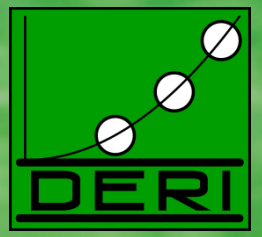
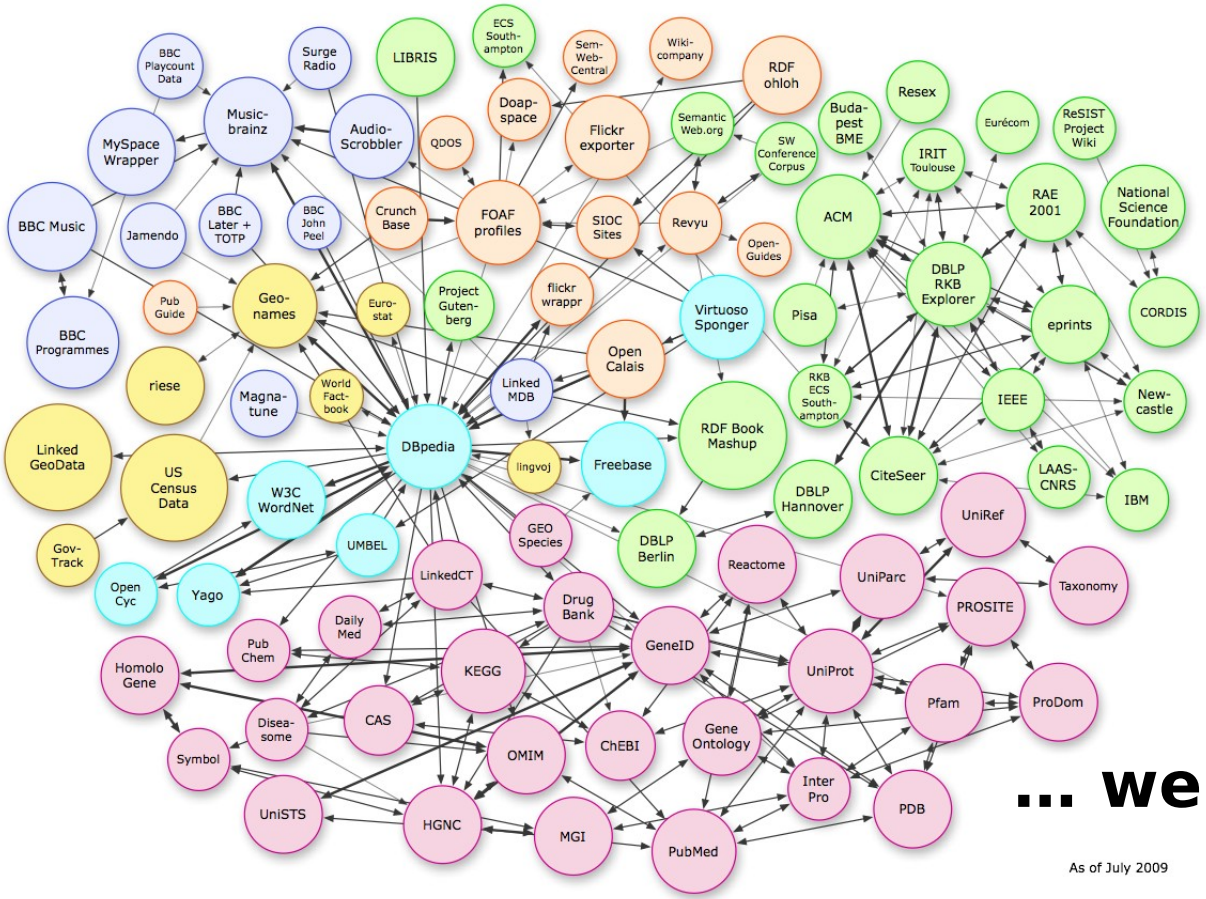


Towards Dataset Dynamics:
Change Frequency of Linked Open Data Sources

J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, S. Decker

Motivation (Situation)



courtesy of Richard Cyganiak and Anja Jentzsch

In late 2009: we have

over **100** open datasets

providing over **13.7 billion RDF triples**,

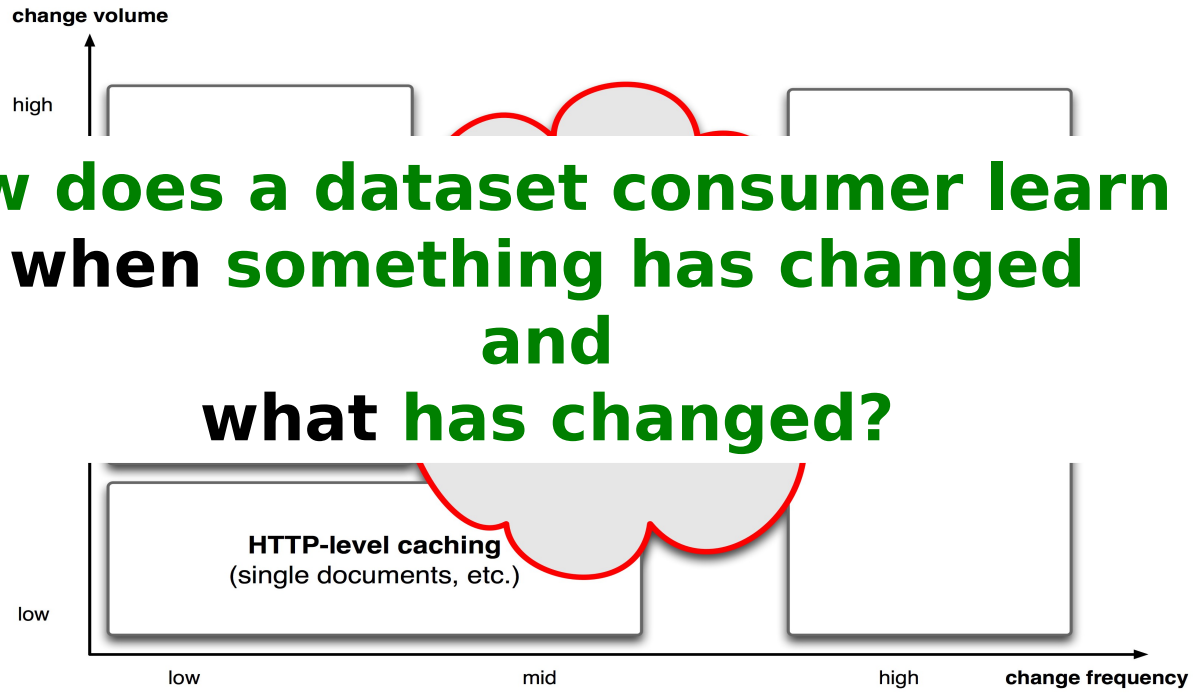
interlinked by some**142 million links …**

… we can expect/have many changes

(*new resources, added links, updates)*

Motivation (Dataset Dynamics)

- Outcomes/Findings/Lessons learned from
  - LDC09: **D**ataset **C**hange **M**anager/**W**atch-Dog demo
  - Dataset Dynamics Group



How does a dataset consumer learn when something has changed and what has changed?

Motivation (Dynamics in Web of Documents)

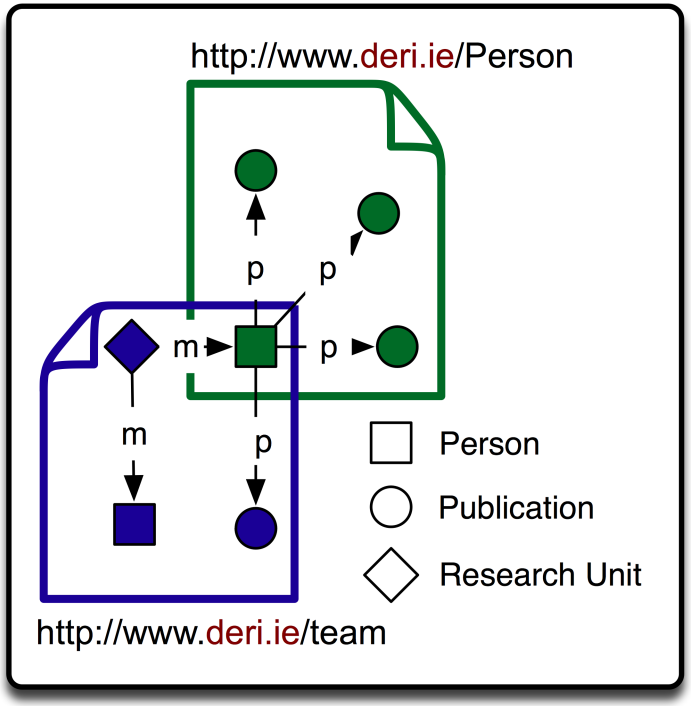
- Web of Documents
  - **Change frequency of Web documents** can be **modeled** as a **Poisson Process** [Cho at al. VLDB 2000]
  - Improvement of Estimators for uncertainty [Cho at al. ACM Journal 2001]
  - What is new on the Web? (search engine perspective) [Ntoulas et al., WWW2004]
- Applications
  - Web crawling and caching
  - Maintaining link integrity
  - replication and synchronisation

  Servicing of continuous queries

How dynamic is the Web of Data ?

Detection Level: A Matter of Granularity

- Document-centric
  - HTTP Get body content
- Entity-centric
  - Entity-per-document change
  - (Global) Entity-change
    - Reuse of entities among sources (*square*)
    - wrt. to specific documents e.g. per namespace, pay-level-dom[illegible]



Change Detection Mechanism

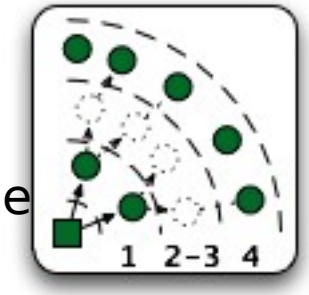
- **Content monitoring**
  fetching the entire content
- **HTTP caching** (response header as of RFC 2616)
- **Notification**
  active notificatio

	Content	HTTP	Notification
availability	+	+/-	+/-
reliability	+	+/-	unknown
costs	high	low	unknown
scalability	high	high	unknown
documents	**yes**	**yes**	**yes**
entities	**no**	**partially**	**yes**

Preliminary Results & Lessons learned

Experimental Setup

- Dataset – Web data - (passive monitoring)
  - weekly snapshots over 24 weeks (start Nov. 08)
  - 4 hop neighborhood from Tim Berners-Lee FOAF file
  - 550K RDF/XML docs, 3.3M unique entities



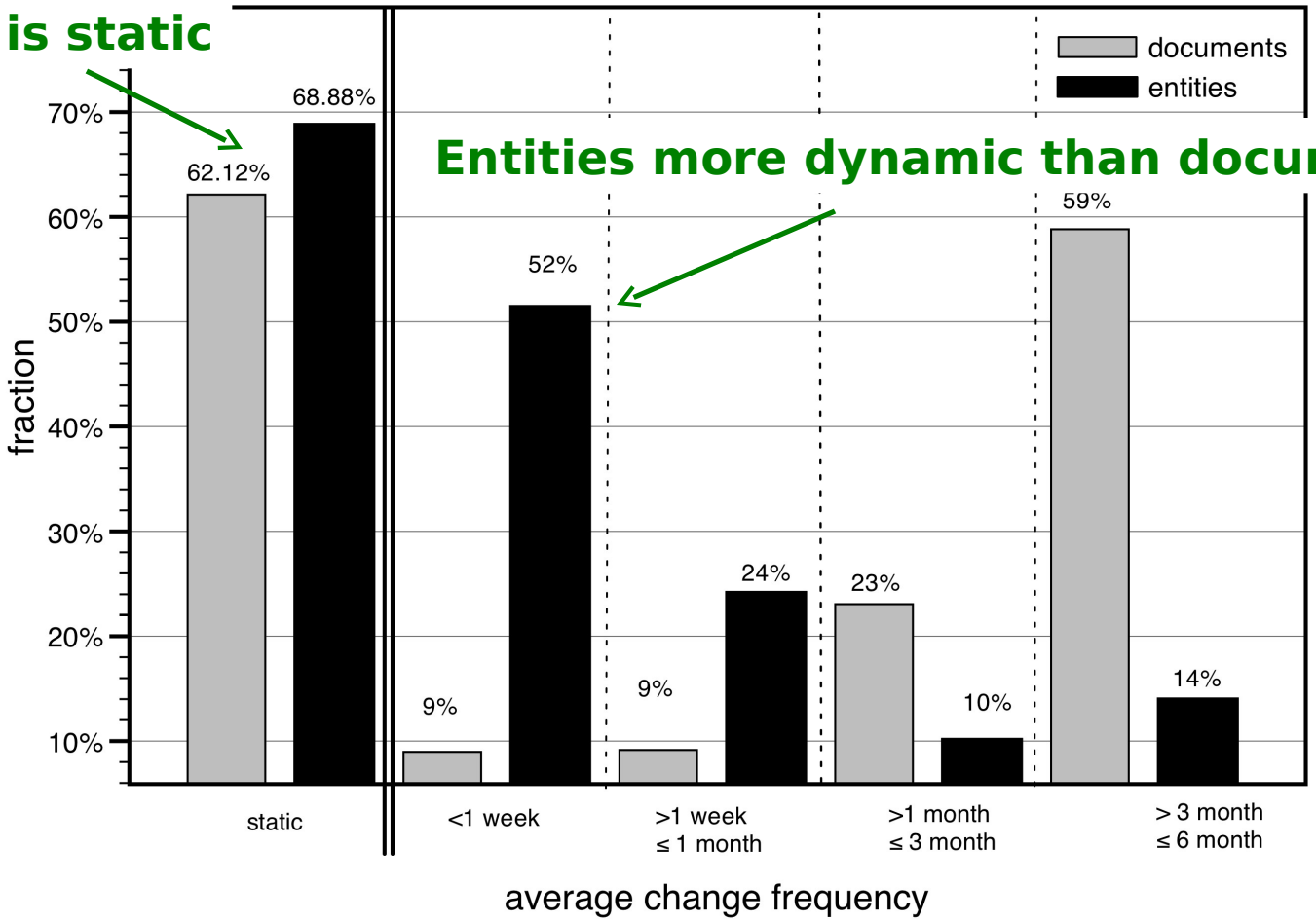
Only ETag	*(7.12%)*
Only Last-Modified	*(8.18%)*
Both	*(16.75%)*
None	*(67.95%)*

- Change detection
  - skolemise blank nodes within a d
  - pairwise comparison of statements by scanning sorted list
- Detection level
  - Document-centric
  - Entity-change (**pay-level-domain**)

Average Change Frequency

>60 % is static

Entities more dynamic than documents

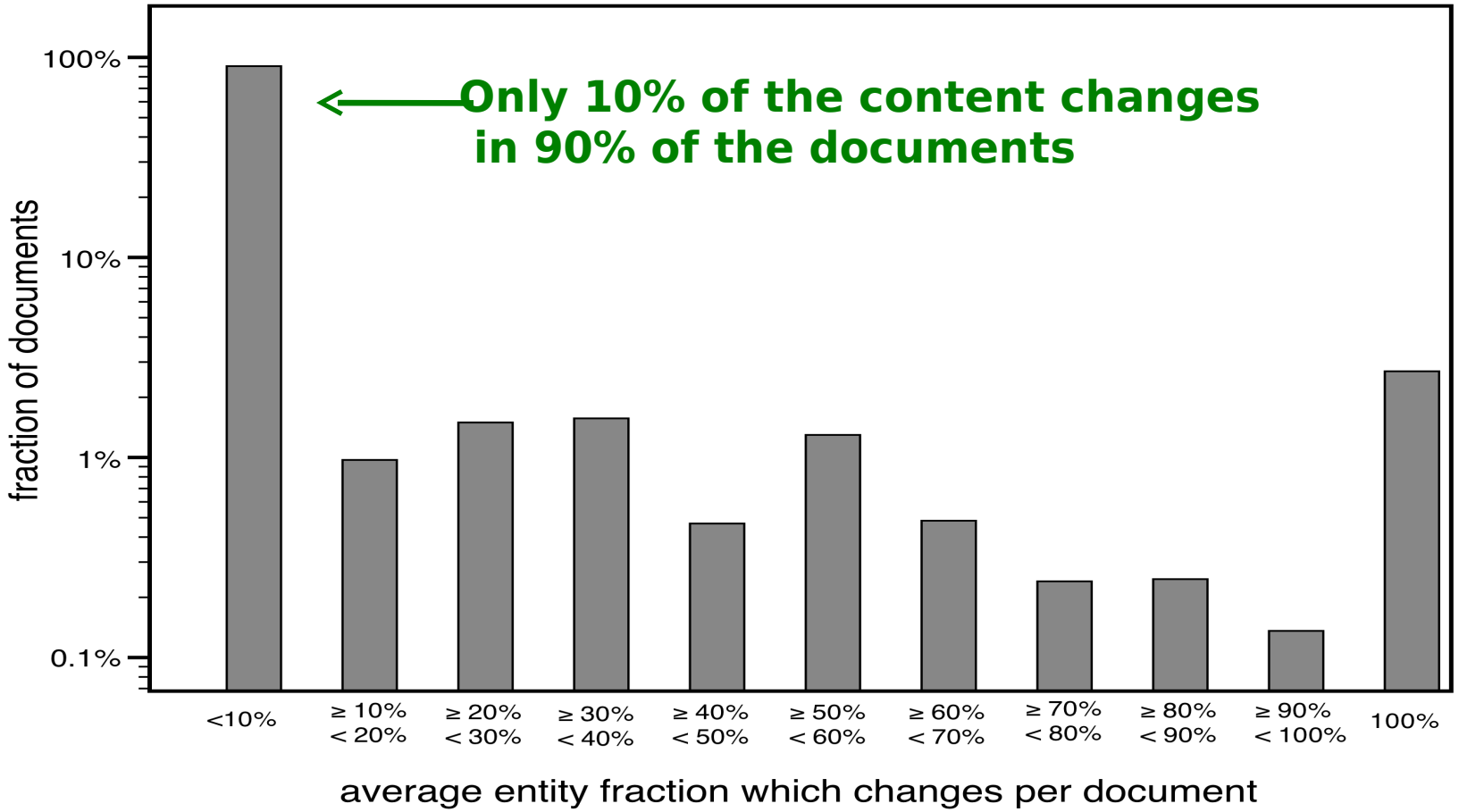


Entity Changes per Document

	U	A	D	(UA | UD | AD)	Total
U	76.88 %	9.46 %	7.08 %	3.87 %	97.29 %
A	9.46 %	0.19 %	2.29 %	3.87 %	15.81 %
D	7.08 %	2.29 %	0.23 %	3.87 %	13.5 %

Legend:
U = Update
A = Add
D = Delete
of an entity

Fraction of Changes per Document



Conclusion

- Web of Data has similar characteristics compared to the Web of Document concerning change frequencies
  - **Change frequency of Web documents** can be **modeled** as a **Poisson Process** [Cho at al. VLDB 2000]
- Document-centric change detection not helpful
  - different focus (entity-centric)
- Still needs more experimental verification

Future Work

- Large scale experiment over a long period
- Study of what and how much changes

- Dataset Dynamics Group:
  http://groups.google.com/group/dataset-dynamics

- Meet-up during WWW2010 on 29 April:
  http://esw.w3.org/Camps:LODCampW3CTrack#breakout

Questions ?