“Ten years back, ten years forward: achievements lessons and the future of digital curation”

IDCC 2015, London, UK

“Designated Communities”: through the lens of the Web

Yunhyong Kim
HATII, University of Glasgow
Glasgow, UK

yunhyong.kim@glasgow.ac.uk

The beginning

'“Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop.”' - Lewis Carroll

from Alice in Wonderland

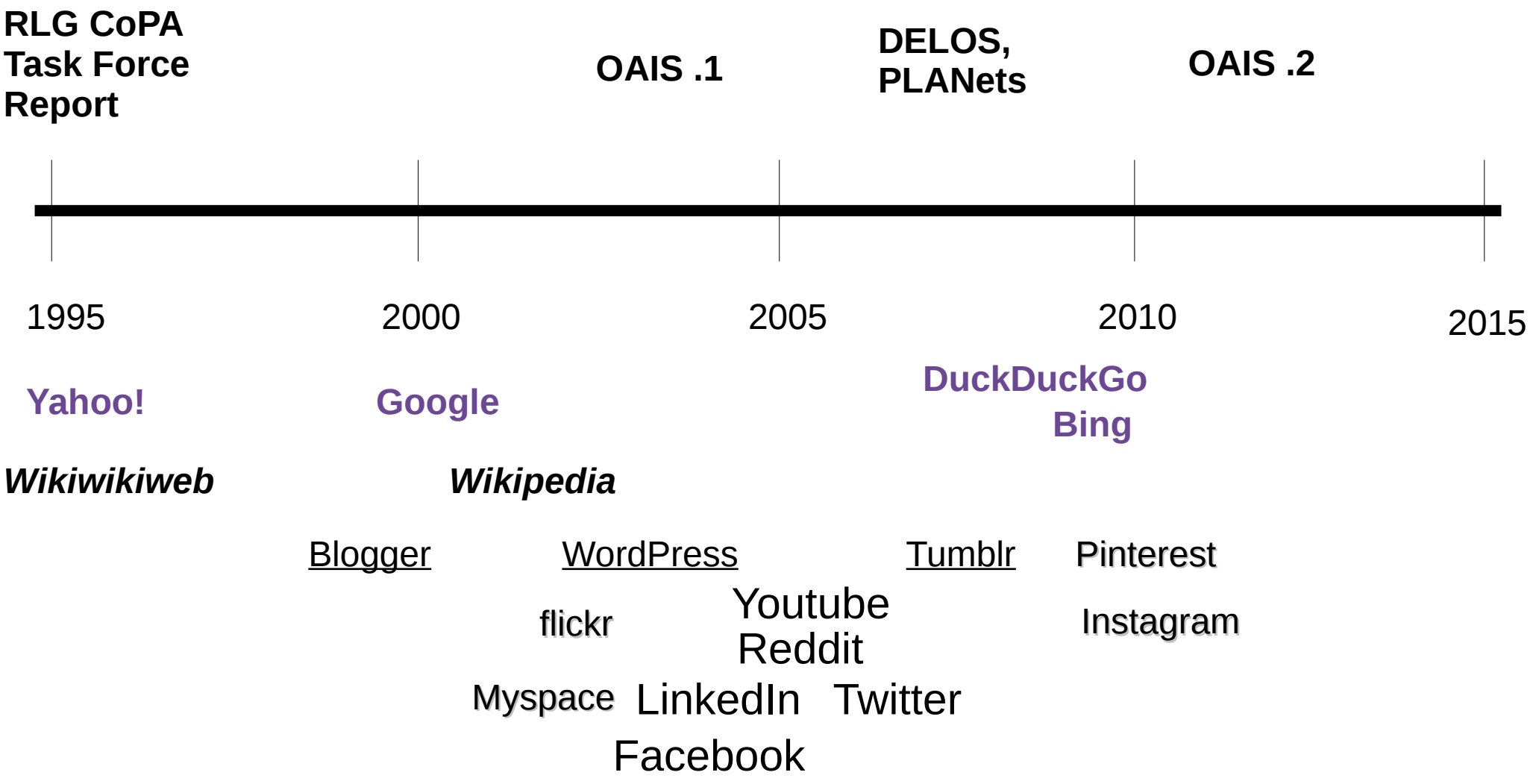
Conclusion

- Data analytics can be useful for profiling a “designated community”
- Social media infrastructure for sharing information has potential to support profiling a target “designated community”.
- Some materials come with web and social media platform infrastructure in place to make community profiling viable.
- Other materials can be provided with such infrastructure by
  - embedding the repository within the social web-like information sharing workflow
  - developing a agreed approach for mapping the “designated community” to an online presence

Benefits for Digital Curation

- Adding community context to **characterise** the target materials to be collected; using an automated approach to doing so
- Providing first steps to **assess the community risks** associated with
  - the variety of adopted technologies and formats
  - changes in concepts with respect to knowledge organisation
  - social impact of information loss given the knowledge exchange network
- studying the community in action as information sharing takes place, to **bridge the gap between archive standards and community needs**

Ten years back: OAIS



“Community” in flux

- Community of Practice: a brief introduction (Wenger 2011)
  - “People who engage in a process of collective learning in a shared domain of human endeavour”
    - Tribe learning to survive.
    - Band of artists seeking new forms of expression.
    - Group of engineers working on similar problems.
    - Clique of pupils defining their identity in the school.
    - Network of surgeons exploring novel techniques
    - Gathering of first-time managers helping each other cope.
  - “Communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly.”

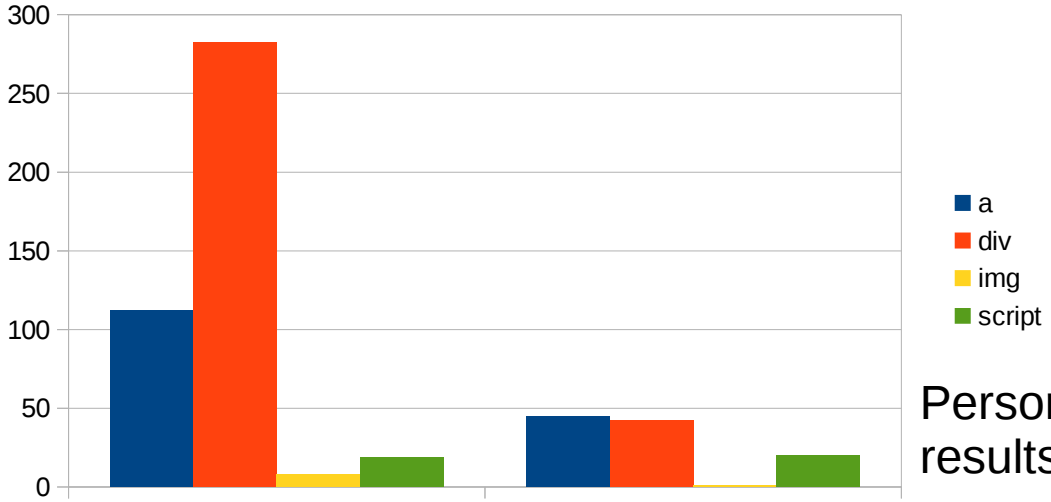
Social media

- Exposes changes in flux at a much finer level of granularity than previous channels of communication
- Allows flexibility to present content as you see fit without mediation
- Offers potential for hierarchical social tagging and annotation
- Mirrors groups offline and vice versa – cf. Bruce Schneier on Security
- Now drives how we interact with the rest of the Web

Fun Interlude

Name	PDF	DOC	DOCX	HTML/HTM	JPG
Person 1	0.1142	0.0025	0.00062	0.88266	0
Person 2	0.0774	0.0094	0.00170	0.91137	2.60766392427344E-005
Person 3	0.0142	0.0003	0.00011	0.98532	2.36635265629476E-005

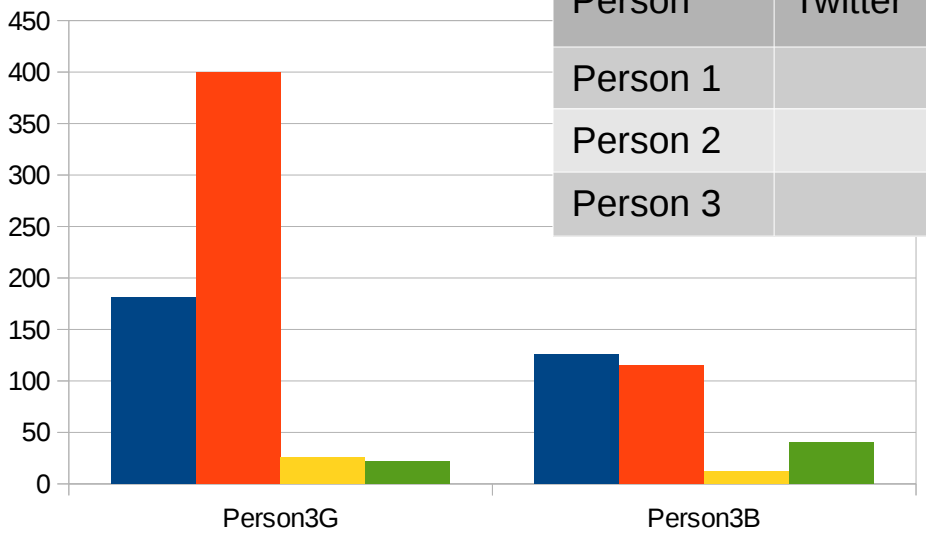
File type distribution for each person using results pages from one search engine



Person 1: First results page

Person	Twitter	Wiki	YouTube	LinkedIn	Facebook	flickr
Person 1	1.3	0	0.3	3	0	0.3
Person 2	1.3	3.3	0.3	1.6	0.6	0.6
Person 3	2.3	9	3.3	0	1.6	0.3

Social media weight for each person using results pages from three search engines



Person 3: First results page

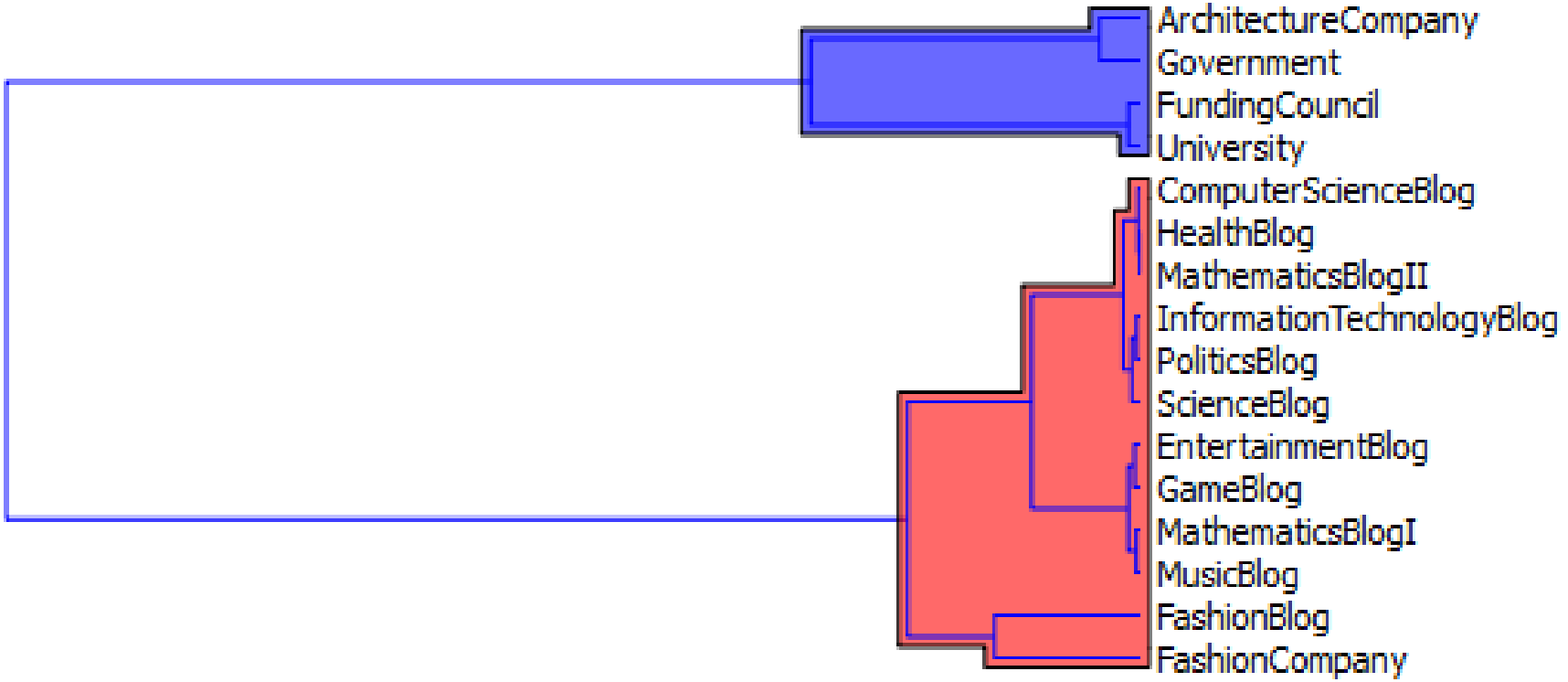
Datasets

Dataset	# of URLs	Number of unique “<!DOCTYPE>” declarations
Spinn3r	223,145	80
ClueWeb09	214,952	1420
BF16Cat	31,690	122

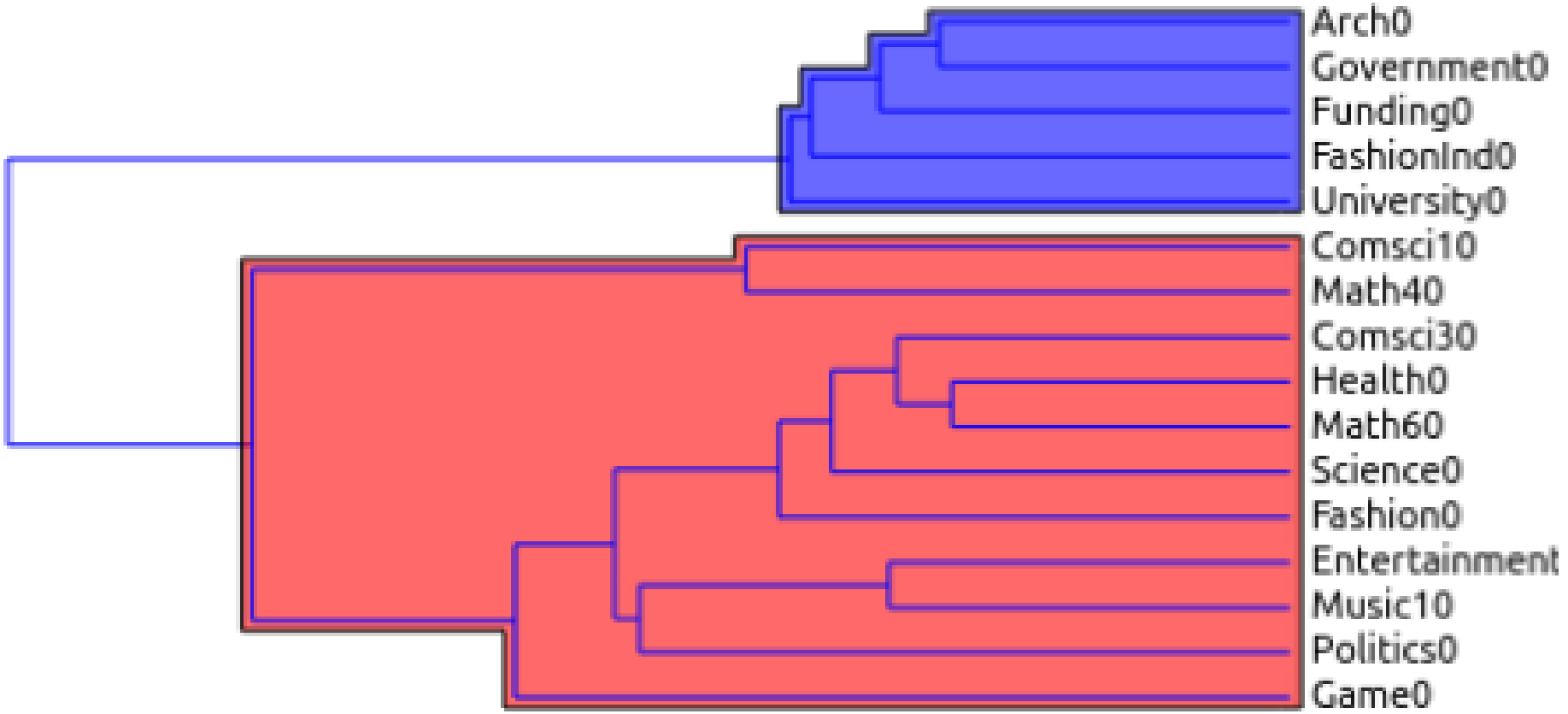
BF16Cat

Type	Subcategory	Size	Source
Blogs	**Computer Science (CS)**	41	StackOverflow
	Information Technology (IT)	138	Technorati “IT” category search
	Entertainment (ET)	110	Technorati “Entertainment” category search
	Fashion (FA)	164	Independent Fashion Bloggers
	Game (GA)	7	University of Glasgow PhD student in Games
	Health Blogs (HB)	130	Technorati “Health” category Search
	Mathematics I (M1)	110	Field's Medalist Terry Tao's Blog
	Mathematics II (M2)	552	Mathblogging.org
	Music (MU)	70	Technorati “Music” category search
	Politics (PO)	107	Technorati “Politics” category search
	Science (SC)	1071	Scienceseeker.org amd Scienceblogging.org
Non-Blogs	**Construction Company (CC)**	27	National Building Specification Website (http://www.thenbs.com/resources/directory/index.asp)
	Fashion Company (FC)	61	http://www.smashingmagazine.com/2009/03/12/showcase-of-beautiful-fashion-websites/ and comments
	Funding Council (FU)	51	Search on google
	Government (GO)	572	http://www.politicsresources.net/official.htm
	University (UN)	100	http://www.guardian.co.uk/news/datablog/2012/mar/15/top-100-universities-times-higher-education

Platforms


ArchitectureCompany
Government
FundingCouncil
University
ComputerScienceBlog
HealthBlog
MathematicsBlogII
InformationTechnologyBlog
PoliticsBlog
ScienceBlog
EntertainmentBlog
GameBlog
MathematicsBlogI
MusicBlog
FashionBlog
FashionCompany

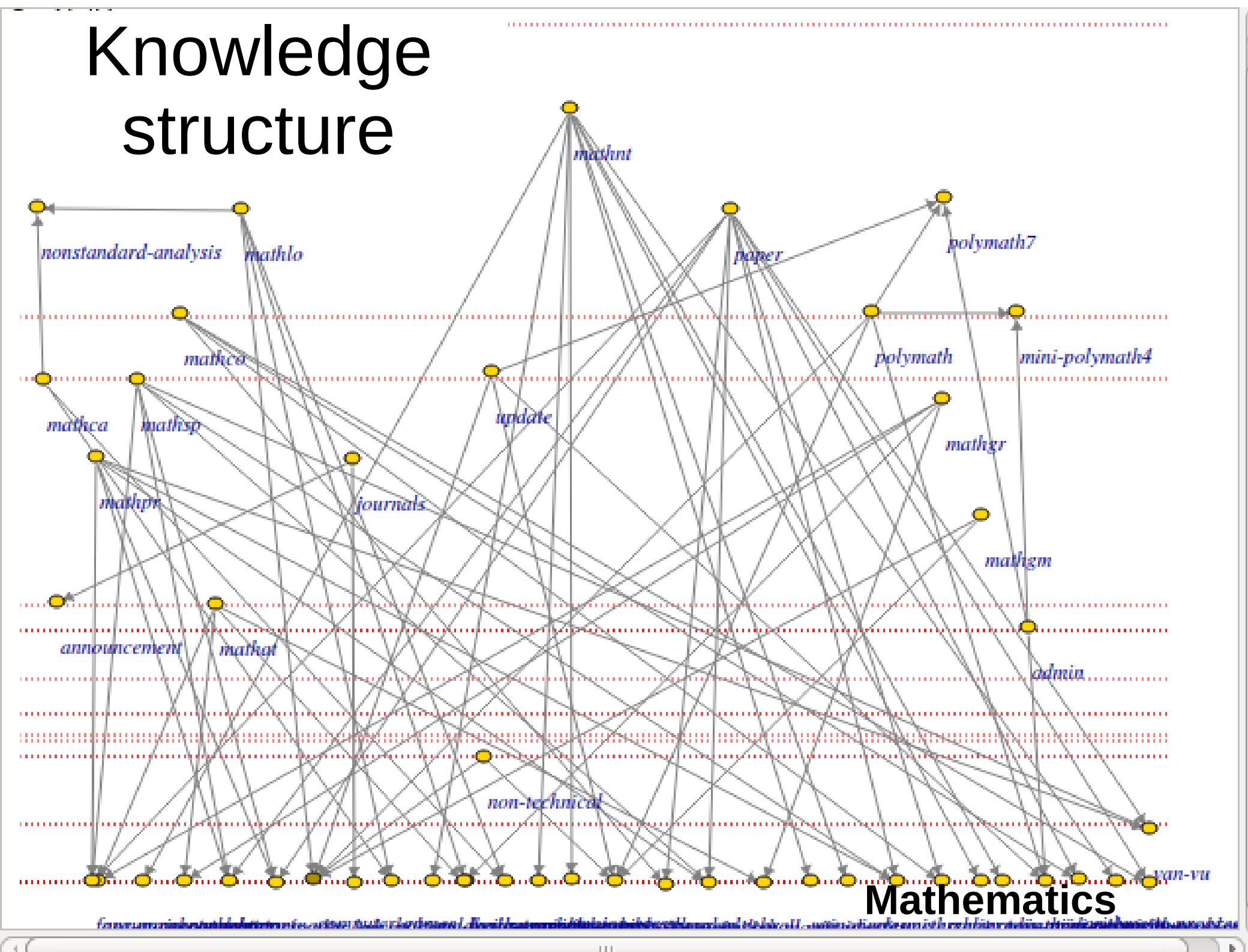
What's in a format?



Link Characteristics

Category	Avg. no. Links per Page (No. All Links)	Distinct Links	No. Non-Self Referential	Self-Referential (%)
Construction Company	41.82 (1129)	843	125	0.8892825509
Computer Science	226.76 (9297)	6823	4773	0.4866085834
Information Technology	219.57 (30300)	21493	10313	0.6596369637
Entertainment	261.21 (28733)	19357	9960	0.6533602478
Fashion Blog	225.24 (36940)	28660	23513	0.3634813211
Fashion Company	105.57 (6440)	4926	1154	0.8208074534
Funding Council	71.84 (3664)	2803	503	0.8627183406
Game Blog	312 (2184)	1479	714	0.6730769231
Government	75.8 (43356)	31524	8464	0.8047790387
Health Blog	224.42 (29175)	21408	14054	0.5182862039
Mathematics Blog I	195.96 (21360)	15251	8977	0.5797284644
Mathematics Blog II	214.62 (118471)	83283	44349	0.6256552236
Music Blog	223.93 (15675)	11357	8959	0.4284529506
Politics Blog	361.08 (38636)	27709	20163	0.4781292059
Science Blog	233.10 (249652)	155420	129754	0.4802605226
University	89.83 (8983)	7369	2095	0.7667816988

Knowledge structure



What next?

- We have profiled communities using:
  - Community links formed through platforms they use
  - Community links defined by formats they share
  - Community links arising from their referencing behaviour
  - Community links established through knowledge structures

We can do much more ...

Network science

- Hubs, centrality, dynamics of change, sel-organisation, heterogeneity, homogeneity.
- Models: for example, Erdös-Renyi random graphs (1951); Watts and Strogatz small world network (1998); Barabási-Albert preferential attachments (1999); Caldarelli et al. fitness network (2002).
- Motifs: repeated patterns of connections.

Ten years forward

- Move away from case studies, bench marks, and standards
- Move towards:
  - Consistency
  - Assessment models
  - End-user validation

The beginning

"In the beginning, there was nothing, which exploded." - Terry Pratchett

from Lords & Ladies