4-23-2013

More Archives, More Better

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

More Archives, More Better

Michael L. Nelson
Old Dominion University
ws-dl.blogspot.com

IIPC General Assembly
Ljubljana, Slovenia
April 23, 2013

Three Easy Pieces

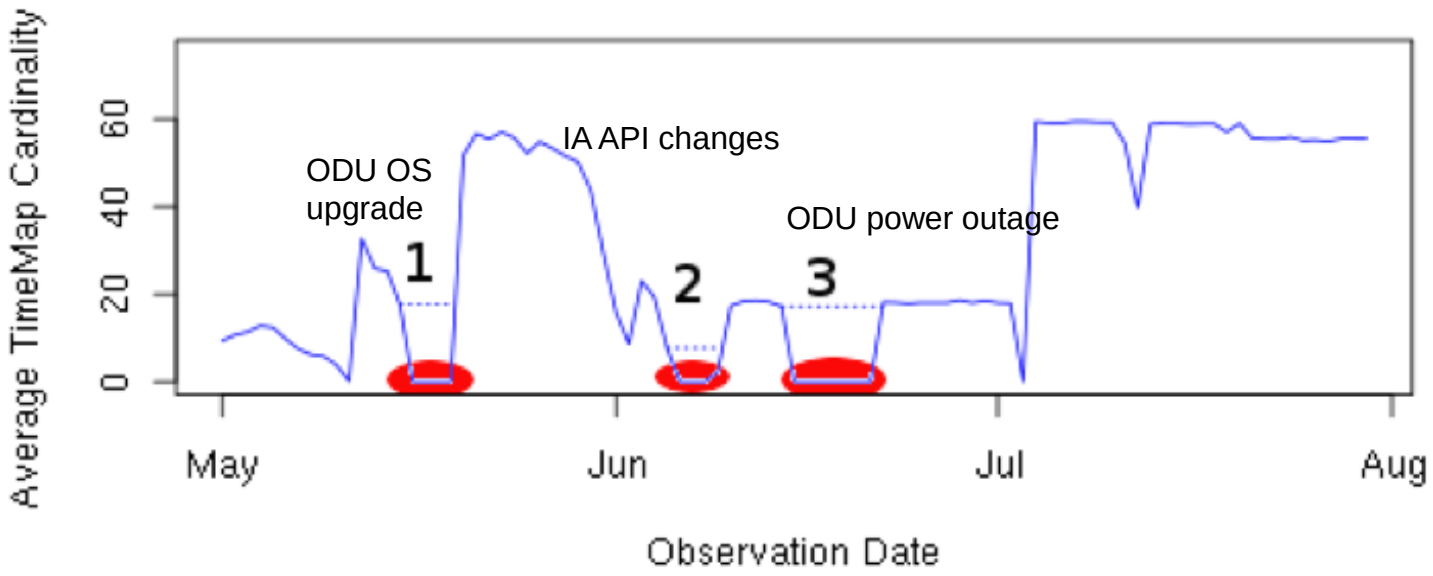
- "An Evaluation of Caching Policies for Memento TimeMaps"
  - 4000 aggregated TimeMaps downloaded daily for 3 months
  - 20% of the time the TimeMaps *shrink*
- "How Much of The Web Is Archived?"
  - 4000 URIs, 9 archives, 3 search engines
  - 16% -- 79% of the web archived
- "Profiling Web Archive Coverage for Top-Level Domain and Content Language"
  - 153329 URIs, 12 archives
  - querying only top 3 archives gives a complete TimeMap 84% of the time (52% of the time even if you exclude the IA)

An Evaluation of Caching Policies for Memento TimeMaps

JCDL 2013

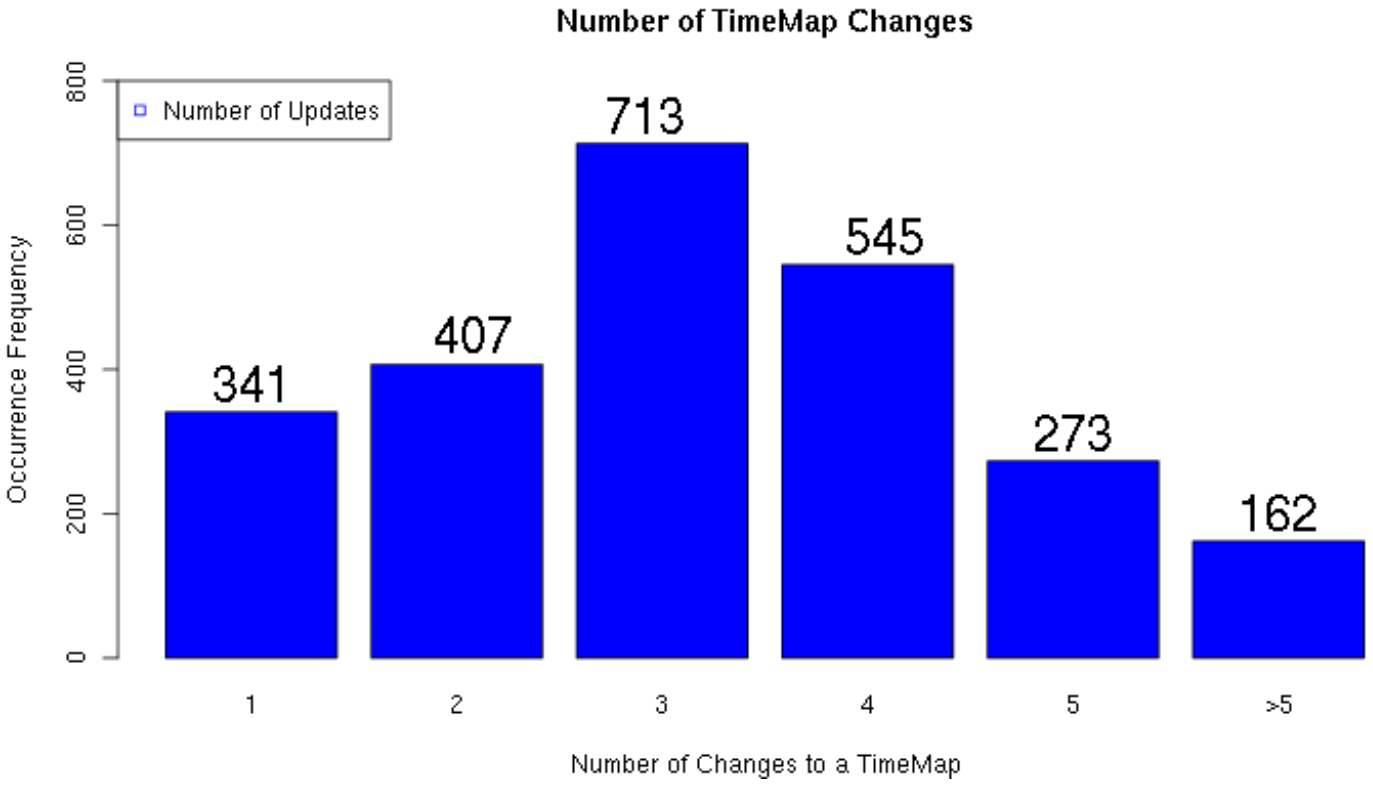
Justin Brunelle, Michael L. Nelson

Mean # Mementos per TimeMap per Day

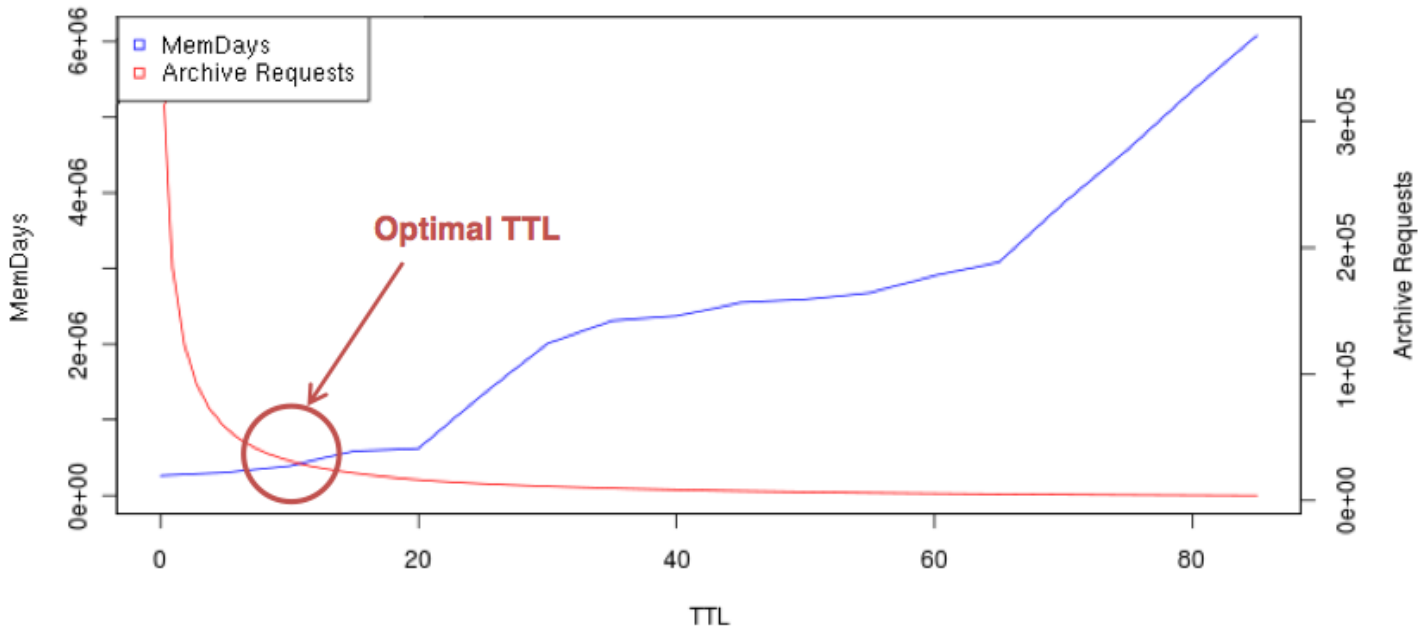


download the same 4000 TimeMaps everyday

Frequency of TimeMap changes over 92 days



Optimal TimeMap Cache TTL=15 days



minimizes queries to archives, minimizes "lost" mementos*days,
will only cache new TimeMap if it is "bigger"
question: can we do this adaptively?

How Much of The Web Is Archived?

JCDL 2011

Scott Ainsworth, Ahmed AlSum, Hany SalahEldeen,
Michele C. Weigle, Michael L. Nelson

Public Archives, ca. late 2010 / early 2011

Three categories of archives

- Internet Archive (classic interface)
- Search engine
- Other archives

bing YAHOO! Google

BRITISH LIBRARY

ARCHIVE-IT

WebCite

archiefweb.eu

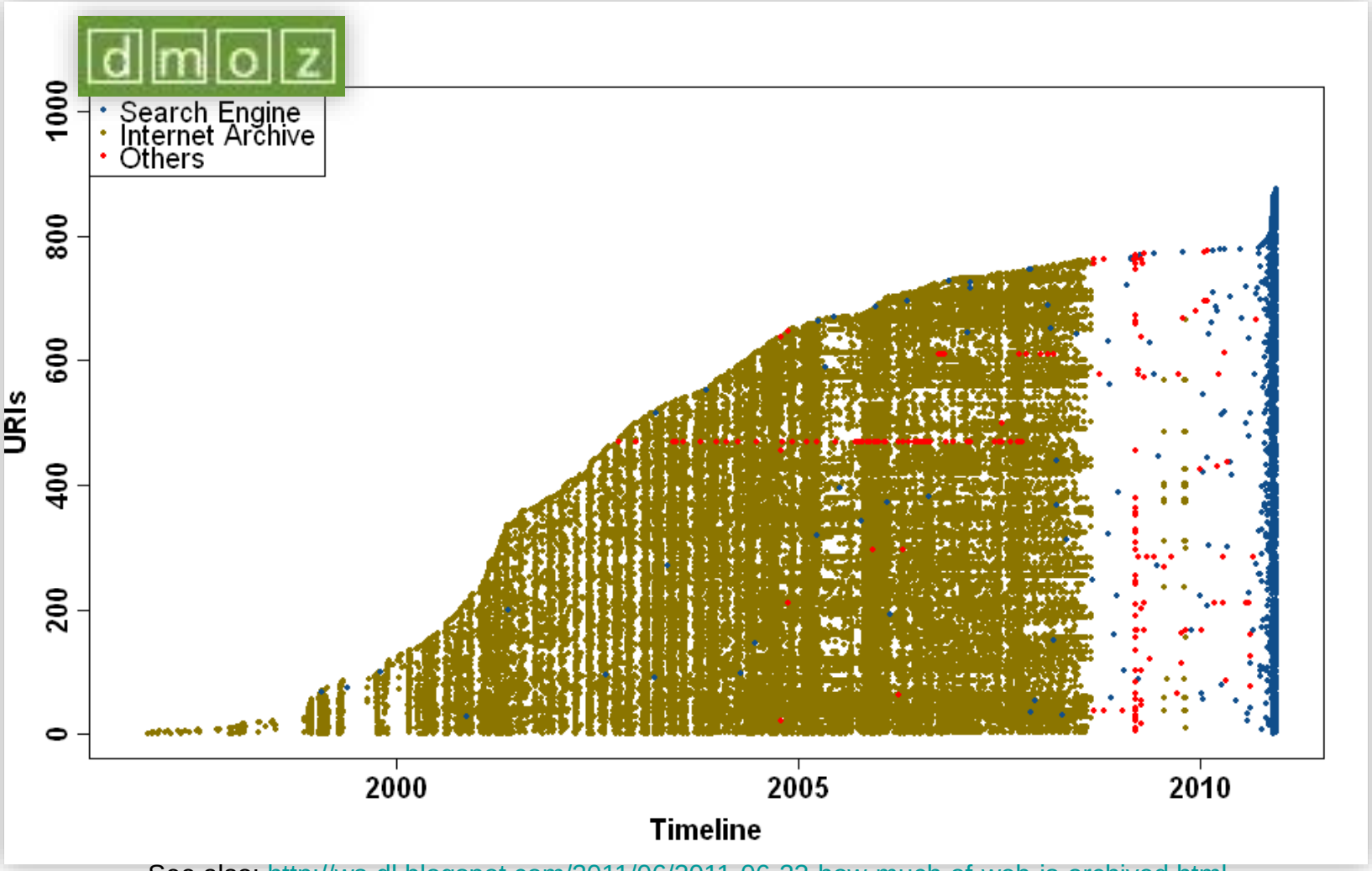
University of California CDL California Digital Library

The National Archives UK

NATIONAL ARCHIVES US

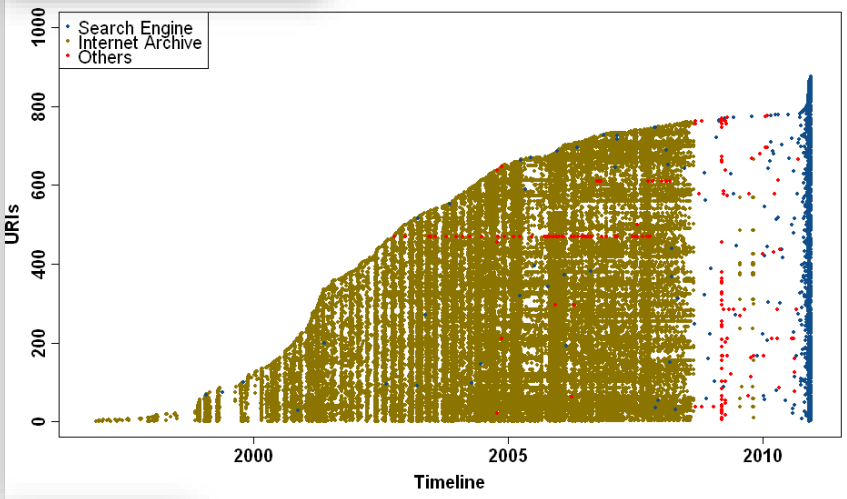
diigo V5.0 beta

1000 URIs, ordered by first observation date

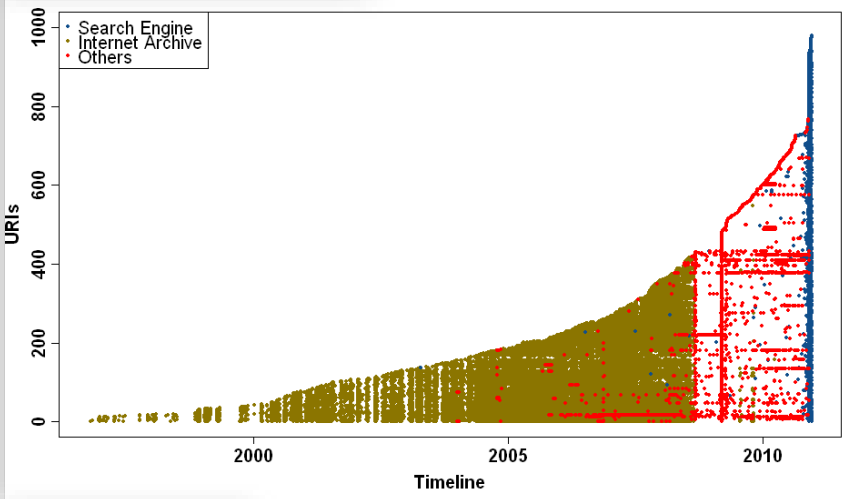


See also: http://ws-dl.blogspot.com/2011/06/2011-06-23-how-much-of-web-is-archived.html

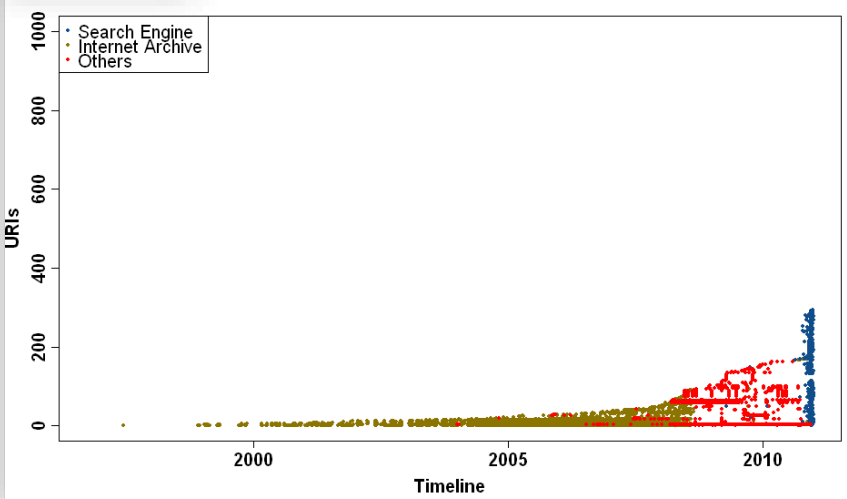
dmoz



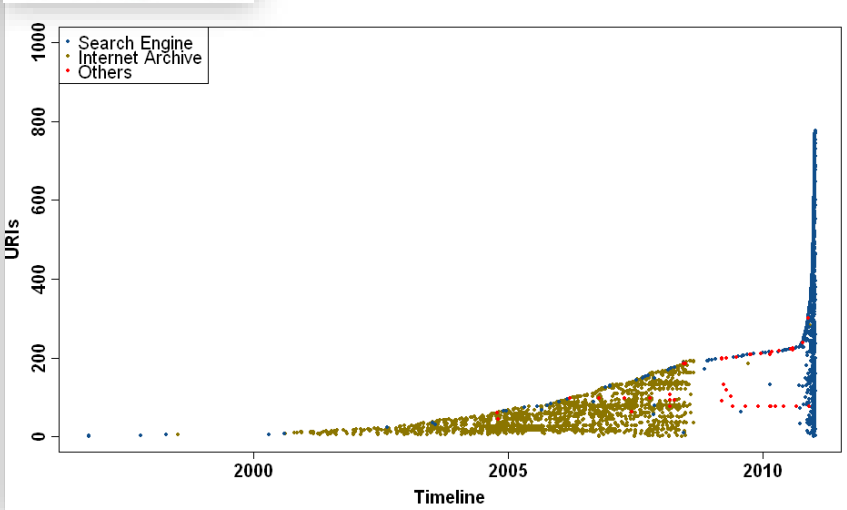
delicious



bitly



SE YAHOO!



see also: http://ws-dl.blogspot.com/2013/04/2013-04-19-carbon-dating-web.html

How Much of the Web is Archived?
It depends on which web…

	Including SE cache	Excluding SE Cache
dmoz	90%	79%
delicious	97%	68%
bitly	35%	16%
SE YAHOO!	88%	19%

Changes since 2011: no more free SE APIs; greatly reduced IA quarantine period

Profiling Web Archive Coverage for Top-Level Domain and Content Language

(submitted for publication)

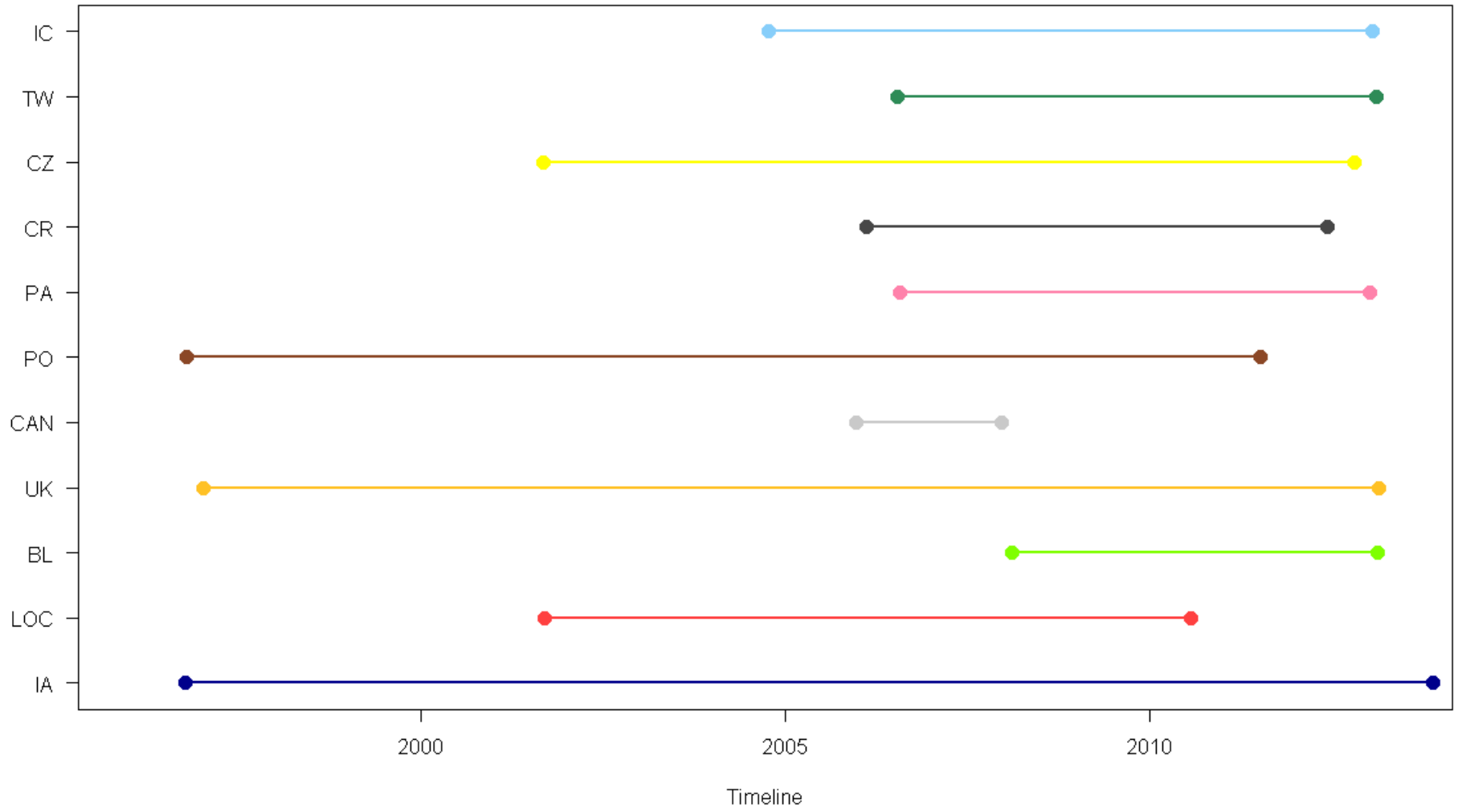
Ahmed AlSum, Michele C. Weigle, Michael L. Nelson, Herbert Van de Sompel

12 (IIPC) Archives

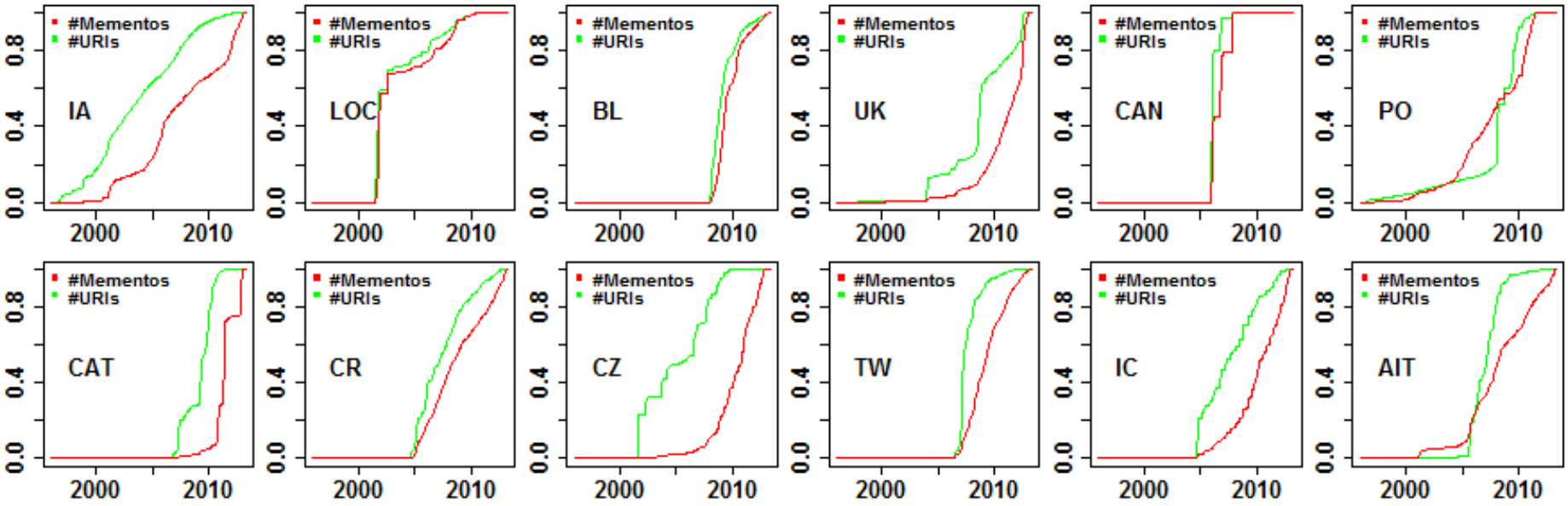
	Archive Name	FullText search	Website
IA	Internet Archive		web.archive.org
LoC	Library of Congress		www.loc.gov/lcwa
IC	Icelandic Web Archive		vefsafn.is
CAN	Library & Archives Canada	x	www.collectionscanada.gc.ca
BL	British Library	x	www.webarchive.org.uk/ukwa
UK	UK Gov. Web Archive	x	webarchive.nationalarchives.gov.uk
PO	Portuguese Web Archive	x	arquivo.pt
CAT	Web Archive of Catalonia	x	www.padi.cat
CR	Croatian Web Archive	x	haw.nsk.hr
CZ	Archive of the Czech Web	x	webarchiv.cz
TW	National Taiwan University	x	webarchive.lib.ntu.edu.tw
AIT	Archive-It	x	www.archive-it.org

153329 URIs from DMOZ, archive fulltext search, IA logs, Memento aggregator logs

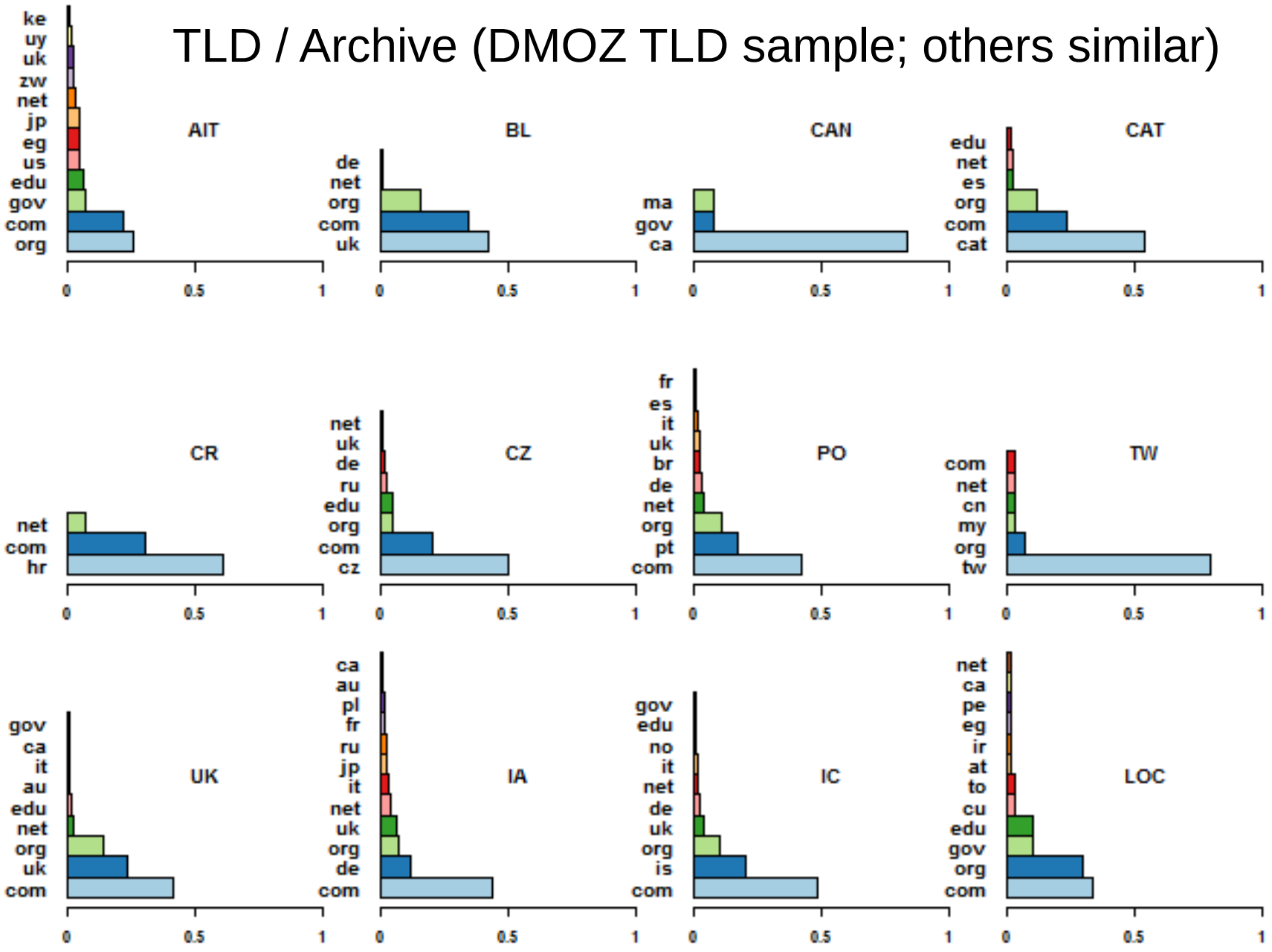
Temporal Spread



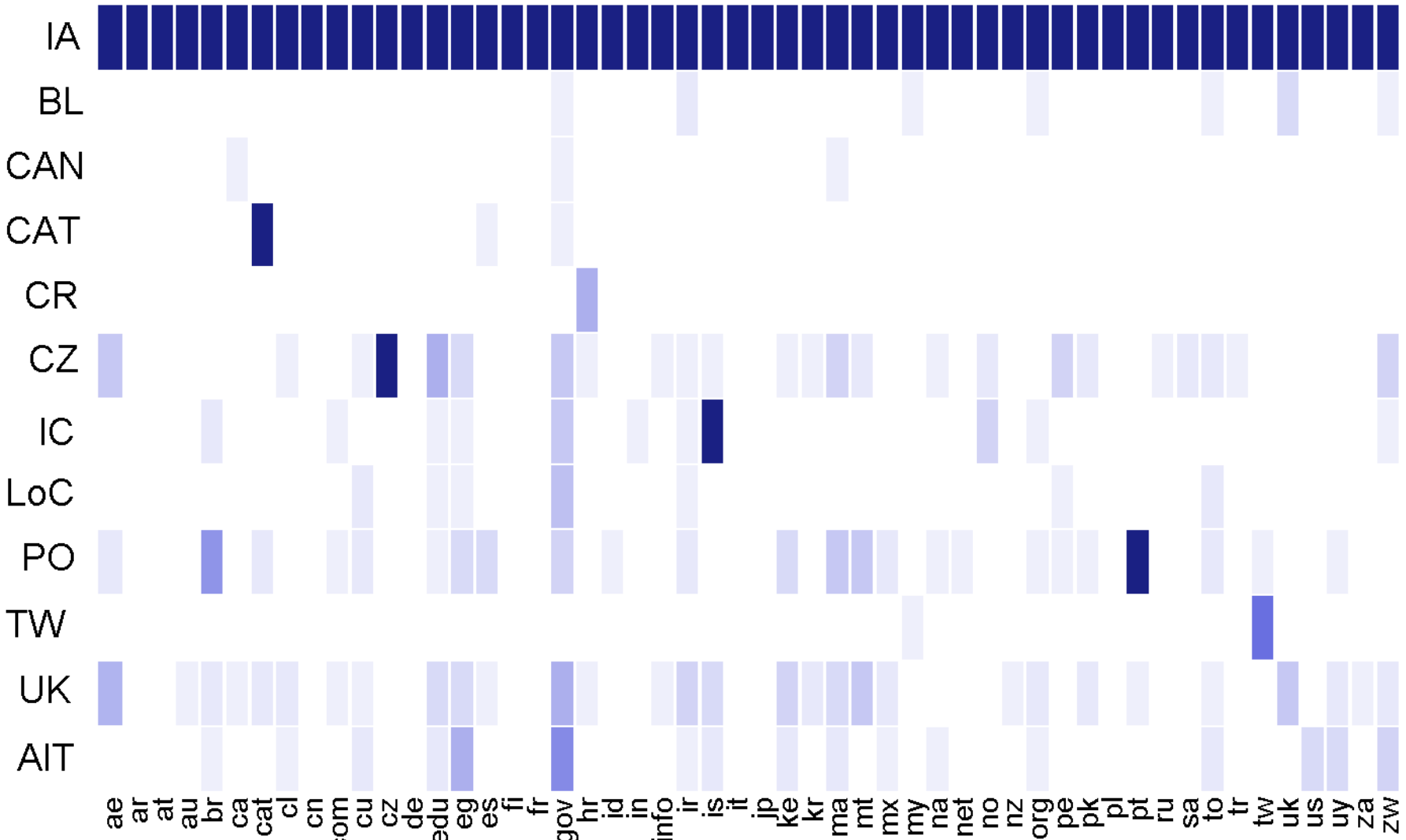
Rate of Acquiring URI-Rs, URI-Ms



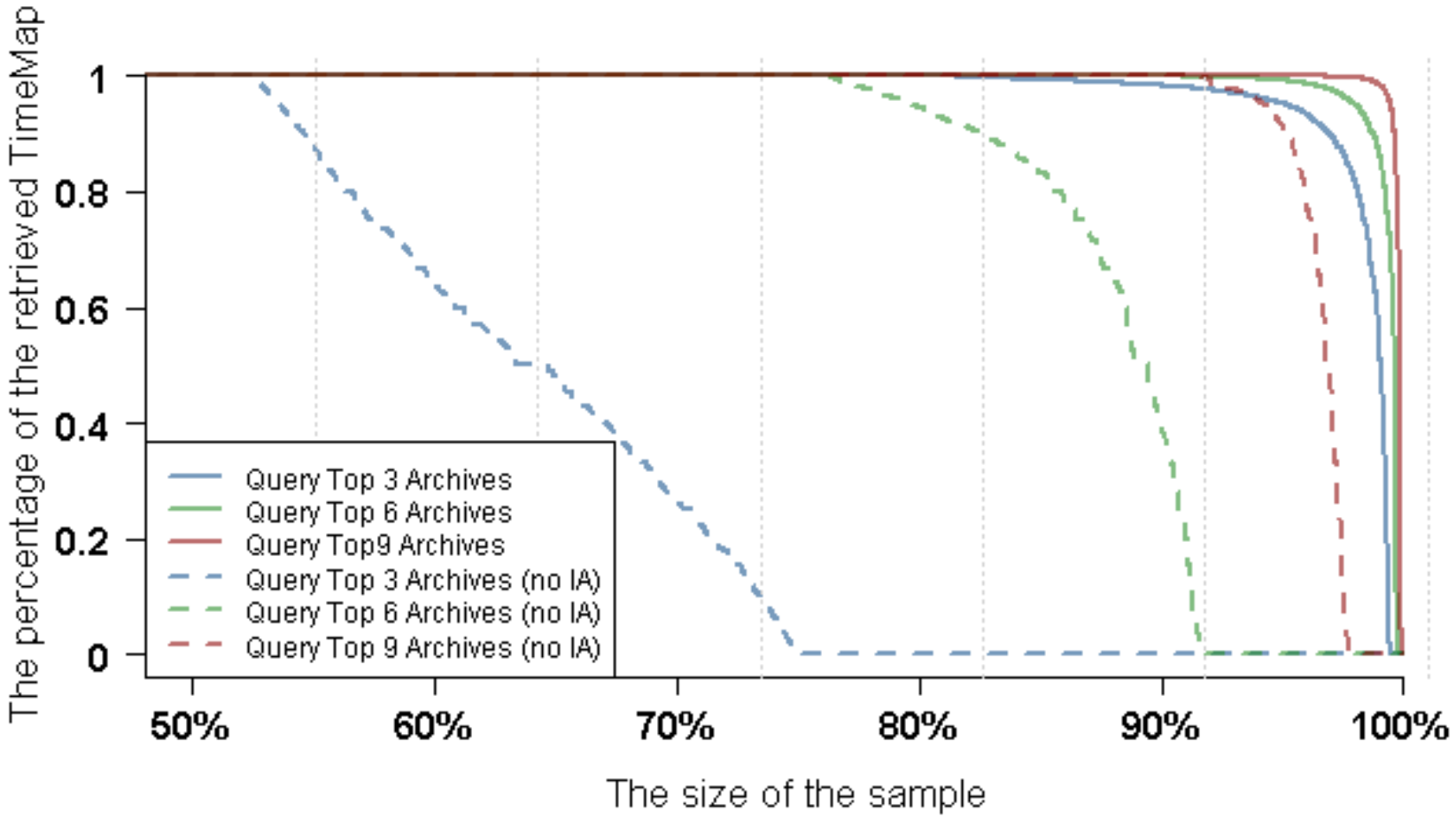
TLD / Archive (DMOZ TLD sample; others similar)



Archive / TLD Heatmap



Using Only Top-k Archives for URI Lookup Yields Good Results



Even when there are 100s of archives, we only need to talk to a few.