Web Archive Analytics
Outline

1. The Global Datasphere
2. The Internet Archive
3. Web Archive Analytics @ Webis
4. Web Archive Processing
5. Webis Archive Research
The Global Datasphere
The Global Datasphere

“A measure of all new data captured, created, and replicated in a single year.”

[IDC, 2018]

“. . . images and videos on mobile phones uploaded to YouTube, digital movies populating the pixels of our high-definition TVs, security footage at airports and major events such as the Olympic Games, subatomic collisions recorded by the Large Hadron Collider at CERN, banking data swiped in an ATM, transponders recording highway tolls, voice calls zipping through digital phone lines, texting as a widespread means of communications, . . . ”

[IDC, 2012]
The Global Datasphere in 2020

Persistent data in data centers (beginning - 2020)

- ca. 59ZB Entire data generated in 2020

- ca. 1ZB Persistent data in data centers

Public access

- ca. 200EB Web pages (< 1EB)
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

Restricted access

- ca. 800EB
  - Data of individuals
  - Data in enterprises
  - Data of public bodies

Transient data

- ca. 59ZB

1GB = 10^9 Bytes
1TB = 10^12 Bytes
1PB = 10^15 Bytes
1EB = 10^18 Bytes
1ZB = 10^21 Bytes

Basis: IDC (2014-20) • Seagate (2018-20) • Cisco Systems (2018) • Statista (2020) • Domo Inc. (2018-20)
The Global Datasphere in 2020

Persistent data in data centers (beginning - 2020)

Web pages (< 1EB)
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

Data of individuals
Data in enterprises
Data of public bodies

Basis: IDC (2014-20) • Seagate (2018-20) • Cisco Systems (2018) • Statista (2020) • Domo Inc. (2018-20)

1GB = 10^9 Bytes
1TB = 10^12 Bytes
1PB = 10^15 Bytes
1EB = 10^18 Bytes
1ZB = 10^21 Bytes
The Global Datasphere in 2020

Relating Data Source Sizes

- ca. 200EB: Public Internet
- ca. 50EB: Google
- ca. 30PB: Web pages @ Internet Archive
- 8PB: Web pages @ Webis
- 300TB: Wikipedia including Wikimedia
- 200GB: English Wikipedia including media
- 30GB: All English Wikipedia article texts

2020

- $10^9$ Giga
- $10^{12}$ Tera
- $10^{15}$ Peta
- $10^{18}$ Exa
- $10^{21}$ Zetta
The Global Datasphere in 2020
Relating Data Source Sizes

- ca. 200EB: Public Internet
- ca. 50EB: Google
- ca. 30PB: Web pages @ Internet Archive
- 8PB: Web pages @ Webis
- 300TB: Wikipedia including Wikimedia
- 200GB: English Wikipedia including media
- 30GB: All English Wikipedia article texts

© WEBIS 2022
The Global Datasphere in 2020
Relating Data Source Sizes

- ca. 200EB: Public Internet
- ca. 50EB: Google
- 300TB: Wikipedia including Wikimedia
- 200GB: English Wikipedia including media
- 30GB: All English Wikipedia article texts
- ca. 30PB: Web pages @ Internet Archive
- 8PB: Web pages @ Webis
- > 500TB: Web archive analytics

10^9 Giga
10^12 Tera
10^15 Peta
10^18 Exa
10^21 Zetta

© WEBIS 2022
The Global Datasphere in 2020

Where is the Data Stored?

Basis: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, May 2020.
The Global Datasphere in 2020

Where is the Data Stored?

Among others:

Basis: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, May 2020.
The Internet Archive
Founded 1996 by Brewster Kahle

For all things digital:
- 477 billion web pages (ca. 30PB) – accessible via the WayBack Machine
- 20 million books and texts
- 4.5 million audio recordings (including 180,000 live concerts)
- 4 million videos (including 1.6 million Television News programs)
- 3 million images
- 200,000 software programs
Founded 1996 by Brewster Kahle

For all things digital:

- 477 billion web pages (ca. 30PB) – accessible via the WayBackMachine
- 20 million books and texts
- 4.5 million audio recordings (including 180,000 live concerts)
- 4 million videos (including 1.6 million Television News programs)
- 3 million images
- 200,000 software programs
Mission: “Universal access to all knowledge.”

- One full copy in San Francisco
- Part at the new Library of Alexandria
- Part in Amsterdam
- Copy representative portion (8PB) to the Digital Bauhaus Lab / Webis group:

  [archive.webis.de]
Web Archive Analytics @ Webis
### Webis Data Center (Digital Bauhaus Lab)

|                  | α-web [2009]     | β-web [2015]     | γ-web [2016 + 2021] | δ-web [2018]     | ε-web [2020]     |
|------------------|------------------|------------------|---------------------|------------------|------------------|
| **Nodes**        | 44               | 135              | 9                   | 78               | 55               |
| **Disk [PB]**    | 0.2              | 4.1              | 0.08                | 12               | 0.1              |
| **Cores**        | 176              | 1,740            | 672 + 227,328       | 1,248            | 1,100            |
| **RAM [TB]**     | 0.8              | 28               | 7.5                 | 10               | 7                |

**Typical research:**

- **α-Web.** Teaching, Staging environment
- **β-Web.** Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- **γ-Web.** Machine learning (embedding, deep learning), Text synthesis, Language modeling
- **δ-Web.** Web archiving, Virtualization (storage)
- **ε-Web.** Search index construction, Argument search
## Webis Data Center (Digital Bauhaus Lab)

|       | α-web [2009] | β-web [2015] | γ-web [2016 + 2021] | δ-web [2018] | ε-web [2020] |
|-------|--------------|--------------|---------------------|--------------|--------------|
| Nodes | 44           | 135          | 9                   | 78           | 55           |
| RAM [TB] | 0.8  | 28            | 7.5                 | 10           | 7            |
| Cores | 176          | 1,740        | 672 + 227,328       | 1,248        | 1,100        |
| Disk [PB] | 0.2  | 4.1           | 0.08                | 12           | 0.1          |

= 3.2 TFLOPs
= 67.4 TFLOPs
= 8 PFLOPs
= 119.8 TFLOPs
= 44 TFLOPs

## Typical research:

- **α-Web.** Teaching, Staging environment
- **β-Web.** Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- **γ-Web.** Machine learning (embedding, deep learning), Text synthesis, Language modeling
- **δ-Web.** Web archiving, Virtualization (storage)
- **ε-Web.** Search index construction, Argument search
## Webis Data Center
**Webis Data Center** (Digital Bauhaus Lab)

|                  | α-web [2009] | β-web [2015] | γ-web [2016 + 2021] | δ-web [2018] | ε-web [2020] |
|------------------|--------------|--------------|--------------------|--------------|--------------|
| **Nodes**        | 44           | 135          | 9                  | 78           | 55           |
| **RAM [TB]**     | 0.8          | 28           | 7.5                | 10           | 7            |
| **Disk [PB]**    | 0.2          | 4.1          | 0.08               | 12           | 0.1          |
| **Cores**        | 176          | 1,740        | 672 + 227,328      | 1,248        | 1,100        |
| **Performance**  | ≅ 3.2 TFLOPs | ≅ 67.4 TFLOPs| ≅ 8 PFLOPs         | ≅ 119.8 TFLOPs| ≅ 44 TFLOPs  |

### Typical research:

- **α-Web.** Teaching, Staging environment
- **β-Web.** Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- **γ-Web.** Machine learning (embedding, deep learning), Text synthesis, Language modeling
- **δ-Web.** Web archiving, Virtualization (storage)
- **ε-Web.** Search index construction, Argument search
### Webis Data Center (Digital Bauhaus Lab)

|                | α-web [2009] | β-web [2015] | γ-web [2016 + 2021] | δ-web [2018] | ϵ-web [2020] |
|----------------|-------------|-------------|---------------------|-------------|-------------|
| **Nodes**      | 44          | 135         | 9                   | 78          | 55          |
| **Disk [PB]**  | 0.2         | 4.1         | 0.08                | 12          | 0.1         |
| **Cores**      | 176         | 1,740       | 672 + 227,328       | 1,248       | 1,100       |
| **RAM [TB]**   | 0.8         | 28          | 7.5                 | 10          | 7           |

**Typical research:**

- **α-Web.** Teaching, Staging environment
- **β-Web.** Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)
- **γ-Web.** Machine learning (embedding, deep learning), Text synthesis, Language modeling
- **δ-Web.** Web archiving, Virtualization (storage)
- **ϵ-Web.** Search index construction, Argument search
# Webis Analytics Stack

| Layer                                           |
|------------------------------------------------|
| Data Consumption Layer                          |
| Data Analytics Layer                             |
| Data Management Layer                           |
| Hardware Layer                                  |
| Data Acquisition Layer                          |
# Webis Analytics Stack

| Layer                          | Vendor stack |
|-------------------------------|--------------|
| Data Consumption Layer        | Prometheus   |
| Data Analytics Layer          | kibana       |
| Data Management Layer         | elasticsearch|
| Hardware Layer                | Consul       |
| Data Acquisition Layer        | SALTStack    |

© WEBIS 2022
## Webis Analytics Stack

| Layer                        | Technology stack                                                                 | Vendor stack |
|------------------------------|---------------------------------------------------------------------------------|--------------|
| Data Consumption Layer       | - Visual analytics<br>- Immersive technologies<br>- Intelligent agents          | Prometheus   |
| Data Analytics Layer         | - Distributed learning<br>- State-space search<br>- Symbolic inference            | Prometheus   |
| Data Management Layer        | - Key-value store<br>- RDF triple store<br>- Graph store<br>- Object store       | Elasticsearch |
| Hardware Layer               | - Orchestration<br>- Parallelization<br>- Virtualization                       | Consul       |
| Data Acquisition Layer       | - Distant supervision<br>- Crowdsourcing<br>- Crawling and archiving             | SaltStack    |
# Webis Analytics Stack

| Data Consumption Layer | Task Stack | Technology stack | Vendor stack |
|------------------------|------------|------------------|--------------|
|                        | - Query and explore | - Visual analytics | Prometheus   |
|                        | - Visualize and interact | - Immersive technologies |            |
|                        | - Explain and justify | - Intelligent agents |              |

| Data Analytics Layer | Task Stack | Technology stack | Vendor stack |
|----------------------|------------|------------------|--------------|
|                      | - Diagnose and reason | - Distributed learning |            |
|                      | - Structure identification | - State-space search |            |
|                      | - Structure verification | - Symbolic inference |              |

| Data Management Layer | Task Stack | Technology stack | Vendor stack |
|-----------------------|------------|------------------|--------------|
|                       | - Provenance tracking | - Key-value store |            |
|                       | - Normalization | - RDF triple store |            |
|                       | - Cleansing | - Graph store |              |
|                       | | - Object store |              |

| Hardware Layer | Task Stack | Technology stack | Vendor stack |
|----------------|------------|------------------|--------------|
|                | - Monitoring | - Orchestration | Prometheus   |
|                | - Replication | - Parallelization |            |
|                |              | - Virtualization |              |

| Data Acquisition Layer | Task Stack | Technology stack | Vendor stack |
|------------------------|------------|------------------|--------------|
|                        | - Replay | - Distant supervision |            |
|                        | - Collect | - Crowdsourcing |            |
|                        | - Log | - Crawling and archiving |              |
# Webis Analytics Stack

| Data Consumption Layer | Task Stack | Technology stack | Vendor stack | Roles |
|------------------------|------------|------------------|--------------|-------|
|                        | - Query and explore | - Visual analytics | | Experts: |
|                        | - Visualize and interact | - Immersive technologies | | - IR |
|                        | - Explain and justify | - Intelligent agents | | - NLP |
| Data Analytics Layer   | - Diagnose and reason | - Distributed learning | | - CSS |
|                        | - Structure identification | - State-space search | | - VA |
|                        | - Structure verification | - Symbolic inference | | |
| Data Management Layer  | - Provenance tracking | - Key-value store | | Data scientist |
|                        | - Normalization | - RDF triple store | | |
|                        | - Cleansing | - Graph store | | |
|                        | | - Object store | | |
| Hardware Layer         | - Monitoring | - Orchestration | | Data engineer |
|                        | - Replication | - Parallelization | | |
|                        | | - Virtualization | | |
| Data Acquisition Layer | - Replay | | | Data scientist |
|                        | - Collect | | | |
|                        | - Log | | | |

Experts:
- IR
- NLP
- CSS
- VA
Web Archive Processing
WARC is a standard format for web archives.

A WARC file consists of a zipped sequence of WARC records. (~1 GiB / file)

A WARC record corresponds to one HTTP request/response for a given URI:
Web Archive Data

Web Archiving

- A web page: Record all HTTP communication between browser and server.
- A browser is simulated to ensure the human-readable version is obtained.
- During web crawling, a web archiver “browses” every crawled page.

![Diagram of web archiving process]

Key:
- Component
- Environment
- File read/write
- HTTP request
- HTTP response
- Browser/DOM control
- Browser/DOM status
Web Archive Data

Web Archiving

- A web page: Record all HTTP communication between browser and server.
- A browser is simulated to ensure the human-readable version is obtained.
- During web crawling, a web archiver “browses” every crawled page.
Given a learning task and ground truth within WARC files, train a model. Only a fraction of the records within the WARC files are ground truth.

Goal: Training at web scale (billions of WARC files)
Given a mining task and a trained (classification) model, collect relevant data. Only a fraction of the records within the WARC files are relevant.

Goal: Mining at web scale (billions of WARC files)
Web Archive Processing
Streamed Model Training Pipeline

- Given a mining task and a trained (classification) model, collect relevant data. Only a fraction of the records within the WARC files are relevant.

- Goal: Mining at web scale (billions of WARC files)

Observations:
- Mining / filtering WARC files is “embarrassingly parallel”.
- Decompressing WARC files, and processing WARC records are CPU bound.
- The mining / preprocessing step results in a variational data flow.
- Training of neural networks is GPU bound and presumes constant data flow.
- WARC storage, parallel processing, and GPU bound processing are on separate clusters.
Web Archive Processing
Streamed Model Training Pipeline

1. PySpark distributes WARC among workers
2. FastWARC decompresses and iterates records
3. First filtering step of records
Web Archive Processing
Streamed Model Training Pipeline

1. PySpark distributes WARC among workers
2. FastWARC decompresses and iterates records
3. First filtering step of records
4. Pickled record streams
1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
3. First filtering step of records
4. Pickled record streams
5. Conversion to Tensorflow datasets and source interleaving
6. Batched processing by a Keras model
Web Archive Processing
Streamed Model Training Pipeline

1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
3. First filtering step of records
4. Pickled record streams
5. Conversion to Tensorflow datasets and source interleaving
6. Batched processing by a Keras model on GPU
7. Second filtering based on classification results
8. Storage of relevant data
Archival support
Argumentation
Language models
Search engines
Social sciences
Text reuse
Text synthesis
Webis Archive Research [publications.webis.de]

- **Web Page Segmentation**
  Goal: Improve reliability of semantic web page segmentation.

- **Web Crawling Quality Analysis**
  Goals: (1) Detect incomplete crawls.
  (2) Improve the web page reconstructability from crawls.

- **Personal Web Archival**
  Goal: Technology for individual web archive creation and search.
Learn Discussion Strategies
Approach: Harvesting talk pages, email repositories, Reddit threads.

Acquire Justification and Reasoning Knowledge
Approach: Construction of a causality graph from causal statements.

Compute Ranking Functions for Arguments
Approach: Analysis of the hyperlink graph of web pages.
Truths and Myths of the Mnemonic Password Advice

Approach: Construction of a position-dependent, higher-order language model, based on word initials of two billion sentences of verified casual language.

Example:

“The quick brown fox jumps over the lazy dog.”

Is “Tqbfjotld” a strong password?
args.me
The first (2017) search engine for arguments on the web.

ChatNoir
Search engine with rank explanation, indexing the ClueWeb and the CommonCrawl.

Netspeak
Phrase search engine for text correction and idiomatic writing.

Picapica
Search engine for text reuse detection.
Detect and Visualize Vandalism in Social Software
Approach: Spatio-temporal analysis of reverted Wikipedia edits.

“Celebrity” Profiling
Goal: Following personal traits on the Internet.

Hyperpartisan News Detection
Goal: Analyzing political bias and illustrating provenance on the Internet.
Webis Archive Research [publications.webis.de]

- **Who Wrote the Web?**
  Applying author identification technology at web-scale.

- **Text Reuse Analytics**
  Goals: (1) Finding Wikipedia text reuse (on the web).
  (2) Quantifying the prevalence of scientific text reuse.

- **Text Reuse Illustration**
  Example: Visualizing article similarities in Wikipedia.

Riemann et al.: Visualizing Article Similarities in Wikipedia. EuroVis 2016
Abstractive Snippet Generation

Approach: Use of anchor contexts to generate abstractive snippets with a pointer-generator network, exploiting ClueWeb09, ClueWeb12, and the DMOZ Open Directory Project.

Learn Automatic Summarization

Approach: Exploit author-provided summaries, taking advantage of the common practice of appending a “TL;DR” to long posts.
Summary

Persistent data in data centers (beginning - 2020)

Public access
- Web pages (< 1EB)
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

Restricted access
- Data of individuals
- Data in enterprises
- Data of public bodies

ca. 12ZB Persistent data in data centers

ca. 59ZB Entire data generated in 2020

ca. 800EB Restricted access

1GB = 10^9 Bytes
1TB = 10^12 Bytes
1PB = 10^15 Bytes
1EB = 10^18 Bytes
1ZB = 10^21 Bytes

Global Datasphere
Summary

Persistent data in data centers (beginning - 2020)

Public access
- Web pages (< 1 EB)
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

Restricted access
- Data of individuals
- Data in enterprises
- Data of public bodies

Entire data generated in 2020
- ca. 59 ZB

1 GB = 10^9 Bytes
1 TB = 10^12 Bytes
1 PB = 10^15 Bytes
1 EB = 10^18 Bytes
1 ZB = 10^21 Bytes

Global Datasphere

Internet Archive
Summary

Persistent data in data centers (beginning - 2020)

Public access
- Web pages (< 1EB)
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

Restricted access
- Data of individuals
- Data in enterprises
- Data of public bodies

Global Datasphere

Internet Archive

Vendor stack

Webis Analytics Stack
Summary

Persistent data in data centers (beginning - 2020)

Public access
- Web pages (< 1EB)
- Books and texts
- Audio recordings
- Videos
- Images
- Software programs

Restricted access
- Data of individuals
- Data in enterprises
- Data of public bodies

Internet Archive

Global Datasphere

Webis Analytics Stack

Vendor stack

Webis Archive Research
Thank You!