How much metadata is too much (and how little is too little)?
Practical Preservation and People
December 3rd 2015, PRONI, Titanic Quarter, Belfast, BT3 9HQ

Metadata and other stories online

or

"Is metadata a love letter to the future?"

Yunhyong Kim

Humanities Advanced Technology and Information Institute (HATII)
School of Humanities
University of Glasgow
Glasgow, UK

yunhyong.kim@glasgow.ac.uk

University of Glasgow | School of Humanities
Sgoil nan Daonnachdan

blog forever

The beginning

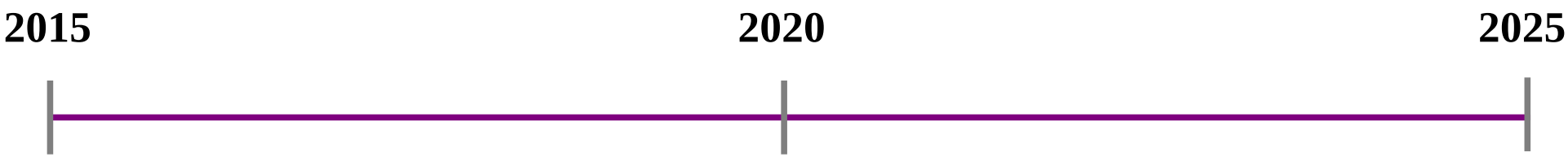
"“Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop.”" - Lewis Carroll

from Alice in Wonderland

Focus: digital preservation

http://britishlibrary.typepad.co.uk/collectioncare/2014/01/digital-preservation-training-programme-snuggling-up-with-oais.html

Ten years forward



Building Models

- Case study
- Data model
- Standards

Testing Models

- Test data
- Defined model
- Benchmarks

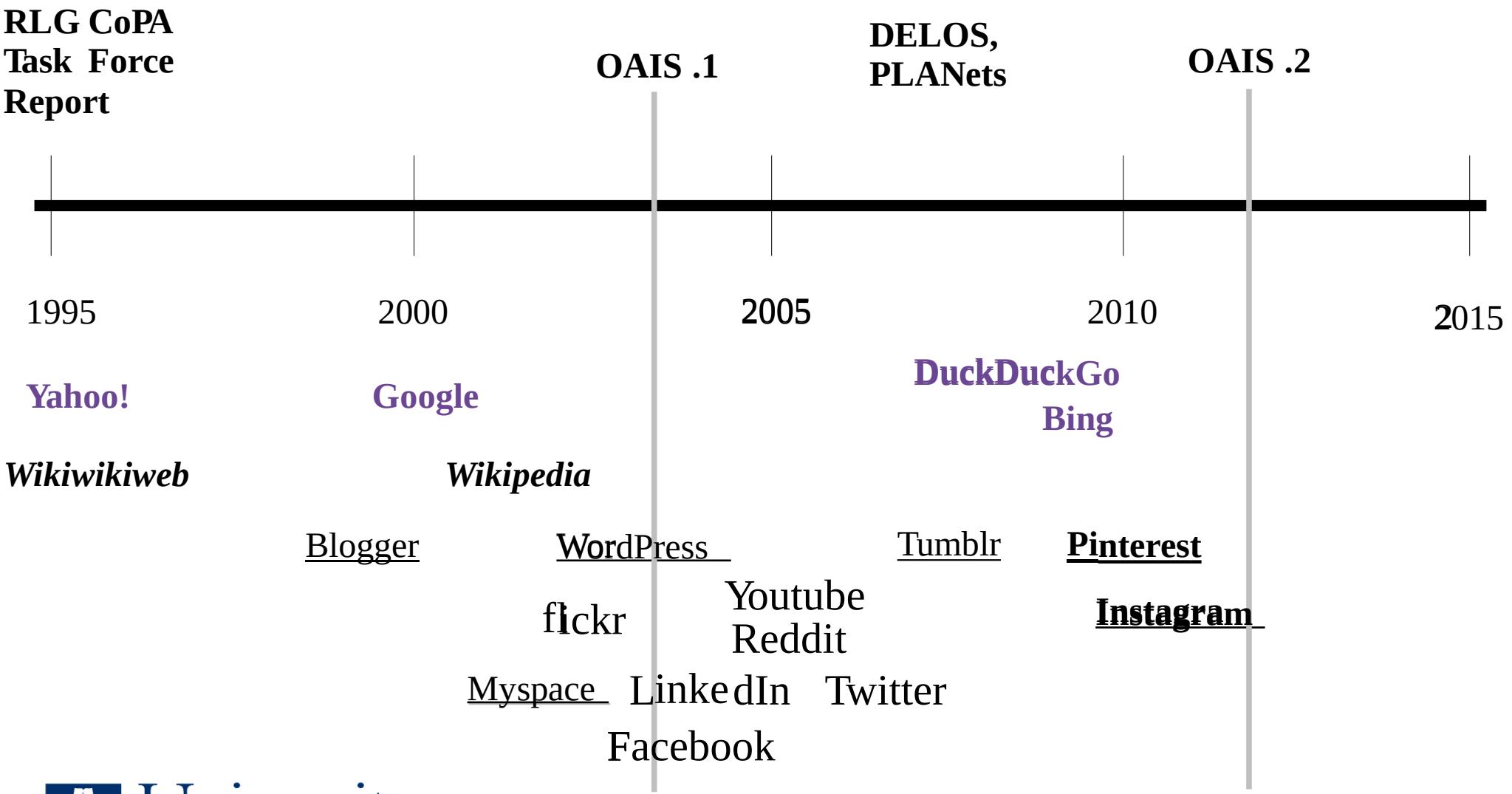
Using Models

- Learning models
- Error detection
- Uncertainty

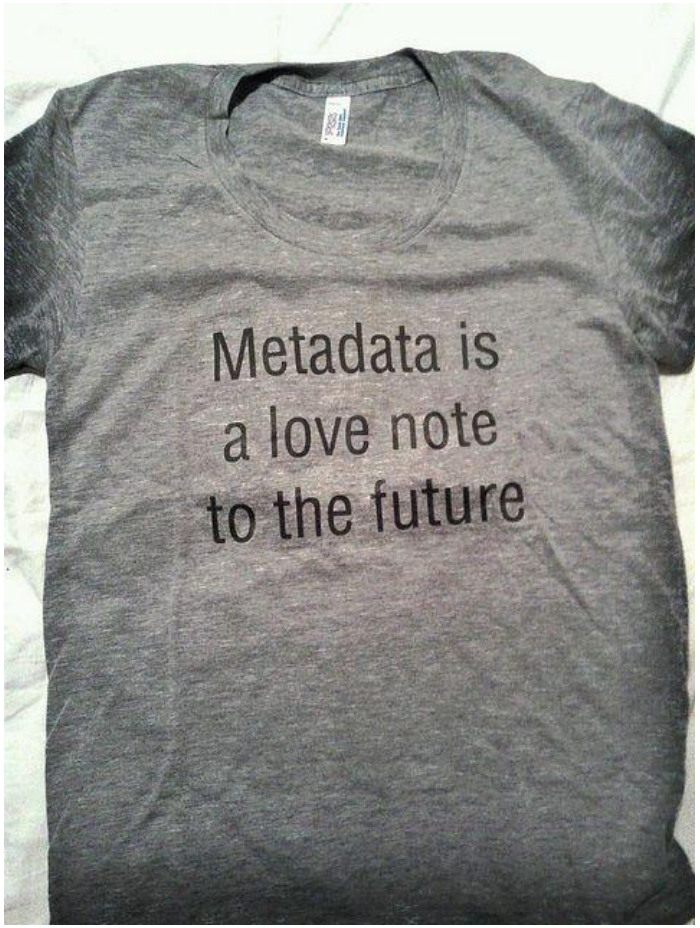
Conclusion

- Metadata standards are only **validated** when it is clear that it supports a "designated community".
- Social media infrastructure for sharing information has potential to support understanding **metadata usage in a community.**
- Data analytics to understand metadata usage can be useful for understanding **preservation complexity.**
- Materials can be **provided with such social media infrastructure** by
  - integrating the repository within such infrastructure
  - mapping the "designated community" to an online presence

Twenty years back: OAIS



Let there be metadata



https://www.pinterest.com/pin/165788830005828904/ (left)
https://www.flickr.com/photos/strongstuff/9762494642 (right)

Social media

- Exposes sharing and networking at a unprecedented **fine level of granularity.**
- Exposes **metadata usage in action**, driving how we interact with others.
- Offers **hierarchical description.**

Community driven standards

- Creation of HTML5
  - Ian Hickson mined data from over a billion web pages, surveying most commonly used attributes. http://code.google.com/webstats/2005-12/classes.html
  - Opera did a similar study, of 3.5 million URLs, calling it MAMA. Smaller set of URL, wider variety of statistics. https://dev.opera.com/articles/view/mama/

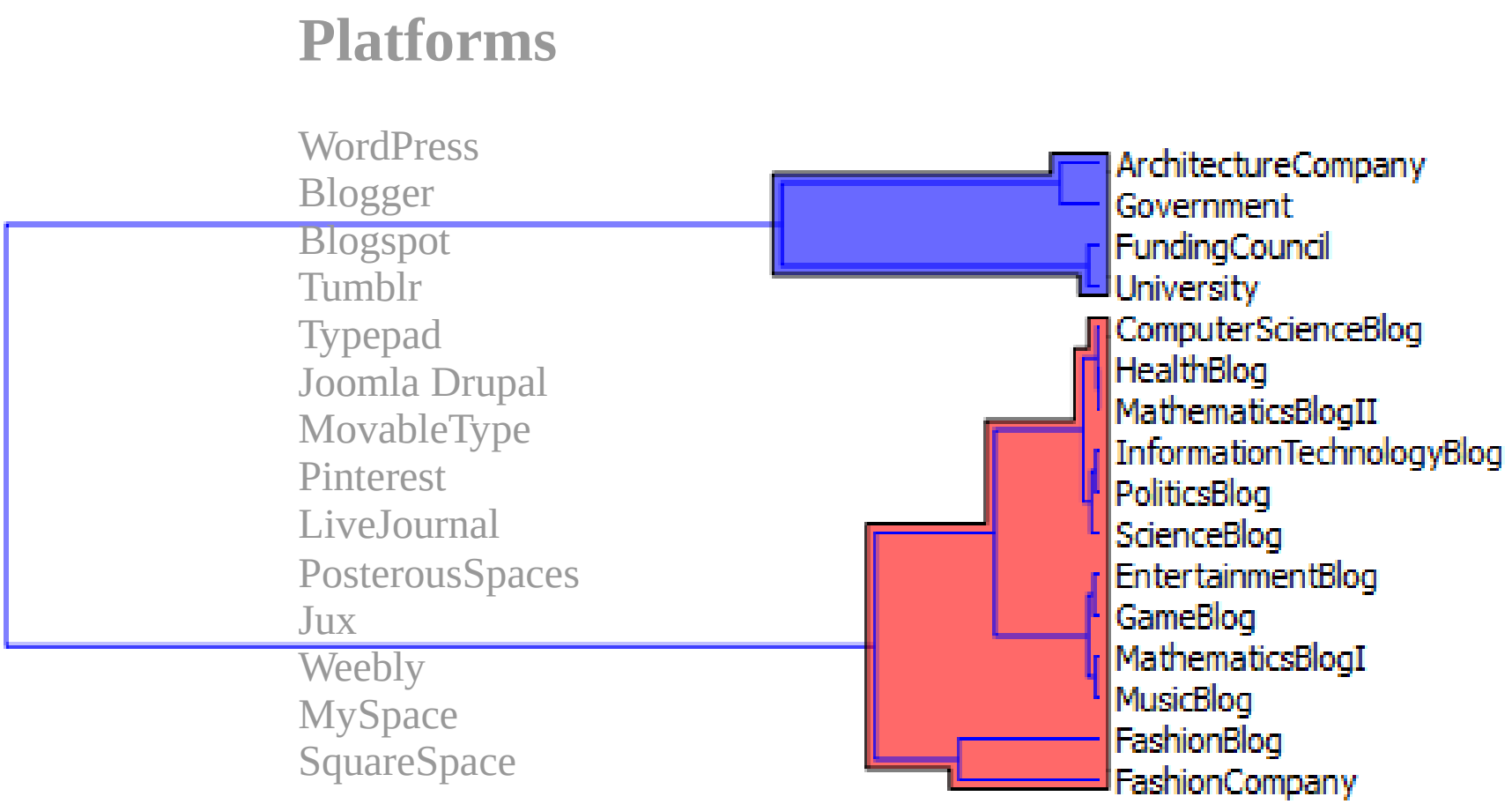
Datasets

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

BF16Cat

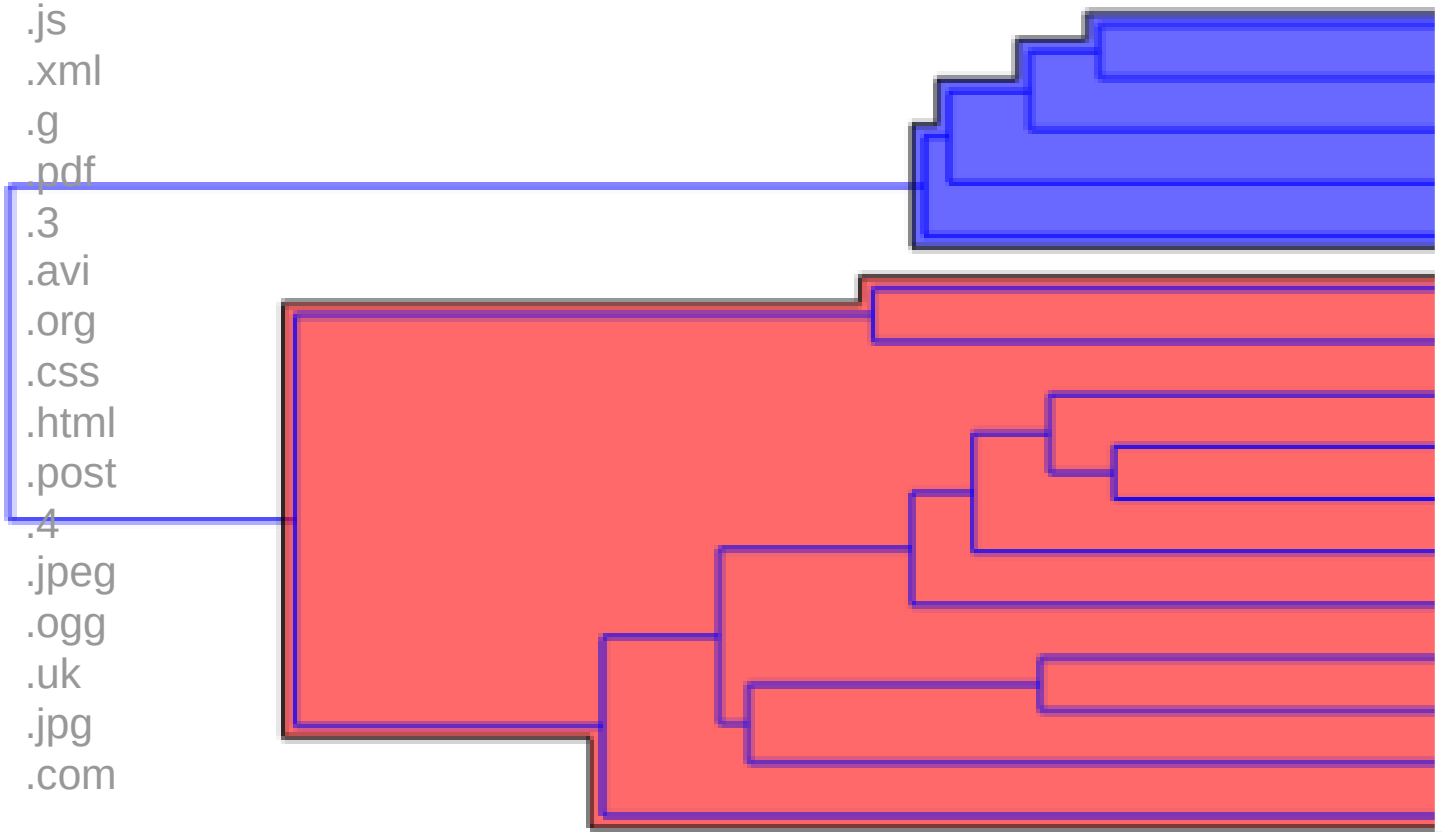
Type	Subcategory	Size	Source
Blogs	**Computer Science (CS)**	41	StackOverflow
	Information Technology (IT)	138	Technorati "IT" category search
	Entertainment (ET)	110	Technorati "Entertainment" category search
	Fashion (FA)	164	Independent Fashion Bloggers
	Game (GA)	7	University of Glasgow PhD student 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


Platforms
WordPress
Blogger
Blogspot
Tumblr
Typepad
Joomla Drupal
MovableType
Pinterest
LiveJournal
PosterousSpaces
Jux
Weebly
MySpace
SquareSpace
ArchitectureCompany
Government
FundingCouncil
University
ComputerScienceBlog
HealthBlog
MathematicsBlogII
InformationTechnologyBlog
PoliticsBlog
ScienceBlog
EntertainmentBlog
GameBlog
MathematicsBlogI
MusicBlog
FashionBlog
FashionCompany

.bmp
.ico
.aspx
.0
.01
.mp3
.stor
.js
.xml
.g
.pdf
.3
.avi
.org
.css
.html
.post
.4
.jpeg
.ogg
.uk
.jpg
.com

What's in a format?



Architecture Company
Government
Funding Council
Fashion Company
University
Computer Science Blog
Mathematics Blog
Information Tech Blog
Health Blog
Mathematics Portal Blog
Science Blog
Fashion Blog
Entertainment Blog
Music Blog
Politics Blog
Game Blog

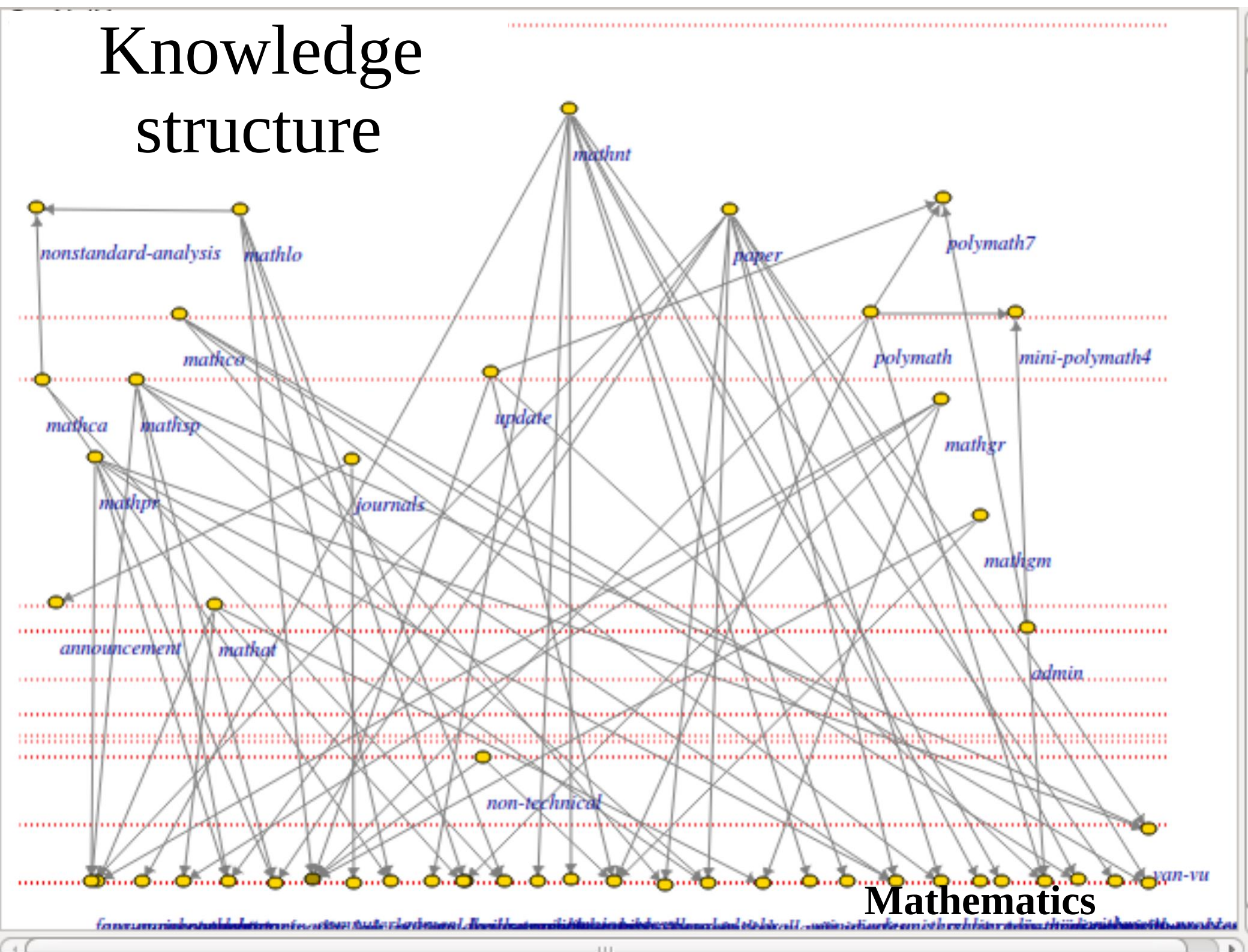
.htm
.cfm
.jsp
.nt
.de
.php
.serv
.href
.1
.swf
.cgi
.fr
.png
.valu
.asp
.shtm
.flv
.6666
.be
.gif
.dele
.gete
.2
.mp4
.net
.es

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

University of Glasgow

Knowledge structure



Mathematics

What next?

- We have tried to understand communities using behaviour online:
  - Platforms they use
  - File formats they share
  - Referencing behaviour
  - Knowledge structures

We can do much more ...

The beginning

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

from Lords & Ladies