Development of Web Archive of the Czech Republic: Data processing and data accessibility in cluster

Mgr. Zdenko Vozár

Lead, Department for support of Applications NK + NDK

National Library of the Czech Republic

Partners:
- Department for Cybernetics, University of West Bohemia
- Institute of Sociology, Academy of Sciences of the Czech Republic
Fortress of knowledge: Siloing written heritage
I. vydání knihy - Snář sebeopoznání

Národní knihovna ČR vybrala tuto knihu jako kvalitní zdroj, který by měl být uchován do budoucna a stát se součástí českého kulturního dědictví.

Jak získat peníze za zpožděný let? Ptejte se našich hostů v Rozstřelu

Sněmovní tisky a ostatní dokumenty

Sněmovní tisky jsou dokumenty, o kterých se ve Sněmovně jedná, debatuje a hlasuje. Všechny jsou k dispozici v elektronické podobě a existuje několik způsobů vyhledávání, které můžete zvolit, abyste se co nejrychleji dostali k textu, který vás zajímá.
Suda (SOL) project:

- Index archived only once
- History of this collaborative effort - who has done what (not once!)
- But - all 28 North Korean webs

Hrm.

The Wayback Machine has not archived that URL.

This page is available on the web!

Help make the Wayback Machine more complete!
Save this url in the Wayback Machine
What is and What should be Web Archive?

- Information siloe: Pile of scrapped content?
  - Elements for reconstruction of the Internet?
  - Human written production? What about IoT?
  - National domain (.cz) without context?
- Archive for dead data?
- LTP repo input?
- Curatorial playground?
- User oriented service ala digital library?
- Scientific instrument?
- Cultural representation of Nation and instrument of cultural diplomacy?
Curve of knowledge

Original idea: Andrew Jackson

Researcher Queries
Jupyter Notebook
SQL Queries, Fulltext Analysis
Link Analysis
RAW data, cluster computation

Data preprocessed stats

Fulltext, Advanced search
URL, Webpages
1. Web Archive: Case of big data?
2. Short Story: WA of the Czech Republic
3. Hierarchical storage
4. World of clusters
5. Some tools
6. Data processing
7. Expected results
Web Archive: Any Big data?

- Data lakes - unstructured data -> Dead lakes?
- ARC / WARC files - finely done on backend
- Could we use them?
  - Is there any catalogue?
  - Sheer volume
    - Data management problems
    - Accessibility and presentation problems
  - Huge indexes - Only URL
- Future - In rank of many Petabytes
- Obvious limitations of public non-technological institutions (also BWA)
- Could we play on Big data?
- Curve of users’ knowledge
- Cheap storage!
Short story: Web Archive of Czech Republic

- 2001 - First crawls, as pilot (NK ČR, MZK Brno and MUNI Brno)
- 2005 - Crawls on regular basis
- 2007 - Member of IIPC
- 2011 - New hierarchical storage facilities
- 2018 - Procurement of machines for testing cluster

Accessibility: Online (under legal agreement) vs. Inhouse (all)

Accession y.: c. 40 TB                      Total: 365 TB

Topic and Serials (15 TB) - 1x or 2x wholedomain .cz (20-30TB)
Actual storage policy - GPFS, HSM

+ Huge volumes of data accessible per year (specific NFS endpoint)
+ Hierarchical Storage Management - lower cost - cheaper media
- Vendor lock in - Enterprise system from IBM
- Costly licenses for direct writing to GPFS Spectrum Scale (using as Fe NFSv3)
+ Good enough performance (depends on the setup of tiers)
+ Retention 45 days - 3 diff. versions, 60 days until file is deleted
+ Created archival copies with bitstream protection
- Computational performance
  - Data analysis (eg. iterative tasks - network bandwidth)
  - Restoration back to production (eg. 2018 incident vs. 2TB p. hour guess)
1. Mitigation of common problem: Hardware failure and risk of data loss
2. Aggregation of big data
3. Manipulate, Process, Recombine and Administration of large data sets
4. Storage and protection of huge data sets across clusters of computers
5. Sharing and Presentation of data sets

+ Open source (no vendor lock-ins), many contributors to development
+ Commodity hardware - Very cheap in its elements and in continuous replanning of infrastructure
+ Great versatility and huge set of analytic tools
+ Technology used by internet giants to store PBs of data
+ Resistant to HW failure - fast to get into production
- Dedicated operator
- Easy to start, but steep learning curve
- Still not replacement to LTP solutions
Data processing WF

1. Establishing a collection: WARC
   Data pump - Storage / Crawl

2. Archives Unleashed Toolkit
   a. Basic preprocessing and data cleansing
   b. Links extraction

3. Advanced processing via python toolset
   a. Near-duplicates analysis
   b. NER extraction
   c. Categorisation with ML
   d. Catalogisation

4. Complementation of intermediary data format

5. Data pump to SOLR and indexation

6. Accessibility and distribution
Results

- Satisfying provision curve of knowledge
- Fast and durable index
- Discovery module
- Export module
- Cluster computing on demand

- Linked content + Network maps
- Search data by category
- Enriched metadata + NER
- Full-text search
- Advanced search + Filters
- Chosen collections’ analysis
- Download results of exploration
- Preprocessed public datasets
- Jupyter NB for further queries
- Data scientists Research possible
Near Future

Differentiation between archival circuit and presentation storage circuits

Advanced analysis of visual content

No more disparate information silos! Data have value only if we could grasp them.

Linked data on demand

Integrated object repositories with data analytic functions and advanced API connectors over complex datastreams

Backbone to “light” client oriented

Petabytes to sift through

World memory on demand
Thank You!

Mgr. Zdenko Vozár - zdenko.vozar@nkp.cz