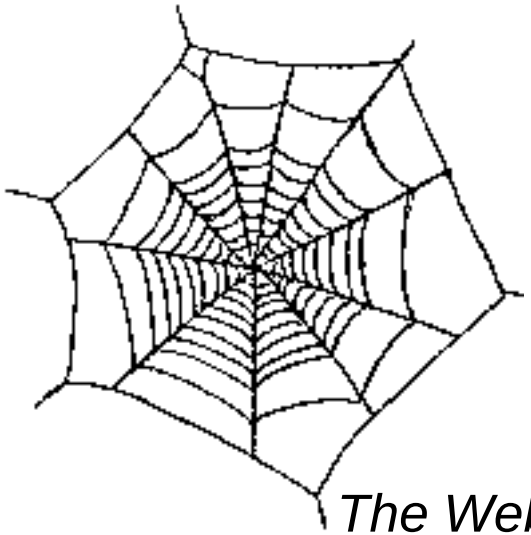
Towards a Dynamic Linked Data Observatory

Tobias Käfer[1], Jürgen Umbrich[2], Aidan Hogan[2], Axel Polleres[3]

WWW2012 Workshop: Linked Data on the Web (LDOW2012)

1) KARLSRUHE INSTITUTE OF TECHNOLOGY, GERMANY 2) DERI, NUI GALWAY, IRELAND 3) SIEMENS AG ÖSTERREICH, VIENNA, AUSTRIA

What‘s this all about?

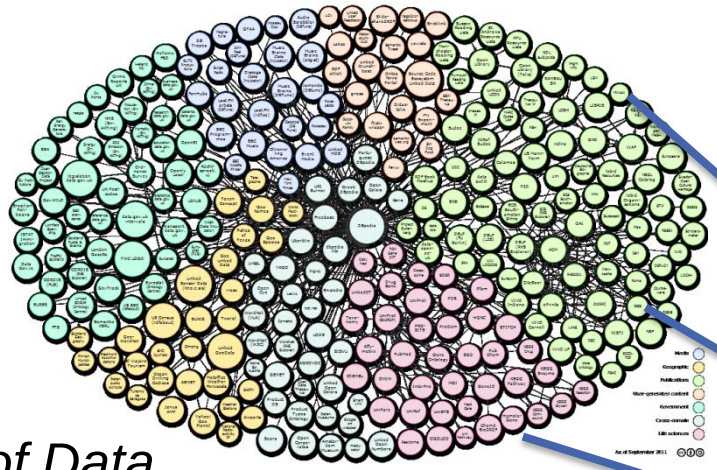


The Web

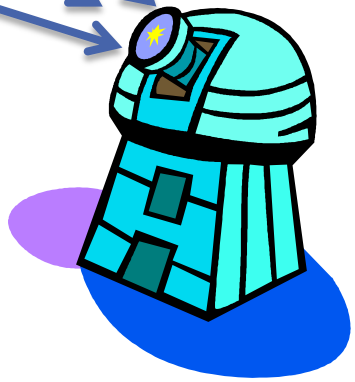
- Dynamic
  - Pages get created
  - Pages get updated
  - Pages get deleted
- Dynamicity causes problems
  - Cache freshness etc.
  - Studied and analysed

Aren‘t we facing similar problems?

What‘s this all about? (Cont‘d)



Periodically download Linked Data



The Web of Data

- Dynamic, too
  - Data gets created, updated, deleted
  - Vocabularies change, predicates are renamed
- Dynamicity influences…
  - Synchronisation of indexes
  - Smart caching of Linked Data content
  - Hybrid search engine architectures
  - …

→ **Creation of a corpus to study the dynamics of Linked Data: The Dynamic Linked Data Observatory**

Building blocks of a Dynamic Linked Data Observatory

In the following

Idea of what to monitor

+

Way of capturing the dimension of time

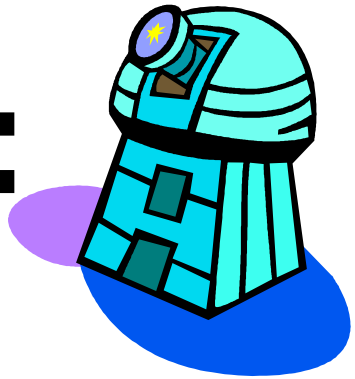
+

Means to create snapshots: **`LDspider`**

+

Bricks (for the sake of the metaphor)

=



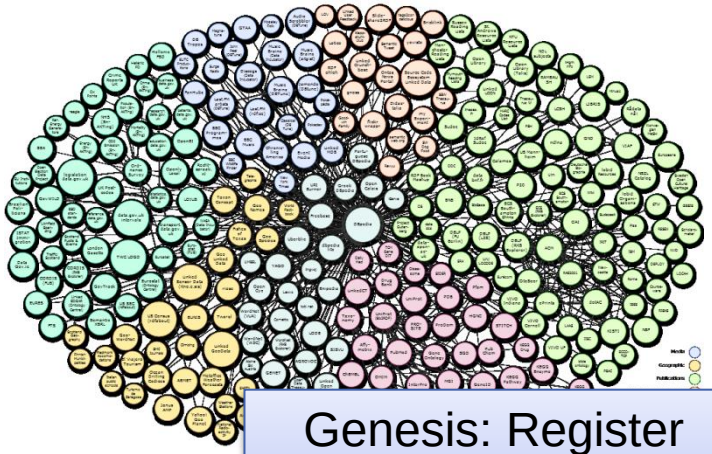
The Dynamic Linked Data Observatory

We need an idea of what to monitor, but:

HOW TO GET A REPRESENTATION OF LINKED DATA?

Requirements for a representation of Linked Data and two candidates

- Coverage
  - Size
  - Diverse data providers
  - Balanced representation of data providers
- Representativeness
  - Study something people consider as LOD



LOD cloud

Genesis: Register dataset, meet requirements

?

Billion Triple Challenge Dataset

Genesis: A crawl

Pros and cons of both datasets

	LOD/CKAN	BTC2011
PROS	Domains pass "quality control" Community validated	Covers more domains* (791) Empirically validated Includes vocabularies Includes decentralised datasets
CONS	Covers fewer domains* (133) Misses vocabularies Misses decentralised datsets like SIOC, GoodRelations	Influence of high-volume domains → unbalanced Misses 47.4% of LOD/CKAN domains

* pay-level domains (PLDs) to be precise

LOD/CKAN vs. BTC2011

WHAT WOULD WE MISS BY CHOOSING EITHER OF THEM?

What sites* would we miss, which would we get? (Top 10 statements)

More fine-grained analysis in the paper!

# of stmts

LOD-Cloud

- linkedgeodata.org
- concordia.ca
- rdfabout.com
- unime.it
- uriburner.com
- sudoc.fr
- viaf.org
- europeana.eu
- moreways.net
- uberblic.org

- tfri.gov.tw
- ontologycentral.com
- legislation.gov.uk
- dbpedia.org
- freebase.com
- bio2rdf.org
- data.gov.uk
- loc.gov
- vu.nl
- bbc.co.uk

BTC2011

- hi5.com
- livejournal.com
- scinets.org
- rdfize.com
- identi.ca
- bibsonomy.org
- opera.com
- archiplanet.org
- rambler.ru
- daml.org

foaf scientific government linking publications

* pay-level domains (PLDs) to be precise

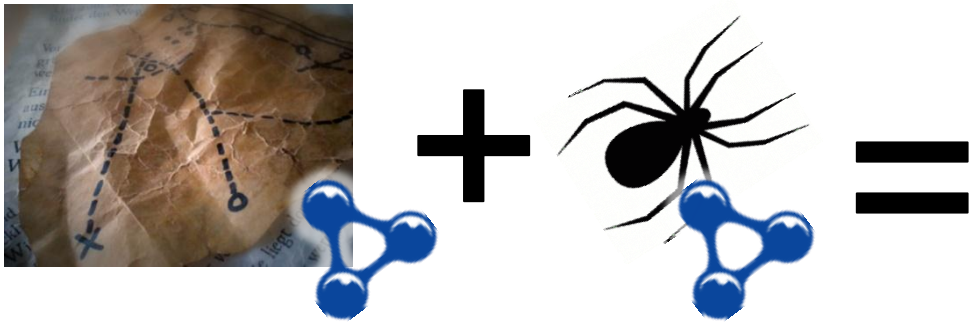
Our conclusion: a compromise

- Combination of CKAN/LOD-Cloud and BTC2011
  - Our sample:
    - 220 example URIs from the LOD-Cloud‘s bubbles
    - 220 highest-ranked (PageRank) URIs from BTC2011*
    - Crawl from there to get a reasonably big seedlist

* Cf. B. Glimm , A. Hogan , M. Krötzsch , A. Polleres: OWL: Yet to arrive on the Web of Data? CoRR abs/1202.0984: (2012)

OUR MONITORING SETUP

Our setup



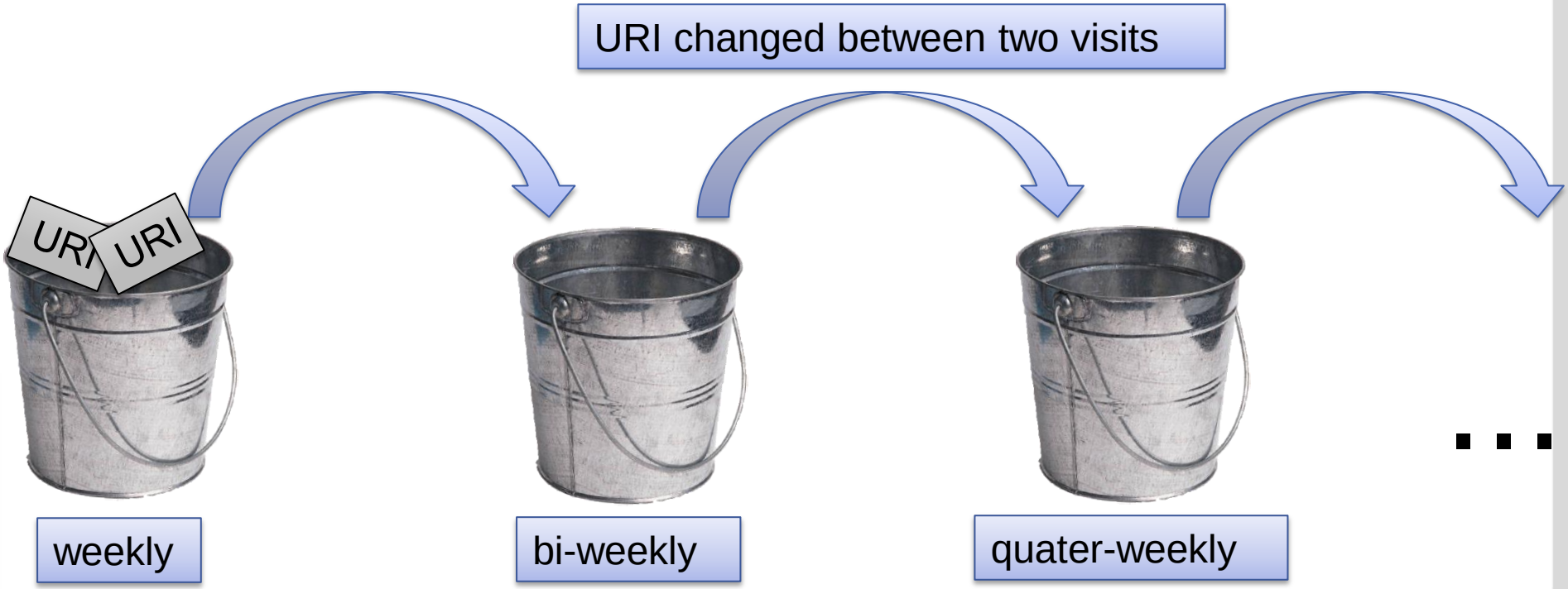
Download seedlist *Crawl*

Published data:

- Seedlist
- The data itself
- access.log
- Frontier of the crawl after each hop

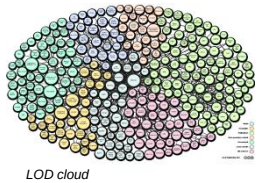
=Taking into account RDF/XML, Turtle, RDFa, N-Triples, Nquads

The Dimension of Time: Sketch of our adaptive revisiting scheme (only for seedlist URIs)



Summary / Q&A

- Summary:
  - Motivated Dataset Dynamics
  - Contrasted CKAN/LOD and BTC2011
  - Described our setup
- Status quo:
  - Close to launch (never been so close)
    - Expected: May 1
  - Web page up
    - http://swse.deri.org/dyldo
  - Google Group up
    - http://groups.google.com/group/dyldo
- Outlook
  - Expected run-time: 1 year
  - Elaborate on publishing issues
  - Interpret data
- Q&A
  - What would be your use-case?
    - Does it need changes to our setup?
  - How do you like our working definition of Linked Data?



vs.

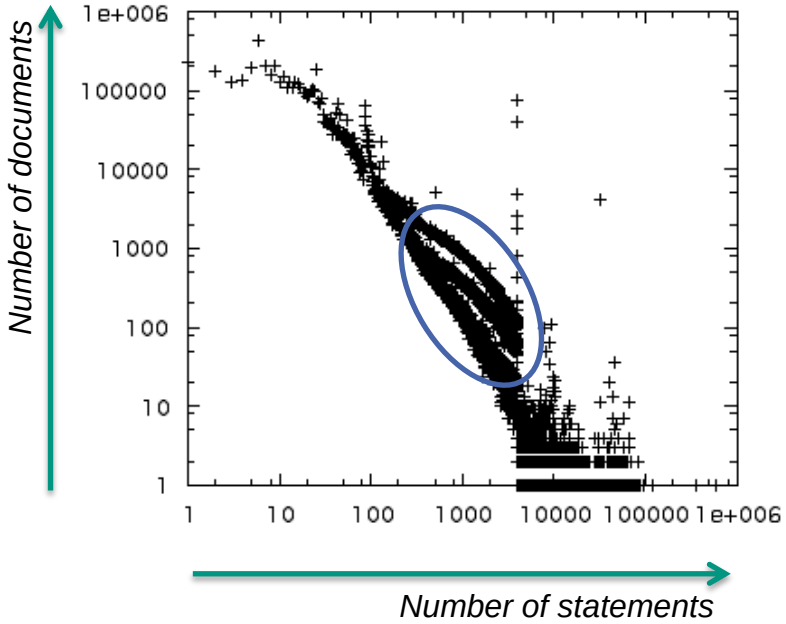
Thanks for your attention!

This presentation is CC BY-SA

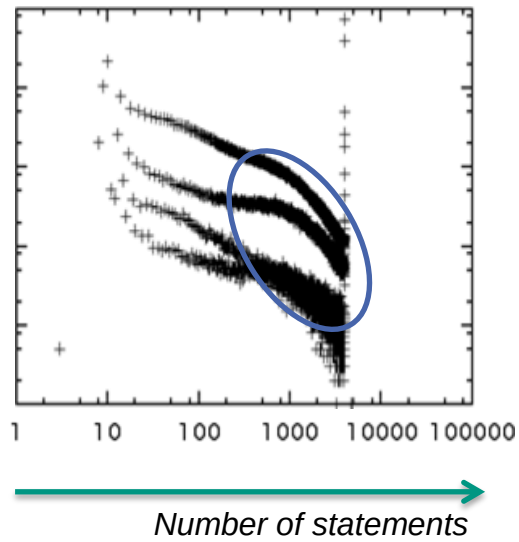
- Picture on title slide based on a picture by A. Sparrow http://www.flickr.com/photos/49937157@N03/
  - CC BY 2.0
- Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/
  - CC BY-SA
- Treasure hunting map by kruxmux http://www.flickr.com/photos/76476049@N00/3946522483/in/photostream
  - CC BY-NC 2.0
- Clock picture by millynet http://www.flickr.com/photos/millynet/134071210/lightbox/
  - CC BY-NC-SA 2.0
- Lens picture by Ben Cooper http://www.flickr.com/photos/cycleologist/1454436980/
  - CC BY-NC-SA 2.0
- Picture on last slide by http://www.flickr.com/photos/stevendepolo/
  - CC BY 2.0

BACKUP

Domination of large exporters in BTC: One provider shapes overall characteristics



BTC2011 dataset



RDF from `http://www.hi5.com` in the BTC2011 dataset

Reasons for largest 10 PLDs in CKAN/LOD not appearing in BTC 2011

PLD	Robots	Http-401	Http-502	Mime	Unreachable	Other
linkedgeodata.org			X	X		
concordia.ca				X		
rdfabout.com				X		
unime.it					X	
uriburner.com	X					
sudoc.fr					X	
viaf.org				X		
europeana.eu						X
moreways.net					X	
uberblic.org		X				

Table 2: Reasons for largest ten PLDs in CKAN/LOD not appearing in BTC 2011.

Excursus: The PLD (pay-level domain)

- Pay money to a Top-level domain registrar
  → get a PLD
- Examples:
  - `http://urq.deri.ie/`
  - `http://www.bbc.co.uk/programmes/b006m10g`
- Same notion, different name:
  - "Site" (Bray, WWW5, 1996)
  - "Top Private Domain" (Google Guava Libraries)

Cf.: Lee et al. Irlbot: Scaling to 6 billion pages and beyond. *ACM Trans. Web*, 3(3):1-34, 2009.