Building Event Collections from Crawling Web Archives

Martin Klein[1]
Lyudmila Balakireva[1]
Herbert Van de Sompel[2]

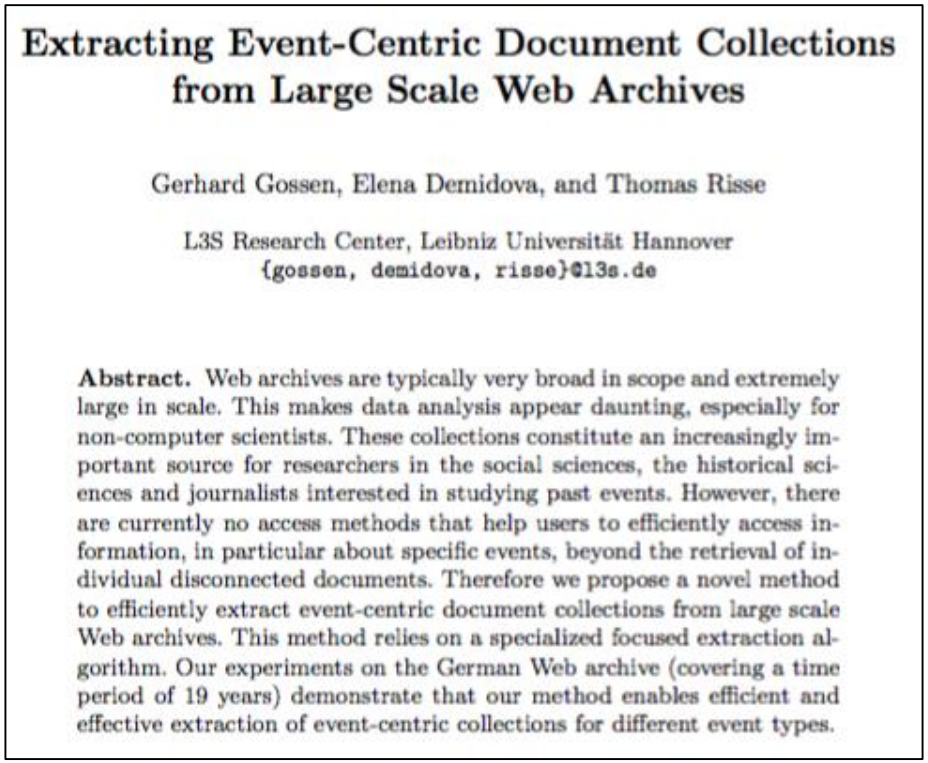
[1]Research Library
Los Alamos National Laboratory

[2]Data Archiving and Networked Services
The Netherlands

CC BY SA

Los Alamos NATIONAL LABORATORY

Inspiration from Previous Work



Extracting Event-Centric Document Collections from Large Scale Web Archives

Gerhard Gossen, Elena Demidova, and Thomas Risse

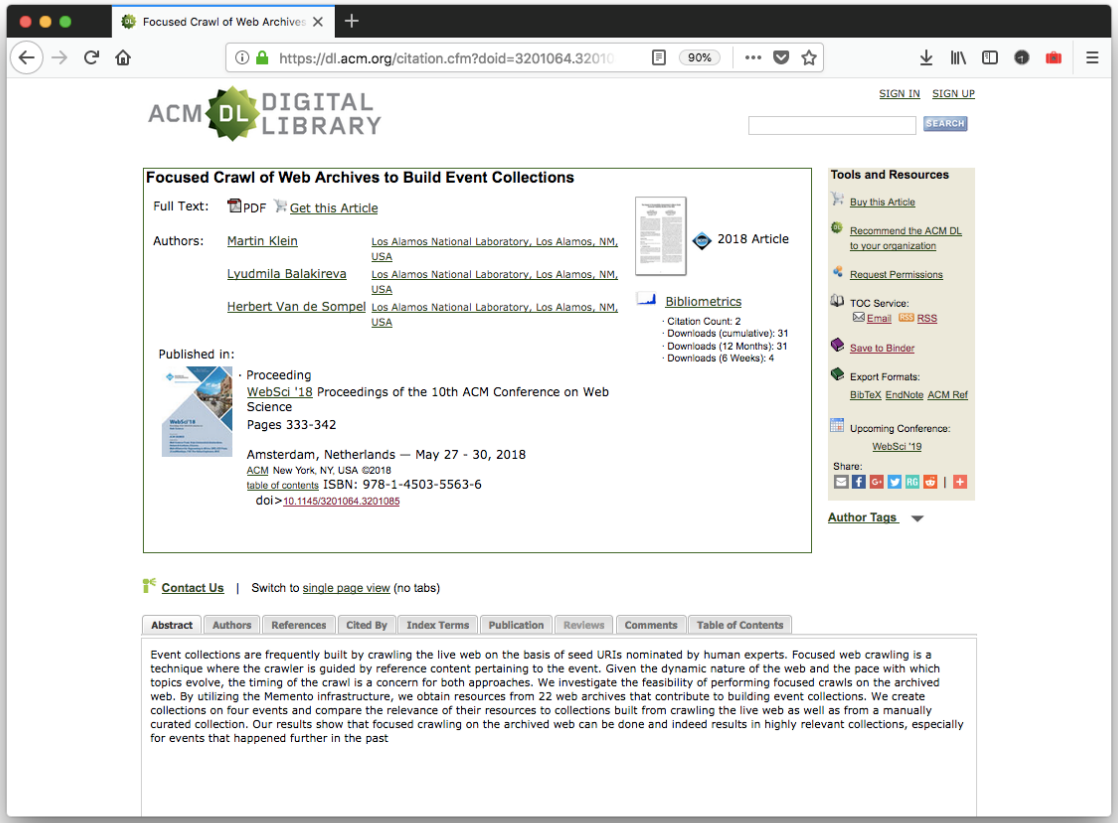
L3S Research Center, Leibniz Universität Hannover
{gossen, demidova, risse}@l3s.de

Abstract. Web archives are typically very broad in scope and extremely large in scale. This makes data analysis appear daunting, especially for non-computer scientists. These collections constitute an increasingly important source for researchers in the social sciences, the historical sciences and journalists interested in studying past events. However, there are currently no access methods that help users to efficiently access information, in particular about specific events, beyond the retrieval of individual disconnected documents. Therefore we propose a novel method to efficiently extract event-centric document collections from large scale Web archives. This method relies on a specialized focused extraction algorithm. Our experiments on the German Web archive (covering a time period of 19 years) demonstrate that our method enables efficient and effective extraction of event-centric collections for different event types.

https://doi.org/10.1007/978-3-319-67008-9_10

CC BY SA

Published at WebSci 2018



https://doi.org/10.1145/3201064.3201085

CC BY SA

Los Alamos NATIONAL LABORATORY

Questions

1. Can we create event collections by focused crawling online-available web archives?

2. How do event collections created from the archived web compare to those created from the live web?

3. How does the amount of time passed since the event affect the collections built from the live and the archived web?

4. How do event collections built from the archived web compare to manually curated collections?

CC BY SA

Background – Event Collection Building

- Often orchestrated by subject matter experts, archivists, special collection librarians, technicians
- Potentially with guidance from institutional collection policy
- Results in a list of seeds (URIs, social media accounts, etc)
- Utilization of crawling services such as Archive-It, Social Feed Manager

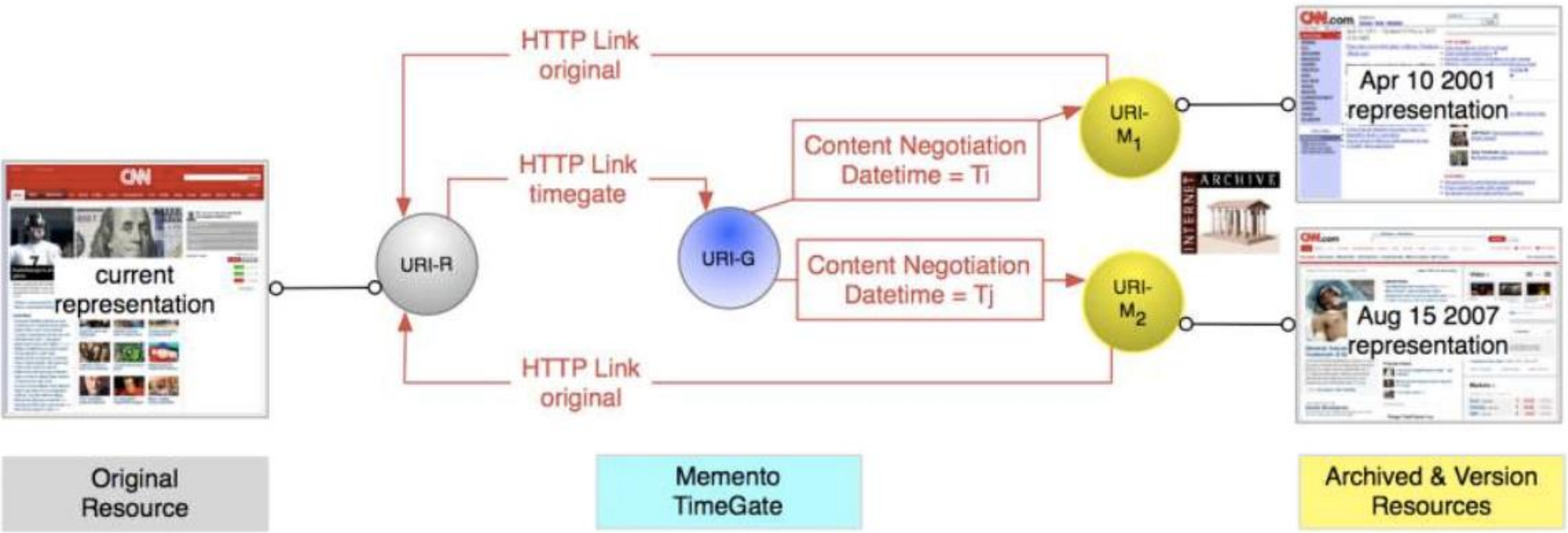
CC BY SA

Problems and our Approach

- Temporal: time passed since event is of concern
  → Use of web archives via Memento infrastructure
- Selection: seeds often picked manually
  → Use of references from Wikipedia pages
- Relevance: seed assessment often done by humans
  → Use of focused crawling with content and temporal relevance assessment

CC BY SA

Memento Framework



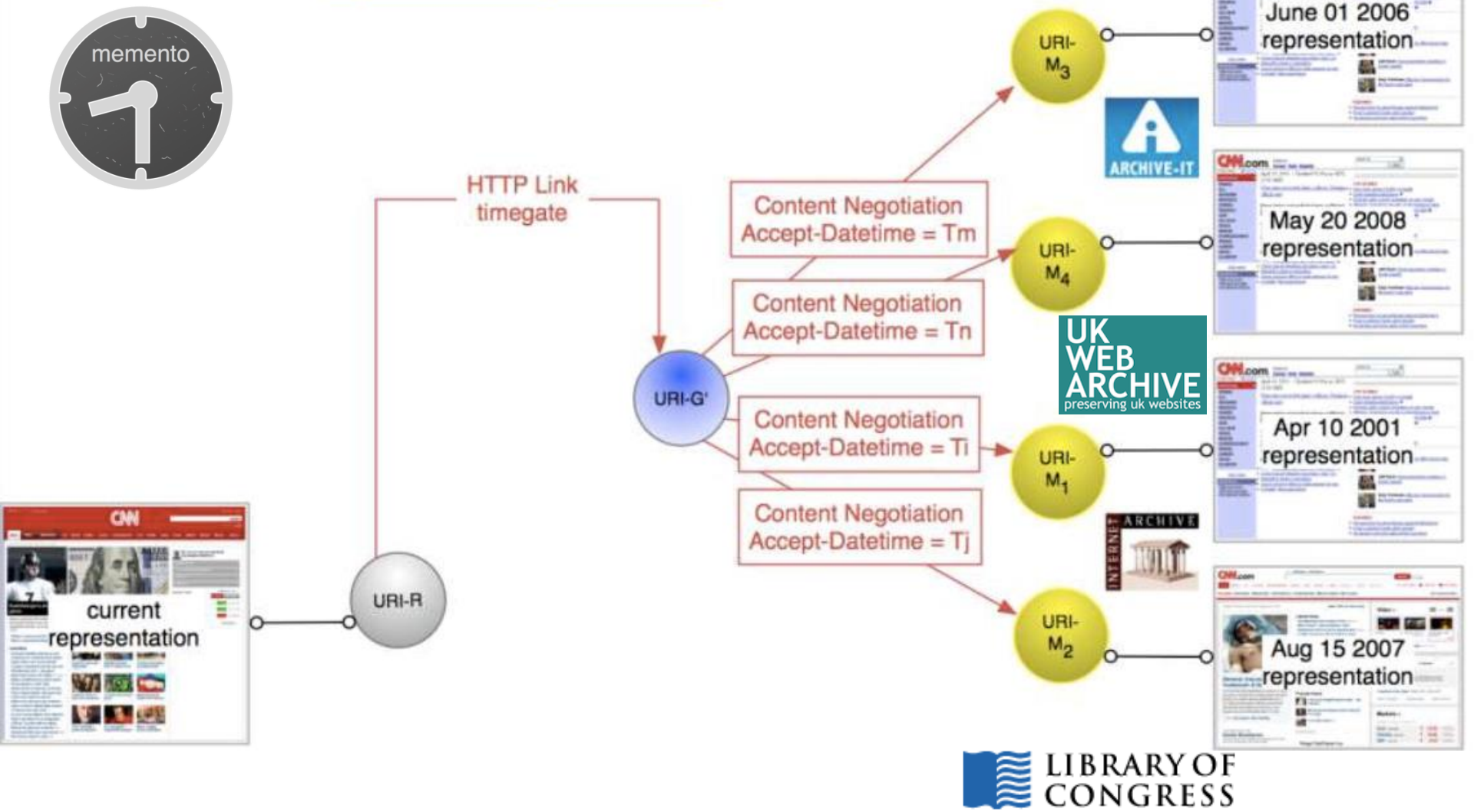
CC BY SA

Los Alamos NATIONAL LABORATORY

Integrating Multiple Archives

more info: http://mementoweb.org/
https://tools.ietf.org/html/rfc7089

archive.is
webpage capture



LIBRARY OF CONGRESS

CC BY SA

Building Event Collections from Crawling Web Archives
@mart1nkle1n
IIPC WAC 2018, 11/13/2018, Wellington, NZL

Los Alamos NATIONAL LABORATORY

Problems and our Approach

- Temporal: time passed since event is of concern
  → Use of web archives
- Selection: seeds often picked manually
  → Use of references from Wikipedia pages
- Relevance: seed assessment often done by humans
  → Use of focused crawling with content and temporal relevance assessment

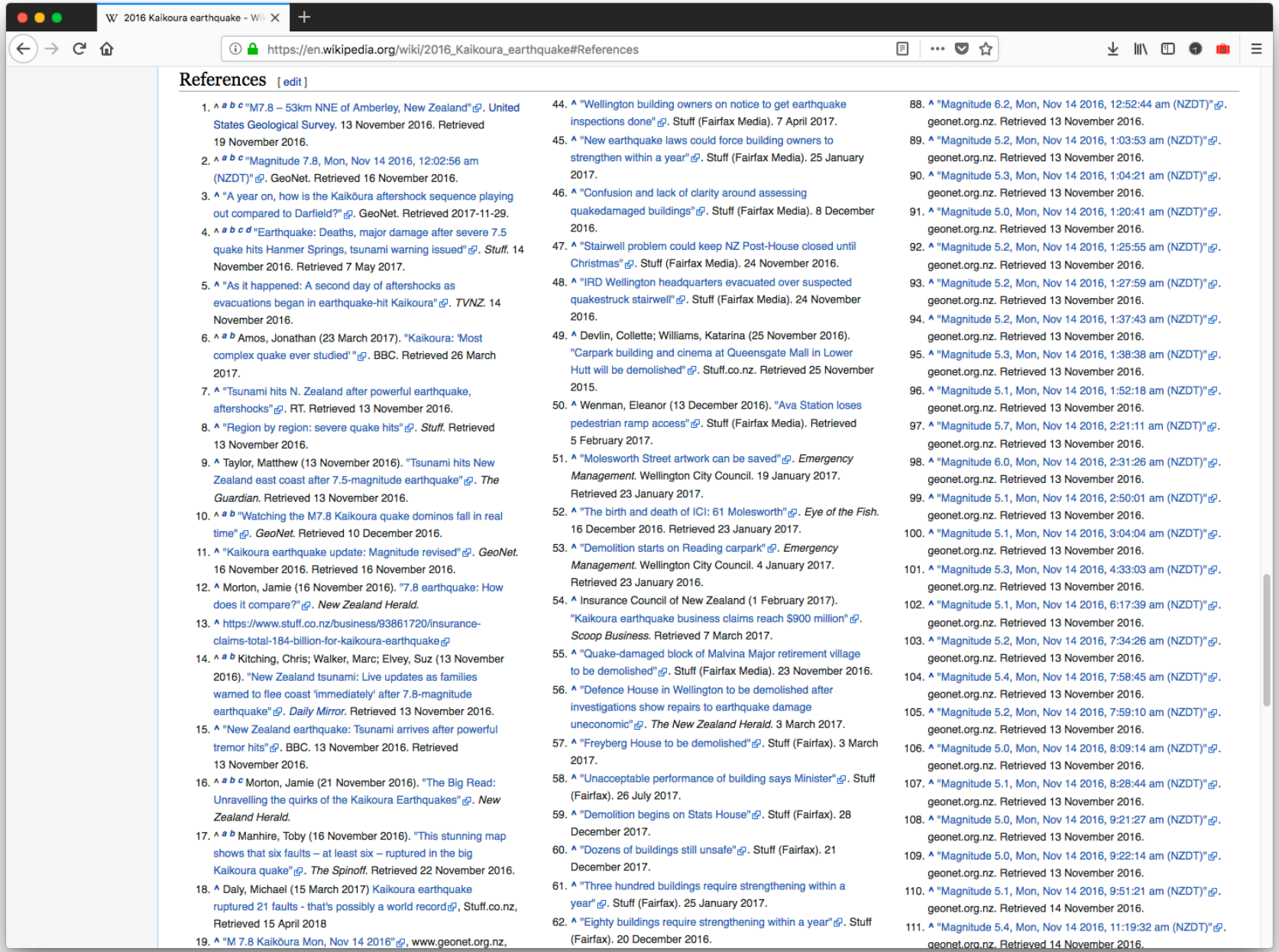
CC BY SA

CC BY SA

Building Event Collections from Crawling Web Archives

@mart1nkle1n

IIPC WAC 2018, 11/13/2018, Wellington, NZL



References [edit]

1. ^ a b c "M7.8 – 53km NNE of Amberley, New Zealand". United States Geological Survey. 13 November 2016. Retrieved 19 November 2016.
2. ^ a b c "Magnitude 7.8, Mon, Nov 14 2016, 12:02:56 am (NZDT)". GeoNet. Retrieved 16 November 2016.
3. ^ "A year on, how is the Kaikōura aftershock sequence playing out compared to Darfield?". GeoNet. Retrieved 2017-11-29.
4. ^ a b c d "Earthquake: Deaths, major damage after severe 7.5 quake hits Hanmer Springs, tsunami warning issued". *Stuff*. 14 November 2016. Retrieved 7 May 2017.
5. ^ "As it happened: A second day of aftershocks as evacuations began in earthquake-hit Kaikoura". *TVNZ*. 14 November 2016.
6. ^ a b Amos, Jonathan (23 March 2017). "Kaikoura: 'Most complex quake ever studied' ". BBC. Retrieved 26 March 2017.
7. ^ "Tsunami hits N. Zealand after powerful earthquake, aftershocks". RT. Retrieved 13 November 2016.
8. ^ "Region by region: severe quake hits". *Stuff*. Retrieved 13 November 2016.
9. ^ Taylor, Matthew (13 November 2016). "Tsunami hits New Zealand east coast after 7.5-magnitude earthquake". *The Guardian*. Retrieved 13 November 2016.
10. ^ a b "Watching the M7.8 Kaikoura quake dominos fall in real time". *GeoNet*. Retrieved 10 December 2016.
11. ^ "Kaikoura earthquake update: Magnitude revised". *GeoNet*. 16 November 2016. Retrieved 16 November 2016.
12. ^ Morton, Jamie (16 November 2016). "7.8 earthquake: How does it compare?". *New Zealand Herald*.
13. ^ https://www.stuff.co.nz/business/93861720/insurance-claims-total-184-billion-for-kaikoura-earthquake
14. ^ a b Kitching, Chris; Walker, Marc; Elvey, Suz (13 November 2016). "New Zealand tsunami: Live updates as families warned to flee coast 'immediately' after 7.8-magnitude earthquake". *Daily Mirror*. Retrieved 13 November 2016.
15. ^ "New Zealand earthquake: Tsunami arrives after powerful tremor hits". BBC. 13 November 2016. Retrieved 13 November 2016.
16. ^ a b c Morton, Jamie (21 November 2016). "The Big Read: Unravelling the quirks of the Kaikoura Earthquakes". *New Zealand Herald*.
17. ^ a b Manhire, Toby (16 November 2016). "This stunning map shows that six faults – at least six – ruptured in the big Kaikoura quake". *The Spinoff*. Retrieved 22 November 2016.
18. ^ Daly, Michael (15 March 2017) Kaikoura earthquake ruptured 21 faults - that's possibly a world record, Stuff.co.nz, Retrieved 15 April 2018
19. ^ "M 7.8 Kaikōura Mon, Nov 14 2016", www.geonet.org.nz,
44. ^ "Wellington building owners on notice to get earthquake inspections done". Stuff (Fairfax Media). 7 April 2017.
45. ^ "New earthquake laws could force building owners to strengthen within a year". Stuff (Fairfax Media). 25 January 2017.
46. ^ "Confusion and lack of clarity around assessing quakedamaged buildings". Stuff (Fairfax Media). 8 December 2016.
47. ^ "Stairwell problem could keep NZ Post-House closed until Christmas". Stuff (Fairfax Media). 24 November 2016.
48. ^ "IRD Wellington headquarters evacuated over suspected quakestruck stairwell". Stuff (Fairfax Media). 24 November 2016.
49. ^ Devlin, Collette; Williams, Katarina (25 November 2016). "Carpark building and cinema at Queensgate Mall in Lower Hutt will be demolished". Stuff.co.nz. Retrieved 25 November 2015.
50. ^ Wenman, Eleanor (13 December 2016). "Ava Station loses pedestrian ramp access". Stuff (Fairfax Media). Retrieved 5 February 2017.
51. ^ "Molesworth Street artwork can be saved". *Emergency Management*. Wellington City Council. 19 January 2017. Retrieved 23 January 2017.
52. ^ "The birth and death of ICI: 61 Molesworth". *Eye of the Fish*. 16 December 2016. Retrieved 23 January 2017.
53. ^ "Demolition starts on Reading carpark". *Emergency Management*. Wellington City Council. 4 January 2017. Retrieved 23 January 2016.
54. ^ Insurance Council of New Zealand (1 February 2017). "Kaikoura earthquake business claims reach $900 million". *Scoop Business*. Retrieved 7 March 2017.
55. ^ "Quake-damaged block of Malvina Major retirement village to be demolished". Stuff (Fairfax Media). 23 November 2016.
56. ^ "Defence House in Wellington to be demolished after investigations show repairs to earthquake damage uneconomic". *The New Zealand Herald*. 3 March 2017.
57. ^ "Freyberg House to be demolished". Stuff (Fairfax). 3 March 2017.
58. ^ "Unacceptable performance of building says Minister". Stuff (Fairfax). 26 July 2017.
59. ^ "Demolition begins on Stats House". Stuff (Fairfax). 28 December 2017.
60. ^ "Dozens of buildings still unsafe". Stuff (Fairfax). 21 December 2017.
61. ^ "Three hundred buildings require strengthening within a year". Stuff (Fairfax). 25 January 2017.
62. ^ "Eighty buildings require strengthening within a year". Stuff (Fairfax). 20 December 2016.
88. ^ "Magnitude 6.2, Mon, Nov 14 2016, 12:52:44 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
89. ^ "Magnitude 5.2, Mon, Nov 14 2016, 1:03:53 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
90. ^ "Magnitude 5.3, Mon, Nov 14 2016, 1:04:21 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
91. ^ "Magnitude 5.0, Mon, Nov 14 2016, 1:20:41 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
92. ^ "Magnitude 5.2, Mon, Nov 14 2016, 1:25:55 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
93. ^ "Magnitude 5.2, Mon, Nov 14 2016, 1:27:59 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
94. ^ "Magnitude 5.2, Mon, Nov 14 2016, 1:37:43 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
95. ^ "Magnitude 5.3, Mon, Nov 14 2016, 1:38:38 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
96. ^ "Magnitude 5.1, Mon, Nov 14 2016, 1:52:18 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
97. ^ "Magnitude 5.7, Mon, Nov 14 2016, 2:21:11 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
98. ^ "Magnitude 6.0, Mon, Nov 14 2016, 2:31:26 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
99. ^ "Magnitude 5.1, Mon, Nov 14 2016, 2:50:01 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
100. ^ "Magnitude 5.1, Mon, Nov 14 2016, 3:04:04 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
101. ^ "Magnitude 5.3, Mon, Nov 14 2016, 4:33:03 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
102. ^ "Magnitude 5.1, Mon, Nov 14 2016, 6:17:39 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
103. ^ "Magnitude 5.2, Mon, Nov 14 2016, 7:34:26 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
104. ^ "Magnitude 5.4, Mon, Nov 14 2016, 7:58:45 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
105. ^ "Magnitude 5.2, Mon, Nov 14 2016, 7:59:10 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
106. ^ "Magnitude 5.0, Mon, Nov 14 2016, 8:09:14 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
107. ^ "Magnitude 5.1, Mon, Nov 14 2016, 8:28:44 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
108. ^ "Magnitude 5.0, Mon, Nov 14 2016, 9:21:27 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
109. ^ "Magnitude 5.0, Mon, Nov 14 2016, 9:22:14 am (NZDT)". geonet.org.nz. Retrieved 13 November 2016.
110. ^ "Magnitude 5.1, Mon, Nov 14 2016, 9:51:21 am (NZDT)". geonet.org.nz. Retrieved 14 November 2016.
111. ^ "Magnitude 5.4, Mon, Nov 14 2016, 11:19:32 am (NZDT)". geonet.org.nz. Retrieved 14 November 2016.

CC BY SA

Building Event Collections from Crawling Web Archives

@mart1nkle1n

IIPC WAC 2018, 11/13/2018, Wellington, NZL

Los Alamos NATIONAL LABORATORY

Problems and our Approach

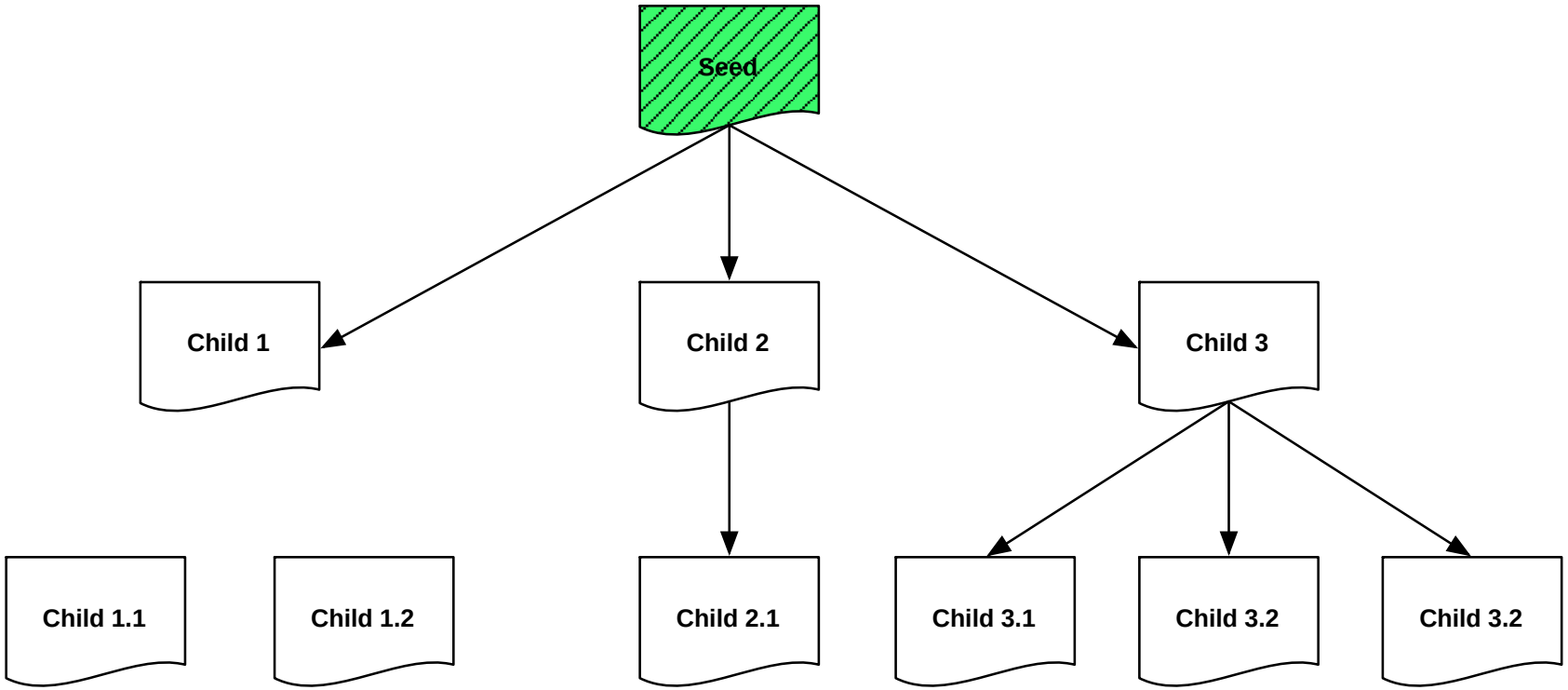
- Temporal: time passed since event is of concern
  → Use of web archives

- Selection: seeds often picked manually
  → Use of references from Wikipedia pages

- Relevance: seed assessment often done by humans
  → Use of focused crawling with content and temporal relevance assessment

CC BY SA

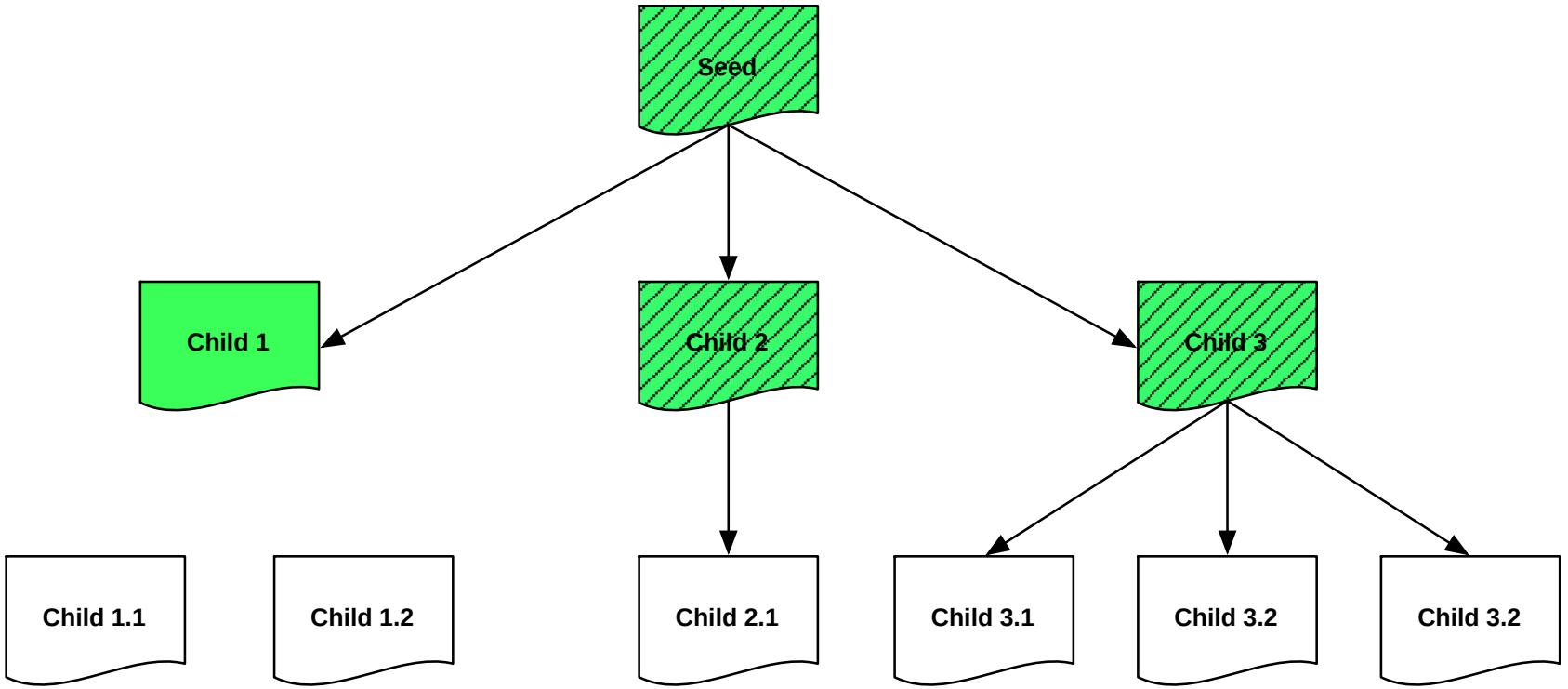
Focused Crawling



CC BY SA

Los Alamos NATIONAL LABORATORY

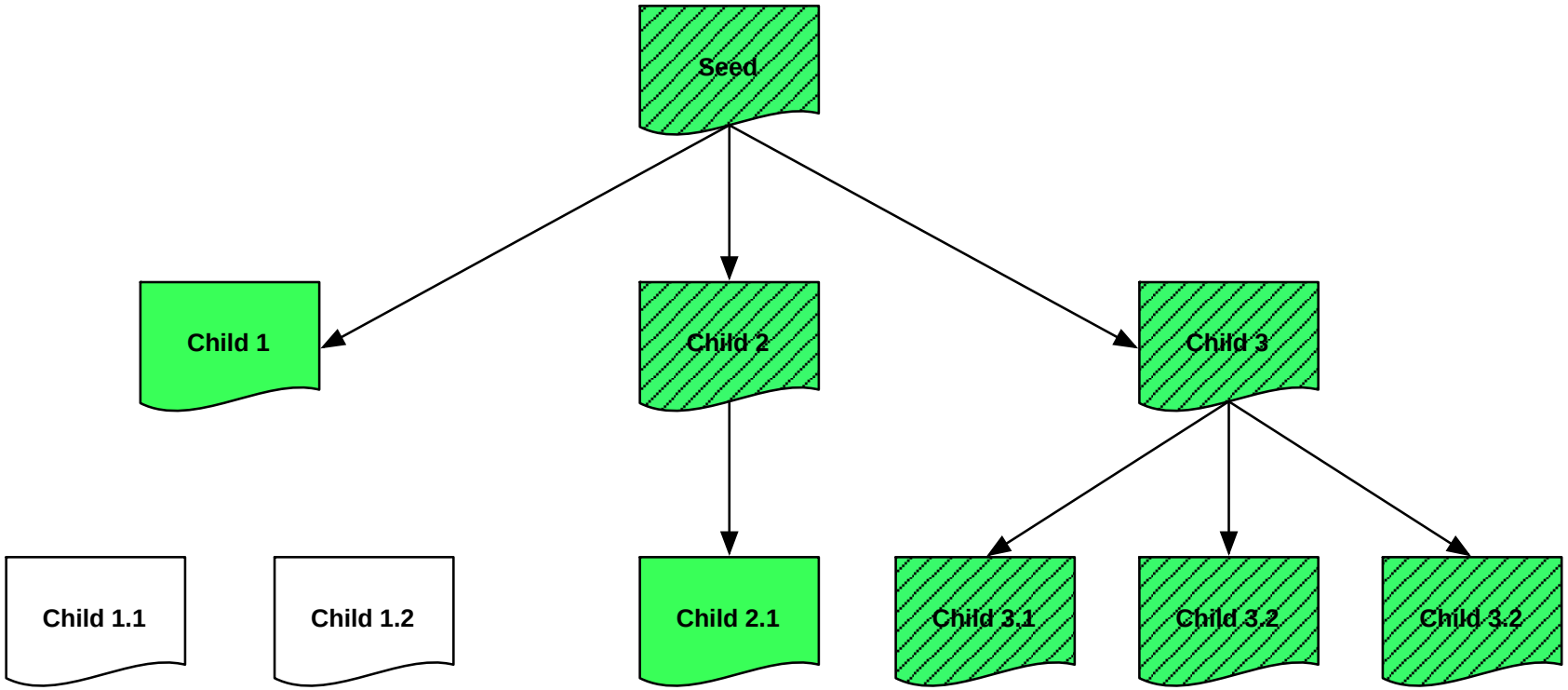
Focused Crawling



CC BY SA

Los Alamos NATIONAL LABORATORY

Focused Crawling



CC BY SA

Los Alamos NATIONAL LABORATORY

Content Relevance

1. Content of Wikipedia page + random 60% of page's references
   - Generate topic vector (TF-IDF of 1grams + 2grams)

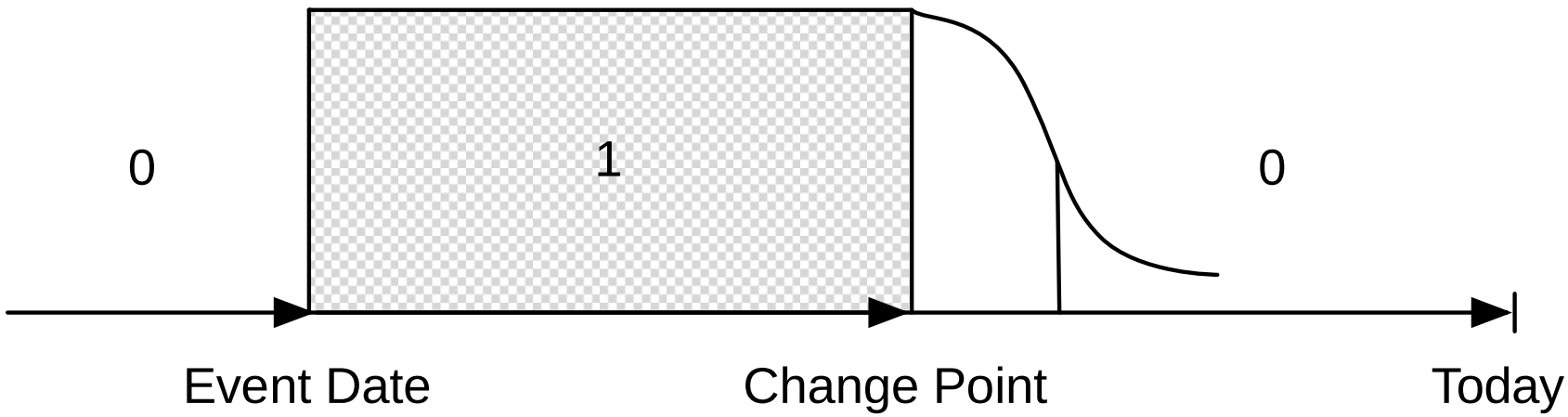
2. Content of remaining 40% of Wikipedia page's references
   - Generate topic vector (TF-IDF of 1grams + 2grams)

- Compute cosine similarity value between vectors 1 and 2
- Run 10 times
- Take average cosine similarity value as content threshold

CC BY SA

Temporal Relevance

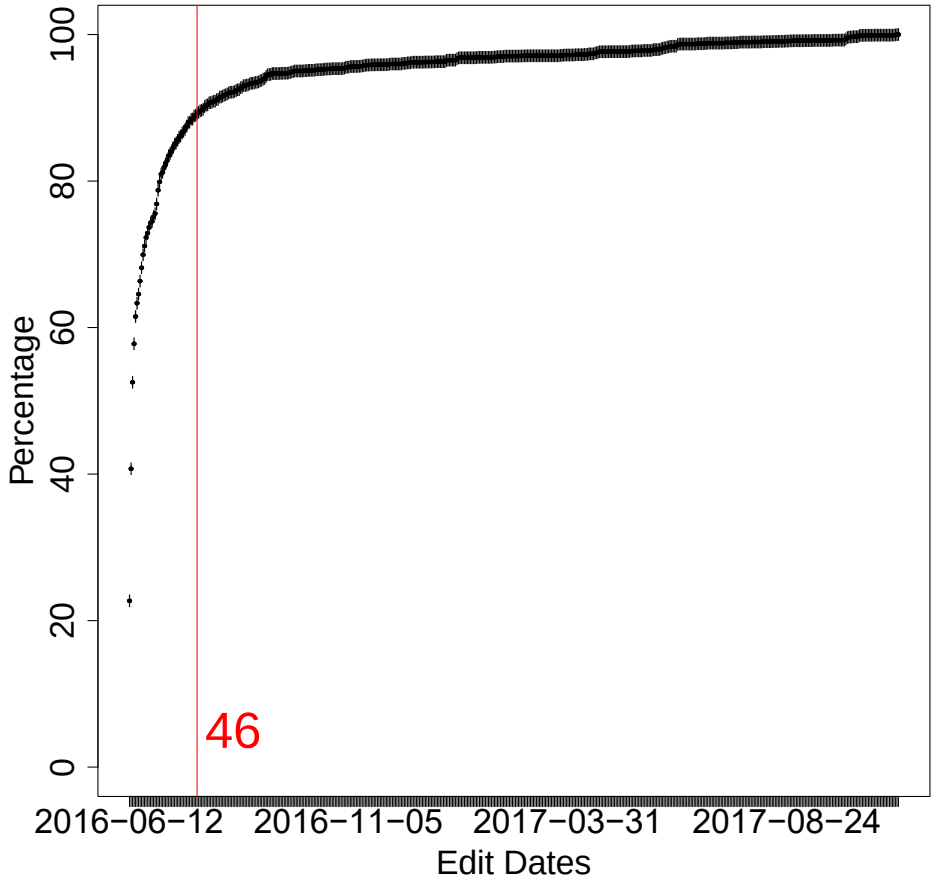
- Define temporal interval for which crawled pages are considered relevant
- Event date extracted from Wikipedia event page



CC BY SA

Los Alamos NATIONAL LABORATORY

Change Point Detection



- Plot number of Wikipedia page edits per day
- Run R's changepoint algorithm
- Detect significant change in curve

https://cran.r-project.org/web/packages/changepoint/index.html

CC BY SA

Datetime Extraction

- Extract datetime from pages via:
  - URI
    http://www.cnn.com/2017/12/09/us/wildfire-fighting-tactics/
  - Meta tags
    <meta property="article:published" itemprop="datePublished" content="2017-12-09T10:14:50-05:00" />
  - ODU's Carbondate tool
    http://carbondate.cs.odu.edu/
  - Memento datetime
  - X-Header

CC BY SA

Experiment Details

- Topics limited to terror attacks and mass shootings in the U.S.
  - From different times in the past
- Take content and temporal relevance into account
  - Equally weighted
- Use events' Wikipedia page as input for focused crawler
  - Version that was live at change point

CC BY SA

Los Alamos NATIONAL LABORATORY

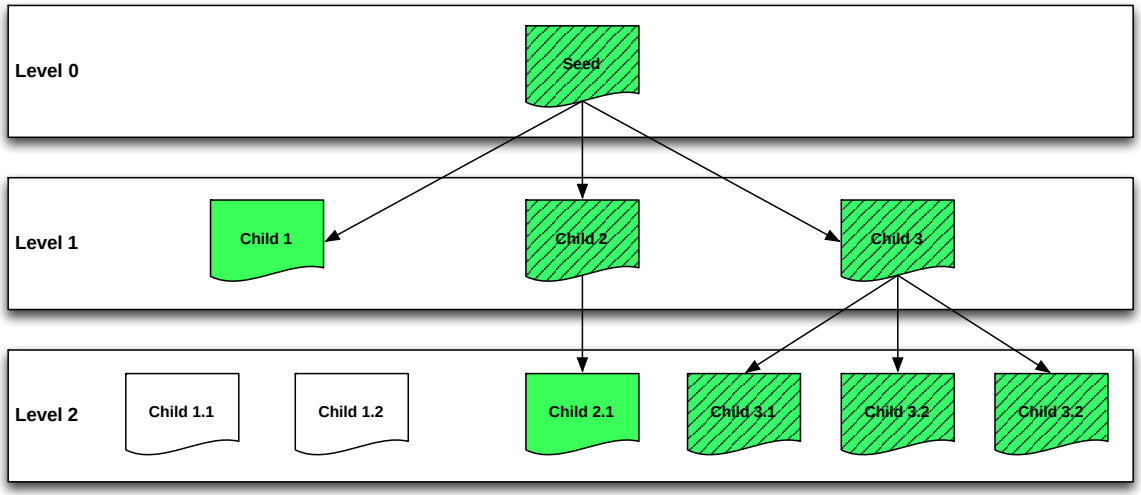
Crawl Details

- Focused crawl of:
  - 22 archives, simultaneously, via Memento infrastructure
  - The live web
- Seeds
  - Memento of Wikipedia page references closest to and after event time
  - Subject to temporal and contextual relevance assessment
- Crawled outlinks
  - Memento of outlinks closest to and after event time
  - Subject to temporal and contextual relevance assessment

CC BY SA

Crawl Details

- Crawl stop conditions:
  - No more relevant documents left
  - 5 levels deep
- Utilized crawl priority queue



CC BY SA

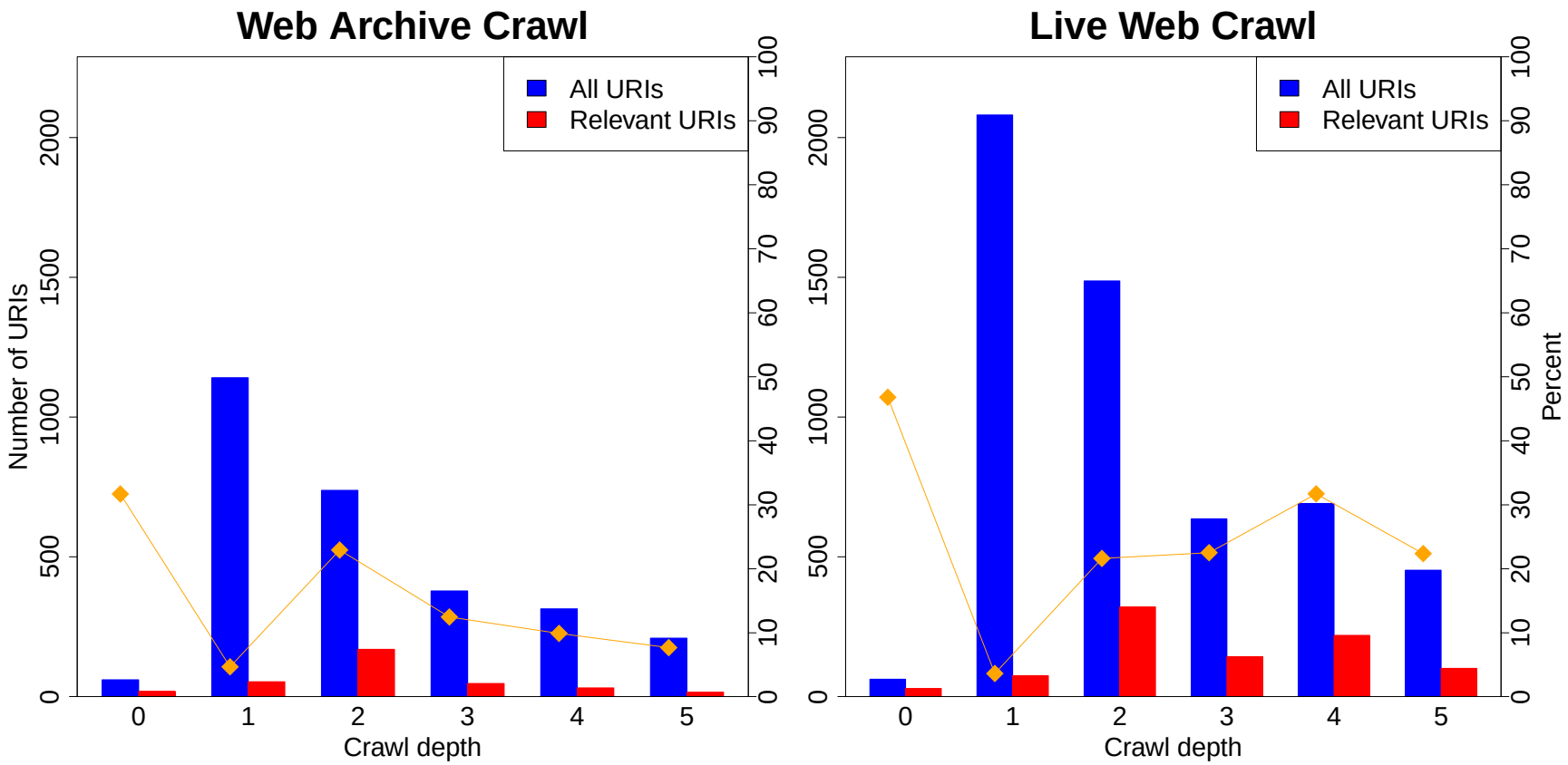
Los Alamos NATIONAL LABORATORY

Collections Crawled (in November 2017)

- **New York City, October 31st 2017**
- Las Vegas, October 1st 2017
- Orlando, June 12th 2016
- San Bernadino, December 2nd 2015
- **Tucson, January 8th 2011**
- Binghampton, April 3rd 2009

CC BY SA

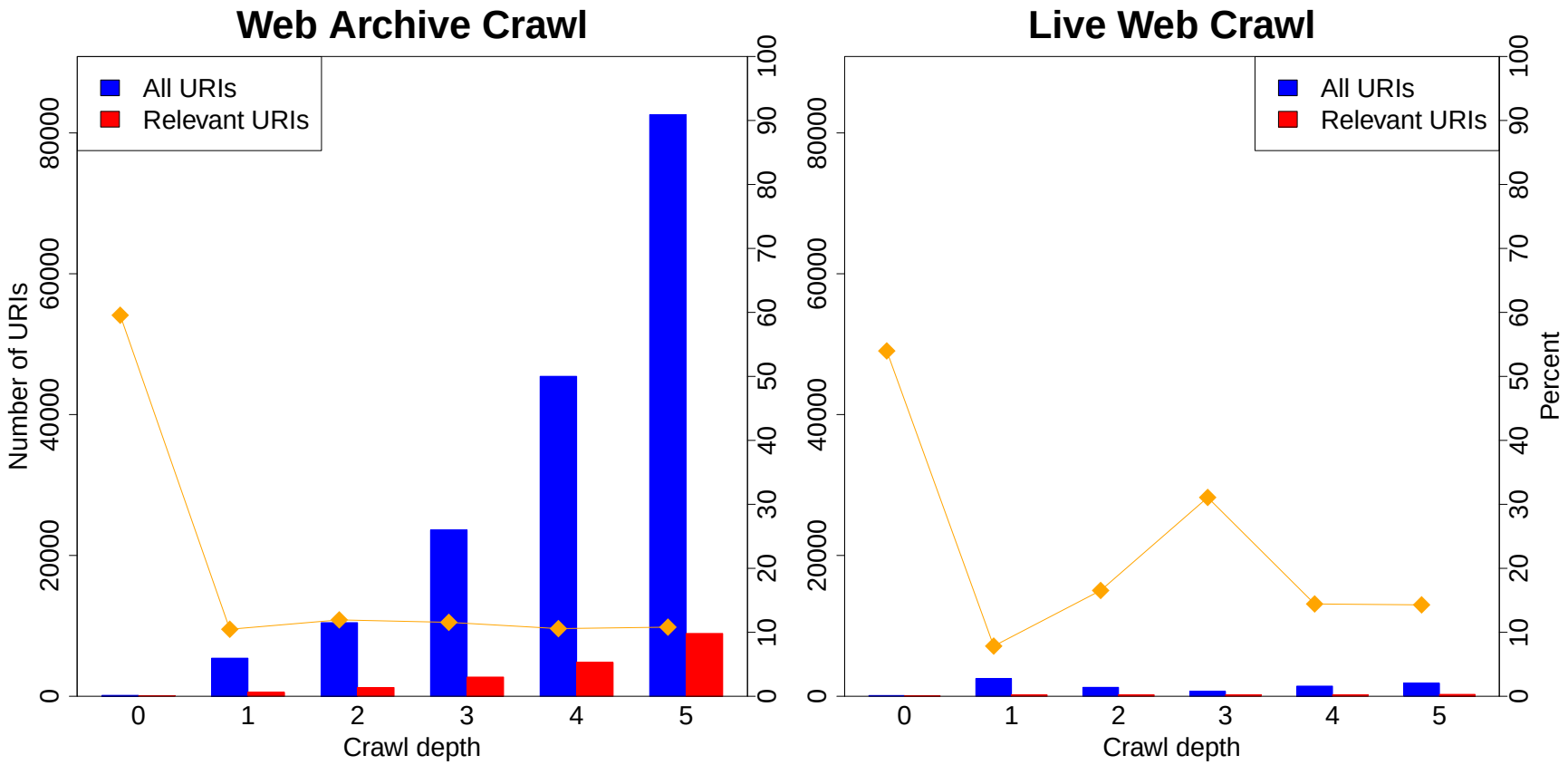
NYC, 10/31/2017 – URIs per Level



CC BY SA

Los Alamos NATIONAL LABORATORY

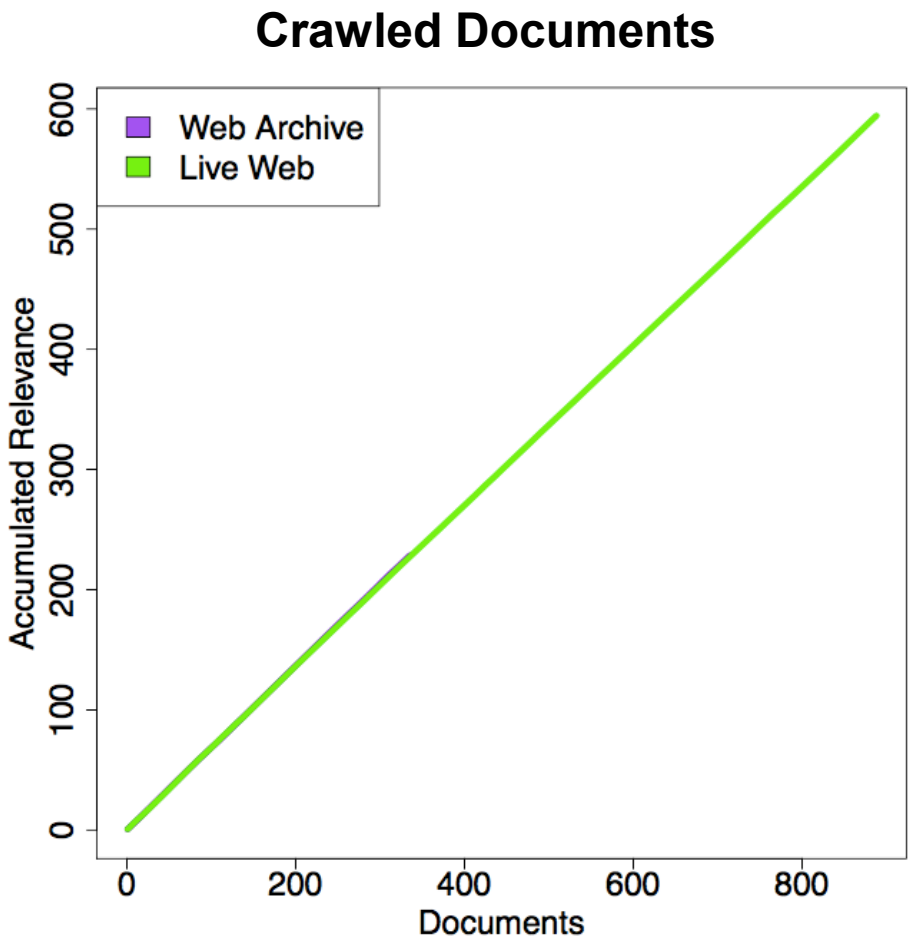
TUC, 01/08/2011 – URIs per Level

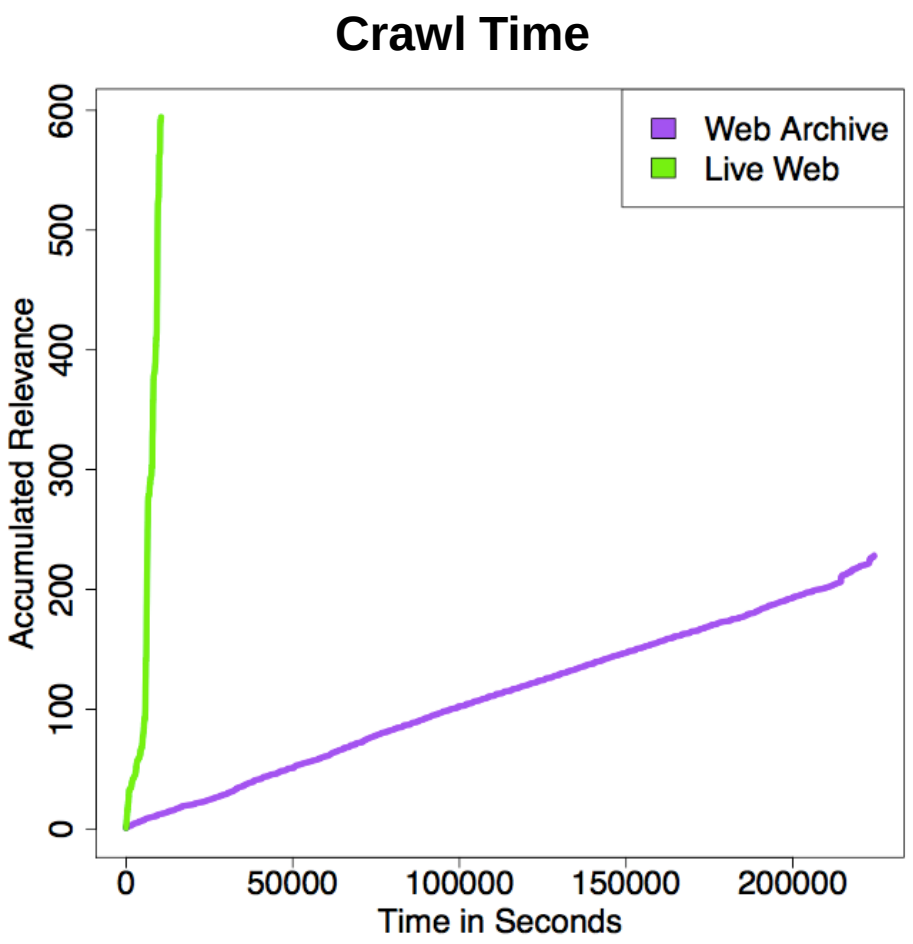


CC BY SA

Los Alamos NATIONAL LABORATORY

NYC, 10/31/2017 – Relevance over…



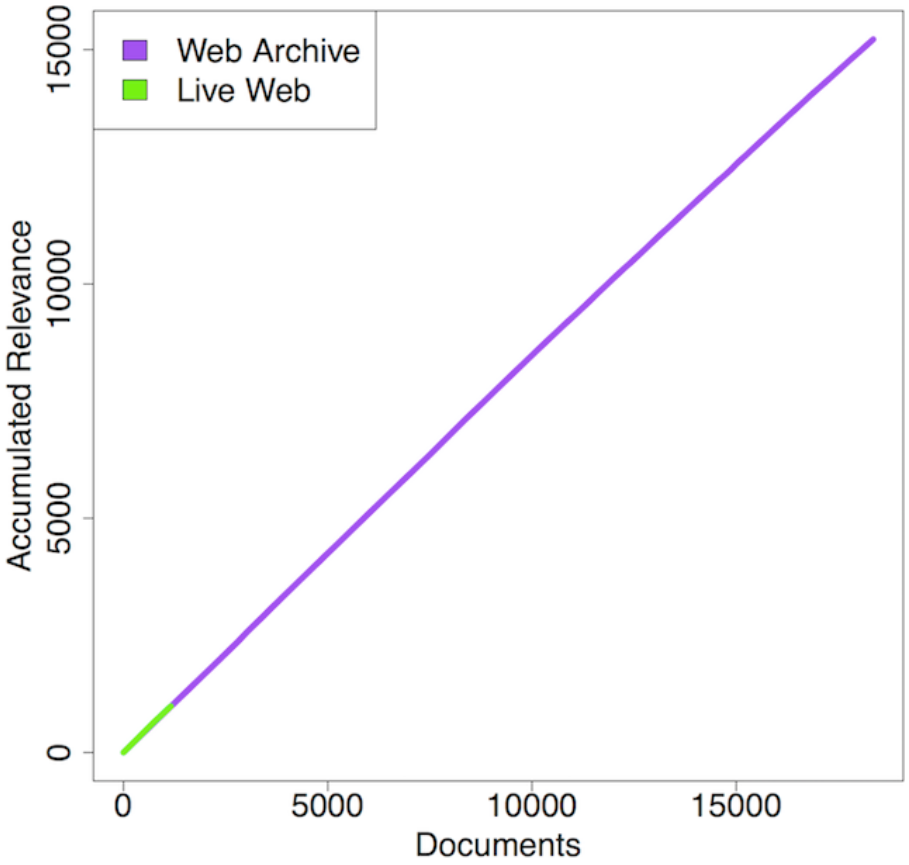


CC BY SA

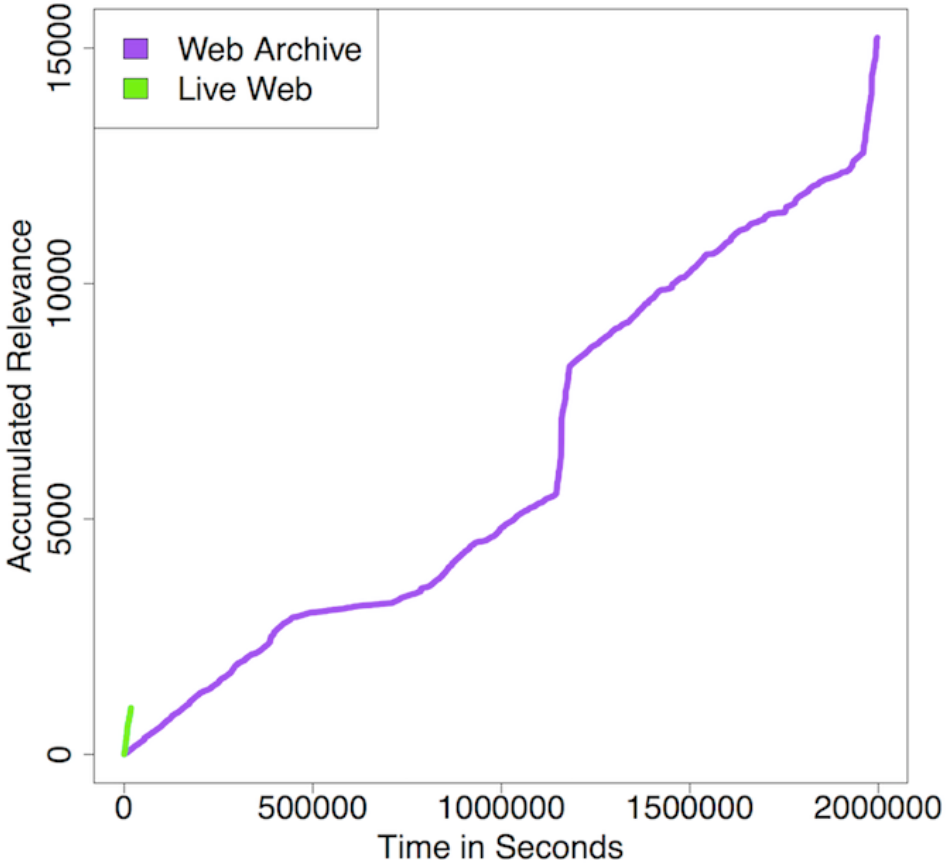
Los Alamos NATIONAL LABORATORY

TUC, 01/08/2011 – Relevance over…

Crawled Documents



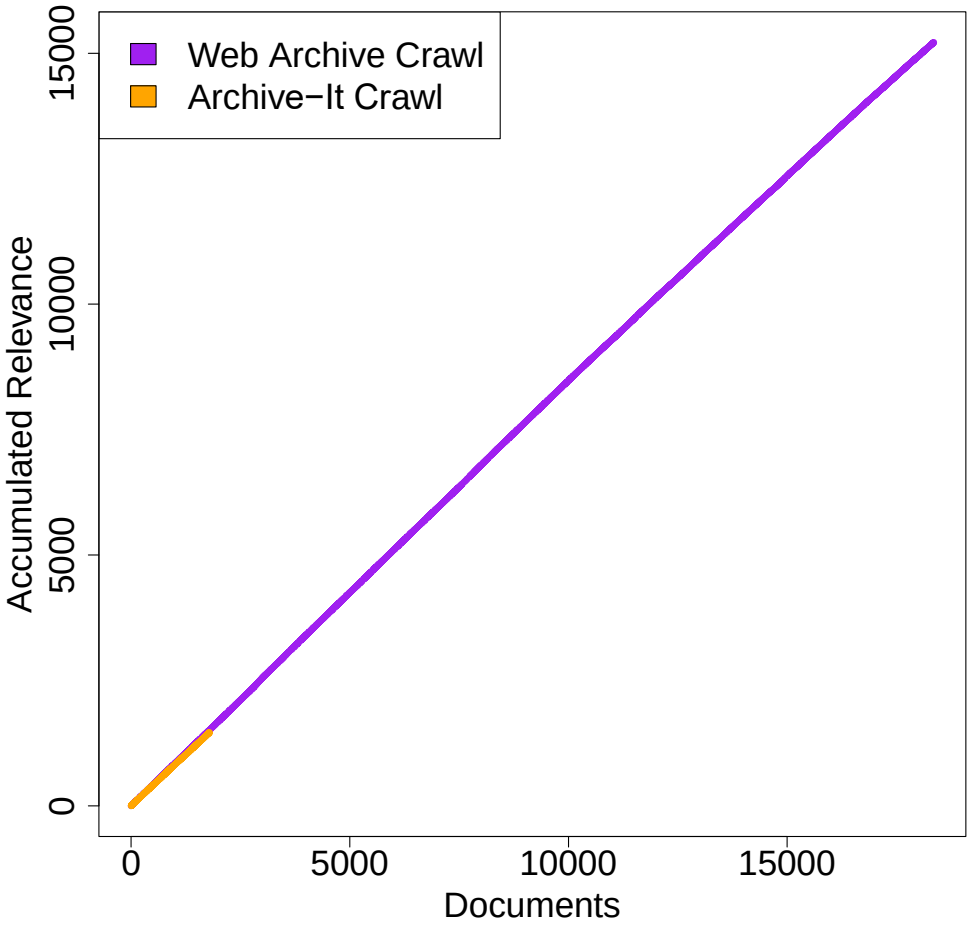
Crawl Time



CC BY SA

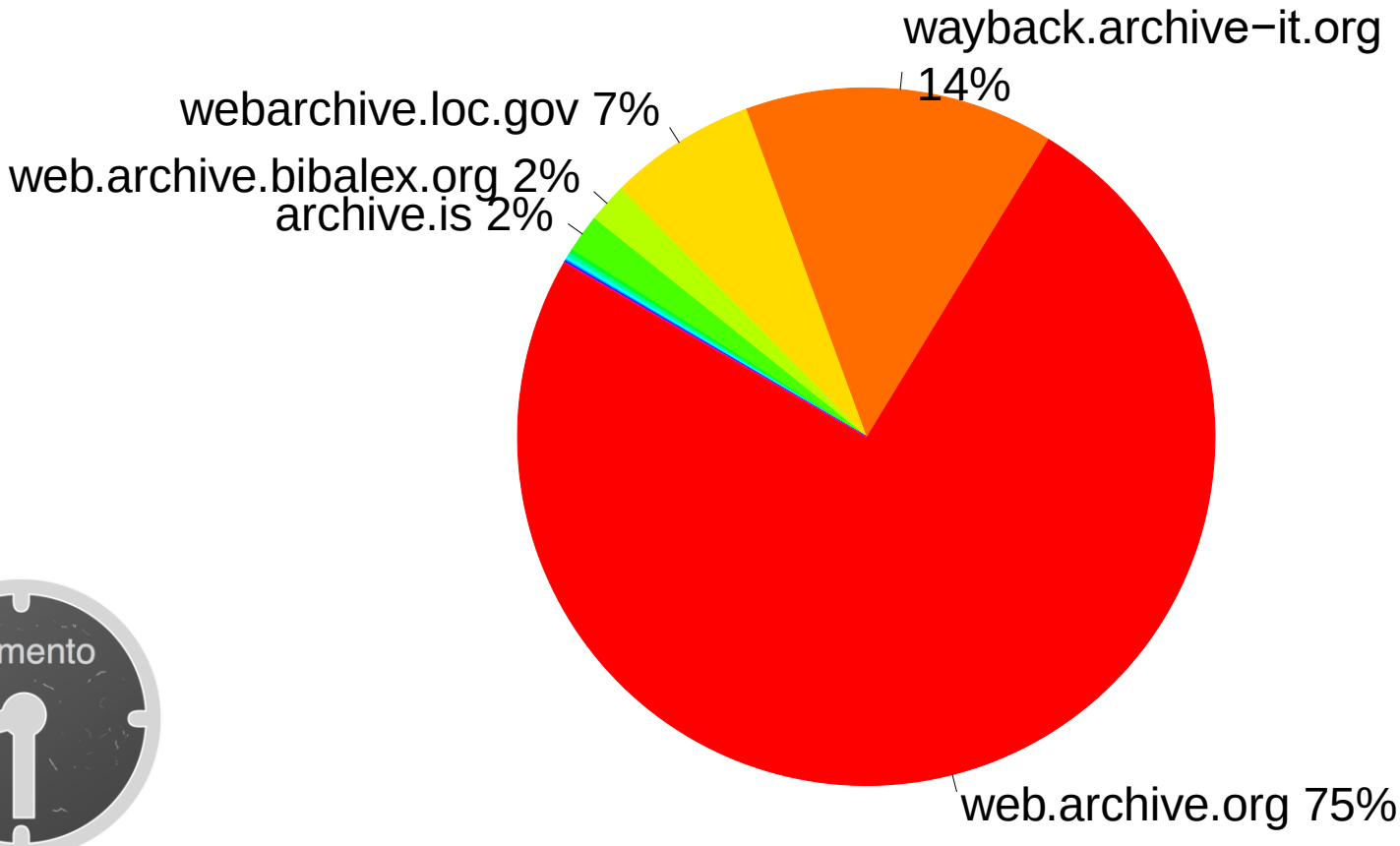
Los Alamos NATIONAL LABORATORY

TUC, 01/08/2011 – Comparison to Archive-IT



CC BY SA

TUC, 01/08/2011 – Web Archive Contributions



CC BY SA

Los Alamos NATIONAL LABORATORY

Takeaways

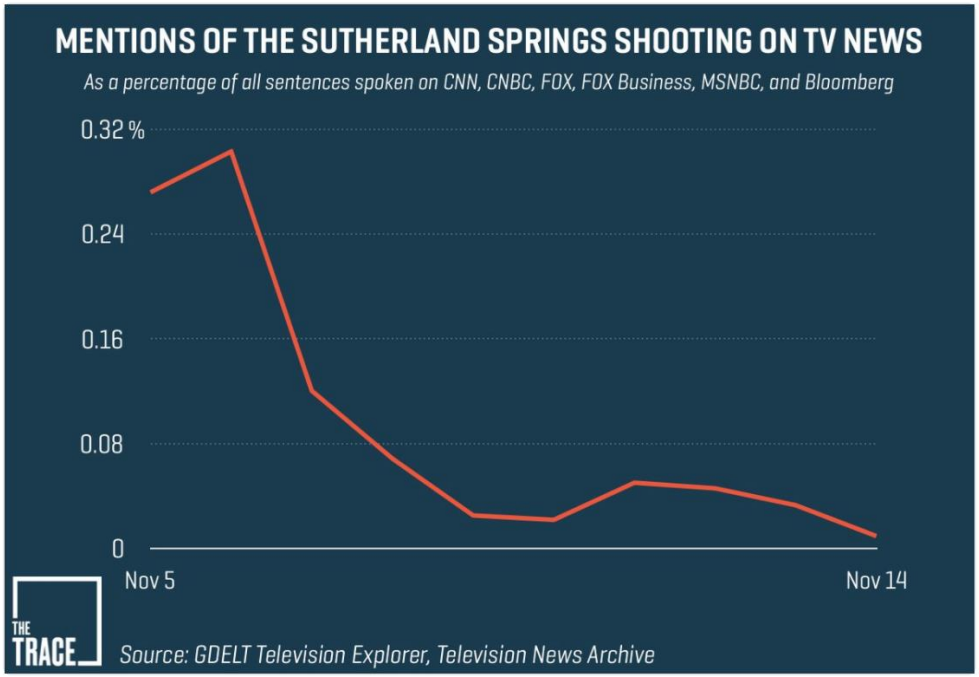
- Web archives are great resources to build event collections of web resources
- Crawling web archives is much slower than the live web
- Collections about very recent events benefit more from the live web than the archived web
  but
- Collections about events from the distant past benefit more from the archived web than the live web
- Utilizing multiple web archives is beneficial for the collection
- Focused crawls have the potential to outperform manual collection building

CC BY SA

TV News Archive
@TVNewsArchive

Follow

"Two days after 26 people were massacred in a Texas church, the incident — one of the worst mass shootings in American history — had nearly vanished from the major cable news networks." @teamtrace thetrace.org/rounds/texas-c …



10:54 AM - 6 Dec 2017

https://web.archive.org/web/20171206181955/https:/twitter.com/TVNewsArchive/status/938466726190096384

CC BY SA

Building Event Collections from Crawling Web Archives

Martin Klein[1]
Lyudmila Balakireva[1]
Herbert Van de Sompel[2]

[1]Research Library
Los Alamos National Laboratory

[2]Data Archiving and Networked Services
The Netherlands

CC BY SA

Los Alamos NATIONAL LABORATORY