

# When Web Archiving Meets Corpus Creation

The Web Harvesting Project of the National laboratory of Digital Heritage

BALÁZS INDIG, ZSÓFIA SÁRKÖZI-LINDNER, MIHÁLY NAGY

EÖTVÖS LORÁND UNIVERSITY, DEPARTMENT OF DIGITAL HUMANITIES  
NATIONAL LABORATORY FOR DIGITAL HUMANITIES

LASTNAME.FIRSTNAME@BTK.ELTE.HU

18 July 2023

# Web archiving: Are we there yet?

**Definition:** *“Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use.”* (IIPC)

- Which portions?
- Who collects?
- Who will have access and when?
- What is it good for?

There are plenty of web archives, because everybody is gone DIY.

# Web Corpus Creation: We're just "archiving"...

**Definition:** *"Web corpus creation is the process of collecting portions of the World Wide Web, preserving the collections in **a custom format (Pile, SPL, etc.), and then training models from them** for access and use."* (Me)

- Which portions?
- Who collects?
- Who will have access and when?
- What is it good for?
- **Is collection, publishing, etc. legal?**

There are plenty of web corpora, because everybody is gone DIY.

# Web archives/corpora are like chocolate



(Mr. Artúr Gombóc is a famous Hungarian character who likes every kind of chocolate)

# Web archives/corpora are like a box of chocolates... ...you never know what you're going to get

## **Huge public archives**

- Archive.org
- Common Crawl

## **Smaller private archives, on varying scale**

- National Libraries
- University research projects
- Indie researchers (e.g. PhD students, non-technical researchers)

After all, they are web archives, despite the substantial differences in their concepts  
But who is the target audience?

**Do they need only need the text from the archive (and sometimes the metadata)?**

# The ecosystem is imbalanced

<b>Producers</b>	<b>Consumers</b>
Huge tech companies	
Libraries	Start-ups
Research groups	Researchers
Individual people (technical)	Individual people (technical and non-technical)

- Using large archives/corpora requires large machinery (and skills to operate)
- Specific data cannot be acquired in larger batches (e.g. distant reading)
- No guarantees at all (e.g. completeness, legal issues, etc.)
- No aim to meet the individual requirements (vs. fragmentation of the ecosystem)

# Small scale vs. Large scale: Is the larger the better?

- Descriptive metadata: manual curation vs. machine learning
- Duplicated content means duplicated load on the server + deduplication
  - Assets should be separated from the HTML files (in the archives)
- Authenticity, reproducibility, presistency issues (Lendák, Indig, and Palkó 2022)
- Completely ignoring content creators
  - The vast majority of websites use Wordpress which could be made more crawler friendly
- Browser automation vs. traditional crawling
  
- Customisability
- Collation of small archives into a big one or filtering them (e.g. scaling up and down)
- Distribution, continuation of the archive/corpus creation process

# Our experience (with web archives as corpus)

- Web archives answer research questions originating from various fields
  - “Personal web archives” could be reused, if they were in the appropriate form
- Questions usually related to descriptive metadata
  - Portal
  - Time period
  - Topic/Author–Source
- People are satisfied with small scale archives
  - Actually creating “micro scale archives” for themselves manually (“personal web archives”)
  - They cannot handle large scale archives as they lack fine-grained search functions
- Almost nobody has machinery or technical skills
  - Recipes like “Install Spark for efficient distributed processing” are mostly a no-go
- Results usually raise more questions
  - Results come from web archives, but web archives are not presented, shared, etc.
  - The same happen with corpora (because legal reasons)

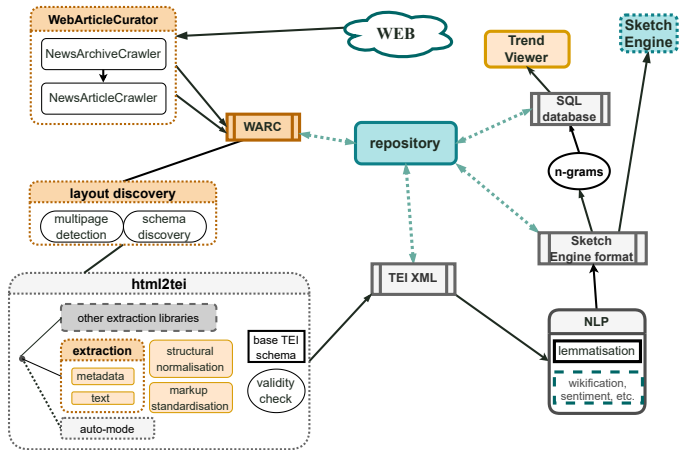


## Our solution: human-centered technology



(High-tech cucumber harvesting machine from Belarus)

## Our solution: from small to middle scale

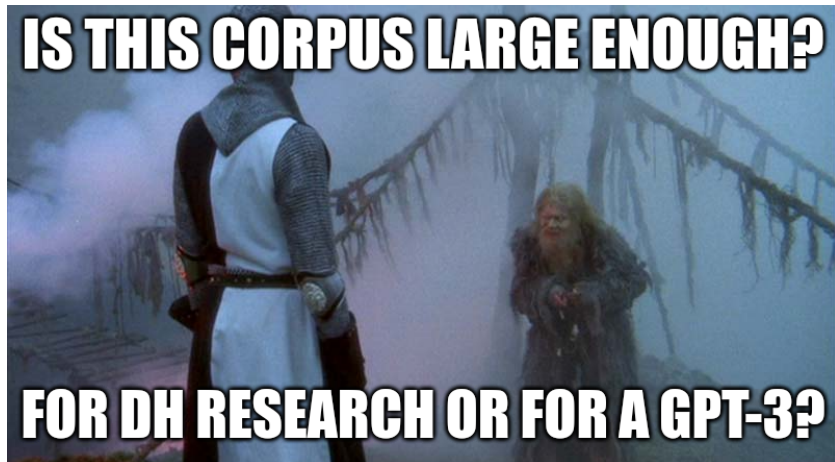


"I have of course done the traditional crawl as well, but this is my other slightly special personal one." (Twisted from Love actually, 2003)

# Our solution: from small to middle scale

- 34 datasets with DOI from 30 distinct portals
- 3 811 845 URL from 3 776 034 articles (and about 1,5 million half-baked not included)
- 1 182 100 108 tokens (about the size of HGC 2.0 (Oravecz, Váradi, and Sass 2014))
- From 22 years (1998-11 to 2021-06)
- Different genres
  - Forums
  - Transylvanian news portals
  - Covid-19 related news updates
  - From the far-right to the left different views
- **Manually curated!**
- In standard TEI XML format with descriptive metadata
- Uploaded into Zenodo repository and have DOI
- Sketch Engine corpus query service
- And more to come...

## King Arthur at the bridge of death



# The problems with our archive/corpus

- Too small/big, you name it!
- “TEI XML?! Just give me the text!”
- “Why don’t you crawl this and that portal or time period?”
- Still need technical skills to use...
- Not interesting as it does not answers questions alone
  - We are creating a GPT for this scenario ;)
  
- Lot of manual work
  - But now, we have the volume of data to experiment with AI and ML methods!
- Fragile: hard to maintain
- Does not work on dynamic Javascript heavy portals
  - Browser automation is the future!
- Cannot be connected to the mainstream methods and archives :(

# A trend viewer that goes against the trend

- Where are the once famous trend viewers? (e.g. Google n-gram viewer 2011)
  - No reliable way to extract temporal metadata automatically at scale from web pages
- We have temporal metadata. Let's visualise the trends in it!
  - We have other metadata too!
  - Why not classify results based on metadata?
- **We created a tool that answers questions, instead of raising them**
- Publication-ready graphs for a variety of questions without the need to write code
- KISS: n-grams and SQLite database to be super lightweight and built to last
  
- Demo scripts included: the easiest way to create custom data
  - Modular tools, which can be replaced if needed
  - Web crawling and NLP, the easy way

## A trend viewer that goes against the trend (cont.)

- Input data (the web archive) is simple as possible
  - Ideal for “personal web archives”
- Standard format: small, independent data sets can be merged or filtered
  - Allows crowd sourcing, sharing and distribution of work
- Imagine a research paper bundled with a trend viewer to support the results
- Easily extendable with new views (future work)
  - Distribution of the results over the year (e.g. annual recurring trending periods)
  - Link graphs (which portal cites which portal)
- No duplication, no assets
  - Traditional archives do not separate assets (images, etc.) from textual content
  - Tedious filtering as the first step of processing, but it could be avoided with **smart archives**

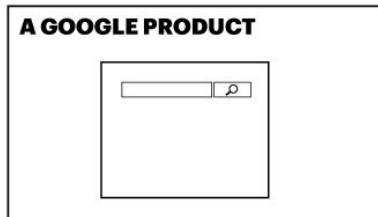
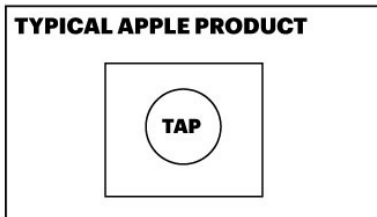
# Demo

<https://meta-trend-viewer.elte-dh.hu/>

(Indig, Sárközi-Lindner, and Nagy 2022)



# Summary and takeaways



**YOUR COMPANY'S APP**

FIRST NAME: <input type="text"/>	ADDR 1: <input type="text"/>	CITY: <input type="text"/>	4 - K AA2 DK9B KKA? CNS AA9 <input type="button" value="NEW"/> <input type="button" value="DEL"/>
LAST NAME: <input type="text"/>	ACCT #: <input type="text"/>	STATE: <input type="text"/>	
SSN: <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> FT/PT: <input checked="" type="checkbox"/>	TYPE CD: <input type="text"/>	ZIP: <input type="text"/> ...	
ID: <input type="text"/>	TPQ STAT: <input type="checkbox"/> <input type="checkbox"/>	ORD 1: ● ○ ○ ?	
PHONE 1: <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="text"/> ...	VER: <input type="text"/>	ORD 2: ● ○ ○ ?	
PHONE 2: <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input checked="" type="checkbox"/> ●	CAT CD: <input type="text"/>	ORD #: ● ○ ○ ?	

**OKAY   APPLY   SAVE   UNDO   HELP   DELETE   EDIT   SELECT   BROWSE   ERRORS**

## Summary and takeaways (cont.)

- It is not likely that anything important will be lost because the lack of archiving
- However, currently it is hard to find anything in the archives or get anything out
- We have to continue working with the consumers and value descriptive metadata
- A good heuristic: Act as early in the pipeline as possible
  - Try keeping the data separated and clean
  - Do not afraid to get your hands dirty with manual curation
  - Clean data is better than noisy. Only DJs used to say: "Everybody, make some noise!"
- Even in small scale there is great potential, but scaling up never hurt anybody



Thank you for your attention!

Questions?

<https://dh-lab.hu/>




<https://elte-dh.hu/>

<https://github.com/elte-dh>

<https://zenodo.org/communities/elte-dh>

<https://meta-trend-viewer.elte-dh.hu/>

# References I

-  Indig Balázs, Zsófia Sárközi-Lindner, and Mihály Nagy, “Use the Metadata, Luke! – An Experimental Joint Metadata Search and N-gram Trend Viewer for Personal Web Archives”, in: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, Taipei, Taiwan: Association for Computational Linguistics, Nov. 2022, pp. 47–52, url: <https://aclanthology.org/2022.nlp4dh-1.7>.
-  Lendák Imre, Balázs Indig, and Gábor Palkó, “WARChain: Consensus-based trust in web archives via proof-of-stake blockchain technology”, in: *Journal of Computer Security* (2022), pp. 1–17, issn: 0926-227X, doi: 10.3233/JCS-210040.
-  Oravecz Csaba, Tamás Váradi, and Bálint Sass, “The Hungarian Gigaword Corpus”, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1719–1723, url: [http://www.lrec-conf.org/proceedings/lrec2014/pdf/681\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf).