9-23-2013

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

Ahmed AlSum
Old Dominion University

Michele C. Weigle
Old Dominion University, mweigle@odu.edu

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

Herbert Van de Sompel

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

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

International Conference on Theory and Practice of Digital Libraries
September 22-26, 2013
Valletta, Malta


LIBRARY OF
CONGRESS


INTERNET ARCHIVE


Czech Web
WebArchiv


臺灣大學網站典藏庫
NTU Web Archiving System


PORTUGUESE
WEB ARCHIVE


NLB
National Library
Singapore


memento


ARCHIVE-IT


BRITISH
LIBRARY


Archive.is


WebCite


archiefweb.eu

Aggregator


The National Archives


Icelandic Web Archive

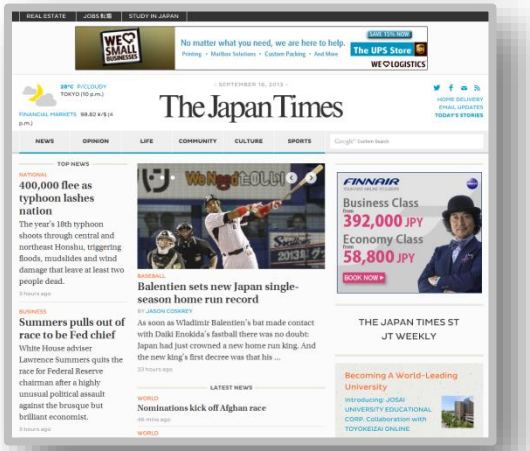

Library and Archives
Canada


padicat
The Web Archive of Catalonia


spletni
ARHIV
NUK


HAW
Hrvatski arhiv weba
Croatian Web Archive

Where to find Mementos for …



http://www.japantimes.co.jp/

NLB National Library Singapore

PORTUGUESE WEB ARCHIVE

The National Archives

臺灣大學網站典藏庫
NTU Web Archiving System

INTERNET ARCHIVE

BRITISH LIBRARY

spletni ARHIV NUK

Icelandic Web Archive

WebCite

ARCHIVE-IT

Where to find Mementos for …

http://www.japantimes.co.jp/

NLB National Library Singapore

PORTUGUESE WEB ARCHIVE

The National Archives

臺灣大學網站典藏庫
NTU Web Archiving System

INTERNET ARCHIVE

BRITISH LIBRARY

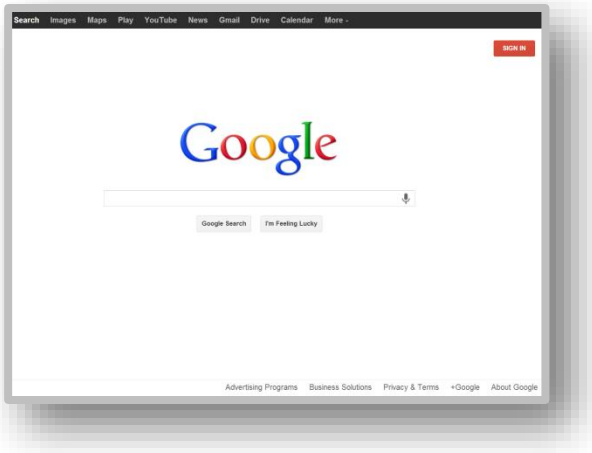
spletni ARHIV NUK

Icelandic Web Archive

WebCite

ARCHIVE-IT

Where to find Mementos for …



http://www.google.com/

NLB National Library Singapore

The National Archives

Icelandic Web Archive

PORTUGUESE WEB ARCHIVE

臺灣大學網站典藏庫 NTU Web Archiving System

WebCite

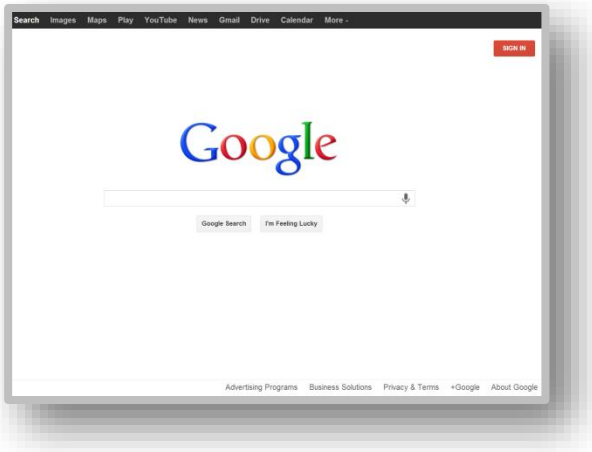
INTERNET ARCHIVE

BRITISH LIBRARY

spletni ARHIV NUK

ARCHIVE-IT

Where to find Mementos for …



http://www.google.com/

NLB National Library Singapore

The National Archives

Icelandic Web Archive

PORTUGUESE WEB ARCHIVE

臺灣大學網站典藏庫 NTU Web Archiving System

WebCite

INTERNET ARCHIVE

BRITISH LIBRARY

spletni ARHIV NUK

ARCHIVE-IT

Research Question

Problem

- Profile public web archives according to the following dimensions:
  - Top-level domains
  - Languages
  - Growth rate
  - Archival date

Motivation

- To determine who is archiving what
- To optimize the query routing for a Memento Aggregator

Web Archives in this Experiment

	Full text	URI-lookup
Internet Archive		√
Library of Congress		√
Icelandic Web Archive		√
Library and Archives Canada	√	√
British Library	√	√
UK National Library	√	√
Portuguese Web Archive	√	√
Web Archive of Catalonia	√	√
Croatian Web Archive	√	√
Archive of the Czech Web	√	√
National Taiwan University	√	√
Archive It	√	√

Experiment Set Up

- Sample URIs from different sources
  - Details coming up
- Retrieve the TimeMap for each URI from all archives
  - A TimeMap lists all Mementos for a given URI
  - A Memento is an archived version of a resource
- Analyze
  - Details coming up

Sampling URIs

Web

1. DMOZ:Random
2. DMOZ:TLD - 2% of each TLD from DMOZ (.com, .org, .jp, etc 52 TLD)
3. DMOZ:Languages - 100 URIs for each Languages (24 lang.)

Web Archives Full Text

4. Top 1-Gram from Bing
5. Top 1000 queries term by Yahoo in 9 languages

User requests

6. IA Wayback Machine Log files
7. Memento aggregator log files

Sampling URIs - DMOZ

1. DMOZ:Random
   - 10,000 URIs randomly sampled from DMOZ directory (~5M URIs).

2. DMOZ:TLD - 2% for each TLD from DMOZ or 100 URIs whichever is greater
   - 52 TLDs (**com** 23,470) (**de** 6,332), (**org** 4,025), (**uk** 3,309), (**net** 2,073), (**it** 1,775), (**jp** 1379), (**ru** 1244), (**fr** 1154), (**pl** 1062), (**au** 764), (**ca** 642), (**at** 438), (**edu** 390), (**cz** 385), (**tr** 334), (**info** 319), (**cn** 278), (**us** 266), (**nz** 265), (**es** 238), (**ar** 213), (**no** 150), (**br** 149), (**tw** 141), (**za** 118), (**fi** 113), (100 URIs for [**ae**, **cat**, **cl**, **cu**, **eg**, **gov**, **id**, **in**, **ir**, **is**, **ke**, **kr**, **ma**, **mt**, **mx**, **my**, **na**, **pe**, **pk**, **pt**, **sa**, **to**, **uy**, **zw**])

3. DMOZ:Languages - 100 URIs for each language
   - 24 languages: Icelandic, Portuguese, Catalan, Afrikaans, Arabic, Indonesian, Chinese (Simplified), Chinese (Traditional), Dutch, Spanish, French, Greek, Hindi, Italian, Japanese, Korean, Norwegian, Persian, Polish , Russian, Turkish, Ukrainian

Sampling URIs – Web Archives Full Text

- Query the fulltext search interface of select web archives with two sets of query terms.

4. Top 1-Gram from Bing
   - Most are English

5. Top 1000 query terms from Yahoo in 9 languages
   - Excluding general keywords such as: Obama, Facebook.

Sampling URIs – Web Archives Full Text

											Yahoo	Bing
Archive with FullText search	AIT										**12617**	**3953**
	BL										**6430**	**3187**
	CAN										**1351**	**1107**
	CR										**1599**	**1201**
	CZ										**6081**	**3360**
	CAT										**8996**	**4241**
	PO										**14126**	**5004**
	TW	357	178	176	165	157	106	7	198	119	**1004**	**354**

Sampling URIs – Web Archives Full Text

		Chinese	English	French	German	Italian	Japanese	Korean	Portuguese	Spanish	Yahoo	
Archive with FullText search	AIT	26	2066	**3512**	**3837**	3321	119	2	2434	2141	**12617**	
	BL	163	2354	**2350**	**2240**	**2068**	225	131	**1940**	**2056**	**6430**	
	CAN	49	800	804	646	601	77	113	580	514	**1351**	
	CR	54	706	697	703	701	74	19	599	600	**1599**	
	CZ	**363**	1782	1578	1695	1519	**577**	114	1310	1278	**6081**	
	CAT	28	2775	2496	2448	2280	209	129	2164	2429	**8996**	
	PO	91	2460	3603	3081	3113	53	69	3267	3177	**14126**	**5004**
	TW	357	178	176	165	157	106	7	198	119	**1004**	**354**

Sampling URIs – User Requests

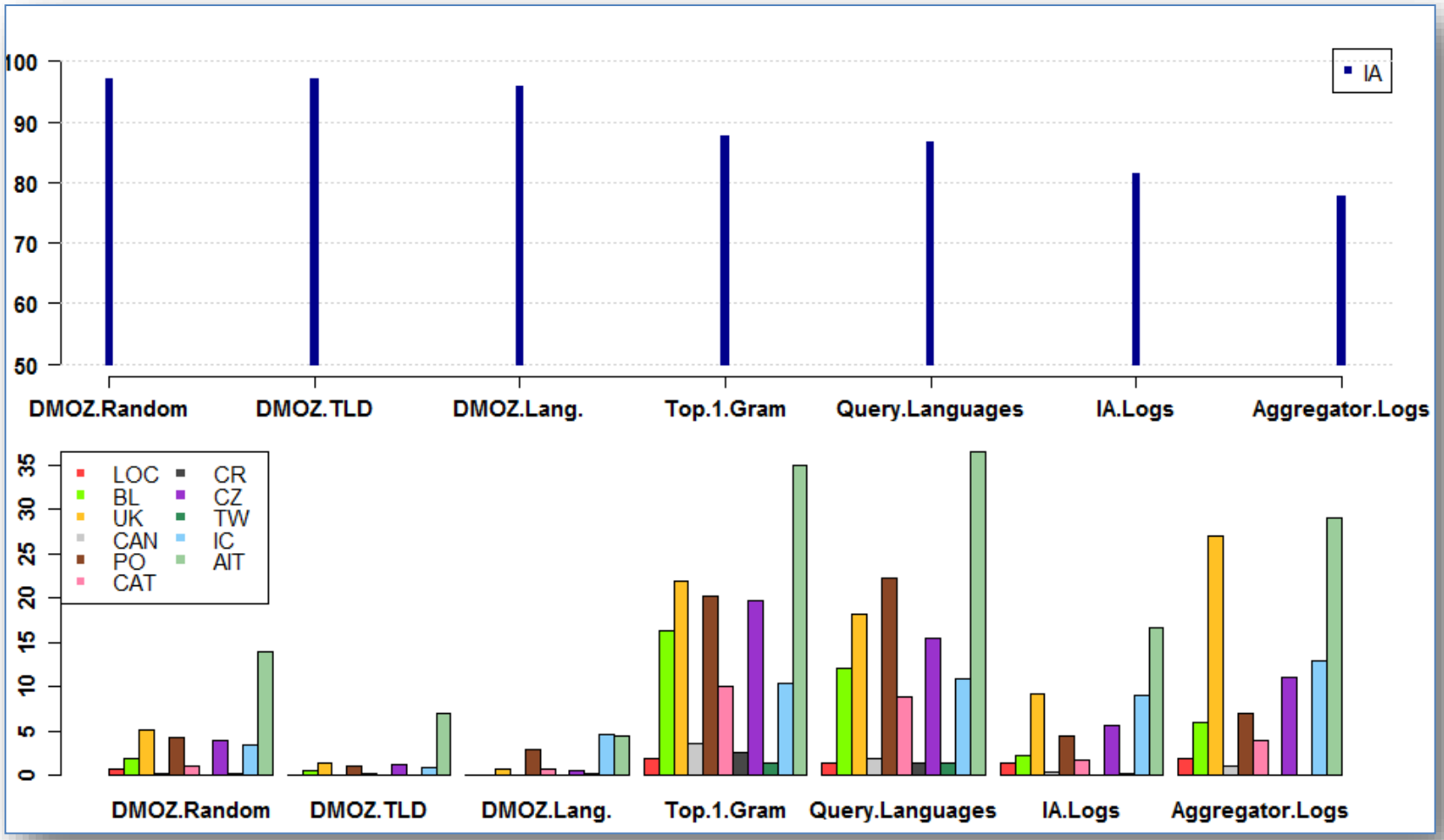
- Sampling from user requests for archived web resources

6. Sample from IA Wayback Machine Log files
   - 1,000 URIs randomly sampled from Feb 22, 2012 to Feb 26, 2012.
7. Sample from Memento Aggregator log files
   - 100 URIs randomly sampled from LANL Memento Aggregator between 2011 to 2013.

Archive Coverage per Sample

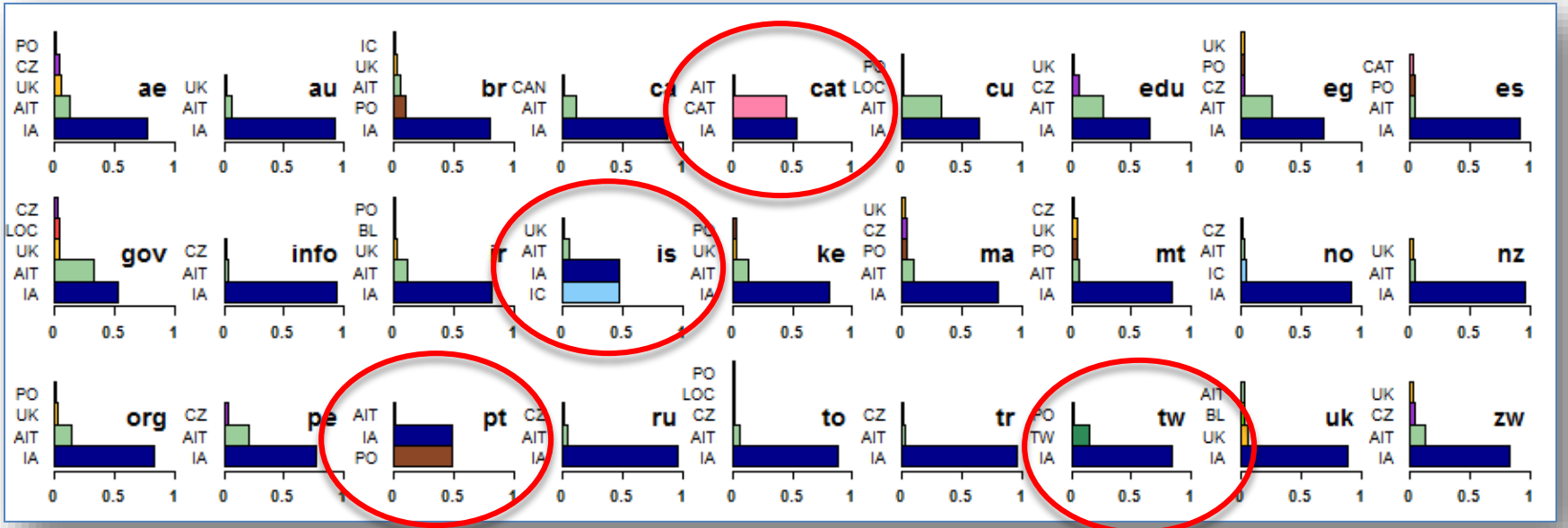
100%

35%



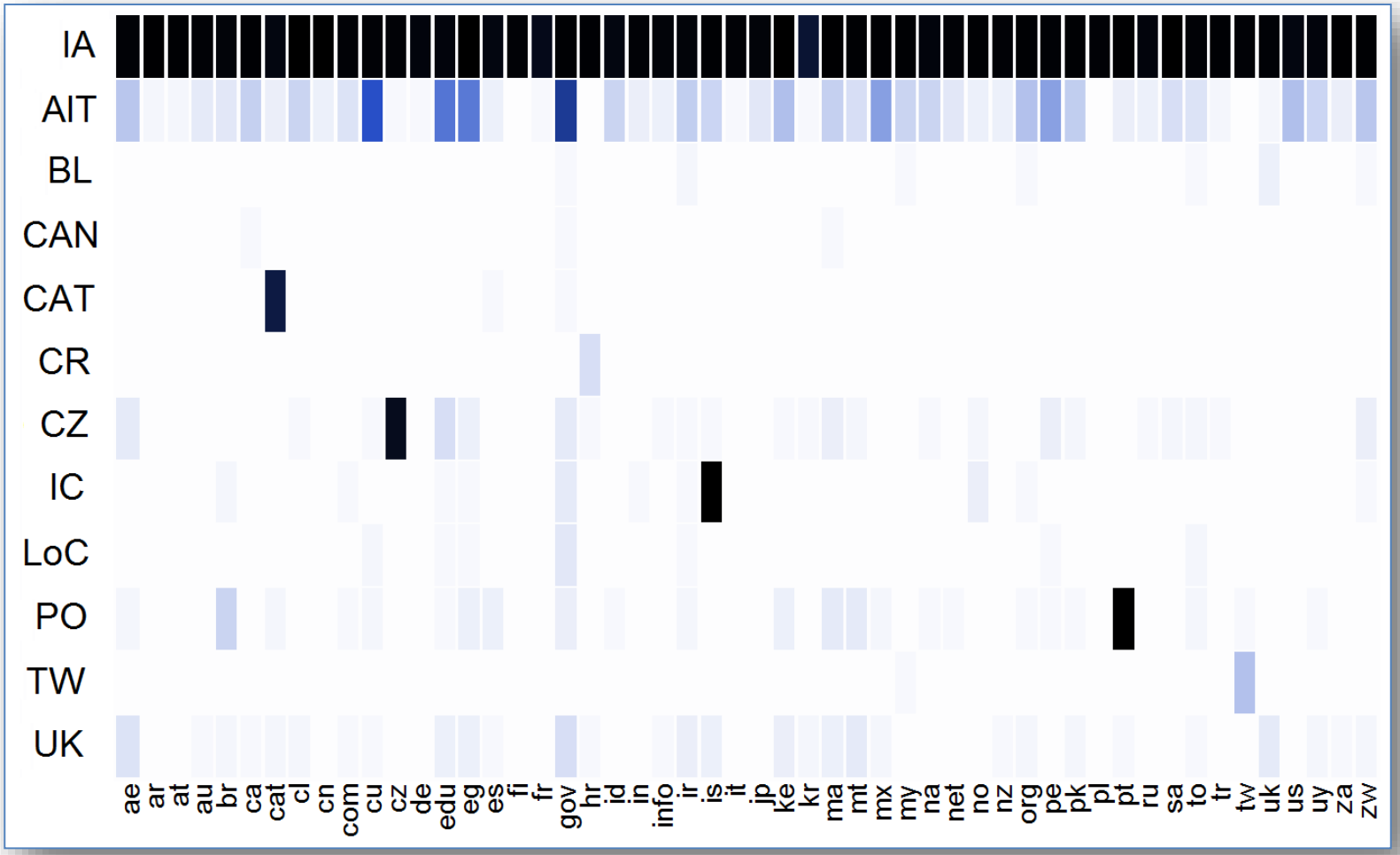
Entire Sample

TLD Coverage across Archives (1)



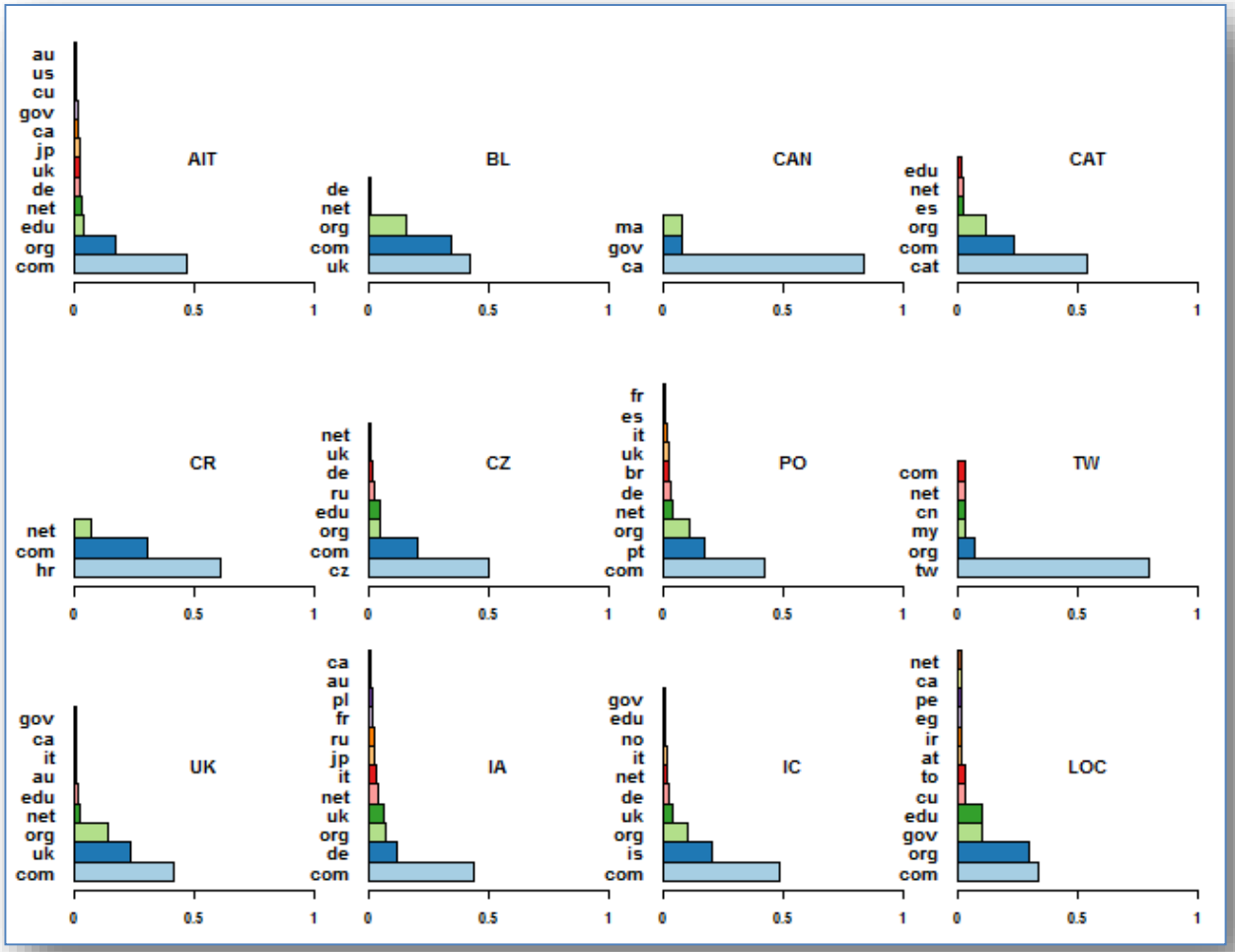
Entire Sample

TLD Coverage across Archives (2)



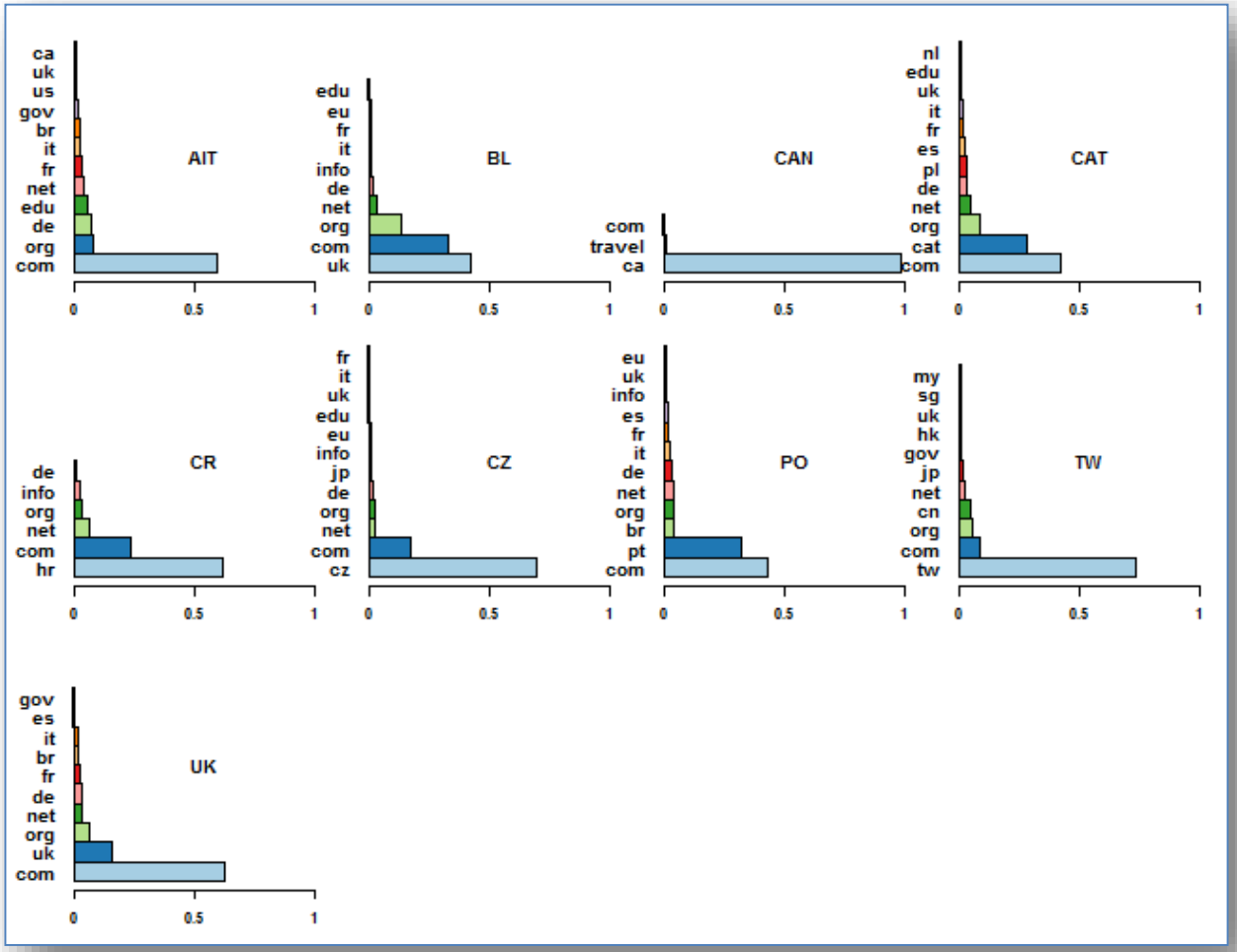
Entire Sample

TLD Distribution per Archive



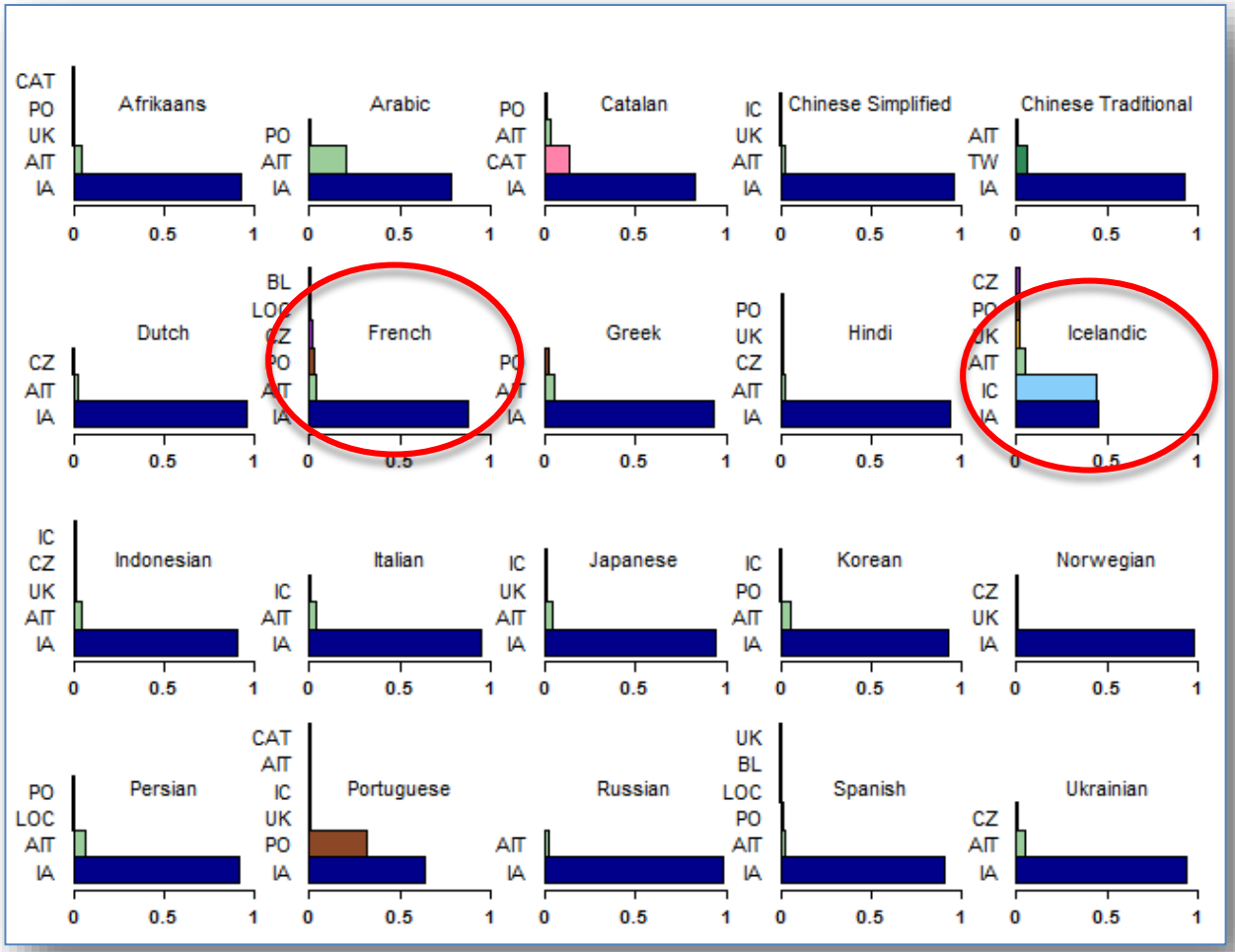
DMOZ:TLD Sample

TLD Distribution per Archive



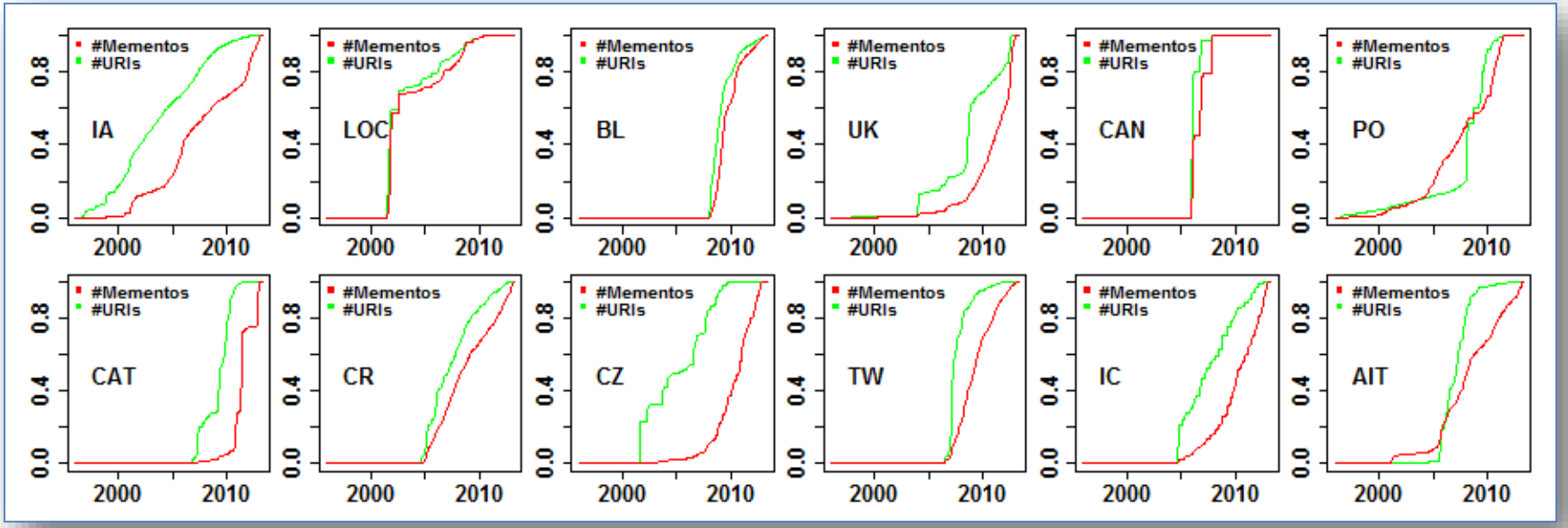
Web Archives Full Text Sample

Language Coverage per Archive



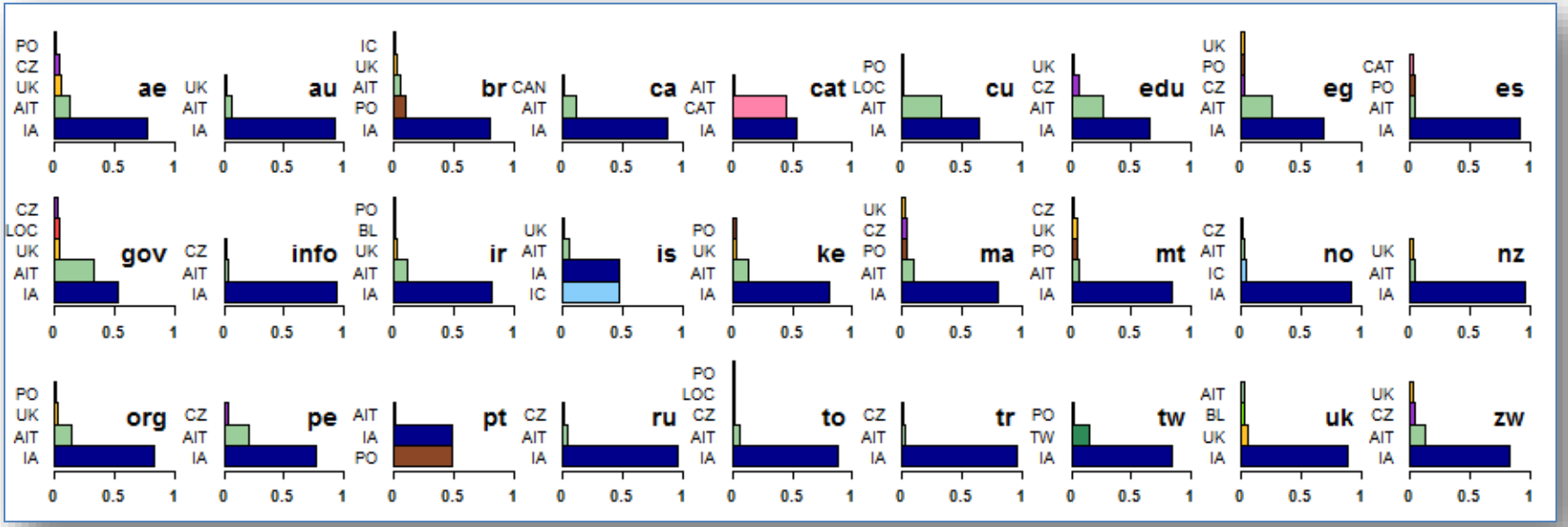
DMOZ Sample

Archive Growth Rate



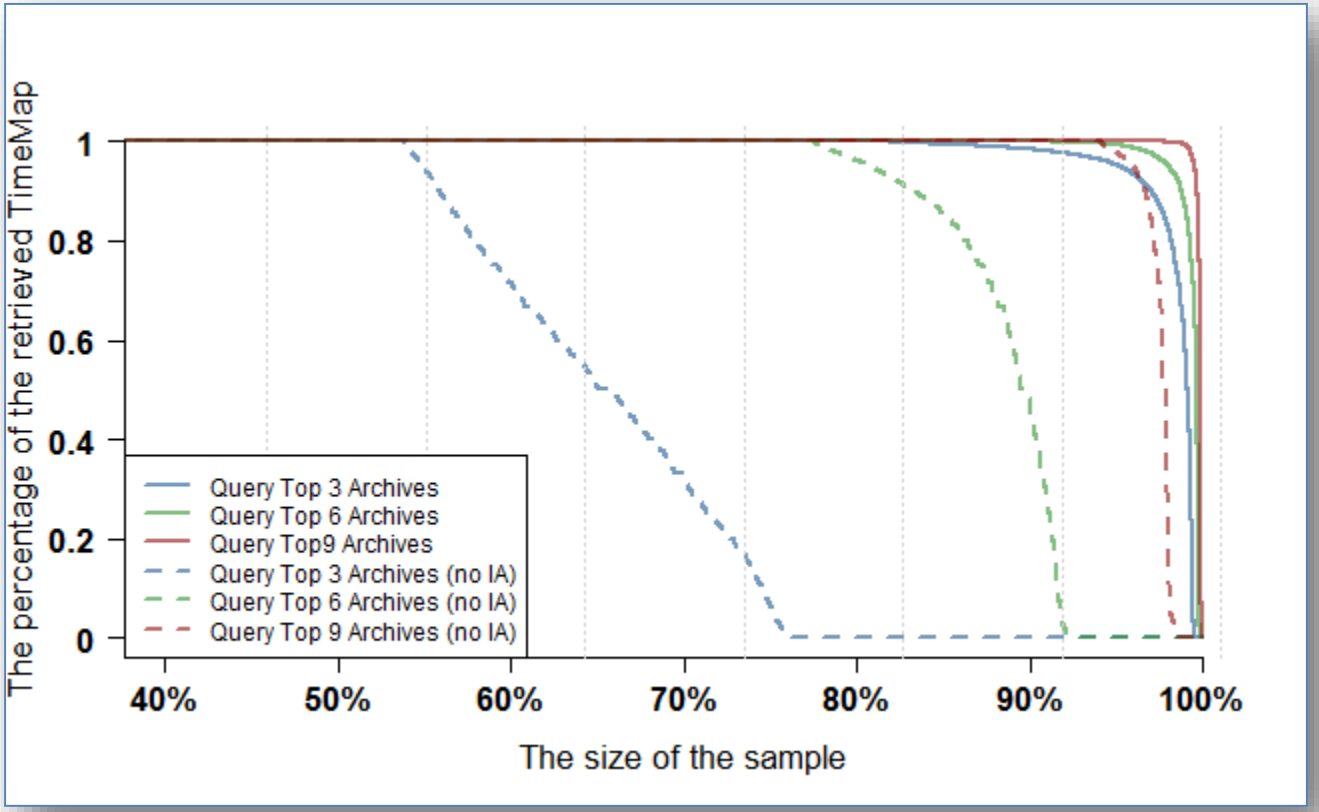
Entire Sample

TLD Coverage across Archives



Entire Sample

Query Routing Evaluation



Conclusions

- Introduced automated methodology to profile web archives using available infrastructure, no privileged access

- Coverage:
  - Internet Archive provides broad coverage
  - National archives have good coverage for their domains
  - Surprising coverage by certain archives

- Query Routing:
  - In 84% of the cases, all existing Mementos for a TLD can be found by using IA and two additional top archives for a TLD
  - In 55% of the cases, all existing Mementos for a TLD can be found by using the top 3 archives for a TLD, excluding IA