5th Linked Data on the Web Workshop (LDOW2012)

April 16th, 2012, Lyon, France

Panel Discussion

Microdata, RDFa, Web APIs, Linked Data: Competing or Complementary?

Ivan Herman (W3C)

Peter Mika (Yahoo!)

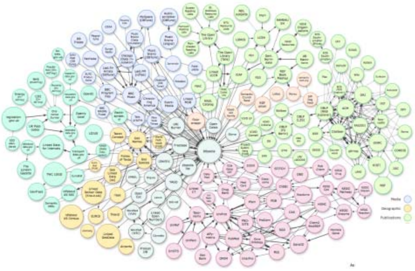
Tim Berners-Lee (MIT/W3C)

Yves Raimond (BBC)

The Web of Data is growing ...

Microformats

RDFa



Linked Data

Microdata

Web APIs

Microformat, Microdata, RDFa Deployment

13% of all HTML pages contain structured data

Source: WebDataCommons.org, Febuary 2012

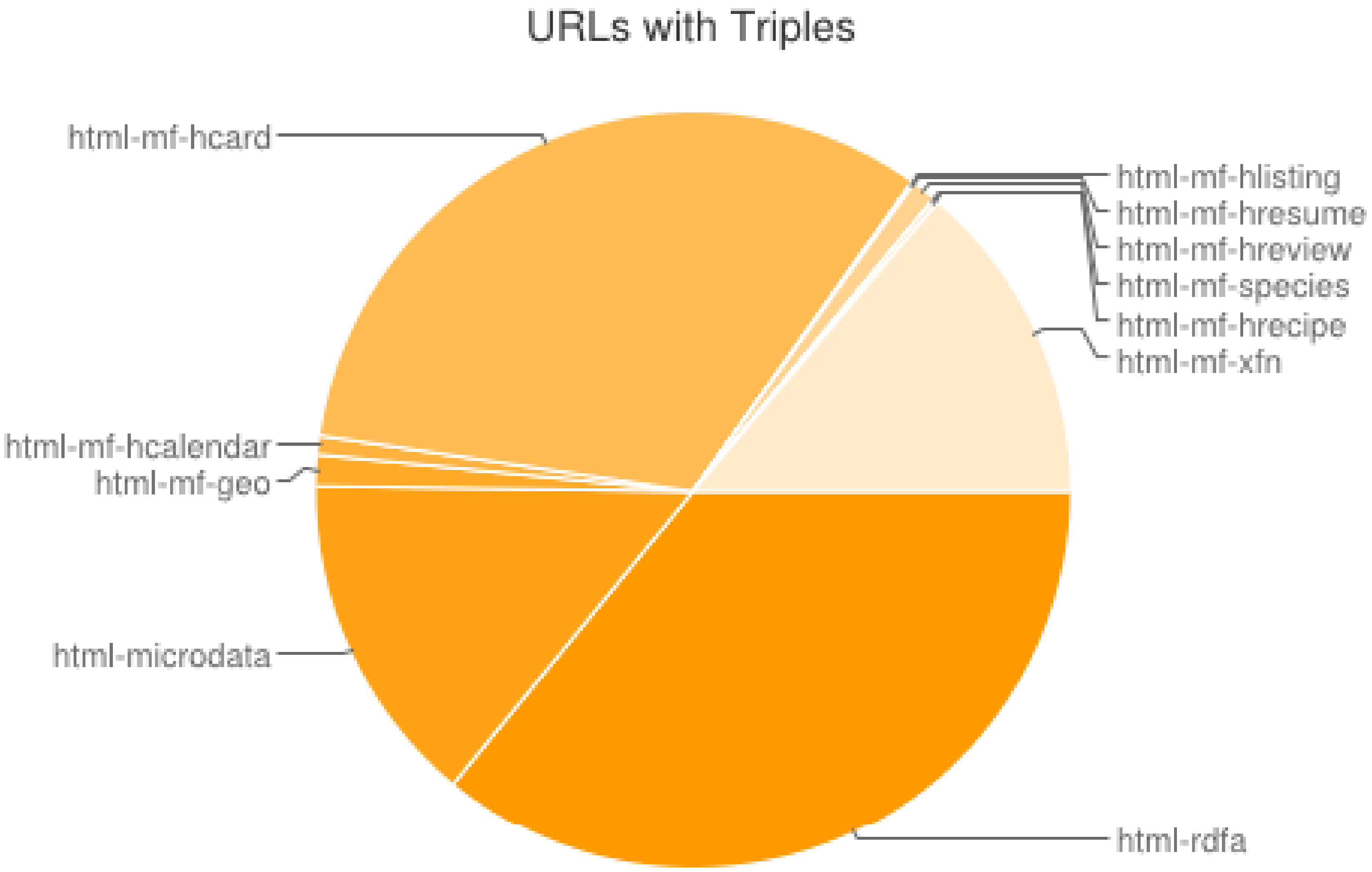
- 1.4 billion HTML pages parsed (Common Crawl corpus)
- 188 million pages contained Microformat, Microdata, RDFa data

30% of all HTML pages contain structured data

Source: Yahoo! Research, January 2012

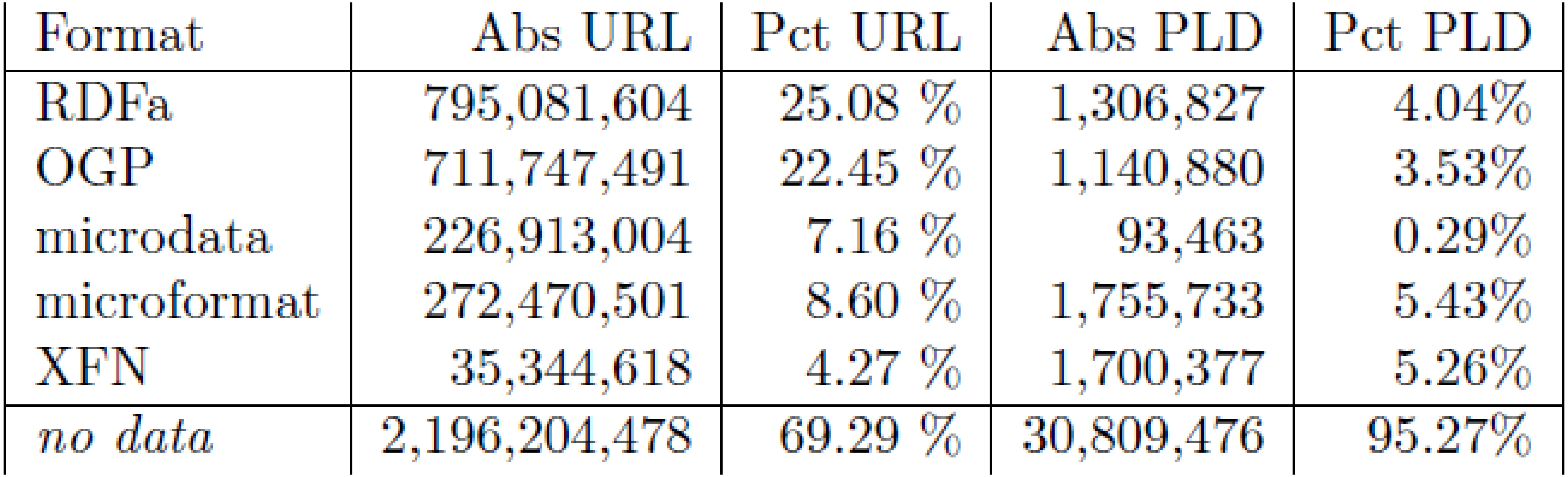
- 3.2 billion HTML pages parsed (Bing Crawl corpus)
- 973 million pages contained Microformat, Microdata, RDFa data

Breakdown by Format



Source: WebDataCommons.org, Feb 2012

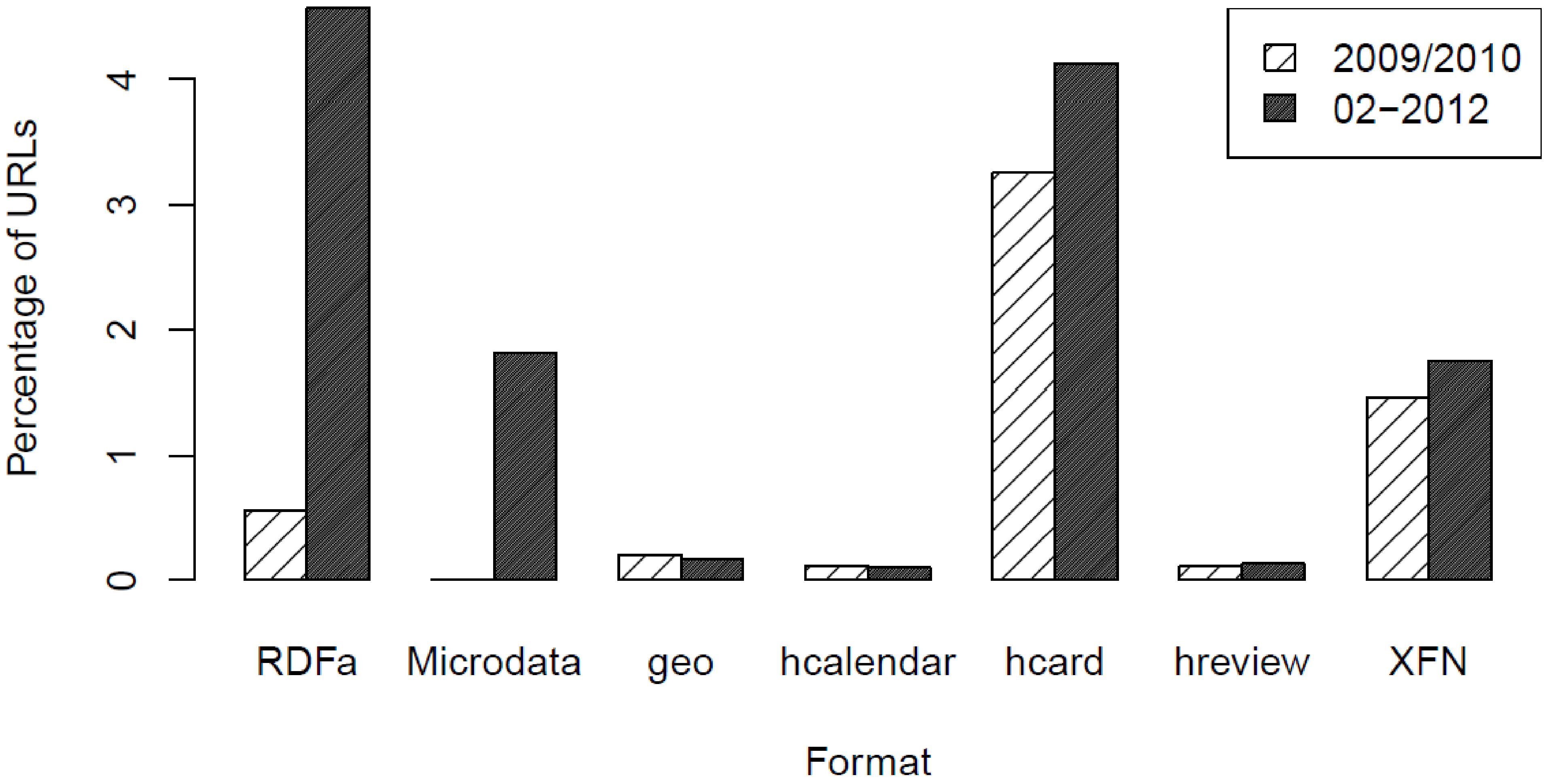
Breakdown by Format



Format	Abs URL	Pct URL	Abs PLD	Pct PLD
RDFa	795,081,604	25.08 %	1,306,827	4.04%
OGP	711,747,491	22.45 %	1,140,880	3.53%
microdata	226,913,004	7.16 %	93,463	0.29%
microformat	272,470,501	8.60 %	1,755,733	5.43%
XFN	35,344,618	4.27 %	1,700,377	5.26%
no data	2,196,204,478	69.29 %	30,809,476	95.27%

Source: Yahoo! Research, Jan 2012

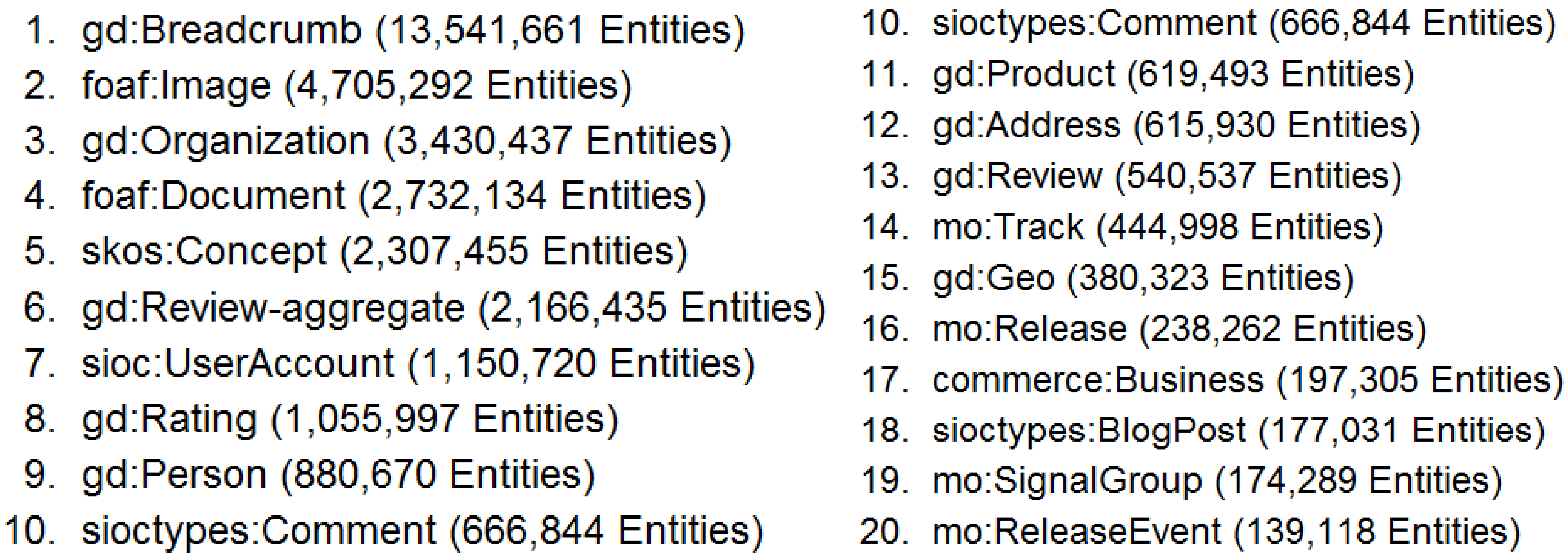
Growth between 2010 and 2012



Source: WebDataCommons.org, Feb 2012

RDFa Topics (2012)

- **Sample size: 49,370,729 instances RDFa from Common Crawl**
- **Top Classes**



1. gd:Breadcrumb (13,541,661 Entities)
2. foaf:Image (4,705,292 Entities)
3. gd:Organization (3,430,437 Entities)
4. foaf:Document (2,732,134 Entities)
5. skos:Concept (2,307,455 Entities)
6. gd:Review-aggregate (2,166,435 Entities)
7. sioc:UserAccount (1,150,720 Entities)
8. gd:Rating (1,055,997 Entities)
9. gd:Person (880,670 Entities)
10. sioctypes:Comment (666,844 Entities)

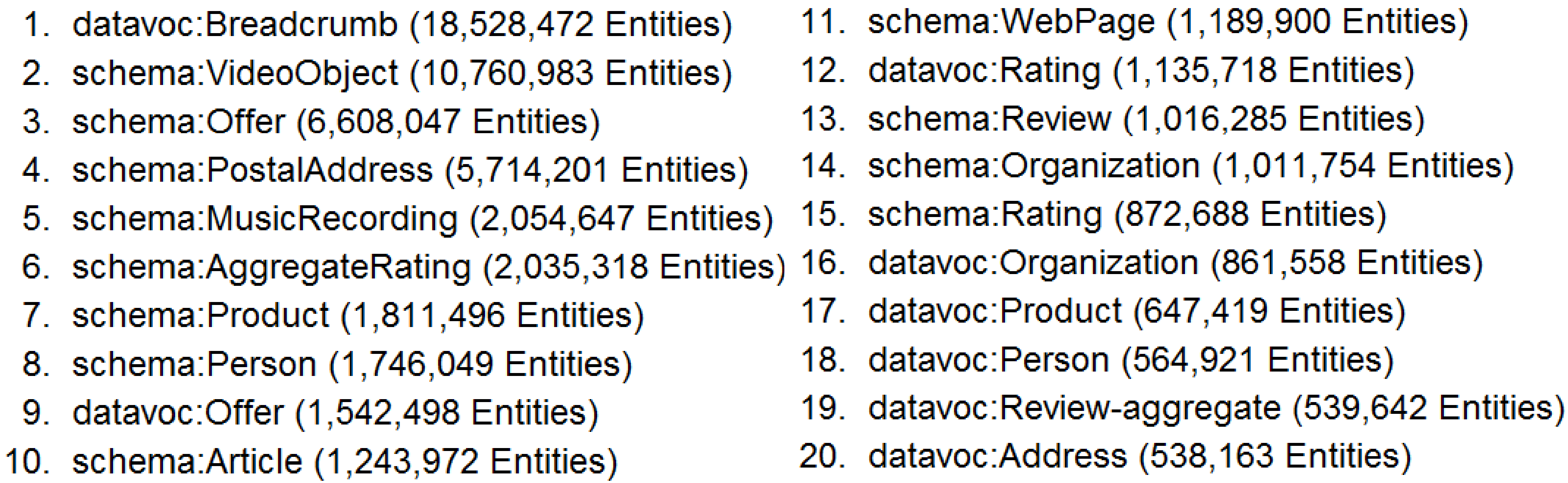
10. sioctypes:Comment (666,844 Entities)
11. gd:Product (619,493 Entities)
12. gd:Address (615,930 Entities)
13. gd:Review (540,537 Entities)
14. mo:Track (444,998 Entities)
15. gd:Geo (380,323 Entities)
16. mo:Release (238,262 Entities)
17. commerce:Business (197,305 Entities)
18. sioctypes:BlogPost (177,031 Entities)
19. mo:SignalGroup (174,289 Entities)
20. mo:ReleaseEvent (139,118 Entities)

gd = Google's Rich Snippet Vocabulary

Source: WebDataCommons.org, Feb 2012

Microdata Topics (2012)

- **Sample size: 90,526,013 Entities from the Common Crawl**
- **Top Classes**



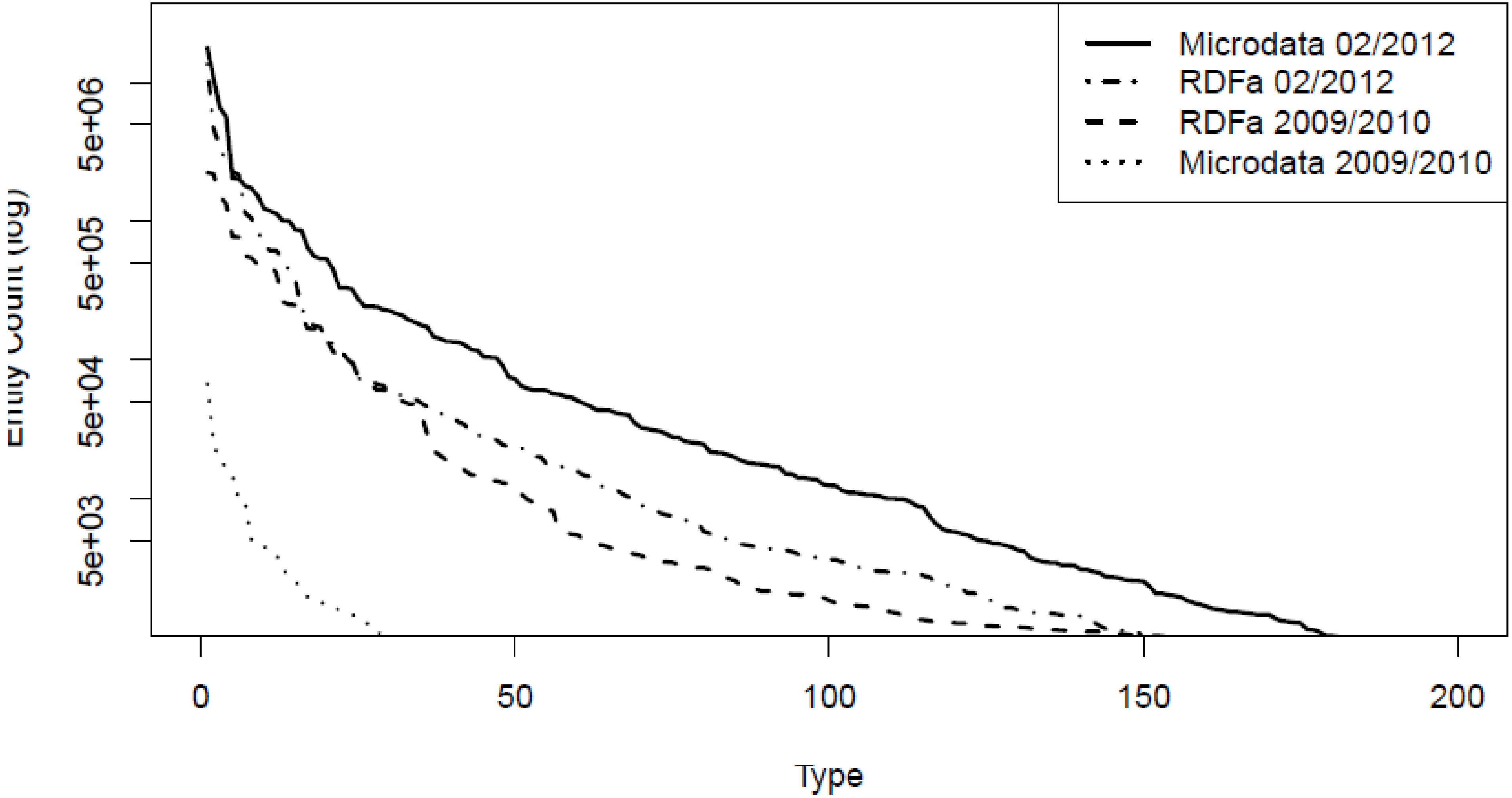
1. datavoc:Breadcrumb (18,528,472 Entities)
2. schema:VideoObject (10,760,983 Entities)
3. schema:Offer (6,608,047 Entities)
4. schema:PostalAddress (5,714,201 Entities)
5. schema:MusicRecording (2,054,647 Entities)
6. schema:AggregateRating (2,035,318 Entities)
7. schema:Product (1,811,496 Entities)
8. schema:Person (1,746,049 Entities)
9. datavoc:Offer (1,542,498 Entities)
10. schema:Article (1,243,972 Entities)
11. schema:WebPage (1,189,900 Entities)
12. datavoc:Rating (1,135,718 Entities)
13. schema:Review (1,016,285 Entities)
14. schema:Organization (1,011,754 Entities)
15. schema:Rating (872,688 Entities)
16. datavoc:Organization (861,558 Entities)
17. datavoc:Product (647,419 Entities)
18. datavoc:Person (564,921 Entities)
19. datavoc:Review-aggregate (539,642 Entities)
20. datavoc:Address (538,163 Entities)

datavoc = Google‘s Rich Snippet Vocabulary
schema = Schema.org

Source: WebDataCommons.org, Feb 2012

Very short tail

- RDFa: 150 classes und 400 properties with 1000+ instances
- Microdata: 182 classes and 690 properties with 1000+ instances

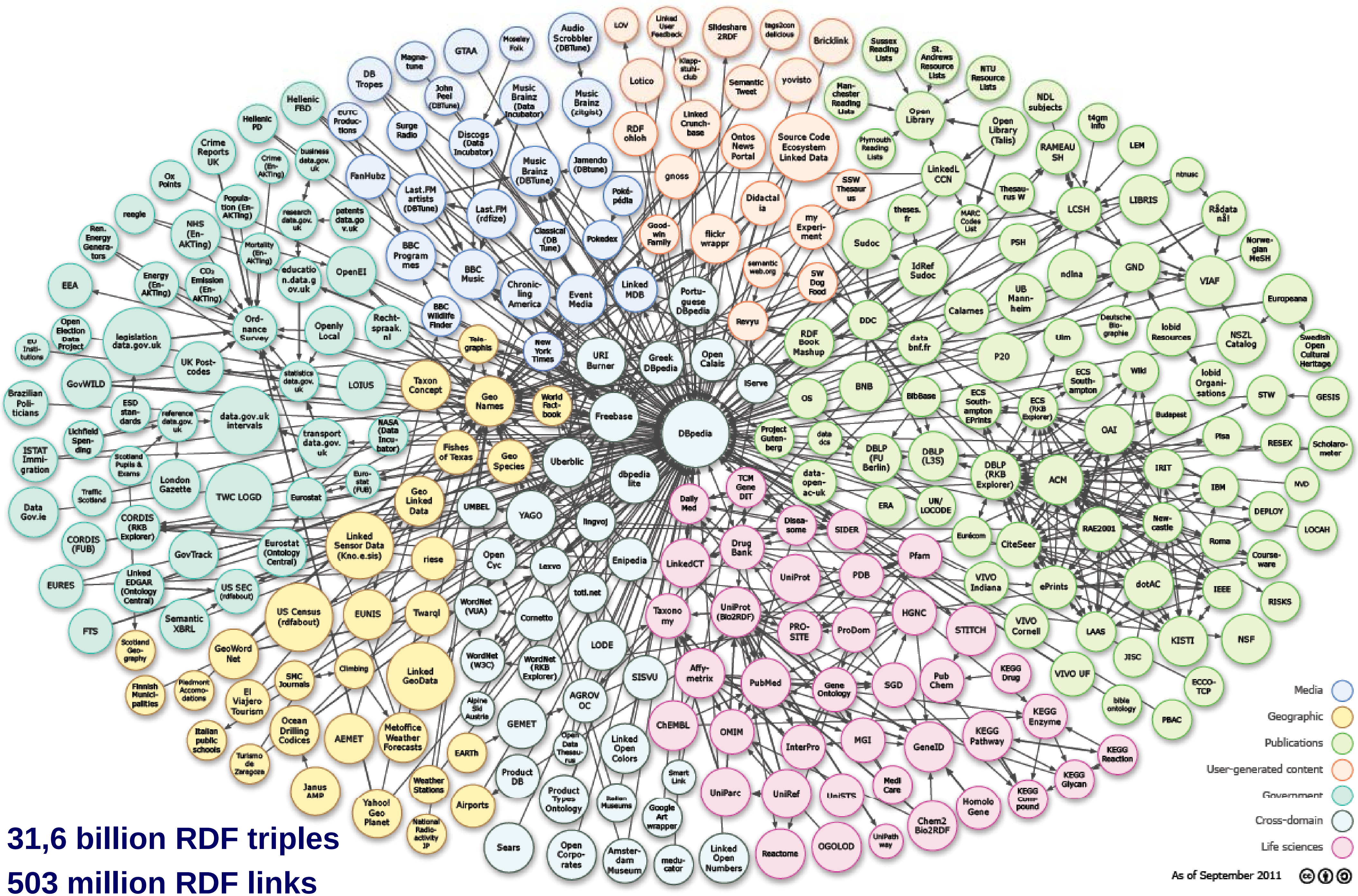


Source: WebDataCommons.org, Feb 2012

Summary: Embedded Data in HTML

- RDFa and Microdata grow, but Microformats are still present
- A rather small set of vocabularies is used
- The content and the vocabularies are very focused towards the mayor consumers (Google, Yahoo, Bing, Facebook)
- Providing structured data has come SEO topic
- The data structures used are rather simplistic (mostly atomar entities, no links between entities)

Linked Data Deployment



- 31,6 billion RDF triples
- 503 million RDF links

The Linked Data Web is heterogeneous

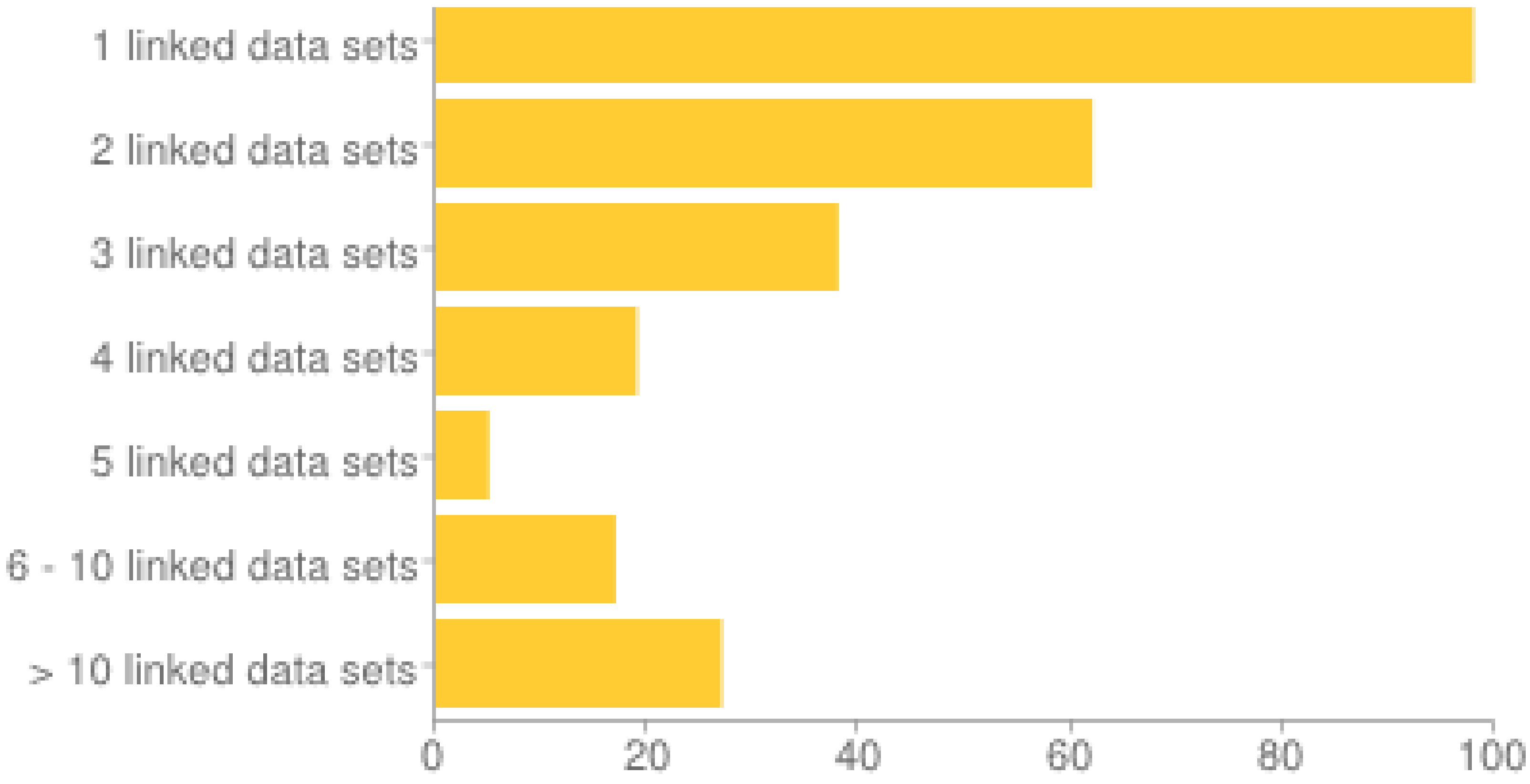
- Use only proprietary vocabularies:
104 (35.25 %) of the 295 sources
- Use some terms from non-proprietary vocabularies:
191 (64.75 %) of the 295 sources
- Common Vocabularies

dc	92 (31.19 %)
foaf	81 (27.46 %)
skos	58 (19.66 %)
geo	25 (8.47 %)
akt	17 (5.76 %)
bibo	14 (4.75 %)
mo	13 (4.41 %)
vcard	10 (3.39 %)
sioc	10 (3.39 %)
cc	8 (2.71 %)

Source: State of the LOD Cloud
http://www4.wiwiss.fu-berlin.de/lodcloud/state/

Sparse Linkage

- RDF links are integration hints for the data consumers
- but setting RDF links is effort for the providers

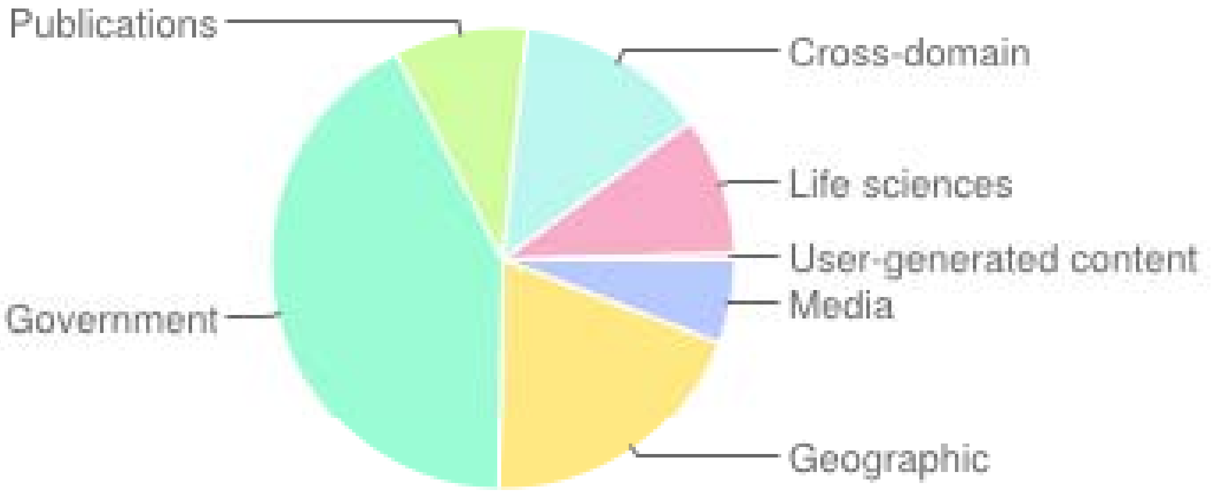


Source: State of the LOD Cloud, Nov 2011

Summary: Web of Linked Data

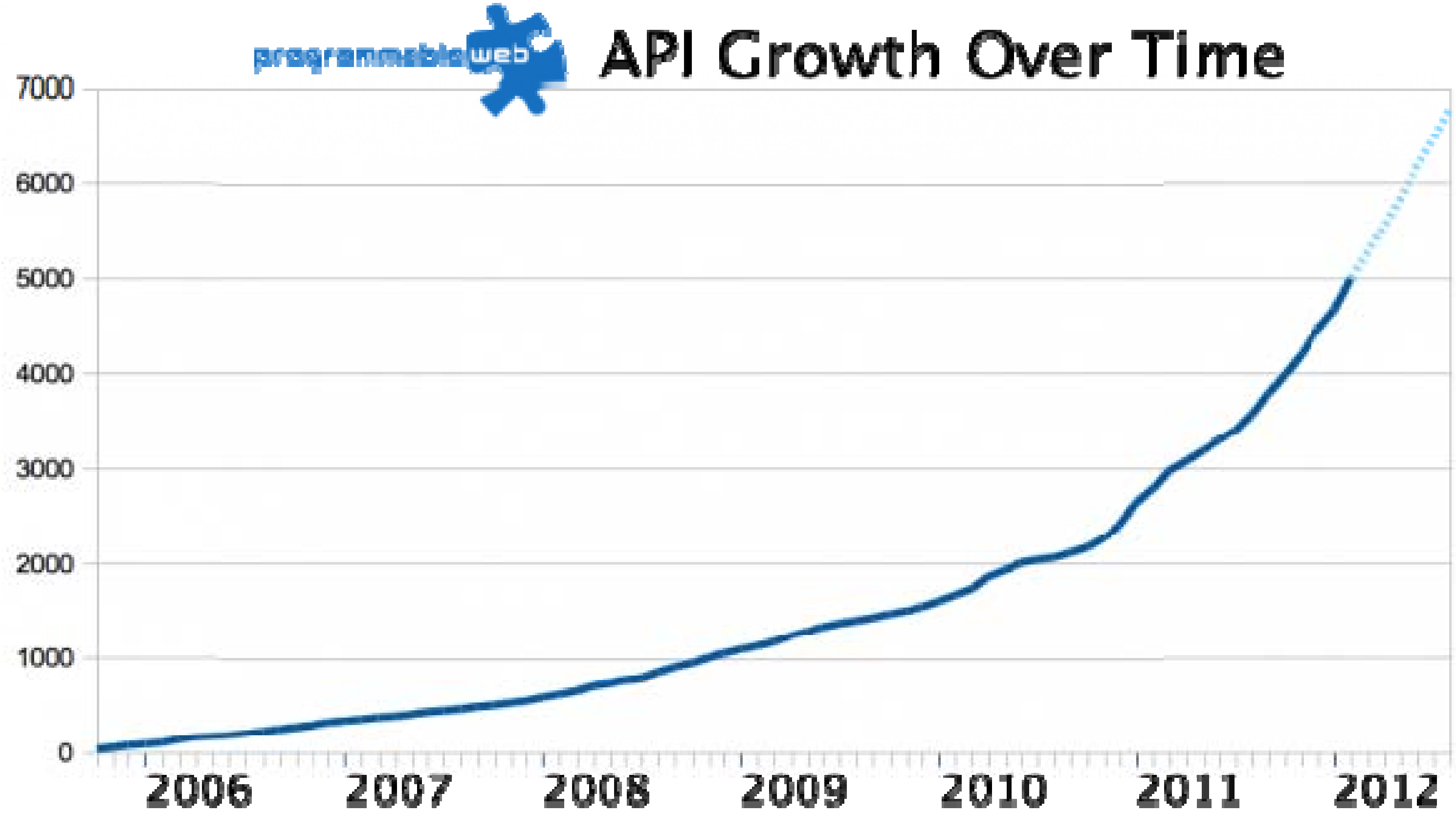
Compared to Microformats, Microdata, RDFa

- the number of data providers is significantly lower
- a wider range of specialized topics is covered
- a wider range of common and proprietary vocabularies is used
- emphasis on setting RDF Links between sources



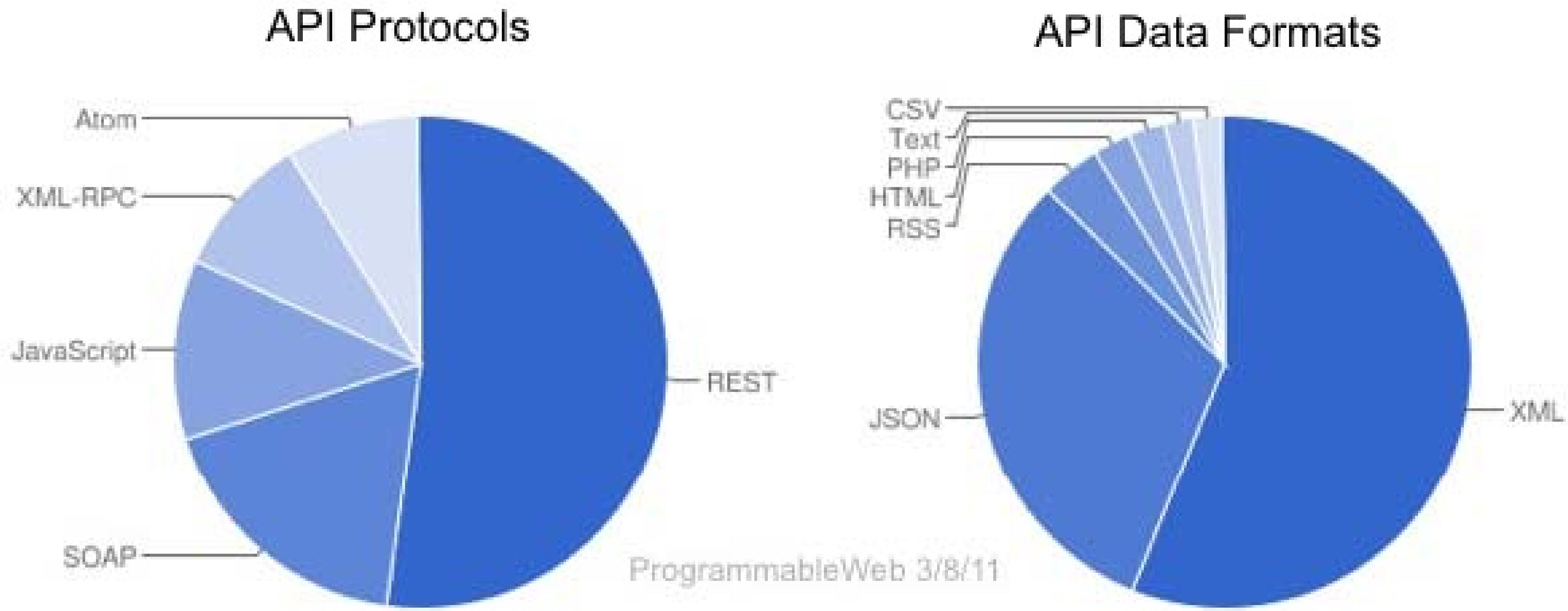
Web API Deployment

- 5600 APIs, 6500 Mashups



- all big Web companies provide APIs
- lots of new APIs published by government bodies in 2011

Web APIs



- **Many different legal terms of use.**

Questions to the Panelists:

1. What are the most important deployment scenarios for each technology and how do the technologies fit these scenarios?

2. How will the deployment of the technologies develop over the next three years?

3. Will the different technologies compete for deployment (in specific scenarios?) or are they complementary?