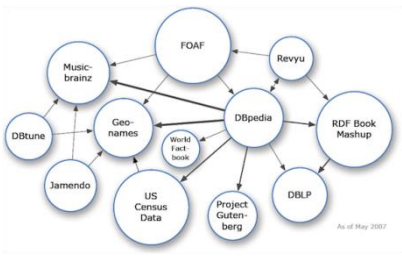
Observing Linked Data Dynamics

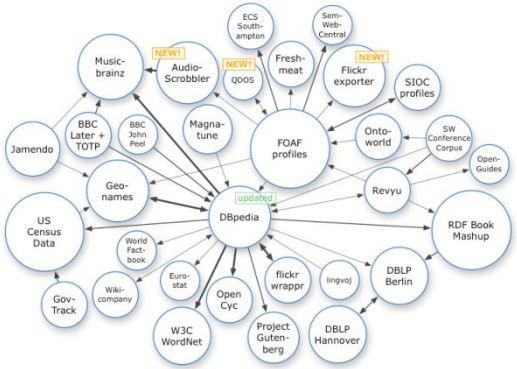
Tobias Käfer[1], Ahmed Abdelrahman[2], Patrick O'Byrne[2], Jürgen Umbrich[2], Aidan Hogan[2]

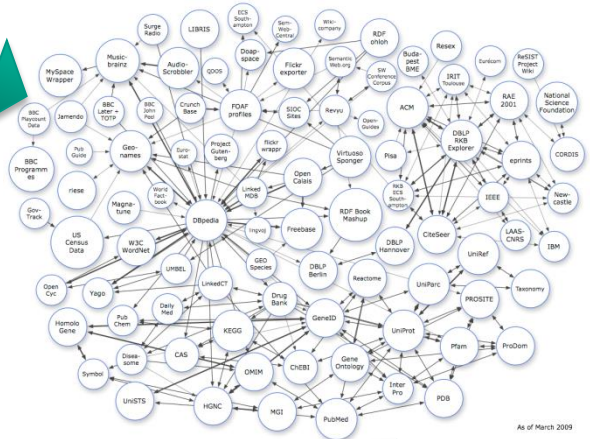
May 30, 2013
Extended Semantic Web Conference (ESWC 2013), Montpellier, France

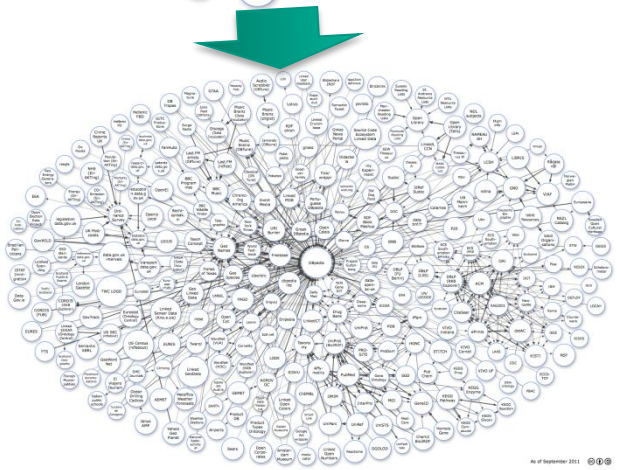
[1] INSTITUTE AIFB, KARLSRUHE INSTITUTE OF TECHNOLOGY, GERMANY; [2] DERI, NATIONAL UNIVERSITY OF IRELAND, GALWAY

Linked Data Dynamics









… more than the growth of the LOD-Cloud

Why you might care:

- As a publisher:
  - Versioning
  - Link Maintenance
- As a consumer:
  - Reasoning
  - Hybrid Linked Data Warehouses

The Dynamic Linked Data Observatory – Part of a Bigger Movement (Web Observatories)

"[…] in order to study the Web, you need to observe what happens on the Web. To do this, one has to study it every day to understand the dynamics of the Web and the interaction with technology, and what people do with it."

Prof. Dame Wendy Hall, 2013

http://www.thehindu.com/sci-tech/internet/web-observatory-for-cybergazing/article4386613.ece

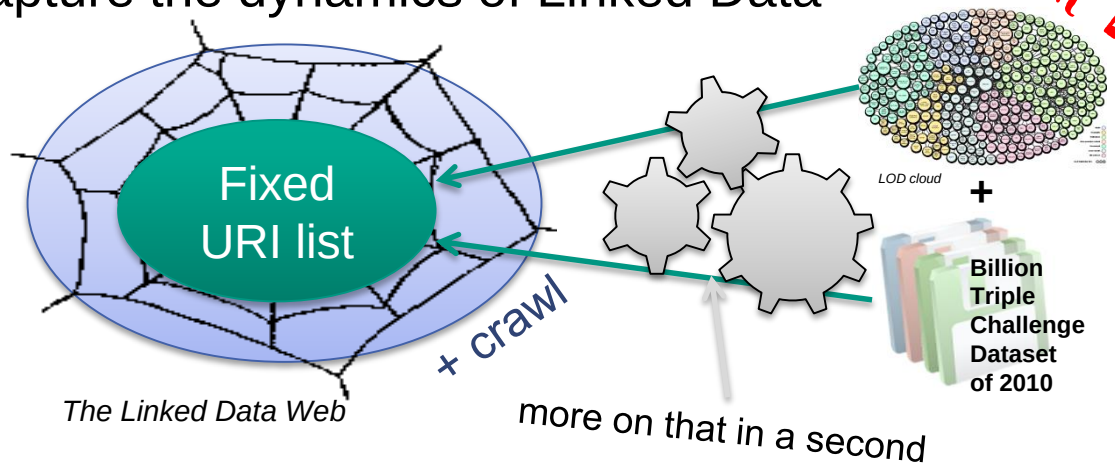
"[…] to create a distributed archive of data on the Web and its activity, and […] mechanisms and tools that will be able to explore its development in the past, to examine its present condition and to establish potential developments in the future."

WebScience Trust: definition of a Web Observatory
A definition of the Web Observatory

The Dynamic Linked Data Observatory

See our Paper at LDOW 2012

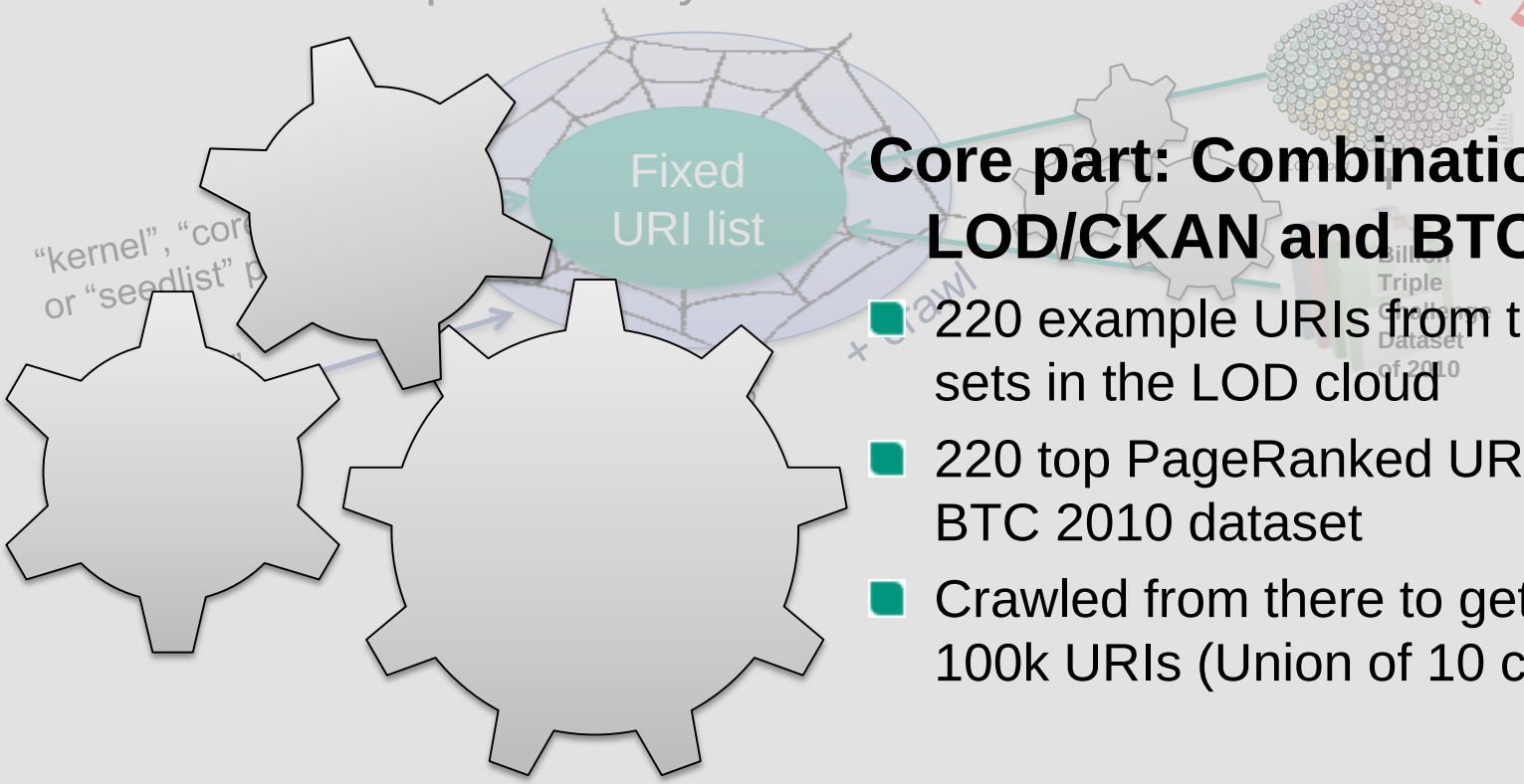
- Mission: To capture the dynamics of Linked Data



The Dynamic Linked Data Observatory

See our Paper at LDOW 2012

- Mission: To capture the dynamics of Linked Data



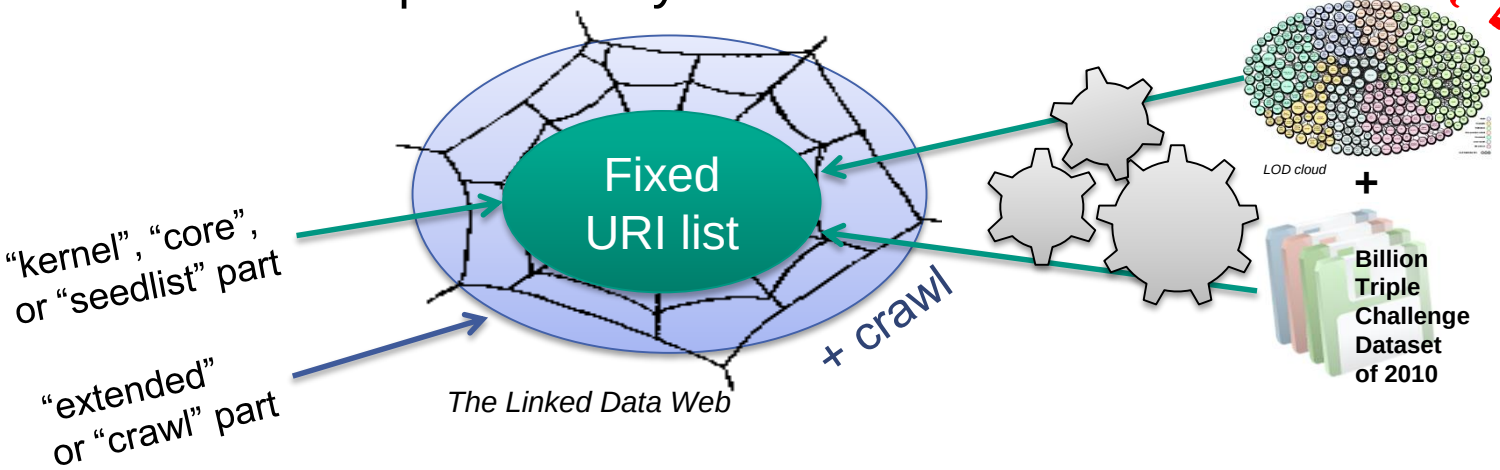
Core part: Combination of LOD/CKAN and BTC

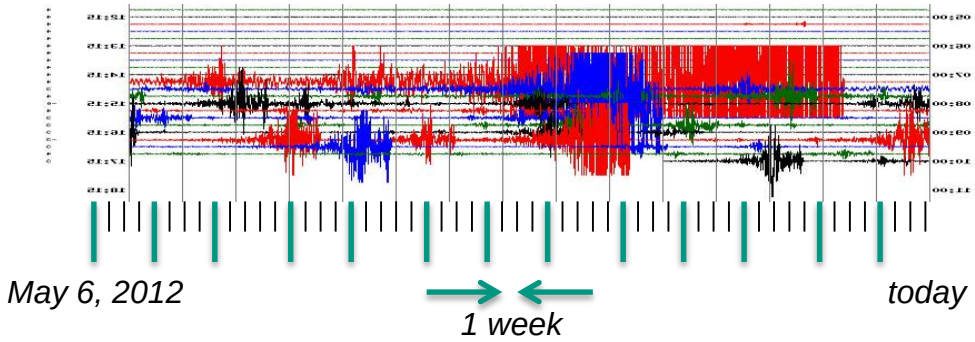
- 220 example URIs from the data sets in the LOD cloud
- 220 top PageRanked URIs from the BTC 2010 dataset
- Crawled from there to get approx. 100k URIs (Union of 10 crawls)

The Dynamic Linked Data Observatory

See our Paper at LDOW 2012

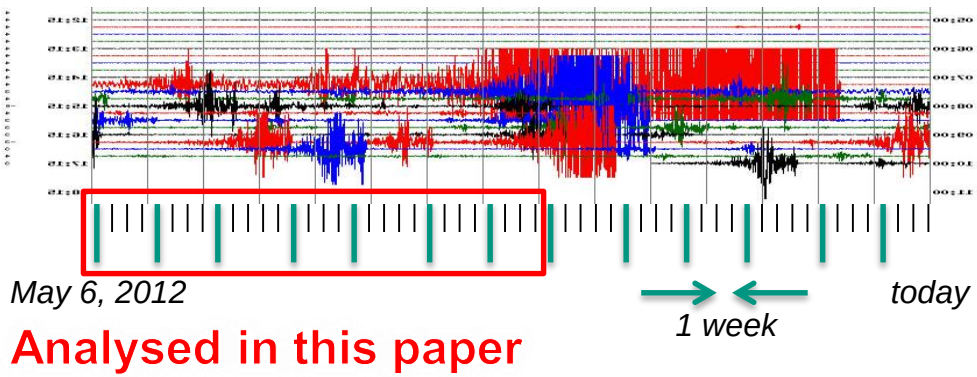
- Mission: To capture the dynamics of Linked Data





→ Weekly snapshots of a URI list derived from the LOD cloud and 2010's Billion triple challenge dataset, chosen for coverage and variety.

This presentation: Findings from the first half year of observation



- Nominal size of a snapshot: 95,737 (Kernel) / 191,474 URIs (Extended)
- May to November 2012: 6 months, 29 (weekly) snapshots
- Statistics on the data basis:

Statistic	Kernel	Extended
Mean pay-level domains	573.6 ± 16.6	1,738.6 ± 218
Mean documents	68,996.9 ± 5,555.2	152,355.7 ± 2,356.3
Mean quadruples	16,001,671 ± 988,820	94,725,595 ± 10,279,806
Sum quadruples	464,048,460	2,747,042,282

Secret questions of a Linked Data geek

How often do links between documents change?

Are document updates mostly additions or mostly deletions?

How frequently does a Linked Data document change?

Can I assume schema data to be static?

What are the most dynamic predicates?

Are there provider-dependent publishing patterns?

→ Call for observations on different levels of abstraction:

(… vs. <html>)

RDF Graphs

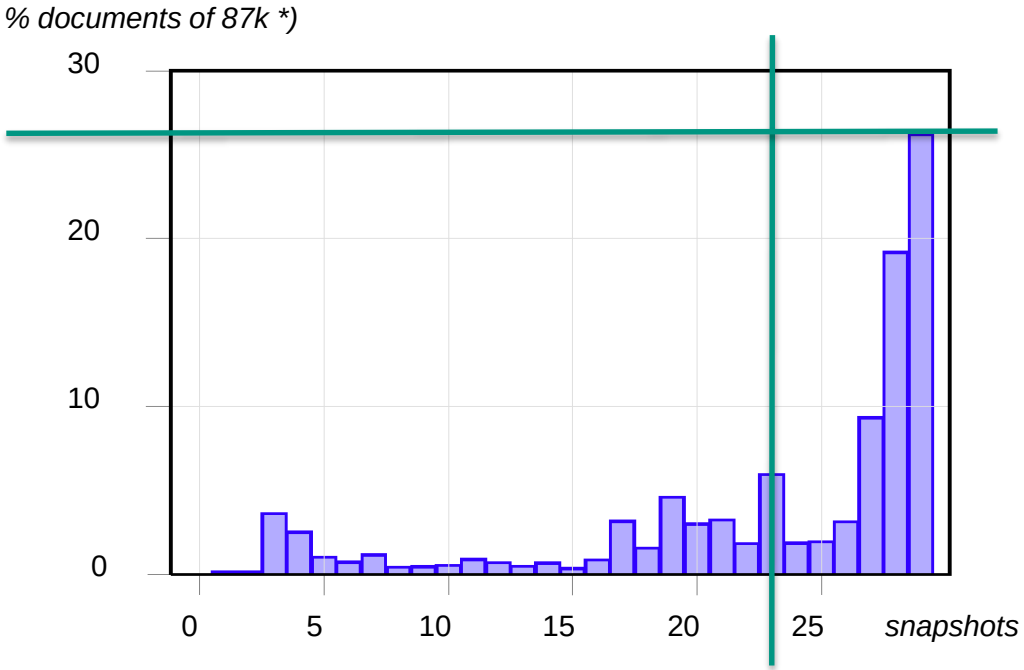
Documents

Hosts (PLD)

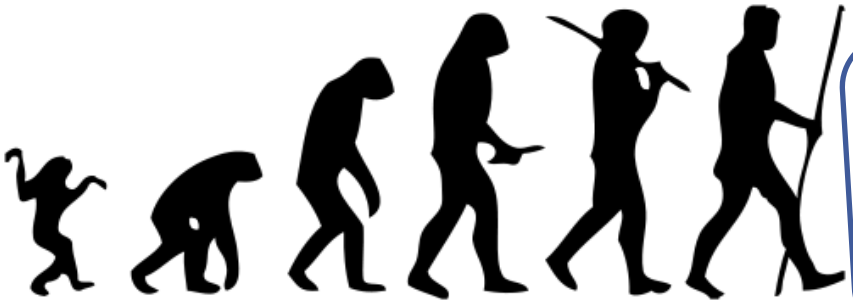
granularity ←

Document-level dynamics: Life (Availability)…

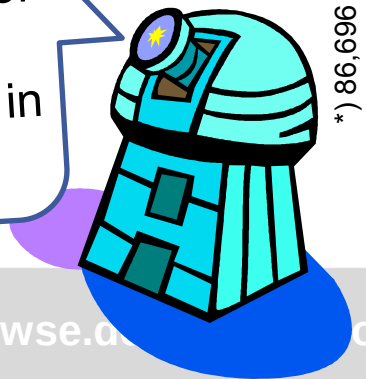
26% URIs available in *all* snapshots



Mean = 23.1 (~80%)

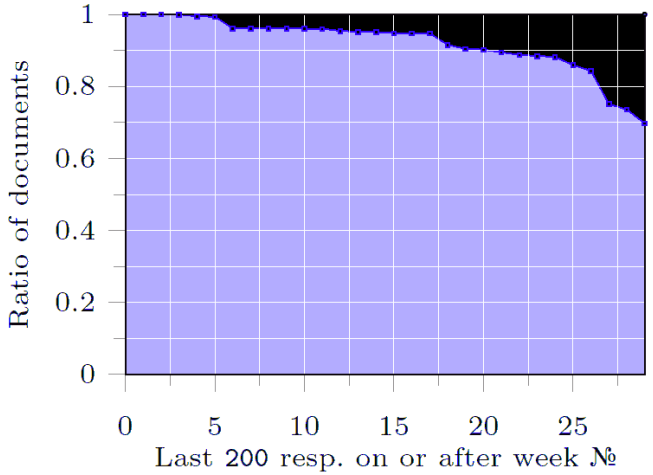


→ You probably miss 20% of the sources in a download (cf. 50% for the HTML web in Fetterly et al. (2003))

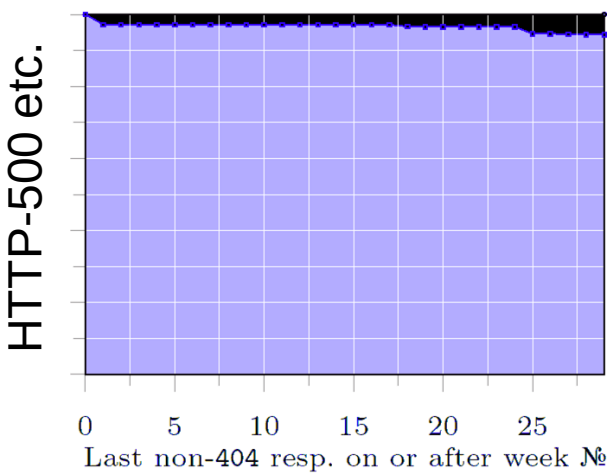


*) 86,696 RDF documents ever appeared in ≥ 1 kernel snapshot

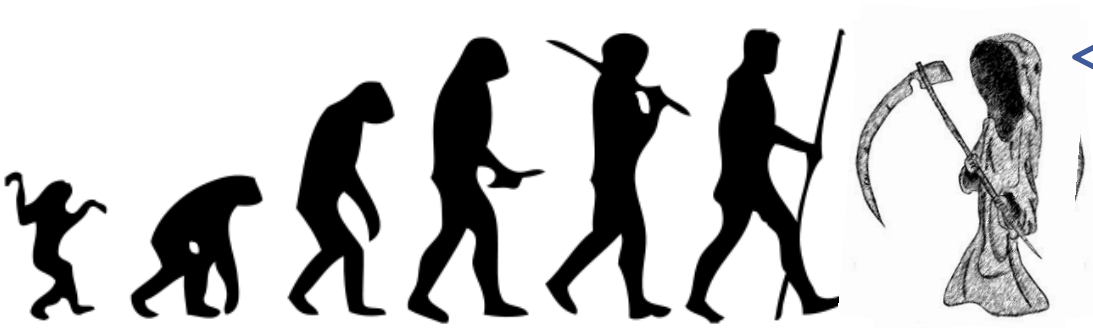
Document-level dynamics: … and Death



Last Heart-Beat:
Overestimates death…

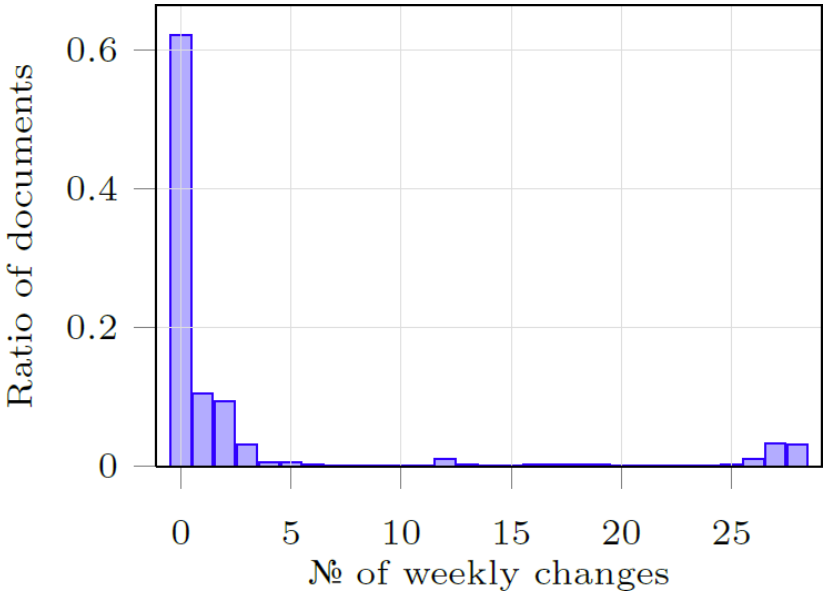


… and death certificate filled:
underestimates death



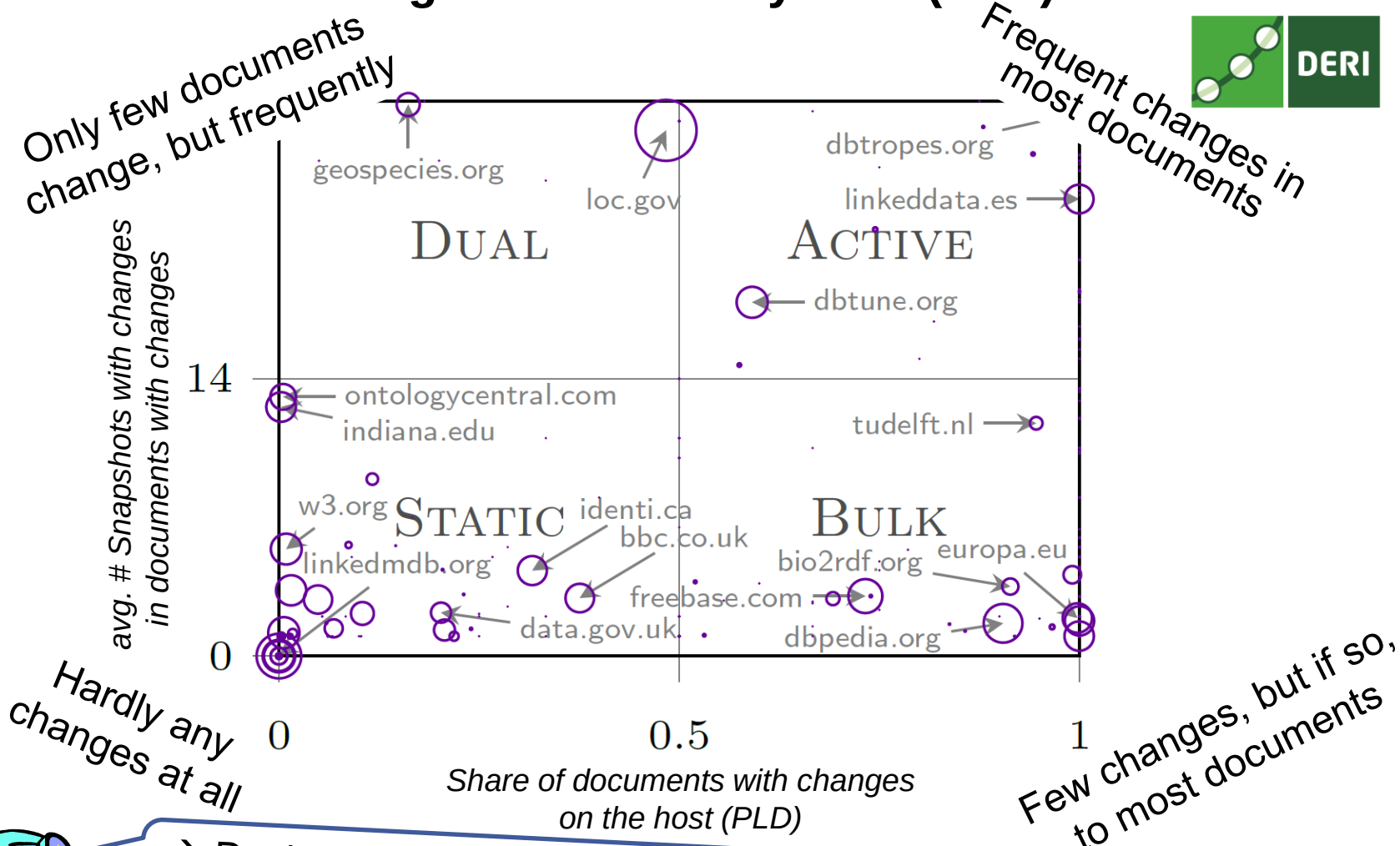
→Of documents, 5% are likely to go dead in 6 months. (cf. 20% and 48% for the HTML web in Koehler (1999) and Ntoulas et al. (2004) resp.)

Document-level dynamics: Changes



→ 62% of all documents were static (cf. 56%, 66%, or 50% reported for the HTML web (Brewington and Cybenko (2000), Fetterly et al. (2003), and Ntoulas et al. (2004)))

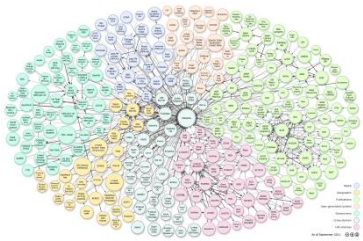
Document-level changes clustered by host (PLD)



→ Decide per host (PLD) on a refreshing strategy (cf. Ntoulas et al. (2004) on per-site HTML change predictions)

Document-level changes per topic and party

- Grouping domains by metadata from the LOD cloud and the DataHub

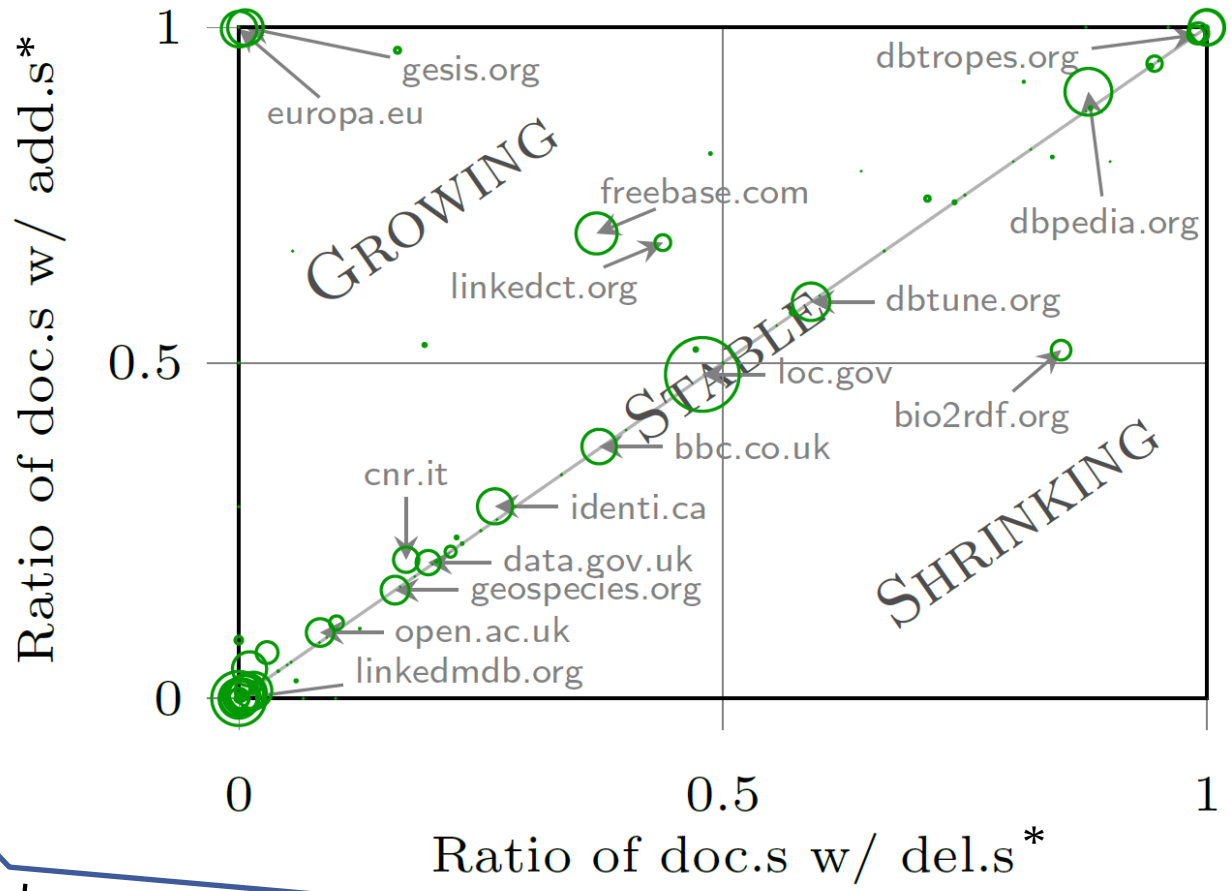


The LOD cloud colour-coded by topic

	Category	Doc №	Dom №	Static		Bulk		Dual		Active	
				№	%	№	%	№	%	№	%
LOD-cloud topic	cross-domain	34,872	33	21	63.64	6	18.18	2	6.06	4	12.12
	geographic	4,693	10	6	60.00	2	20.00	1	10.00	1	10.00
	government	5,544	14	10	71.43	3	21.43	0	0.00	1	7.14
	life-sciences	2,930	4	2	50.00	2	50.00	0	0.00	0	0.00
	media	8,104	10	6	60.00	2	20.00	0	0.00	2	20.00
	publications	14,666	35	24	68.57	8	22.86	2	5.71	1	2.86
	user-generated	7,740	12	7	58.33	0	0.00	0	0.00	5	41.67
	unknown	8,147	502	246	49.00	159	31.67	1	0.20	96	19.12
Party	first-party	22,649	50	38	76.00	8	16.00	2	4.00	2	4.00
	third-party	29,078	61	37	60.66	12	19.67	1	1.64	11	18.03
	both	27,520	23	13	56.52	6	26.09	2	8.70	2	8.70
	unknown	7,449	486	234	48.15	156	32.10	1	0.21	95	19.55
	total	86,696	620	322	51.94	182	29.35	6	0.97	110	17.74

RDF-level dynamics: triples

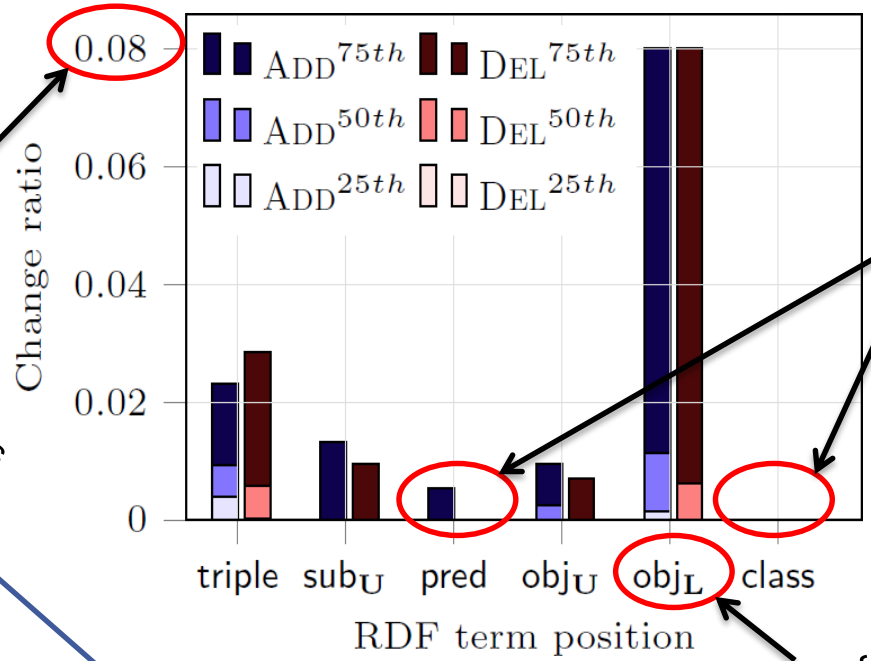
- Only 27,6% of the documents updated values for terms (i.e. one per triple)
- 24% monotonic additions



→Deletions and additions almost always balance out, which calls for efficient data revision strategies in Linked Data Warehouses

* given there are changes at all

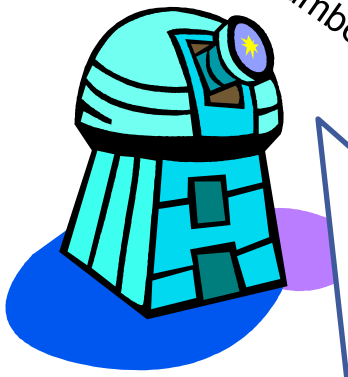
RDF-level dynamics: terms



We're talking small numbers

Static schema signature of a document

Most active; cf. most active predicates



→ Because of the static schema structure of documents, void descriptions don't need to be updated frequently.

RDF-level dynamics: The most dynamic predicates

№	Predicate	Total	+	−
1	dbtont:parsed	35,911	0.94	0.94
2	sioc:has_discussion	3,171	0.87	0.99
3	sioc:content	107,387	0.87	0.98
4	dbtont:fetched	34,894	0.53	0.53
5	swivt:creationDate	35,295	0.53	0.53
6	media:image	1,377	0.49	0.49
7	prv:performedAt	16,706	0.45	0.45
8	xhtml:bookmark	17,852	0.45	0.44
9	linkedct:p.t.u*	2,652	0.42	0.42
10	linkedct:p.t.a*	2,652	0.42	0.42

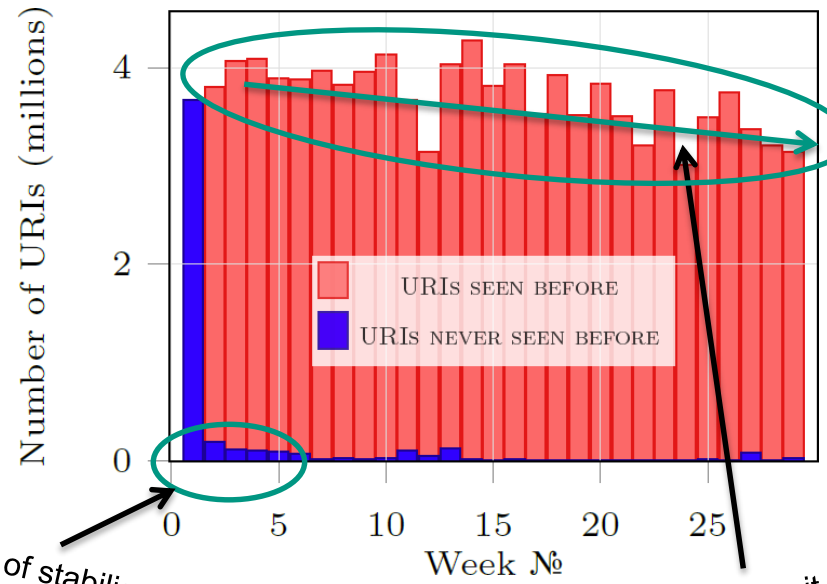
Indicating a timestamp

*) *provenance time updated*, and *provenance time added* respectively

Dynamics of the RDF link structure

- Outward links from the kernel to other documents

Cf. Ntoulas et al. (2004): 25% new links each week (in a growing HTML data set)



Cf. non-200 HTTP responses

Period of stabilisation (cf. inavailability of documents)

If there is a trend, then it is decreasing (cf. dying documents)

Not many new links introduced

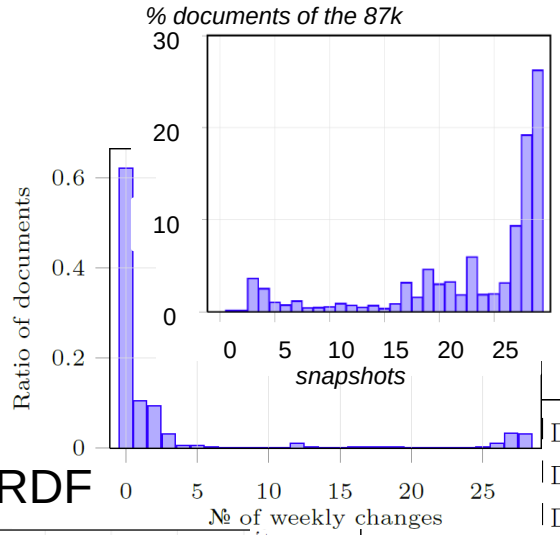
Exceptions (cf. publishing patterns)

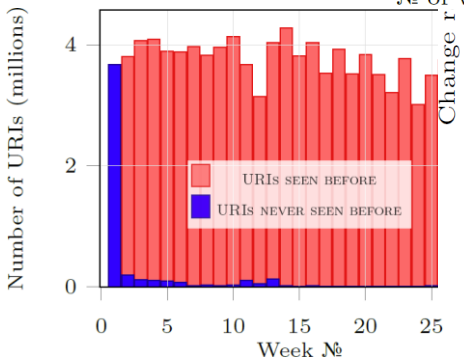
Low-volume but constant stream of fresh outward links : **`sec.gov`**, **`identi.ca`**, **`zitgist.com`**, **`dbtropes.org`**, **`ontologycentral.com`**, **`freebase.com`**

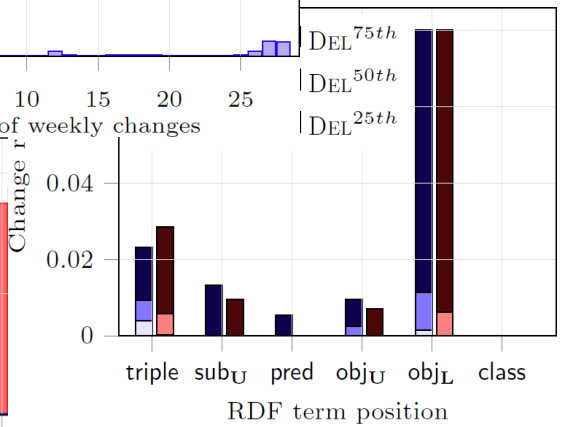
New links in batches: **`bbc.co.uk`**, **`bnf.fr`**, **`dbpedia.org`**, **`linkedct.org`**, **`bio2rdf.org`**

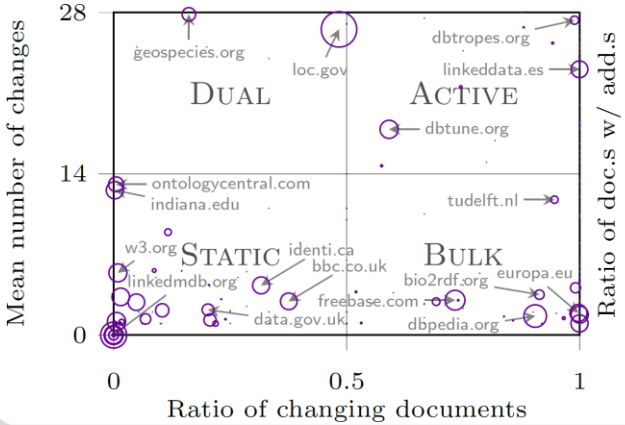
Summary and Q&A

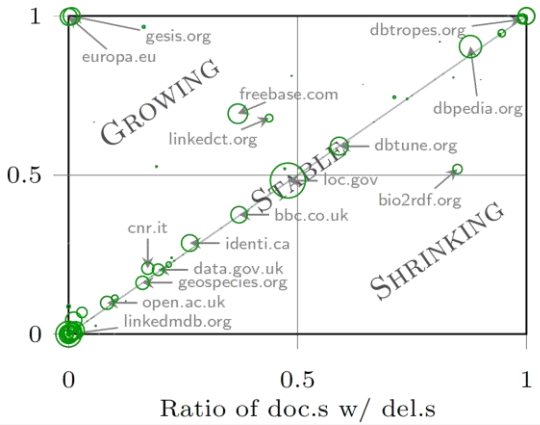
- Analyses from first half year
- Data collection is continuing
- Future work:
  - More sources & analyses, results as RDF
- We appreciate your feed-back and speculations
- What would you look for in the data?
- Thanks for your attention











Our home page with
- more details,
- a google group,
- the data for download,
- and an UI to play around with the data:

http://swse.deri.org/dyldo/

This presentation is CC BY SA – picture credits

- 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
- Evolution http://commons.wikimedia.org/wiki/File:Human_evolution_scheme.svg
  - CC BY SA
- Death http://commons.wikimedia.org/wiki/File:Death.jpg
  - CC BY SA 3.0
- Seismogram http://www.flickr.com/photos/brettneilson/2281403809/
  - CC BY