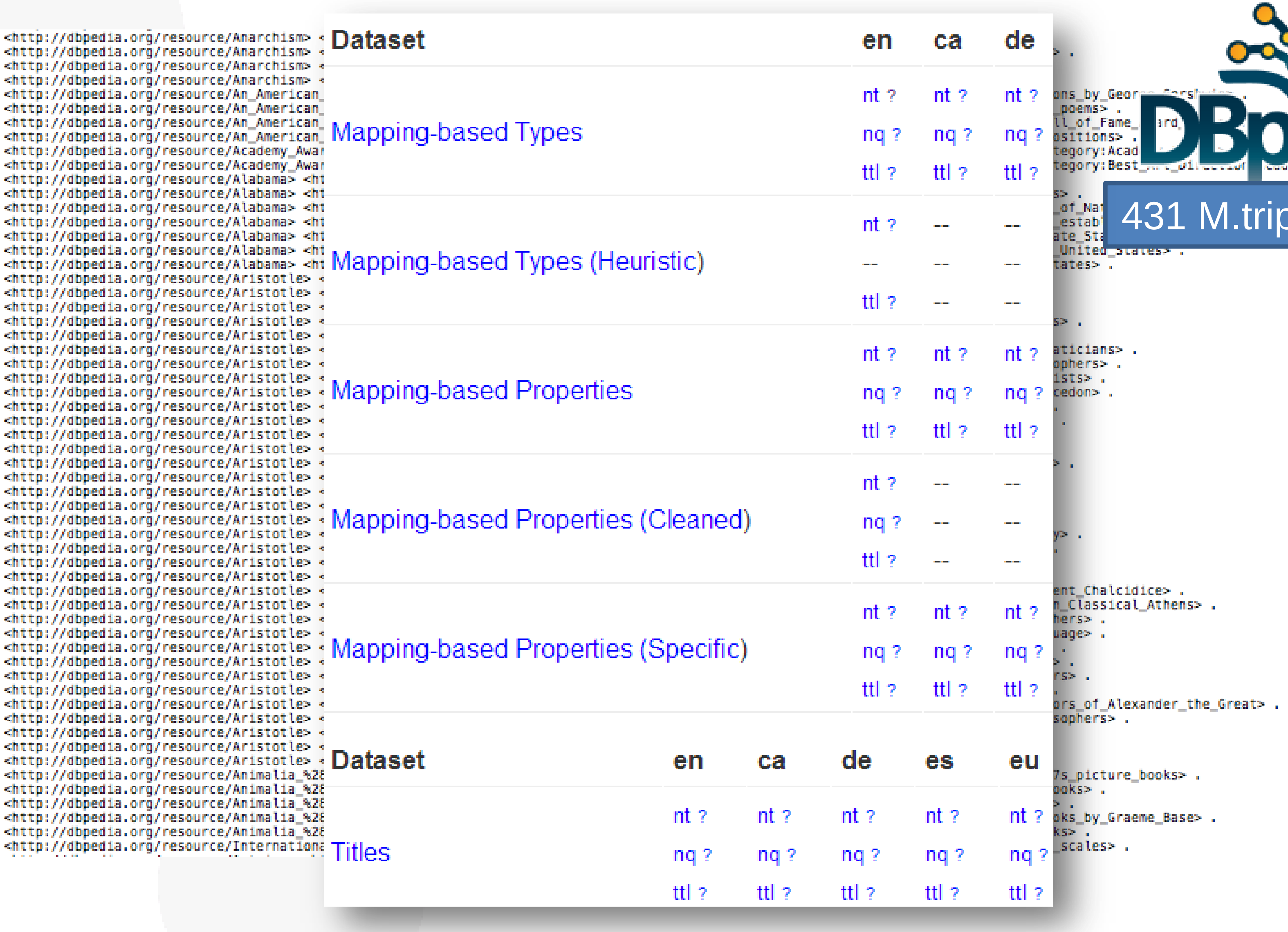


1. PUBLICATION???

2. EXCHANGE???

3. CONSUMPTION???

Publication & Exchange: Is this the best option?



```
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/Academy_Awar
<http://dbpedia.org/resource/Academy_Awar
<http://dbpedia.org/resource/Alabama> <ht
<http://dbpedia.org/resource/Alabama> <ht
<http://dbpedia.org/resource/Alabama> <ht
<http://dbpedia.org/resource/Alabama> <ht
<http://dbpedia.org/resource/Alabama> <ht
<http://dbpedia.org/resource/Alabama> <ht
<http://dbpedia.org/resource/Alabama> <ht
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Internationa
```

Dataset	en	ca	de
Mapping-based Types	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?
Mapping-based Types (Heuristic)	nt ? -- ttl ?	-- -- --	-- -- --
Mapping-based Properties	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?
Mapping-based Properties (Cleaned)	nt ? nq ? ttl ?	-- -- --	-- -- --
Mapping-based Properties (Specific)	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?

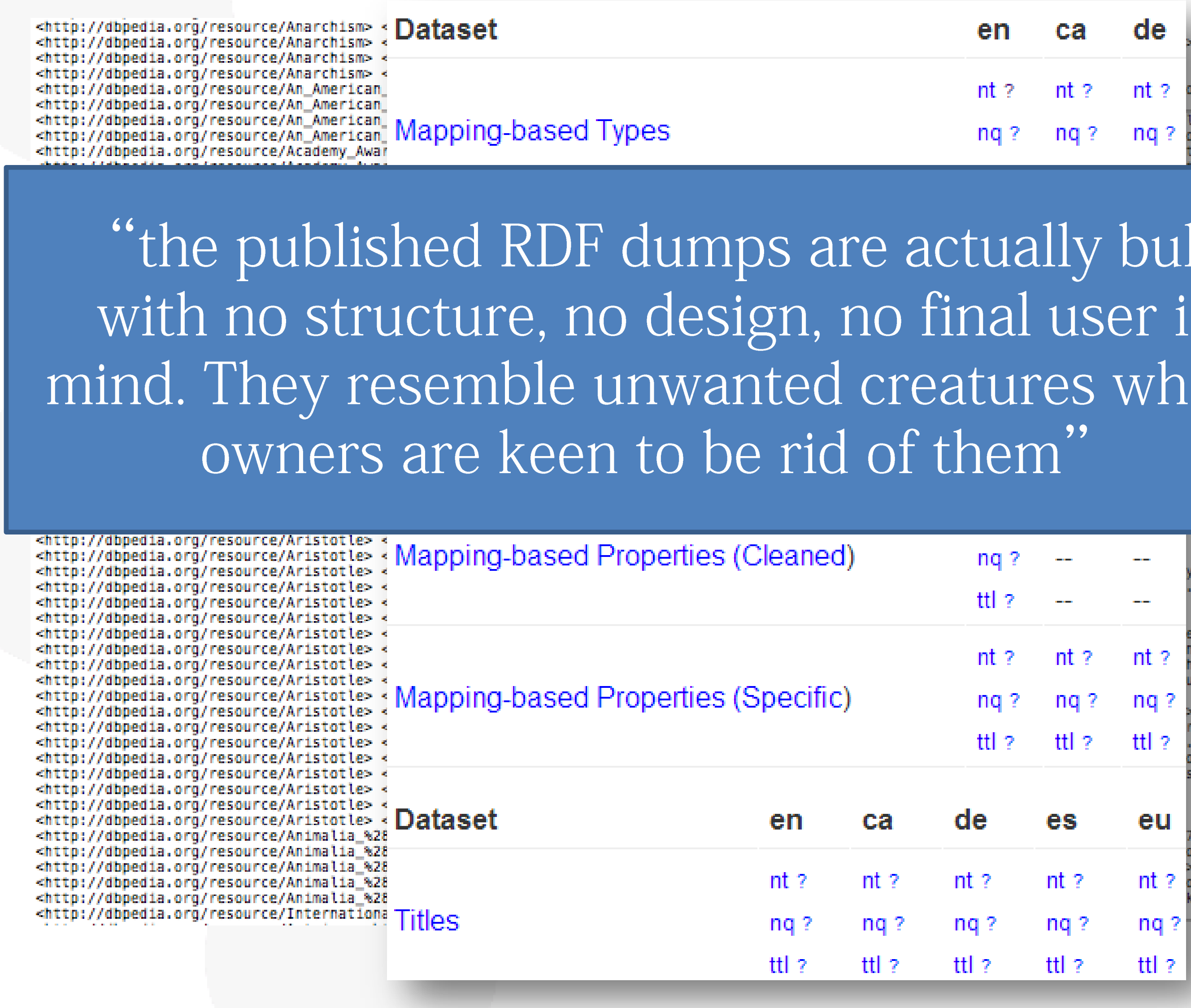
Dataset	en	ca	de	es	eu
Titles	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?

```
> .
ons_by_Geor
_poems> .
ll_of_Fame_
ositions> .
tegory:Acad
tegory:Best_
s> .
_of_Nat
_estab
ate_Sta
_United_States> .
tates> .
s> .
aticians> .
ophers> .
ists> .
cedon> .
.
 .
> .
y> .
.
ent_Chalcidice> .
n_Classical_Athens> .
hers> .
uage> .
 .
> .
rs> .
.
ors_of_Alexander_the_Great> .
sophers> .
7s_picture_books> .
ooks> .
> .
oks_by_Graeme_Base> .
ks> .
_scales> .
```

DBpedia

431 M.triples~ 63 GB

Publication & Exchange: Is this the best option?



```
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/Anarchism> <
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/An_American_
<http://dbpedia.org/resource/Academy_Awar
```

```
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Aristotle> <
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Animalia_%28
<http://dbpedia.org/resource/Internationa
```

Dataset	en	ca	de
Mapping-based Types	nt ? nq ?	nt ? nq ?	nt ? nq ?
Mapping-based Properties (Cleaned)	nq ? ttl ?	-- --	-- --
Mapping-based Properties (Specific)	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?

Dataset	en	ca	de	es	eu
Titles	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?	nt ? nq ? ttl ?

```
> .
ons_by_George_Gershwin> .
_poems> .
ll_of_Fame_Award
ositions> .
tegory:Acad
```

```
y> .
.
ent_Chalcidice> .
n_Classical_Athens> .
hers> .
uage> .
 .
> .
rs> .
.
ors_of_Alexander_the_Great> .
sophers> .
7s_picture_books> .
ooks> .
> .
oks_by_Graeme_Base> .
ks> .
_scales> .
```

> "the published RDF dumps are actually bulks with no structure, no design, no final user in mind. They resemble unwanted creatures whose owners are keen to be rid of them"

B

Claudio Gutiérrez

Drawbacks of Current RDF Serialization

- Very verbose
  - Designed for human readability (not for machines)
  - HUGE → text compression/decompression

- Lack of (standard) metadata
  - "What is this?" phenomenon

- Search offline
  - Scan the whole exchanged dump.
  - (decompress)+ index the file + search

Drawbacks of Current RDF Serialization

- Very verbose
  - for machines)
- Lac
  - “Wh
  - ge
  - dex the fil

In this RDF deluge,
if RDF is meant to be machine processable,
why are we using plain RDF??

My thesis hypothesis

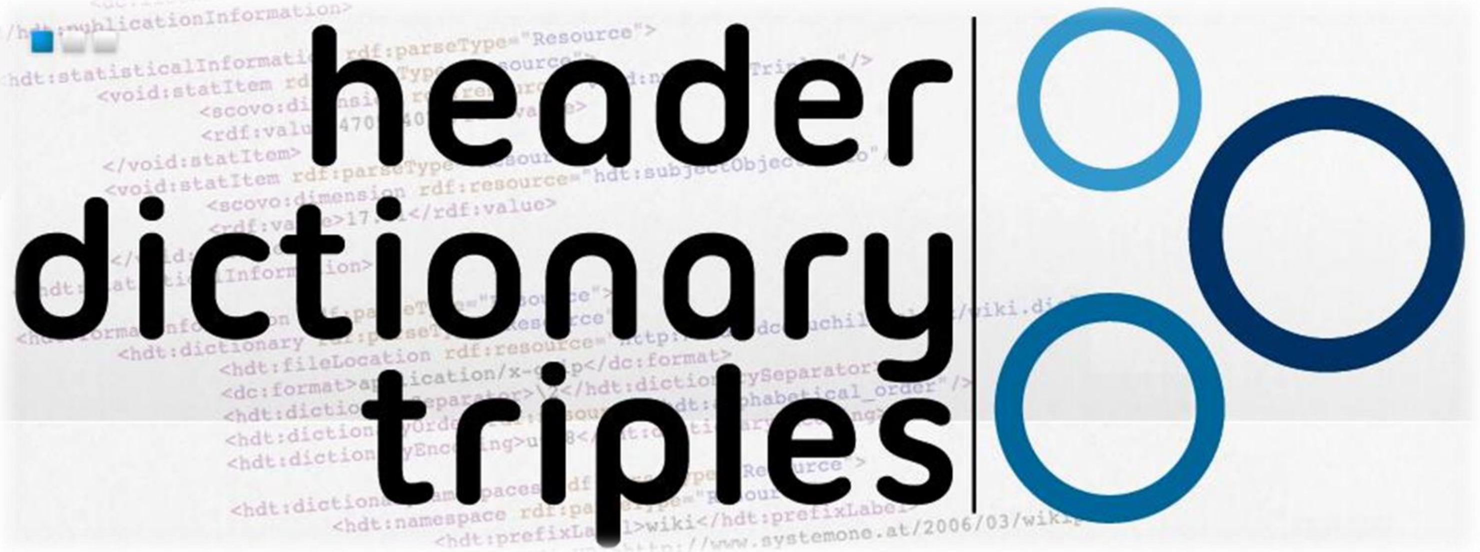
Given an RDF dataset, potentially huge, a lightweight binary RDF can encode the data leveraging the skewed structure of RDF graphs for the purposes of

1. Large spatial savings,
2. easy and modular data-centric publication and parsing and
3. data retrieval.

Applications:

- Publish a large dataset on the Web.
- Transfer between two servers.
- Distributed RDF Data Management.
- Fast In-Memory Query Engine.

What have I done so far?



- Binary Serialization of RDF
- Highly compact
- Includes indexes to solve SPARQL Triple Patterns once it is loaded in main memory
- W3C Submission. http://www.w3.org/Submission/2011/03/

Alcatel·Lucent

cisco

OPENLINK SOFTWARE

profium

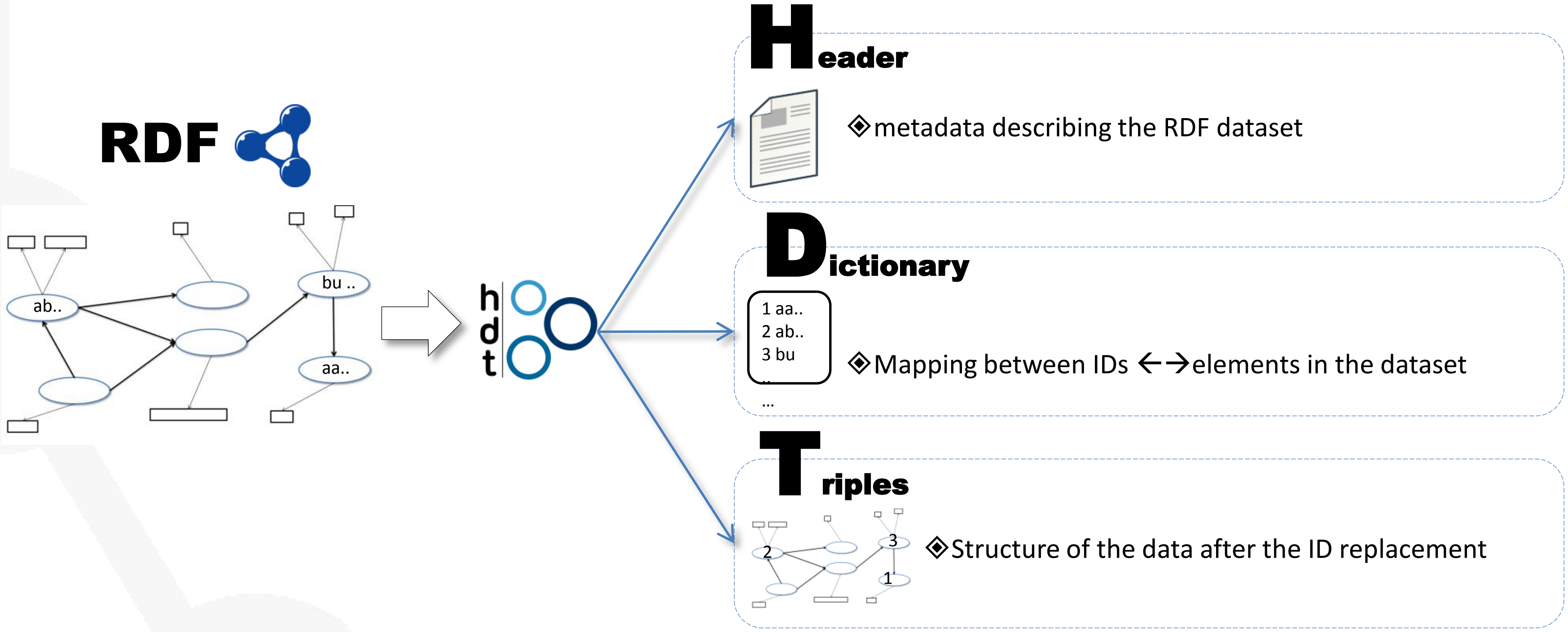
Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen · Bolzano

The Open University

Universidad Politécnica Madrid

DERI

HDT (Header-Dictionary-Triples) Overview



Header

◇metadata describing the RDF dataset

Dictionary

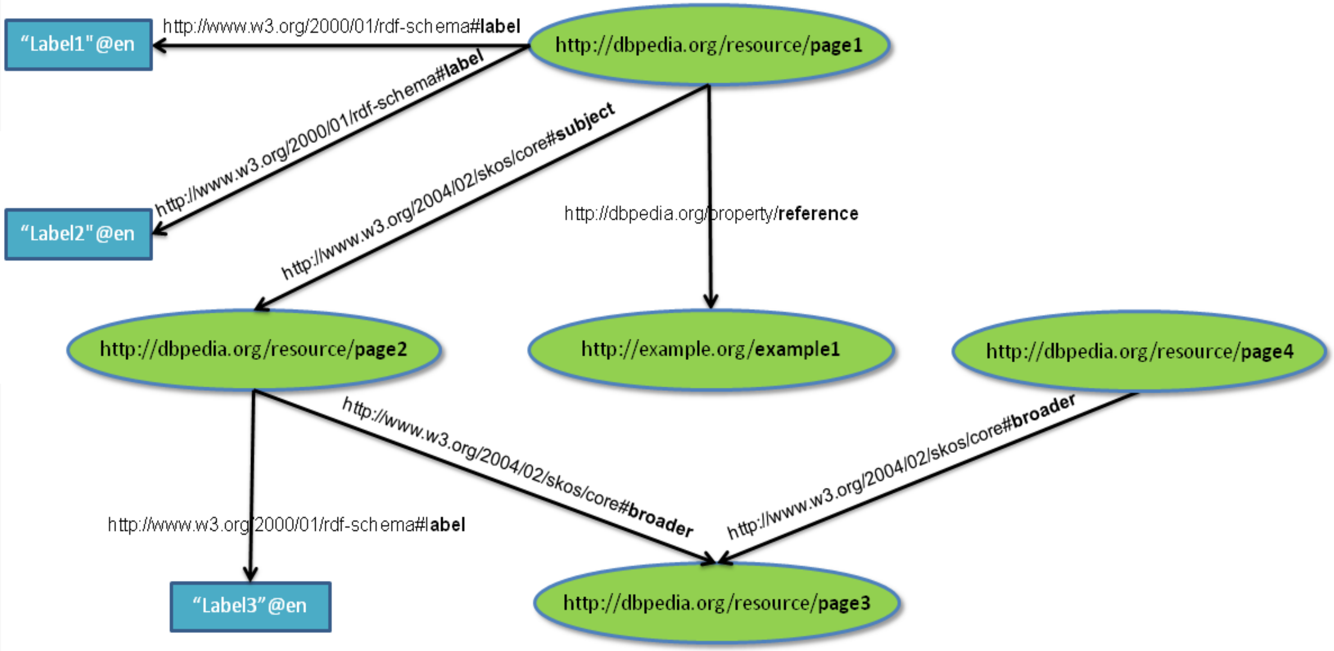
1 aa..
2 ab..
3 bu
…

◇Mapping between IDs ←→elements in the dataset

Triples

◇Structure of the data after the ID replacement

Dictionary


"Label1"@en
http://www.w3.org/2000/01/rdf-schema#label
http://dbpedia.org/resource/page1
http://www.w3.org/2000/01/rdf-schema#label
"Label2"@en
http://www.w3.org/2004/02/skos/core#subject
http://dbpedia.org/property/reference
http://dbpedia.org/resource/page2
http://example.org/example1
http://dbpedia.org/resource/page4
http://www.w3.org/2004/02/skos/core#broader
http://www.w3.org/2004/02/skos/core#broader
http://www.w3.org/2000/01/rdf-schema#label
"Label3"@en
http://dbpedia.org/resource/page3

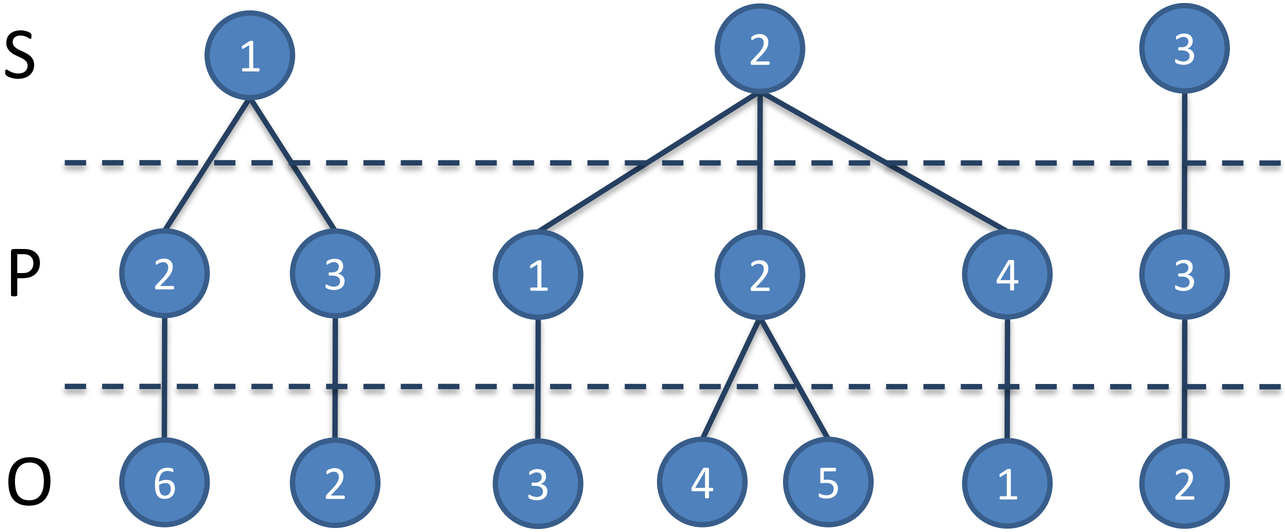
ID		
1	<../page2>	S-O
2	<../page1>	S
3	<../page4>	
2	<../page3>	O
3	<../example1>	
4	"Label1"@en	
5	"Label2"@en	
6	"Label3"@en	
1	<../reference>	P
2	<../#label>	
3	<../#broader>	
4	<../#subject>	

- Mapping of strings to correlative IDs. {1..n}
- Lexicographically sorted, no duplicates.
- Front Coding for each section.

Triples

Triples

1 2 6
1 3 2
2 1 3
2 2 4
2 2 5
2 4 1
3 3 2



Array Y	2	3	1	2	4	3
Bitmap Y	1	0	1	0	0	1

Array Z	6	2	3	4	5	1	2
Bitmap Z	1	1	1	1	0	1	1

Demo

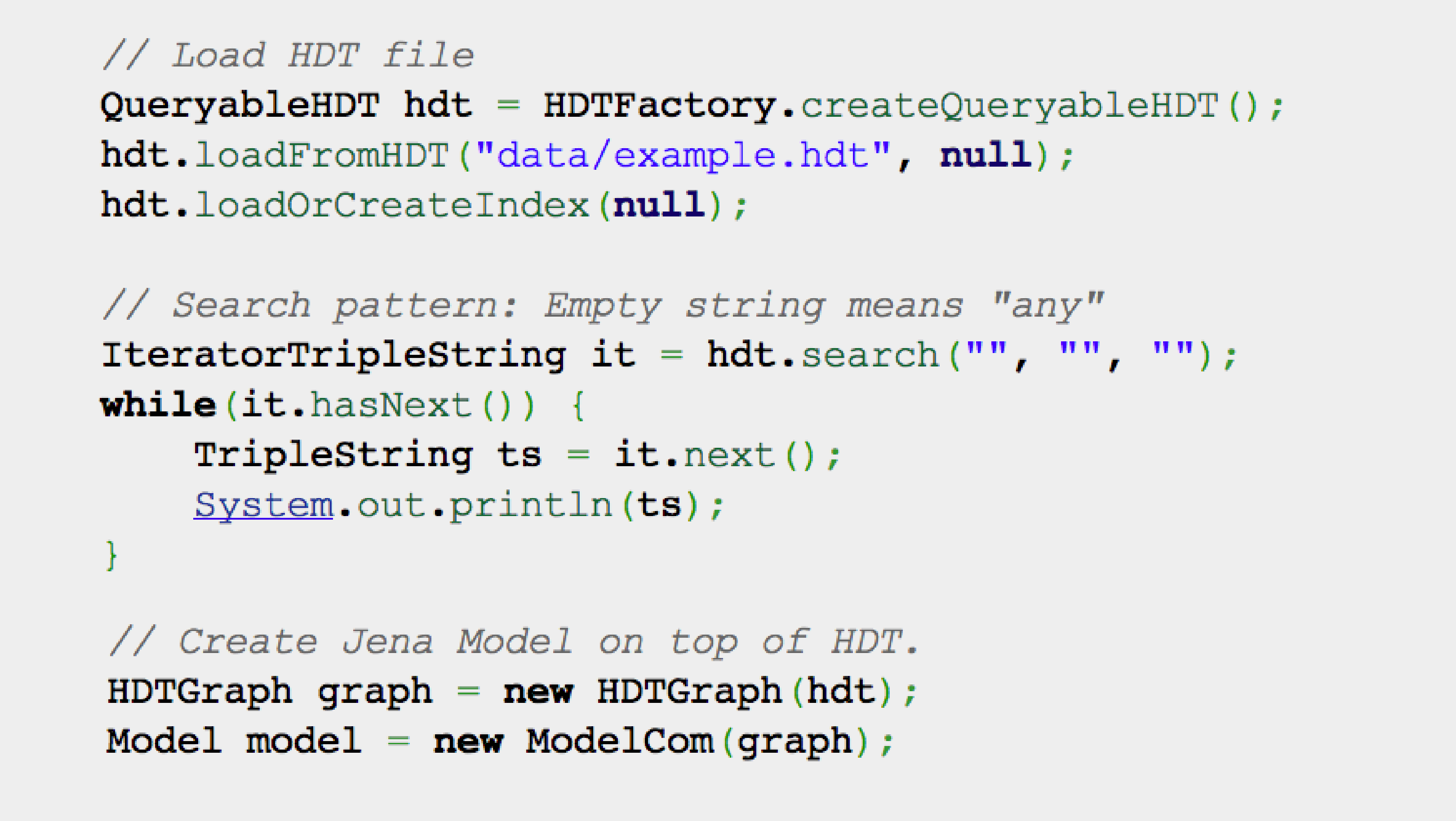
FOUR current ways of consuming HDT

rdfhdt.org

- [1] Command line Tool:
  - Export/Import
    - $ rdf2hdt file.nt output.hdt
    - $ hdt2rdf file.hdt output.nt
  - Query
    - $ hdtsearch file.hdt

FOUR current ways of consuming HDT

- [2] C++/Java Library -> Use within Jena!



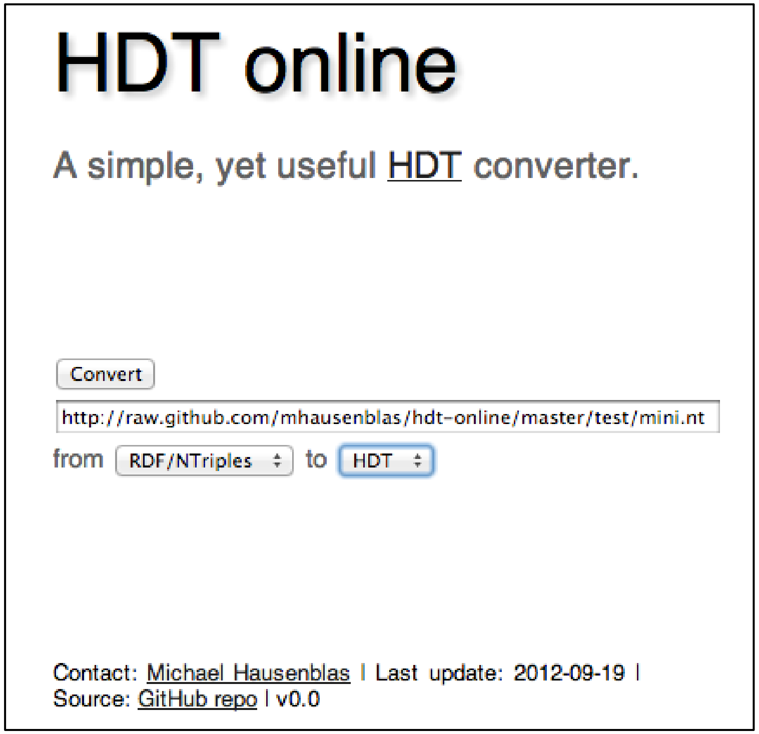
```
// Load HDT file
QueryableHDT hdt = HDTFactory.createQueryableHDT();
hdt.loadFromHDT("data/example.hdt", null);
hdt.loadOrCreateIndex(null);

// Search pattern: Empty string means "any"
IteratorTripleString it = hdt.search("", "", "");
while(it.hasNext()) {
    TripleString ts = it.next();
    System.out.println(ts);
}

// Create Jena Model on top of HDT.
HDTGraph graph = new HDTGraph(hdt);
Model model = new ModelCom(graph);
```

FOUR current ways of consuming HDT

- [3] Web Service:
  - Import into HDT
    - http://srvgal85.deri.ie/hdt-online/
    - Thanks to Michael Hausenblas.



HDT online

A simple, yet useful HDT converter.

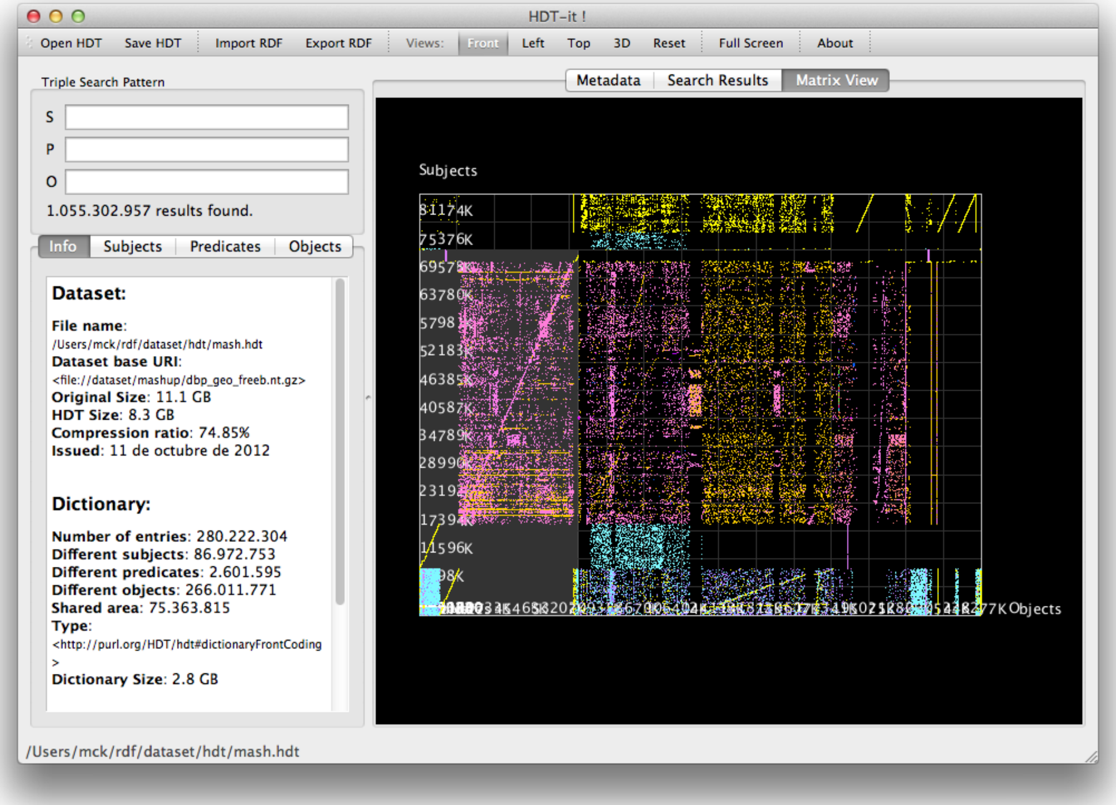
http://raw.github.com/mhausenblas/hdt-online/master/test/mini.nt

from RDF/NTriples to HDT

Contact: Michael Hausenblas | Last update: 2012-09-19 |
Source: GitHub repo | v0.0

FOUR current ways of consuming HDT

- [4] Desktop tool HDT-it!
  - Thanks to Mario Arias (DERI)



1.055.302.957 results found.

Dataset:

File name:
/Users/mck/rdf/dataset/hdt/mash.hdt
Dataset base URI:
<file://dataset/mashup/dbp_geo_freeb.nt.gz>
Original Size: 11.1 GB
HDT Size: 8.3 GB
Compression ratio: 74.85%
Issued: 11 de octubre de 2012

Dictionary:

Number of entries: 280.222.304
Different subjects: 86.972.753
Different predicates: 2.601.595
Different objects: 266.011.771
Shared area: 75.363.815
Type:
<http://purl.org/HDT/hdt#dictionaryFrontCoding
>
Dictionary Size: 2.8 GB

/Users/mck/rdf/dataset/hdt/mash.hdt

Results

- Data is ready to be consumed 10-15x faster.
  - Exchange time reduced.
  - Indexing burden on server = Lightweight client processing.
- Competitive query performance.
  - Very fast on triple patterns.
  - Joins on the same scale of existing solutions.
- This is useful for applications that…
  - need a fast, compact read-only in-memory RDF store.
  - consider a static view of RDF datasets
  - want to share self-queryable RDF dumps.
  - need fast download & query.

What am I currently doing?

Toward Efficient Interchange of RDF

Data Streams

Running example. Preliminary hypothesis & Research Questions

BRONX
QUEENS
BROOKLYN
HUDSON RIVER
EAST RIVER
CENTRAL PARK
Please check our website www.mta.info often for latest service changes.
MTA New York City Transit
Manhattan
LEGEND
All Day Service (Every day 7AM - 10PM)
Part-time Service

- 468 stations
- 4.3 M users/day

PH1.- Given a set of RDF data streams, it is possible to define an RDF interchange format that optimizes the space and time for the data exchange and parsing.

PH2.- Given an RDF streaming engine, and a set of SPARQL queries, a RDF interchange format can be tuned to offer better performance in data exchange among processing nodes and query resolution.

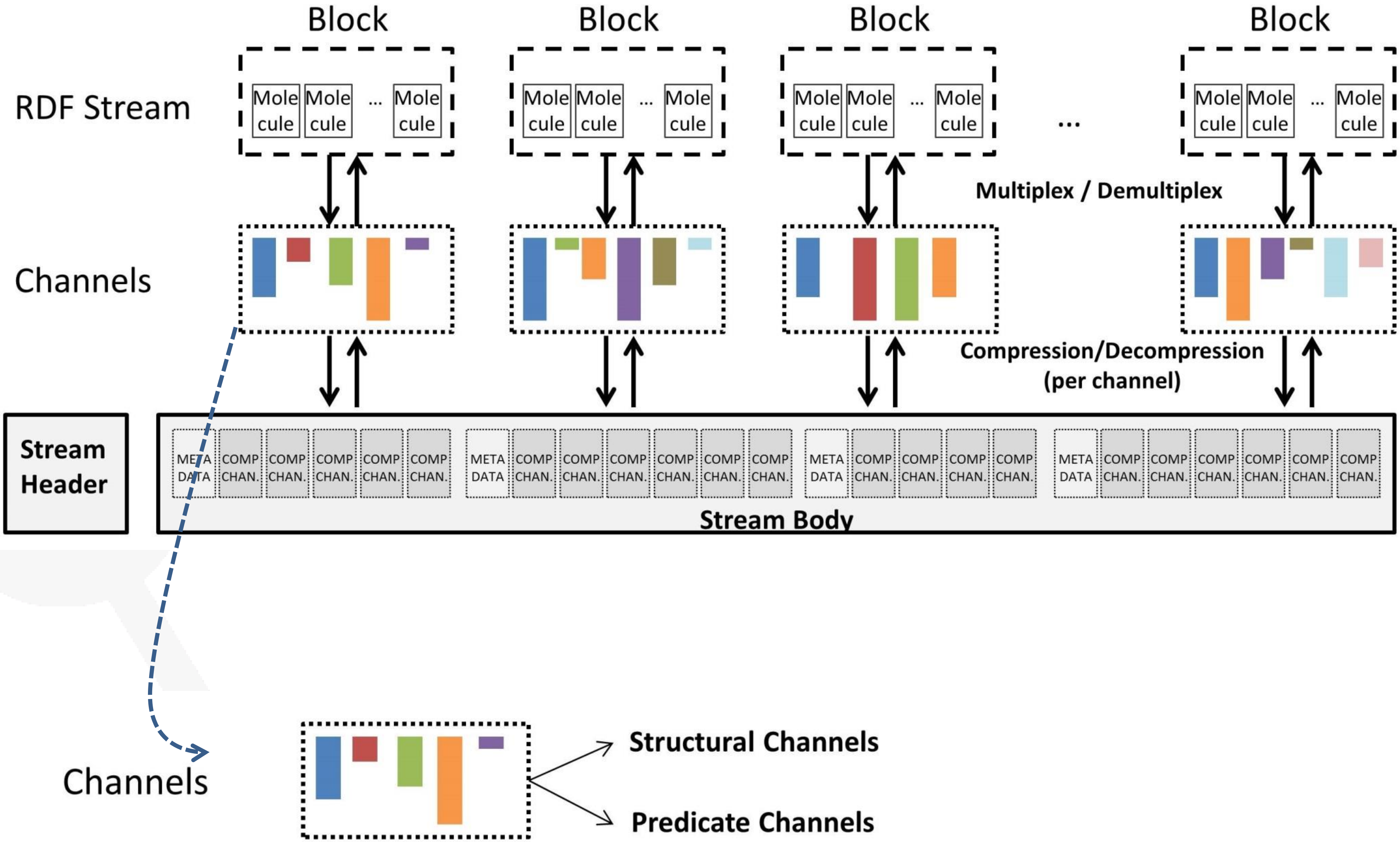
RQ1.- Is HDT a good solution for these dynamic data? NO
RQ2.- Which are the particularities of RDF data streams?
RQ3.- Can an RDF interchange format be parallelizable for compression and decompression (parsing)

Preliminary proposal. Molecules

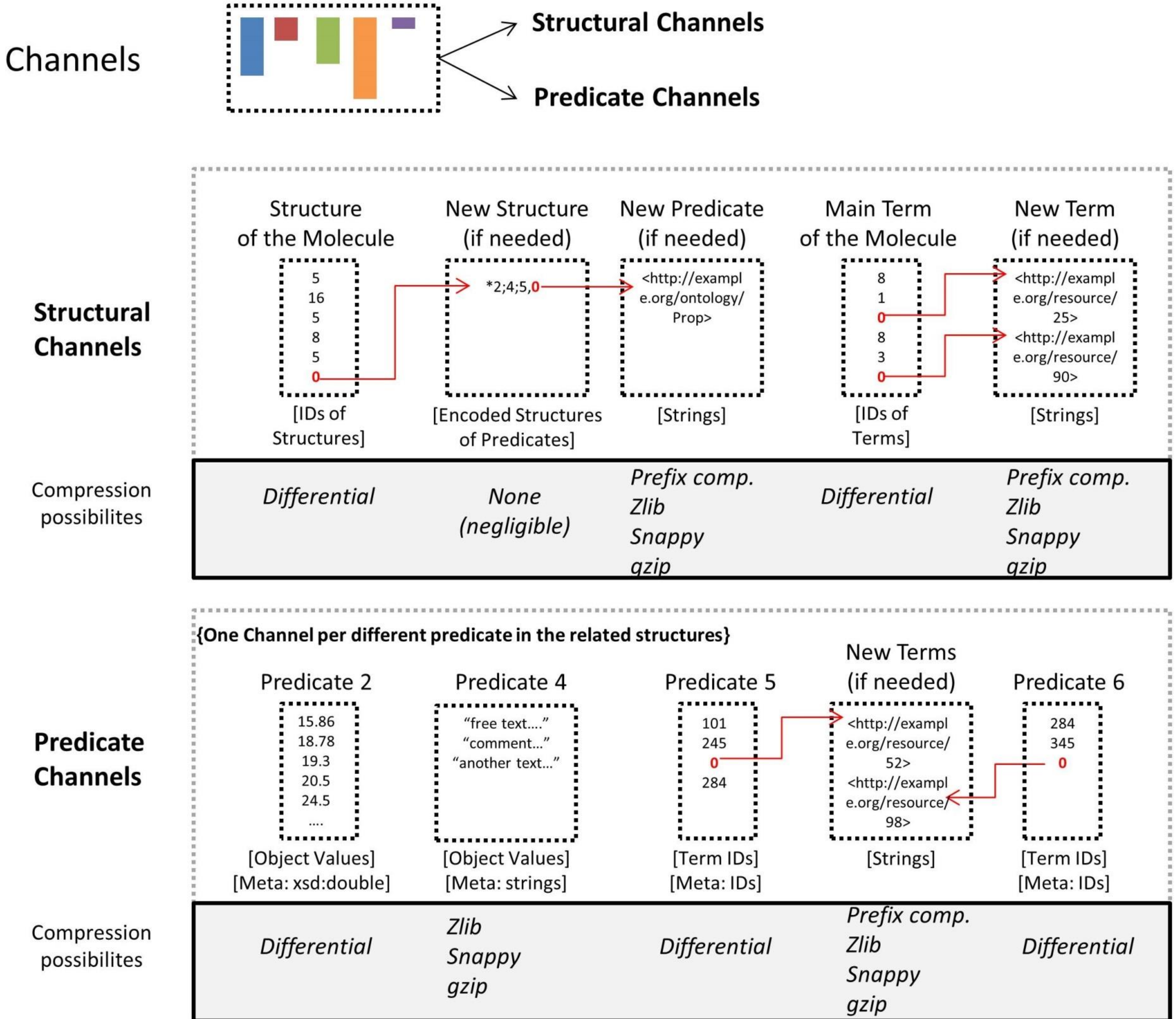
```
sens-obs:Observation_AirTemperature_4UT01_2003_3_31_22_15_00
    a   weather:TemperatureObservation ;
    om-owl:observedProperty weather:_AirTemperature ;
    om-owl:procedure sens-obs:System_4UT01 ;
    om-owl:result sens-obs:MeasureData_AirTemperature_4UT01_2003_3_31_22_15_00 ;
    om-owl:samplingTime sens-obs:Instant_2003_3_31_22_15_00 .
```

```
sens-obs:Observation_WindGust_4UT01_2003_3_31_18_25_00
    a     weather:WindSpeedObservation ;
    om-owl:observedProperty weather:_WindGust ;
    om-owl:procedure sens-obs:System_4UT01 ;
    om-owl:result sens-obs:MeasureData_WindGust_4UT01_2003_3_31_18_25_00 ;
    om-owl:samplingTime sens-obs:Instant_2003_3_31_18_25_00 .
```

Preliminary proposal



Based on: Efficient XML Interchange (EXI) format



Channels

Structural Channels

Predicate Channels

Structural Channels

Structure of the Molecule	New Structure (if needed)	New Predicate (if needed)	Main Term of the Molecule	New Term (if needed)
5 16 5 8 5 0	*2;4;5,0	<http://example.org/ontology/Prop>	8 1 0 8 3 0	<http://example.org/resource/25> <http://example.org/resource/90>
[IDs of Structures]	[Encoded Structures of Predicates]	[Strings]	[IDs of Terms]	[Strings]

Compression possibilites

Differential	None (negligible)	Prefix comp. Zlib Snappy gzip	Differential	Prefix comp. Zlib Snappy gzip

Predicate Channels

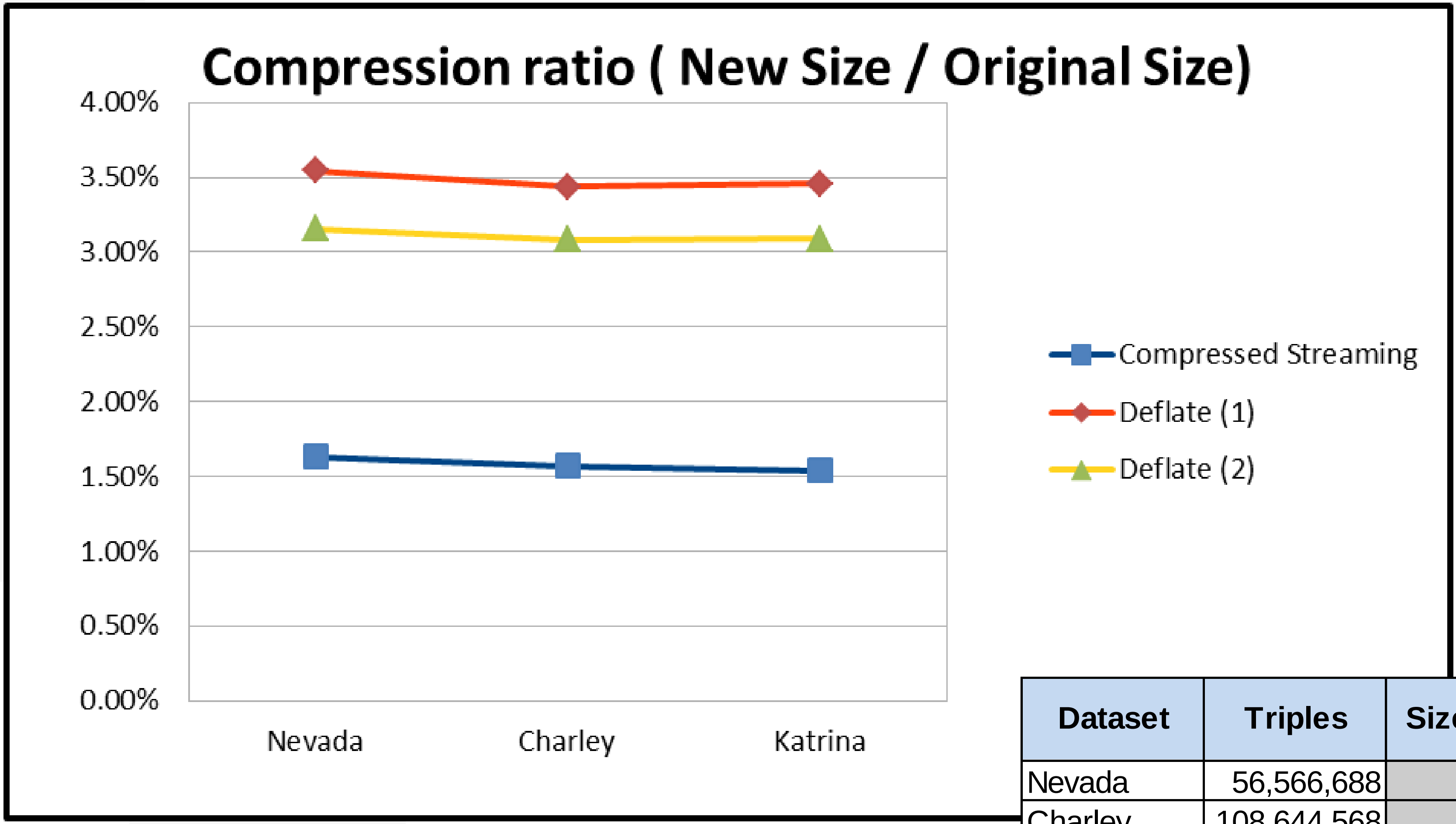
{One Channel per different predicate in the related structures}

Predicate 2	Predicate 4	Predicate 5	New Terms (if needed)	Predicate 6
15.86 18.78 19.3 20.5 24.5	"free text...." "comment..." "another text..."	101 245 0 284	<http://example.org/resource/52> <http://example.org/resource/98>	284 345 0
[Object Values] [Meta: xsd:double]	[Object Values] [Meta: strings]	[Term IDs] [Meta: IDs]	[Strings]	[Term IDs] [Meta: IDs]

Compression possibilites

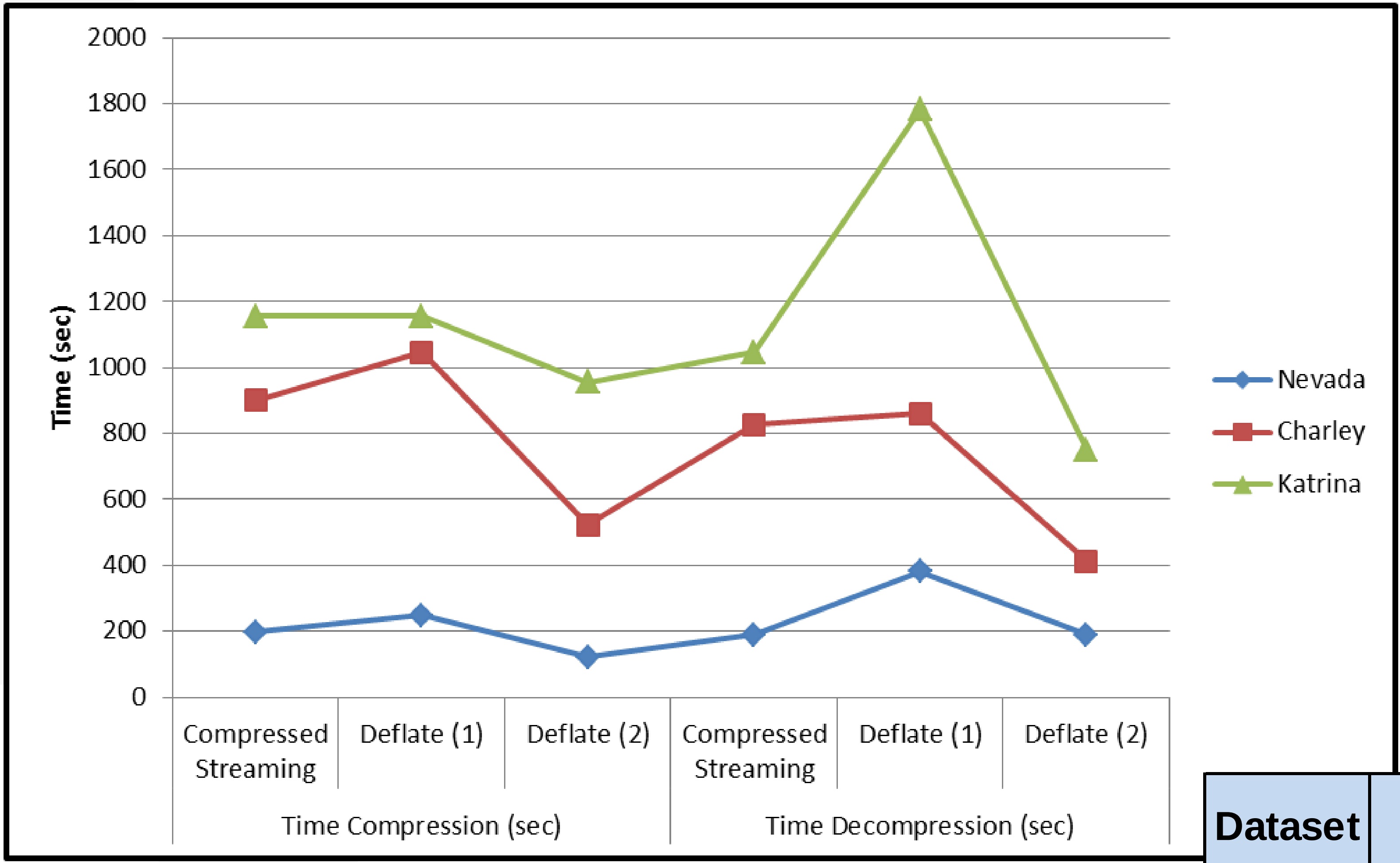
Differential	Zlib Snappy gzip	Differential	Prefix comp. Zlib Snappy gzip	Differential

Preliminary results Compressed Streaming



Dataset	Triples	Size (MB)
Nevada	56,566,688	7,495
Charley	108,644,568	21,470
Katrina	179,128,407	35,548

Preliminary results Compressed Streaming



Dataset	Triples
Nevada	56,566,688
Charley	108,644,568
Katrina	179,128,407

Next steps

- Finish and test the proposal with different data streams
  - Sensor and other data streams → data is welcome!
  - Release the library (Java) → feedback is welcome!
  - ISWC paper
- Parallel compression/decompression
  - preliminary proposal on Storm
- Integration within RDF streaming Engines
  - e.g. morph-streams, CQELS Cloud
  - 3 purposes:
    - scaling to higher input data rates
    - minimizing the data exchange among processing nodes
    - serving a small set of operators on the compressed data

Scalable RDF Management in the Web of Data

Toward Efficient Interchange of RDF Data Streams

Javier D. Fernández

Ontology Engineering Group
Departamento de Inteligencia Artificial
Facultad de Informática
Universidad Politécnica de Madrid

Date: 24/04/2014