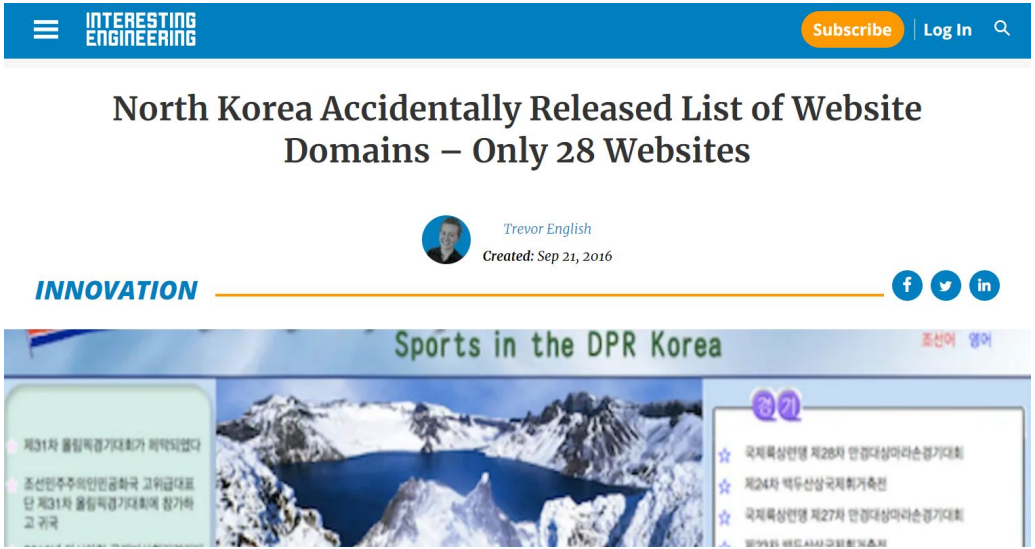
The Internet Archive and the socio-technical construction of historical facts

Anat Ben-David and Adam Amran

Internet Histories, 2018

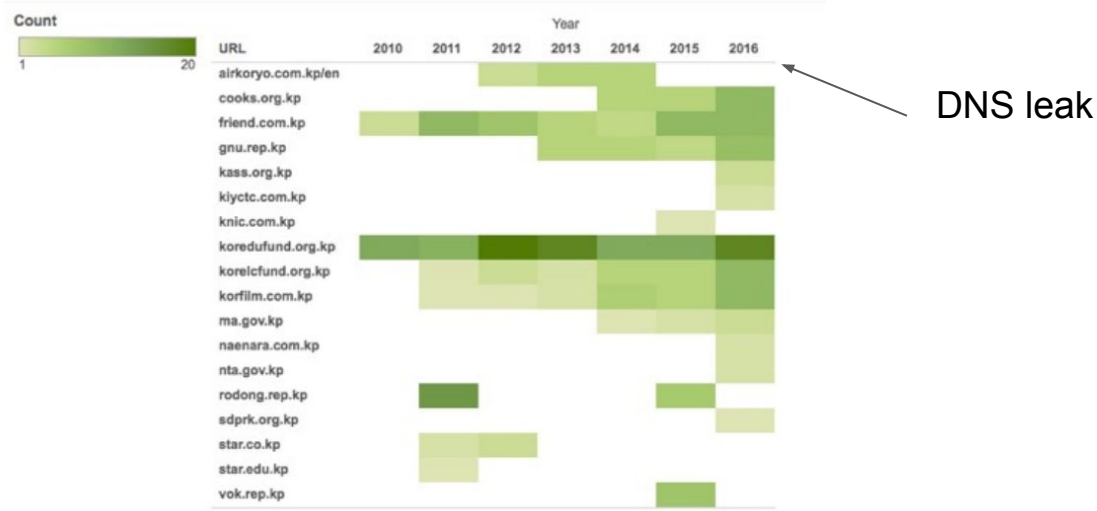
Presented by Lesley Frew
November 7, 2022

In 2016, a nameserver error revealed the comprehensive list of North Korean domains



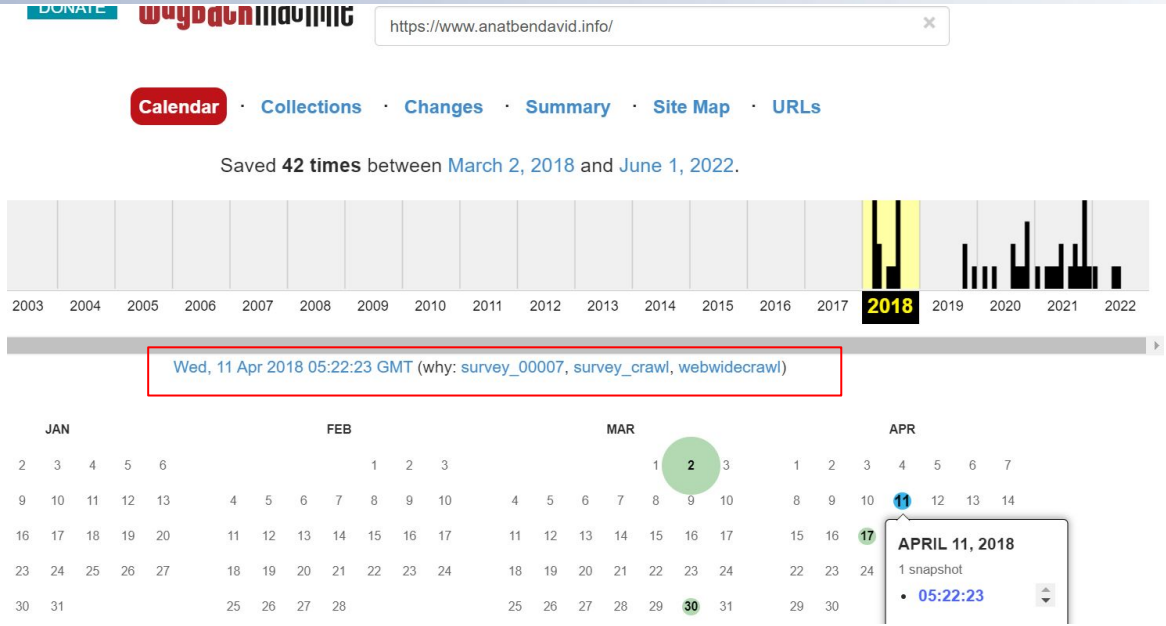
https://interestingengineering.com/innovation/north-korea-accidentally-released-list-of-website-domains-only-28-websites

How did the North Korean webpages get archived long before the public knew about them?



Ben-David et al., Figure 1

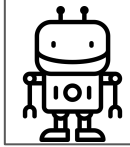
Can the black box of how URIs get added to the Wayback Machine be reverse engineered?



https://web.archive.org/web/20180801000000*/https://www.anatbendavid.info/

Humans and computers contribute URIs to the Wayback Machine

Wayback crawls



+

On-demand snapshots

https://commons.wikimedia.org/wiki/File:Robot.svg
https://commons.wikimedia.org/wiki/File:Person_(1102860)_-_The_Noun_Project.svg
https://commons.wikimedia.org/wiki/File:HFCA_1607_Special_Events_And_VIPS_Volume_1_113.jpg_(77c777ff7b0a47d18c318ae9af9eb557).jpg

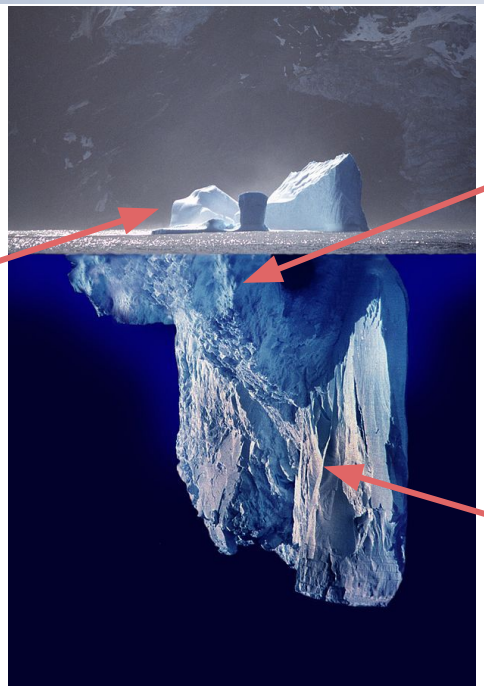
The Wayback Machine discovers new URIs from two types of crawls

<u>Survey crawls</u>

Top level sites already in the Wayback Machine

Not used to discover new URIs

Some crawl details (like seed lists) are not public



https://commons.wikimedia.org/wiki/File:Iceberg.jpg

<u>Shallow crawls</u>

Pages already in the Wayback machine plus a crawl one level deep

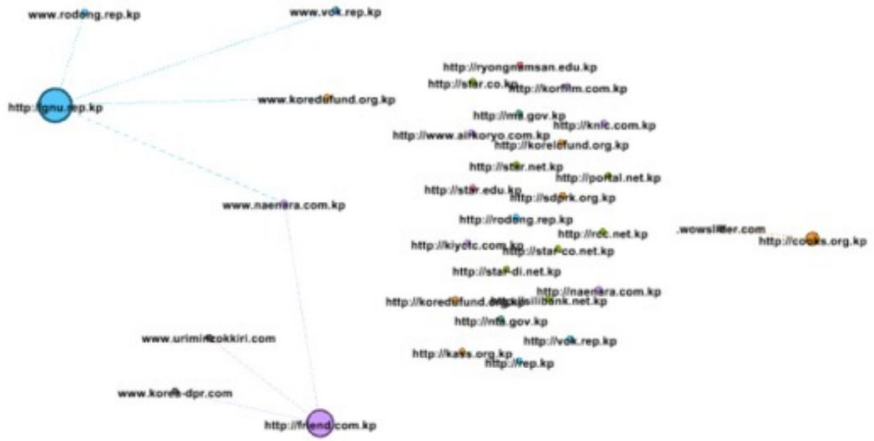
New content discovered from outlinks

<u>Worldwide crawls</u>

Seed lists crawled to a predetermined depth

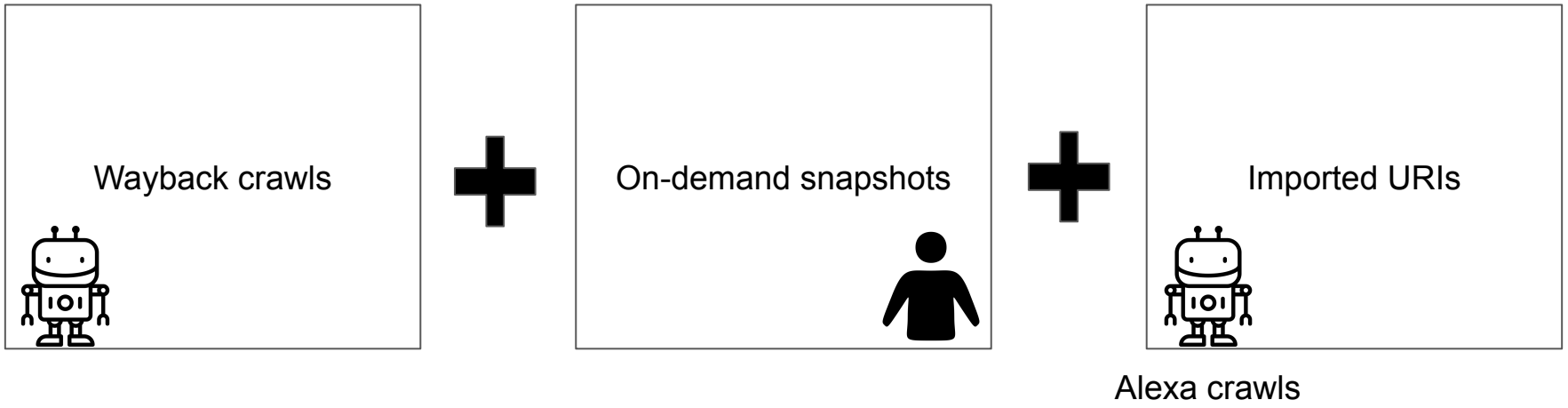
New content discovered from seed lists and outlinks

North Korean websites do not link to each other

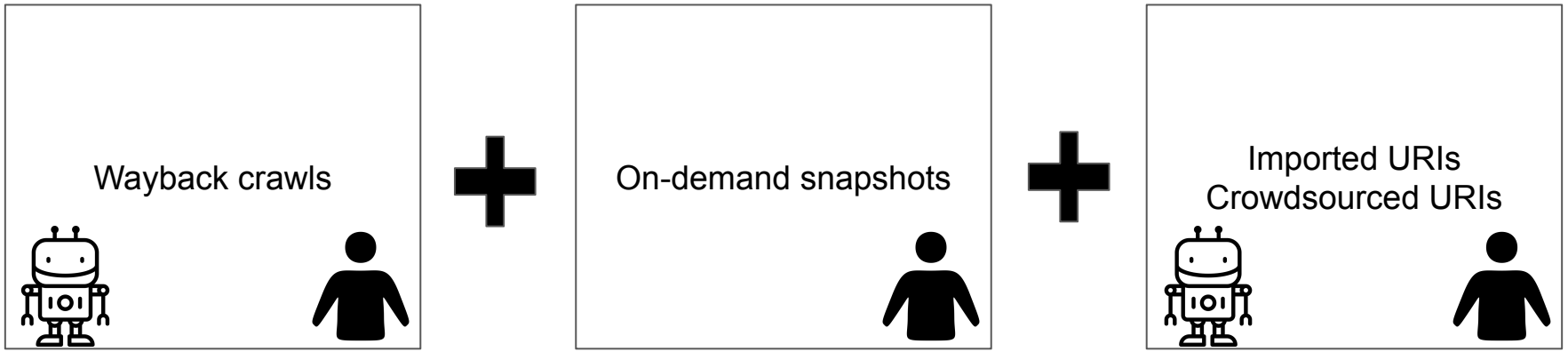


Ben-David et al., Figure 2

Imported URIs also contribute to the contents of the Wayback Machine



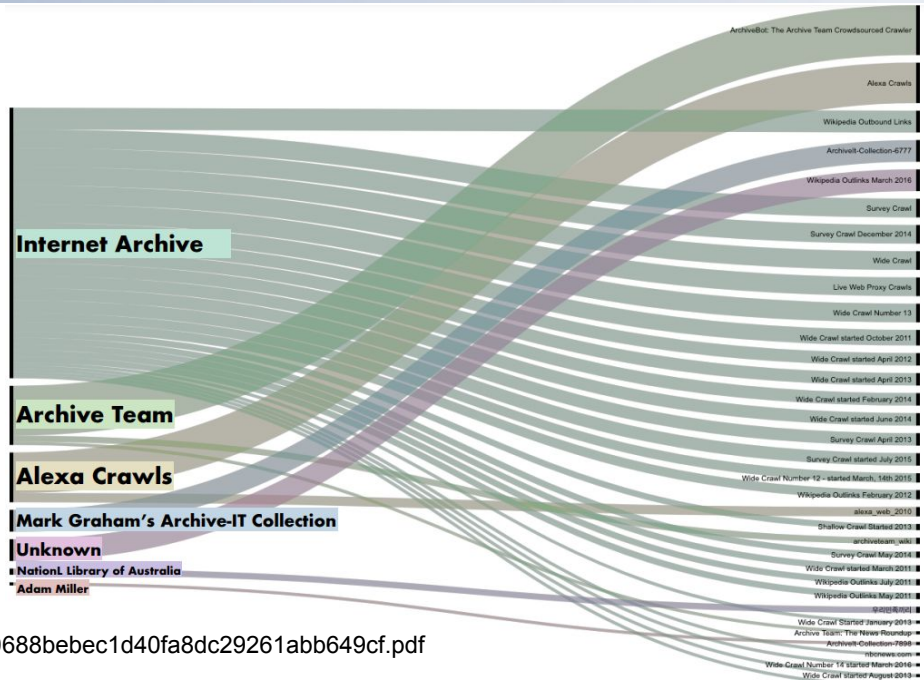
There are underlying social factors that influence how URIs get added to the Wayback Machine



Adding poignant seeds
(Personal collections)

Alexa browsing history
Wikipedia outlinks

Many different sources from before 2016 contain North Korean mementos



http://static.wixstatic.com/ugd/bfb070_70688bebec1d40fa8dc29261abb649cf.pdf

Mark Graham actively archives North Korean websites

https://www.archive-it.org/collections/6777
https://twitter.com/MarkGraham/status/1056992161244147712/photo/1
https://www.reddit.com/r/IAmA/comments/9sgf4z/hey_reddit_we_are_the_folks_behind_the_internet/

North Korea

Collected by: Mark Graham

Archived since: Jan, 2016

Description: A crawl of web sites about/from North Korea

Subject: Government - Counties

internet_archive **OP** · 4 yr. ago

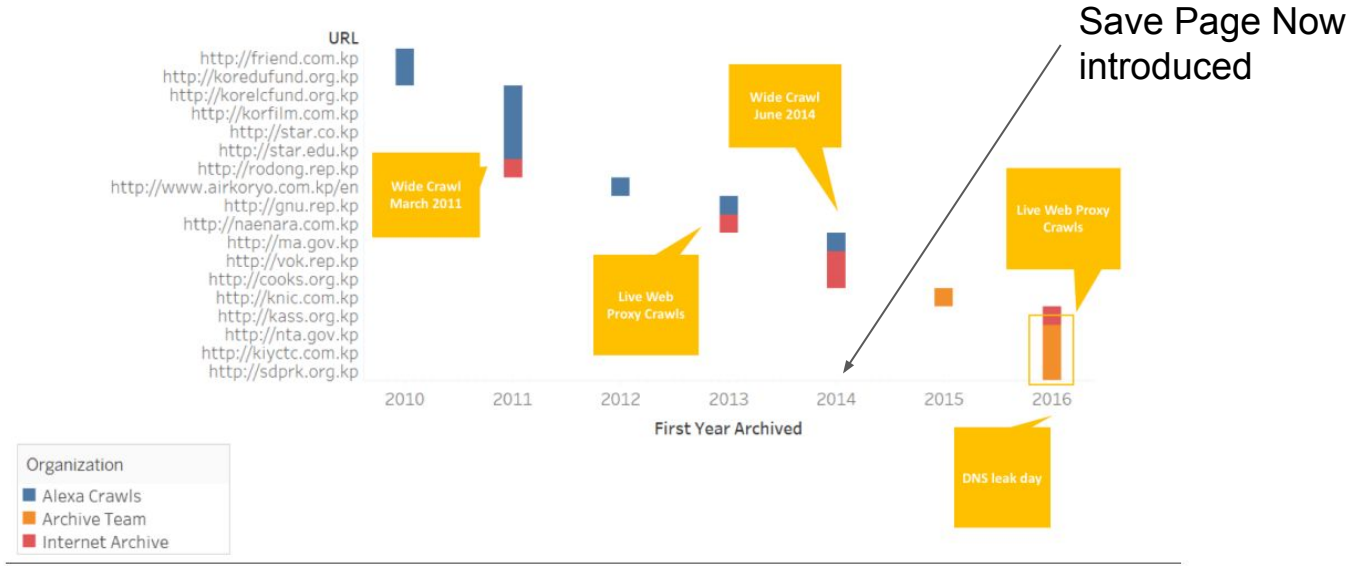
I am especially passionate about this archive of web content from and about North Korea:
https://archive-it.org/collections/6777

- Mark

Archive Team is a community that anticipates websites at-risk of shutdown and archives them

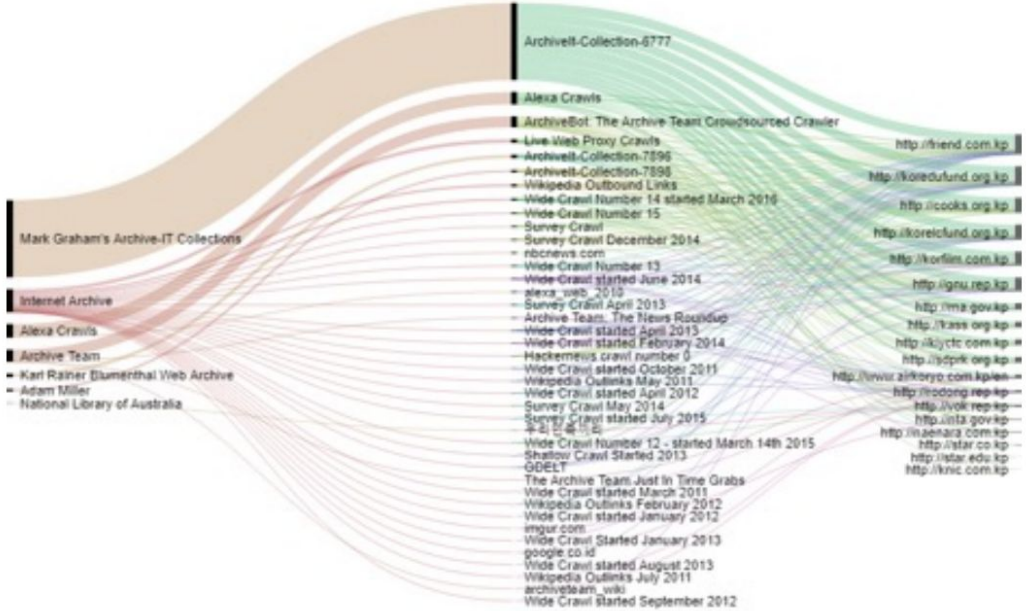
https://www.youtube.com/watch?v=-2ZTmuX3cog

North Korean websites with recent first mementos were archived proactively



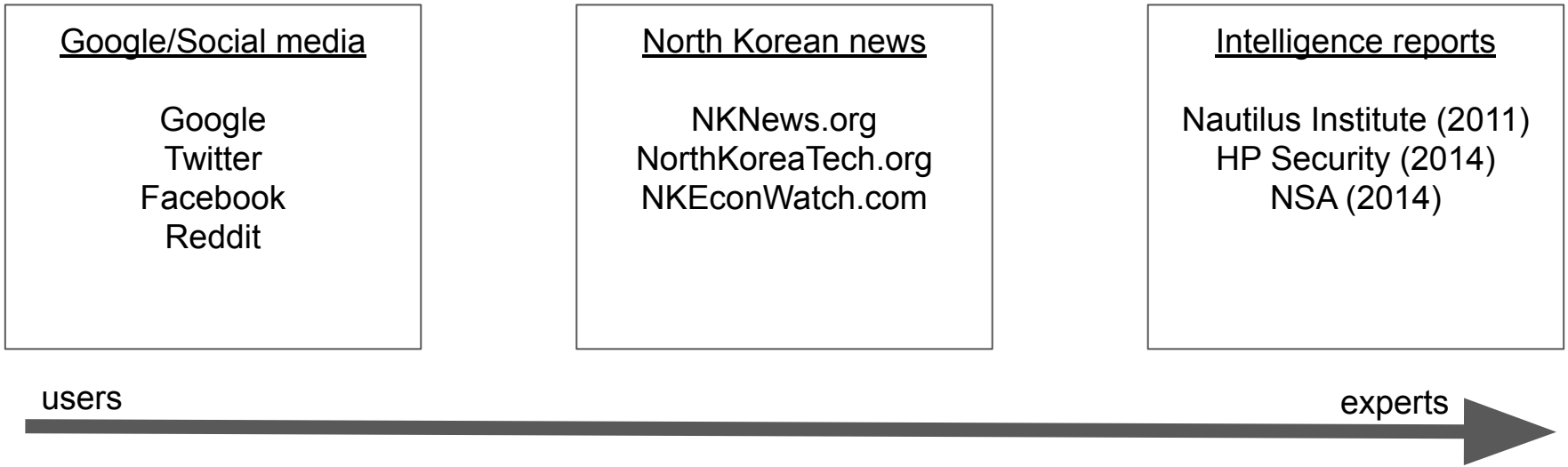
Ben-David et al., Figure 4

Humans were the main contributors of North Korean mementos before 2016



Ben-David et al., Figure 3

How did humans discover the North Korean websites before the 2016 leak?



http://www.nautilus.org/wp-content/uploads/2011/12/DPRK_Digital_Transformation.pdf
https://time.com/wp-content/uploads/2014/12/hpsr_securitybriefing_episode16_northkorea.pdf

Not all of the websites were archived before the leak

First IA snapshot at leak

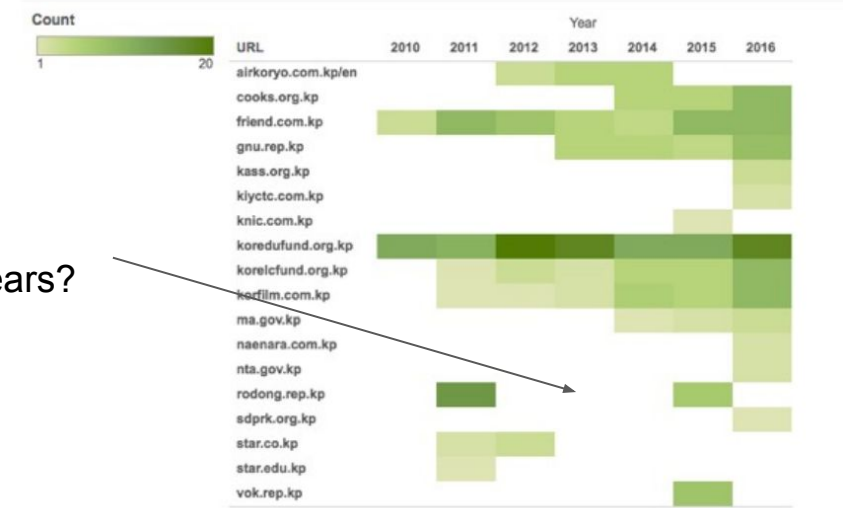
Archive.today 2015

Transient domain names

Ben-David et al., Figure 1

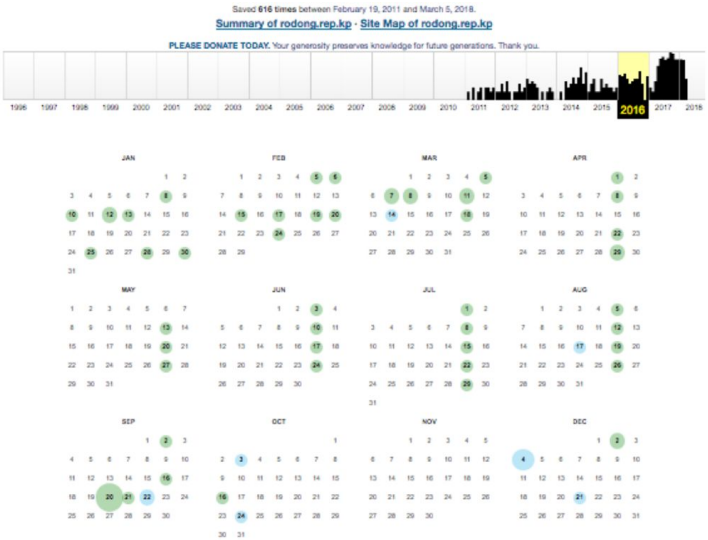
Pages with pre-2016 mementos were all discoverable from shallow crawls

Why are there gaps spanning multiple years?



Ben-David et al., Figure 1

Why do so many North Korean website mementos have non-successful HTTP status codes?



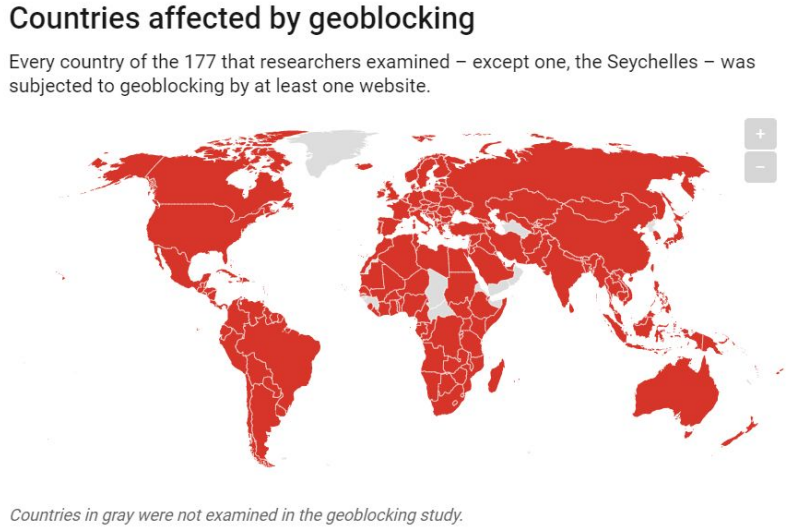
Ben-David et al., Figure 7

Non-successful HTTP response codes can be indicative of country-level censorship

403 ERROR

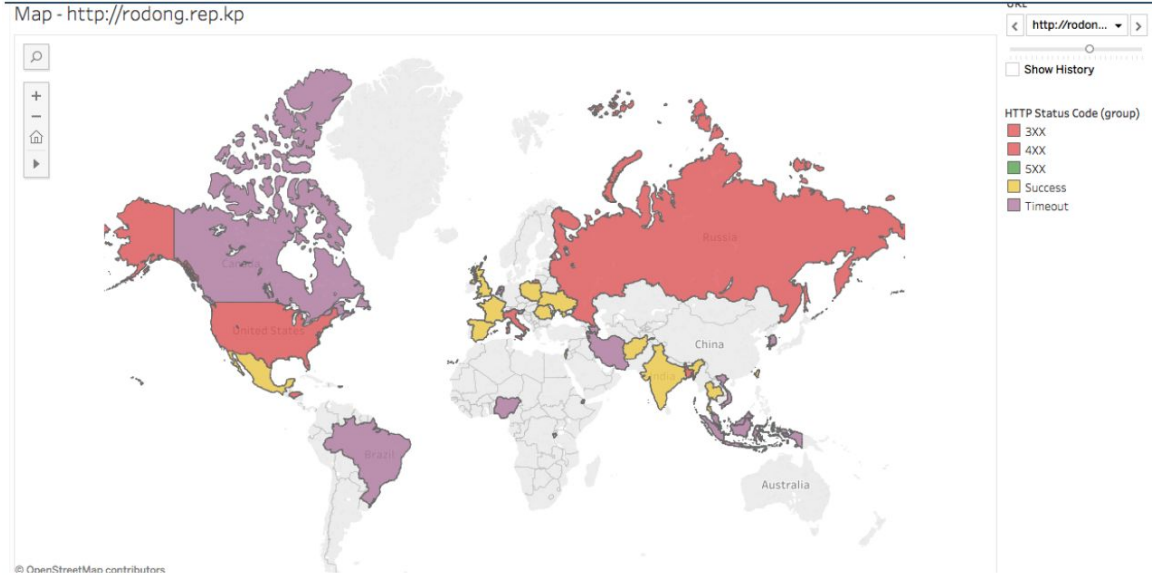
The request could not be satisfied.

The Amazon CloudFront distribution is configured to block access from your country.



https://theconversation.com/the-web-really-isnt-worldwide-every-country-has-different-access-106739

North Korean websites have different visibilities in different countries



Ben-David et al., Figure 6

Both internal and external country-level policies contribute to the visibility of North Korean websites

Geo Analysis

Country	Proxy type	Success	Request timeout	HTTP response code							
				302	401	403	404	500	502	503	504
Azerbaijan	Transparent	23					3				
Great Britain (UK)	Anonymous	14	3	1					1	7	
Russian Federation	Transparent	11	3				2		10		
Korea (South)	Transparent	7	16							1	2
Italy	Transparent	5				20		1			
Brazil	Transparent	2	22				2				
Iran	Anonymous	2	17							7	
United States	Transparent	2	19				1				

https://www.slideshare.net/anatbd/the-internet-archive-and-the-sociotechnical-construction-of-historical-facts

Conclusion: Humans and robots together make the Wayback Machine greater than the sum of its parts

- The provenance of a memento is important at both the micro and macro levels
- Humans proactively archive historically important content more rapidly than it can be discovered via crawls
- Web archiving is affected by geopolitics