Exploring our Digital Past: Web Archive Analytics

Martin Potthast
Universität Leipzig
webis.de

ARQUS Forum on Digital Humanities · Leipzig · June 15, 2022

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

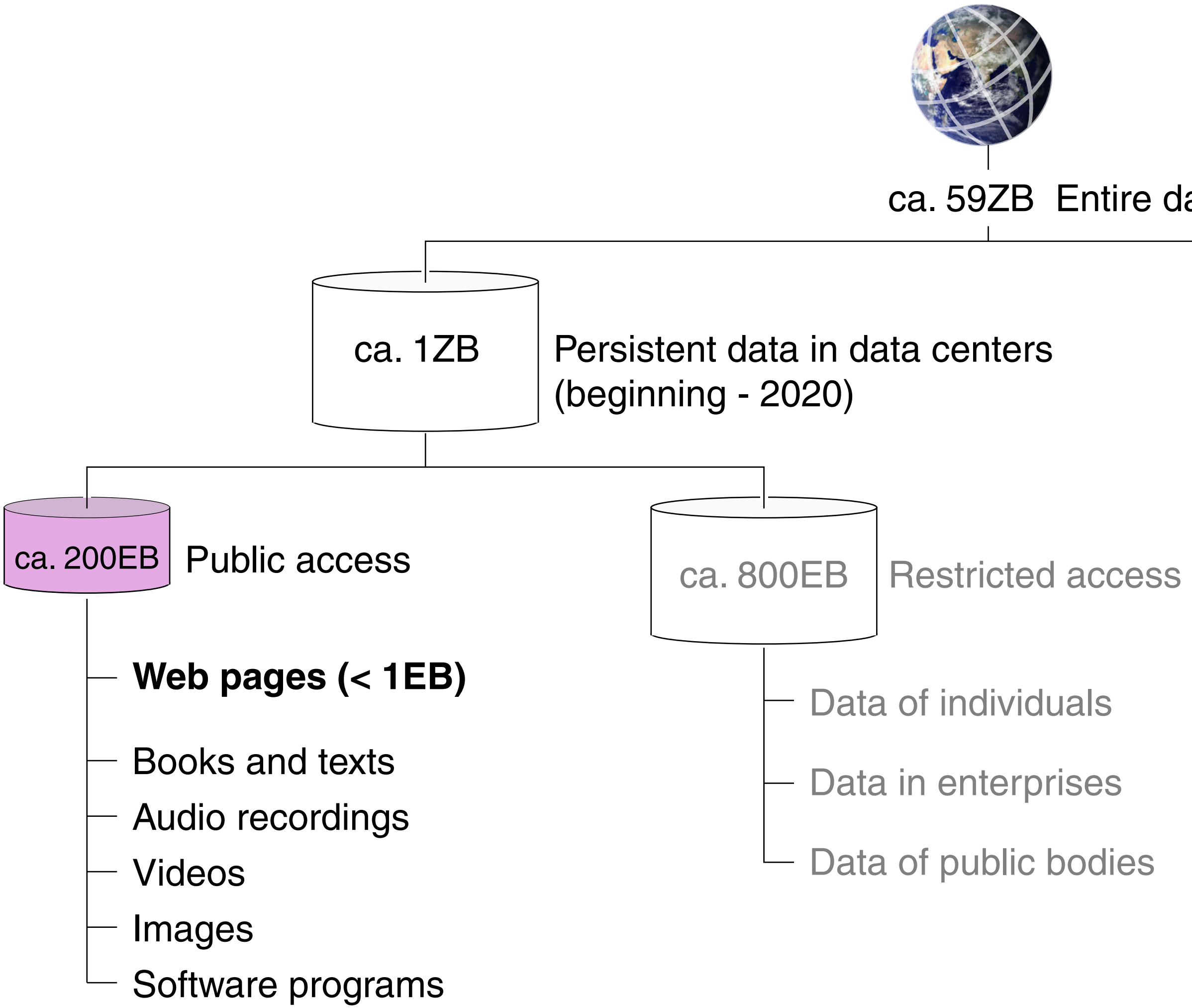
“A measure of all new data captured, created, and replicated in a single year.”

[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



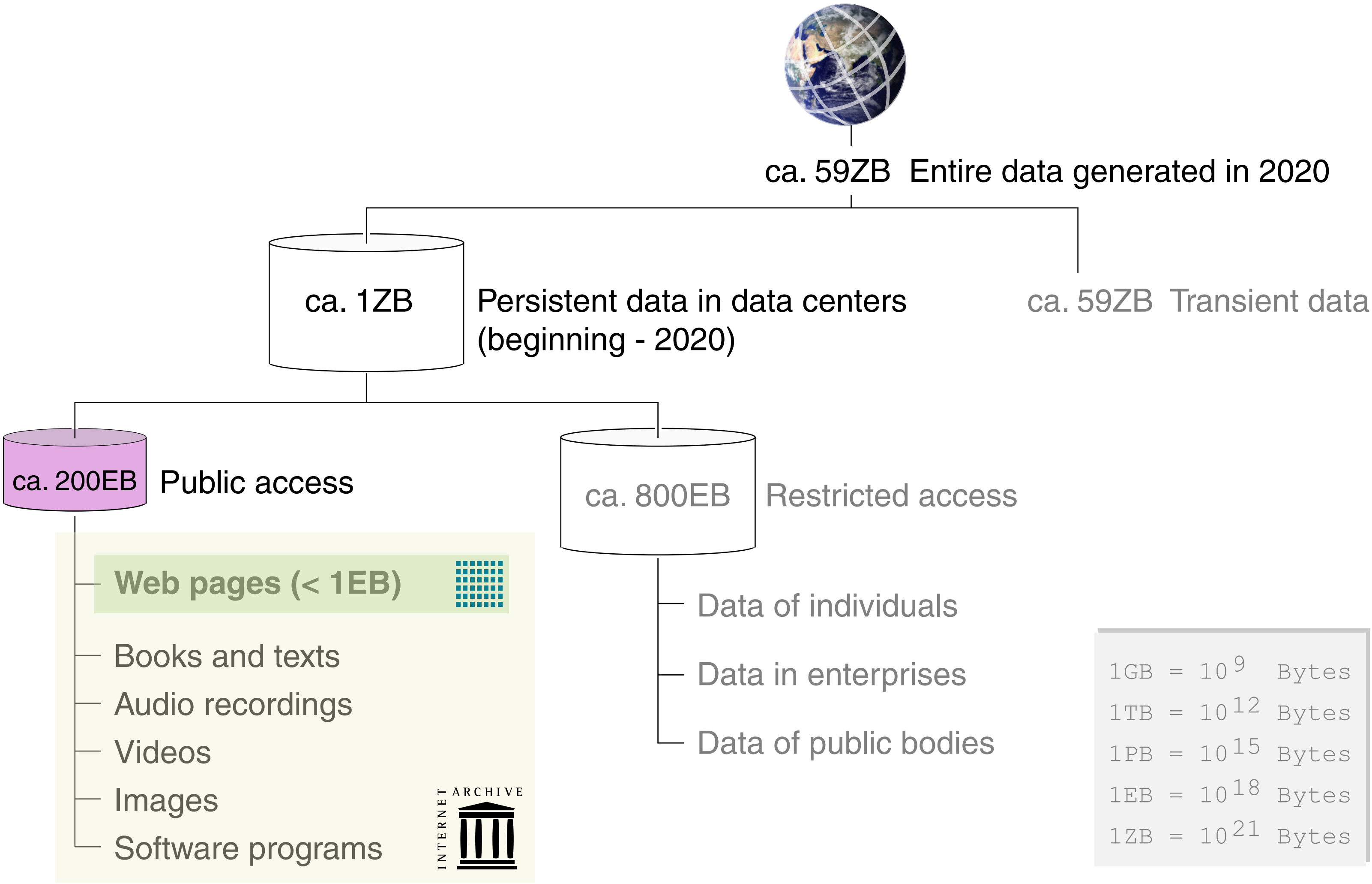
ca. 59ZB Entire data generated in 2020

- ca. 1ZB Persistent data in data centers (beginning - 2020)
  - ca. 200EB Public access
    - **Web pages (< 1EB)**
    - Books and texts
    - Audio recordings
    - Videos
    - Images
    - Software programs
  - ca. 800EB Restricted access
    - Data of individuals
    - Data in enterprises
    - Data of public bodies
- ca. 59ZB Transient data

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



ca. 59ZB Entire data generated in 2020

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 59ZB Transient data

ca. 200EB Public access

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

ca. 800EB Restricted access

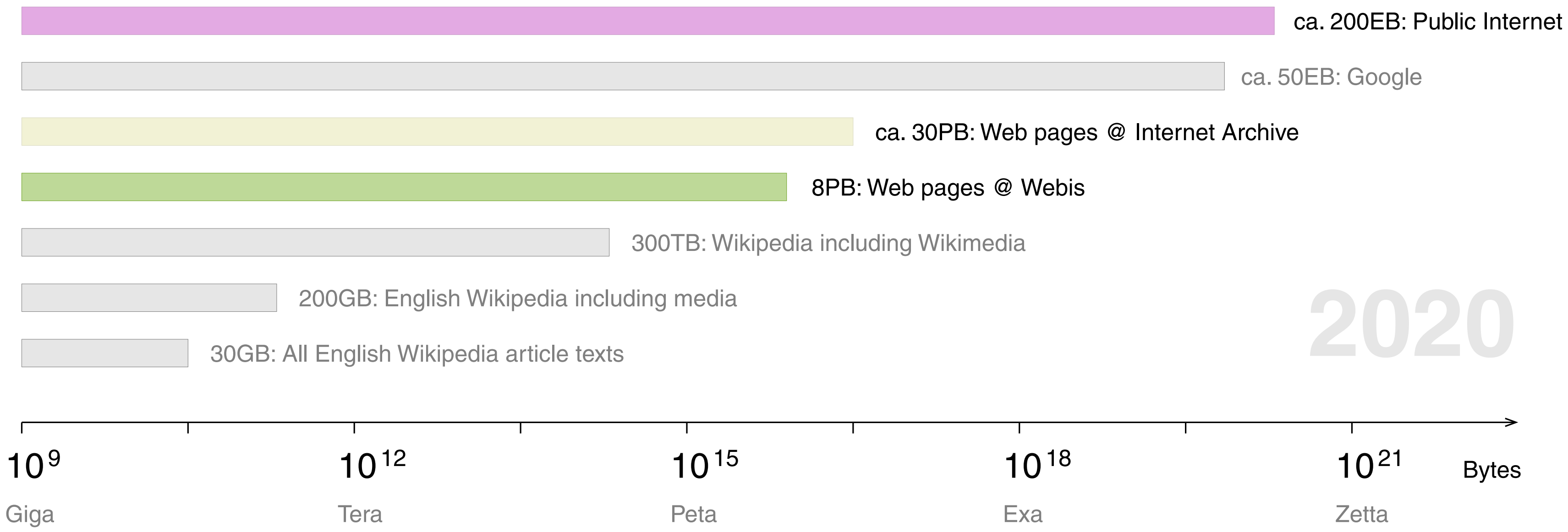
- Data of individuals
- Data in enterprises
- Data of public bodies

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)

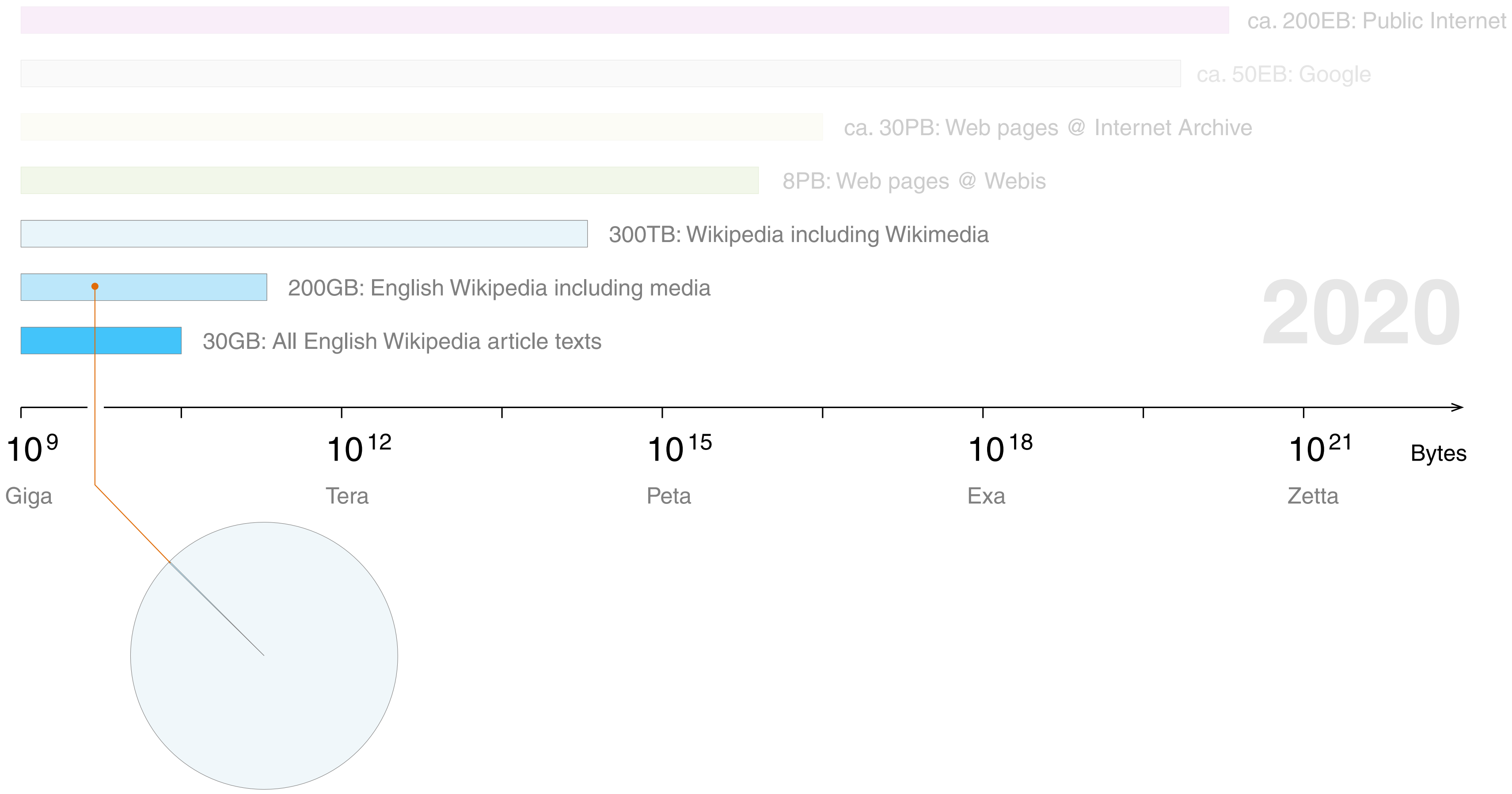
The Global Datasphere in 2020

Relating Data Source Sizes



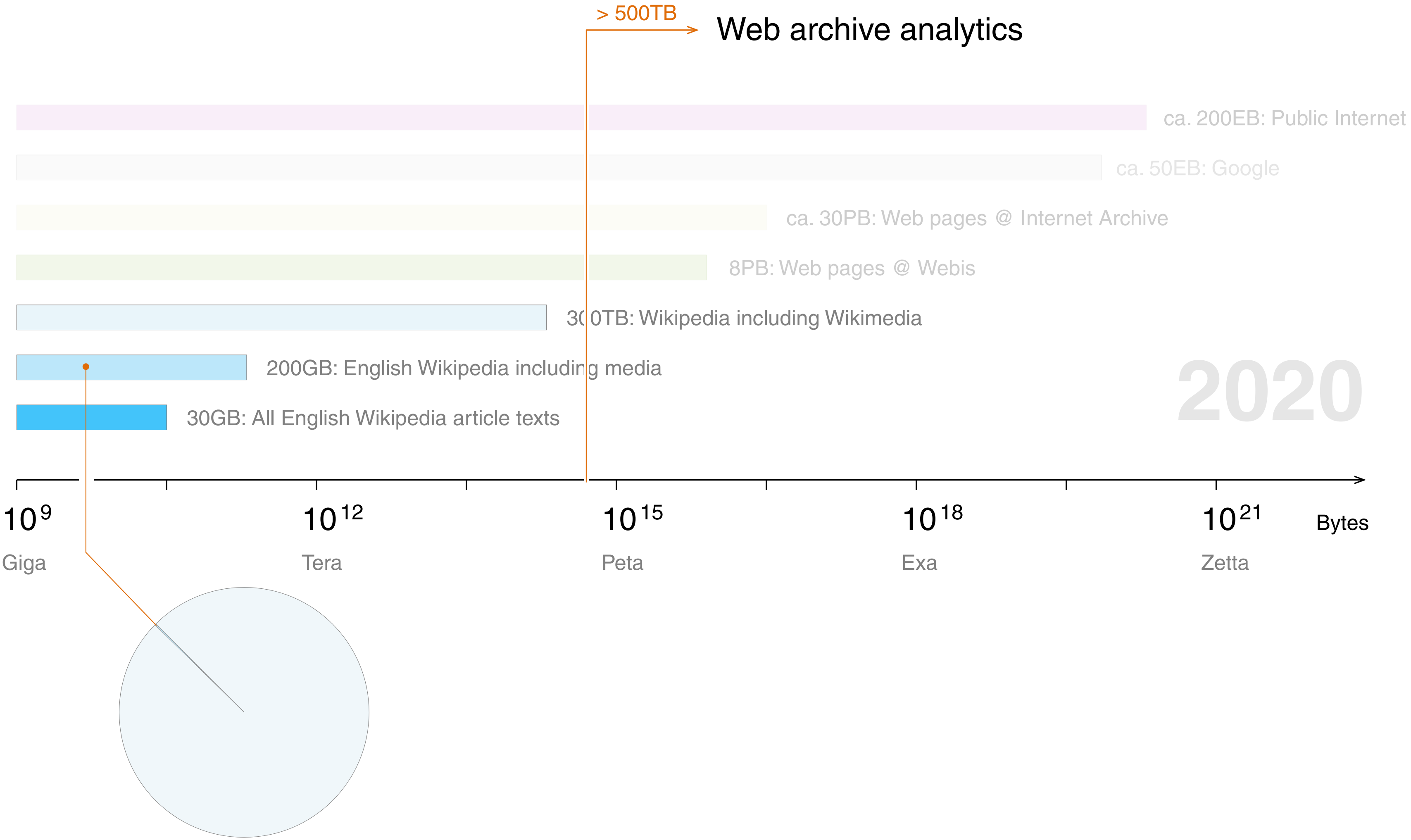
The Global Datasphere in 2020

Relating Data Source Sizes



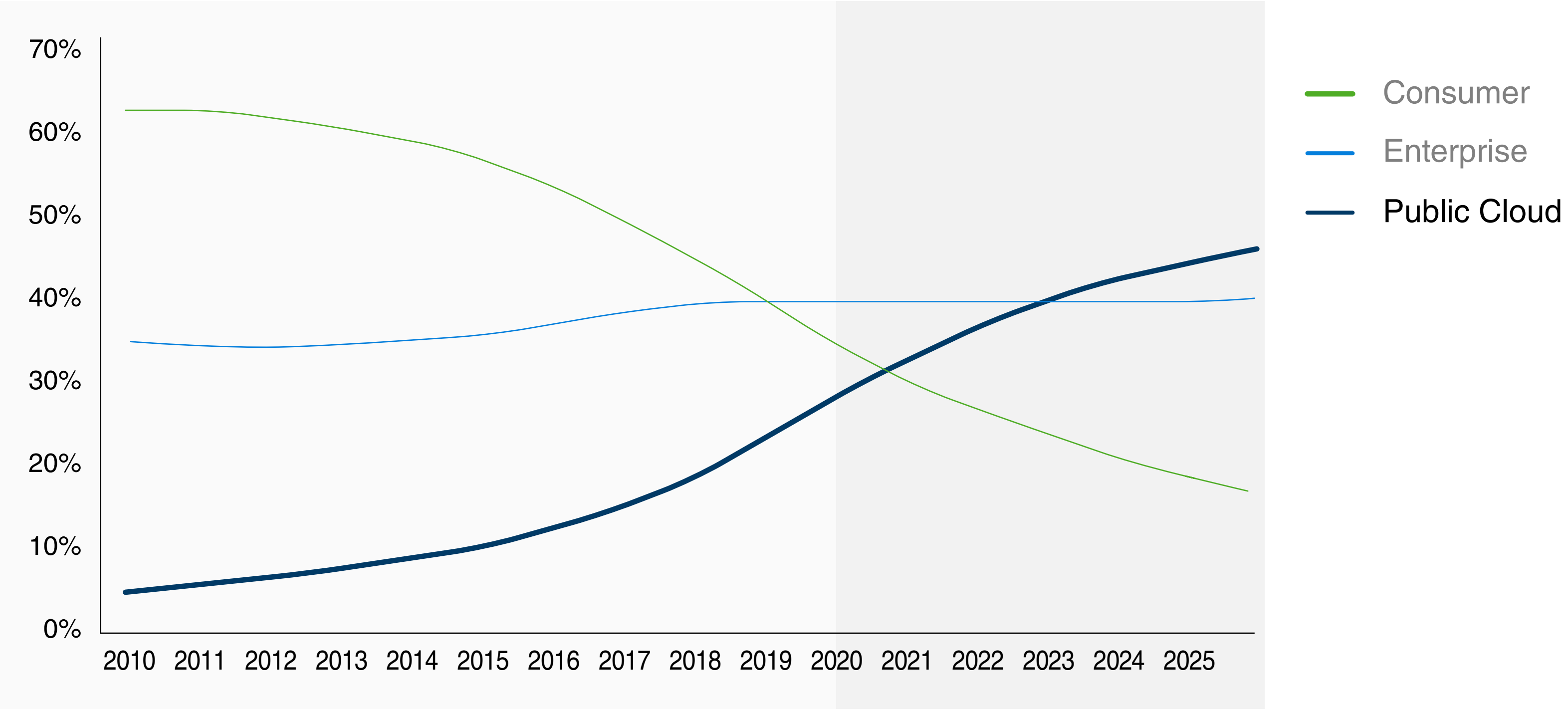
The Global Datasphere in 2020

Relating Data Source Sizes



The Global Datasphere in 2020

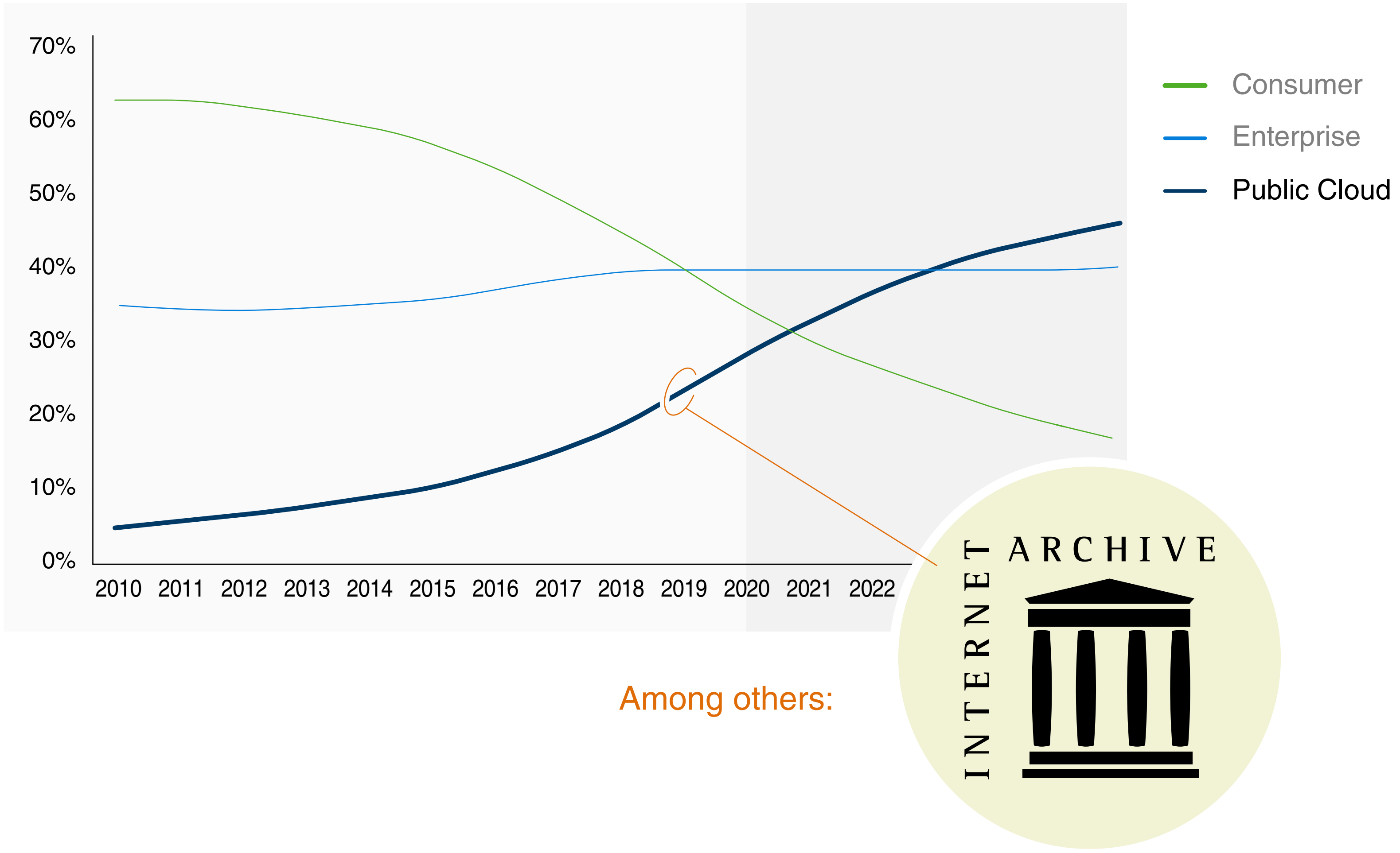
Where is the Data Stored?



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.

②

The Internet Archive

INTERNET ARCHIVE

- Founded 1996 by Brewster Kahle
- For all things digital:
  - 477 billion web pages (ca. 30PB) – accessible via the INTERNET ARCHIVE 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

INTERNET ARCHIVE

- Founded 1996 by Brewster Kahle
- For all things digital:
  - 477 billion web pages (ca. 30PB) – accessible via the INTERNET ARCHIVE 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

INTERNET ARCHIVE

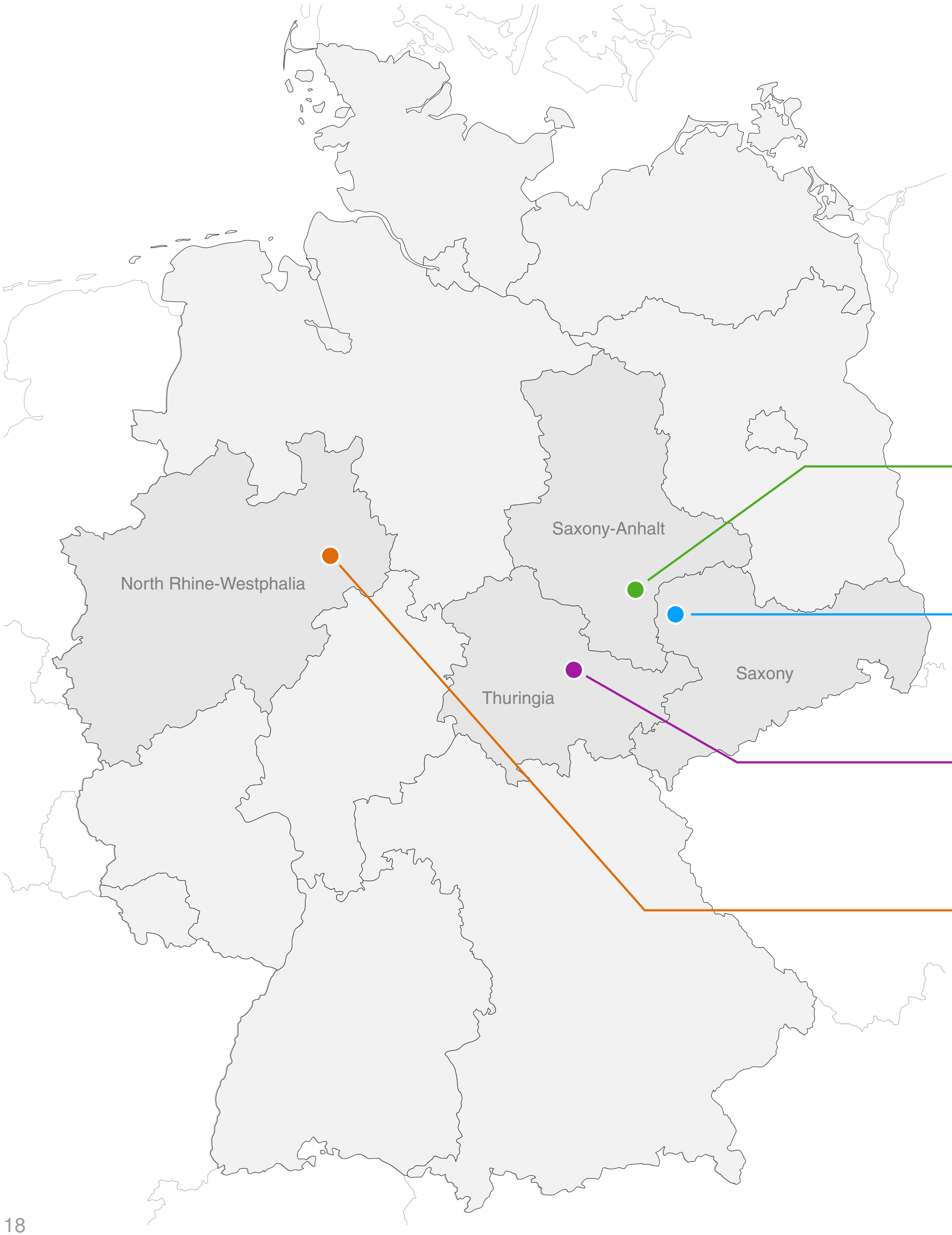
Mission: “Universal access to all knowlege.”

- ❑ 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]

③

Web Archive Analytics @ Webis



MLU Halle-Wittenberg

Prof. Dr. Matthias Hagen

Leipzig University

Prof. Dr. Martin Potthast

Bauhaus-Universität Weimar

Prof. Dr. Benno Stein

Paderborn University

Prof. Dr. Henning Wachsmuth

Webis Data Center (Digital Bauhaus Lab)

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 ≃ 3.2 TFLOPs	1,740 ≃ 67.4 TFLOPs	672 + 227,328 ≃ 8 PFLOPs	1,248 ≃ 119.8 TFLOPs	1,100 ≃ 44 TFLOPs
RAM [TB]	0.8	28	7.5	10	7

Typical research:

α-Web. Teaching, Staging environment

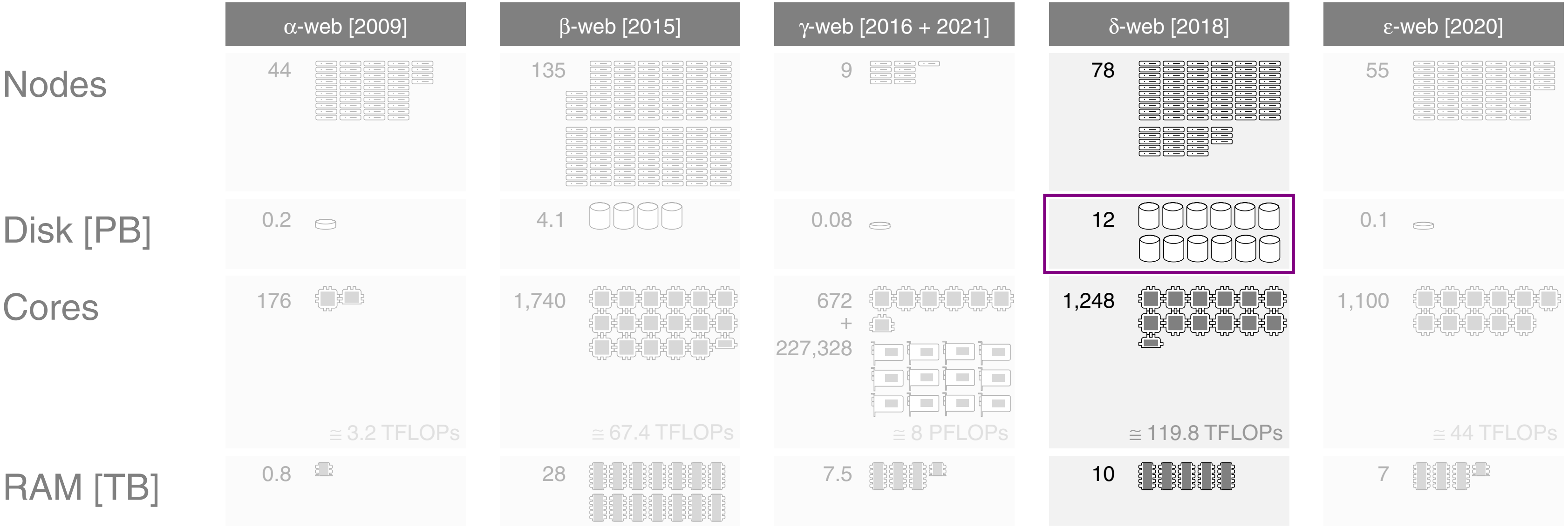
β-Web. Web mining (map reduce), Authorship analytics, Virtualization (compute, web services)

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
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176 ≅ 3.2 TFLOPs	1,740 ≅ 67.4 TFLOPs	672 + 227,328 ≅ 8 PFLOPs	1,248 ≅ 119.8 TFLOPs	1,100 ≅ 44 TFLOPs
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
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176 ≅ 3.2 TFLOPs	1,740 ≅ 67.4 TFLOPs	672 + 227,328 ≅ 8 PFLOPs	1,248 ≅ 119.8 TFLOPs	1,100 ≅ 44 TFLOPs
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

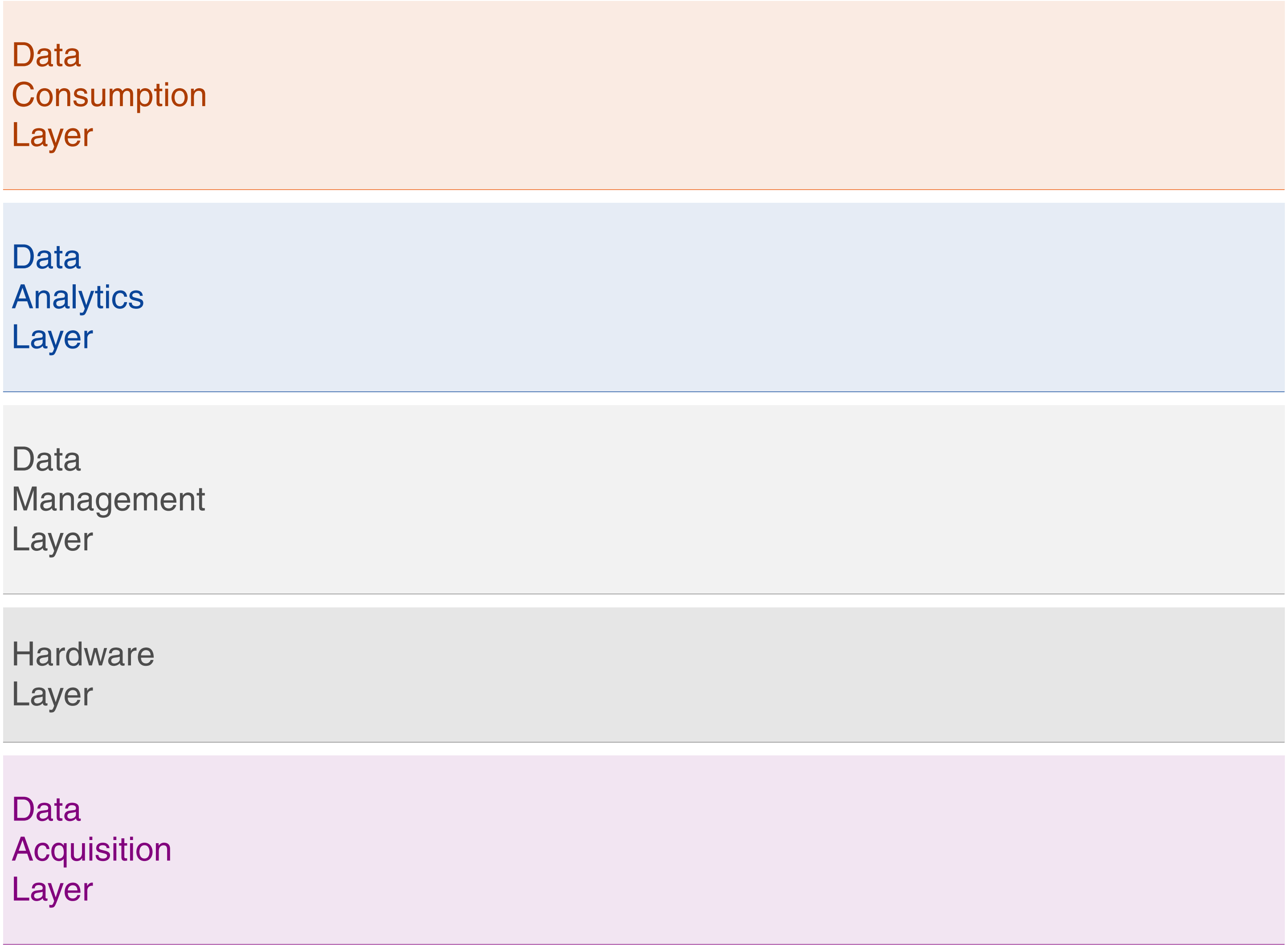
Webis Data Center (Digital Bauhaus Lab)

	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44	135	9	78	55
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176 ≅ 3.2 TFLOPs	1,740 ≅ 67.4 TFLOPs	672 + 227,328 ≅ 8 PFLOPs	1,248 ≅ 119.8 TFLOPs	1,100 ≅ 44 TFLOPs
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 Science Stack



Data Consumption Layer

Data Analytics Layer

Data Management Layer

Hardware Layer

Data Acquisition Layer

Webis Data Science Stack

Data Consumption Layer

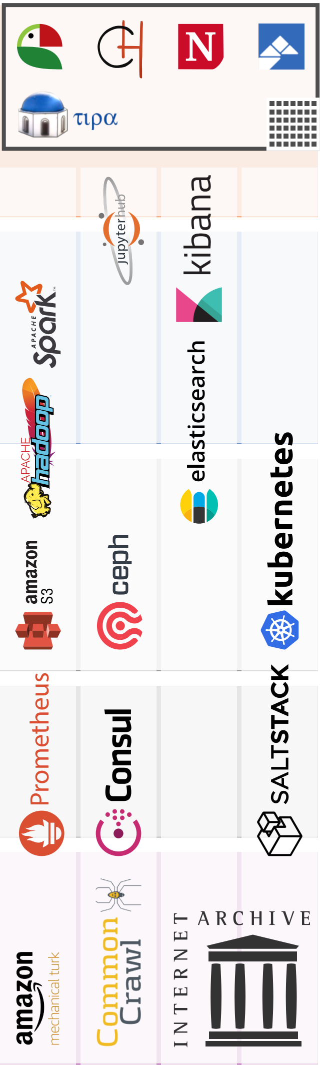
Data Analytics Layer

Data Management Layer

Hardware Layer

Data Acquisition Layer

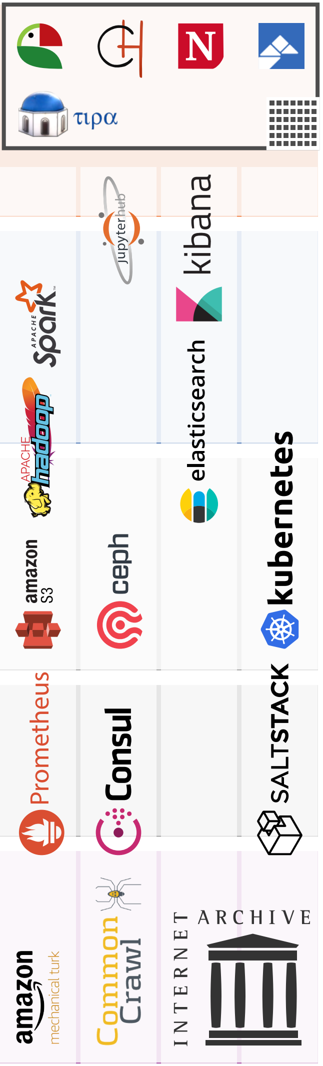
Vendor stack



Webis Data Science Stack

	Technology stack
Data Consumption Layer	- Visual analytics - Immersive technologies - Intelligent agents
Data Analytics Layer	- Distributed learning - State-space search - Symbolic inference
Data Management Layer	- Key-value store - RDF triple store - Graph store - Object store
Hardware Layer	- Orchestration - Parallelization - Virtualization
Data Acquisition Layer	- Distant supervision - Crowdsourcing - Crawling and archiving

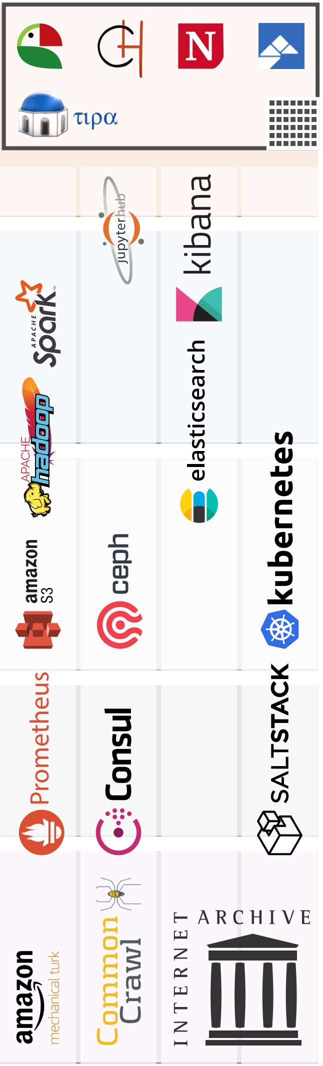
Vendor stack



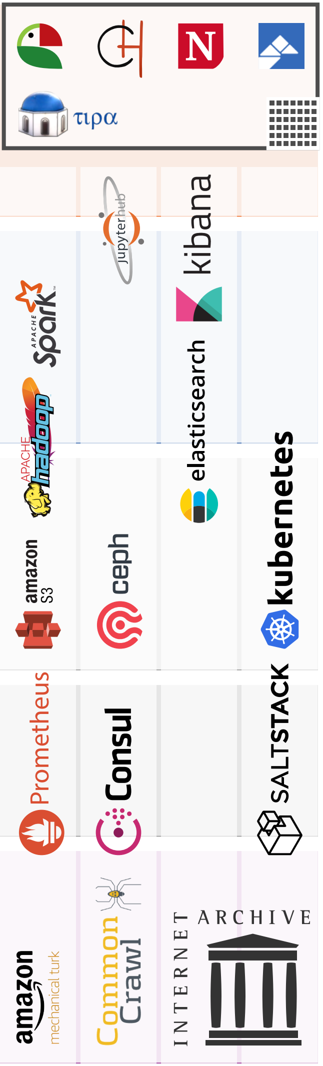
Webis Data Science Stack

	Task Stack	Technology stack
Data Consumption Layer	- Query and explore - Visualize and interact - Explain and justify	- Visual analytics - Immersive technologies - Intelligent agents
Data Analytics Layer	- Diagnose and reason - Structure identification - Structure verification	- Distributed learning - State-space search - Symbolic inference
Data Management Layer	- Provenance tracking - Normalization - Cleansing	- Key-value store - RDF triple store - Graph store - Object store
Hardware Layer	- Monitoring - Replication	- Orchestration - Parallelization - Virtualization
Data Acquisition Layer	- Replay - Collect - Log	- Distant supervision - Crowdsourcing - Crawling and archiving

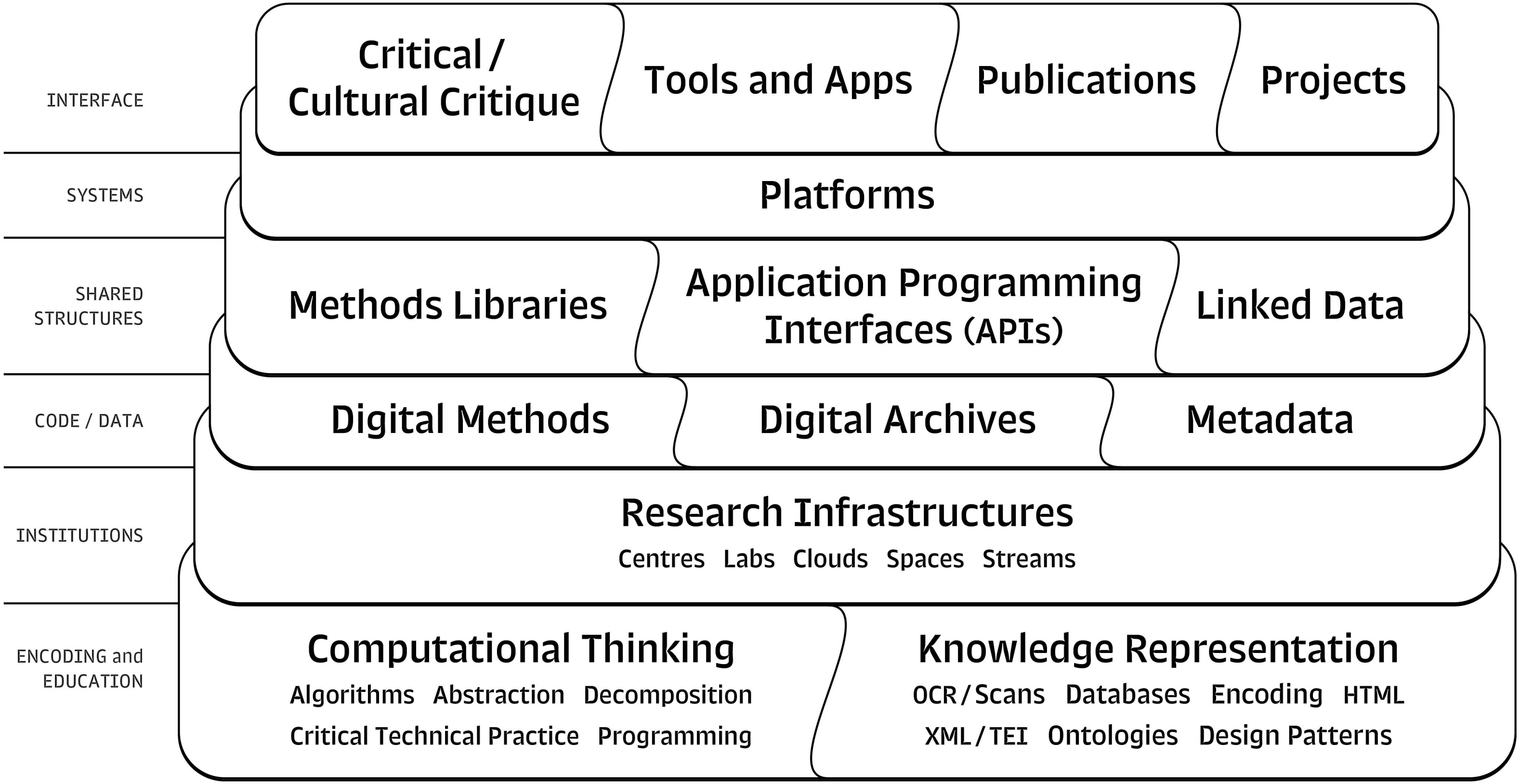
Vendor stack



Webis Data Science Stack

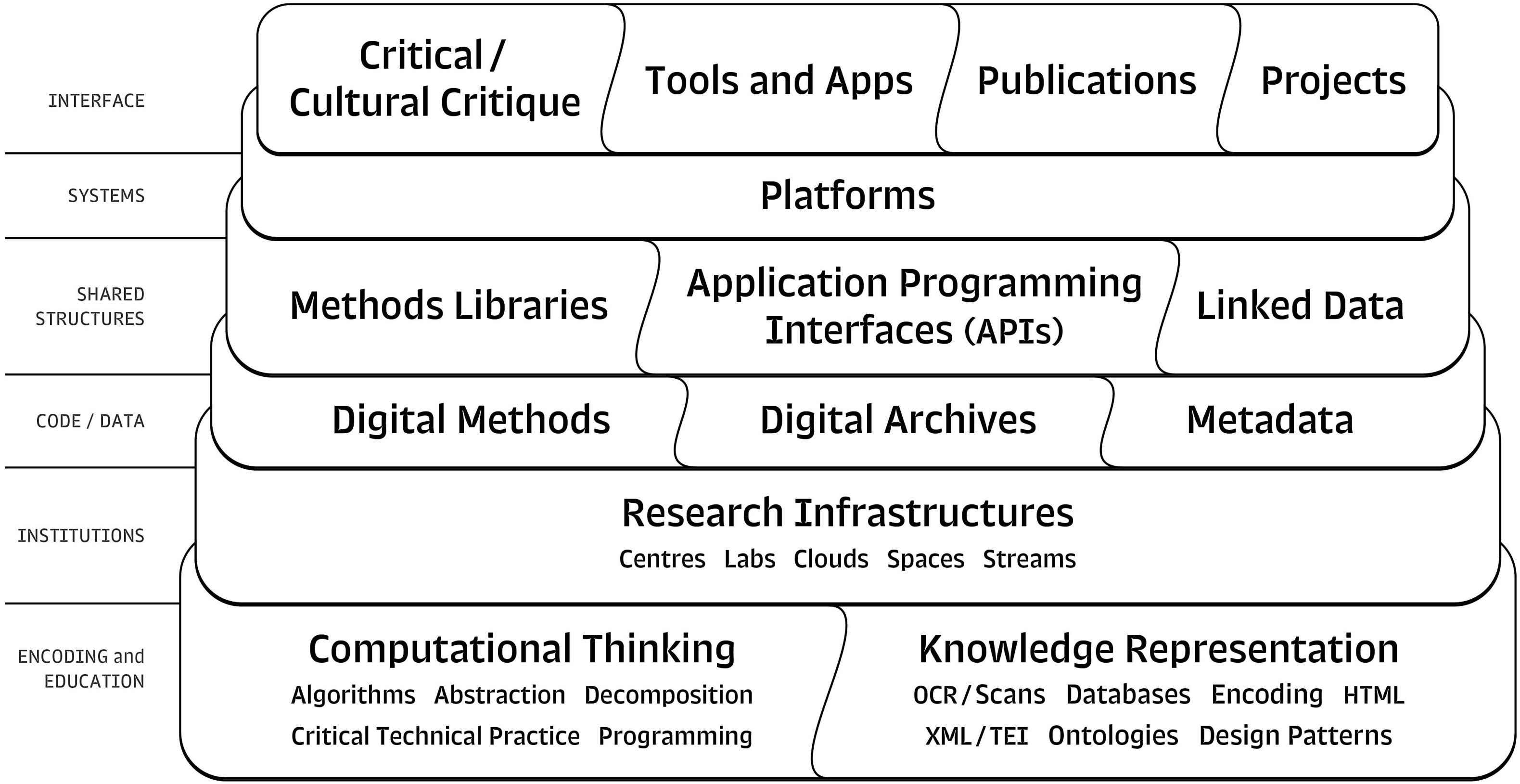
	Task Stack	Technology stack	Vendor stack	Roles
Data Consumption Layer	- Query and explore - Visualize and interact - Explain and justify	- Visual analytics - Immersive technologies - Intelligent agents		Experts: - IR - NLP - CSS - VA
Data Analytics Layer	- Diagnose and reason - Structure identification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

Digital Humanities Stack



[Berry and Fagerjord, 2017]

Digital Humanities Stack



INTERFACE: Critical / Cultural Critique | Tools and Apps | Publications | Projects

SYSTEMS: Platforms

SHARED STRUCTURES: Methods Libraries | Application Programming Interfaces (APIs) | Linked Data

CODE / DATA: Digital Methods | Digital Archives | Metadata

INSTITUTIONS: Research Infrastructures — Centres Labs Clouds Spaces Streams

ENCODING and EDUCATION: Computational Thinking — Algorithms Abstraction Decomposition Critical Technical Practice Programming | Knowledge Representation — OCR / Scans Databases Encoding HTML XML / TEI Ontologies Design Patterns

[Berry and Fagerjord, 2017]

Digital Humanities is a Data Science

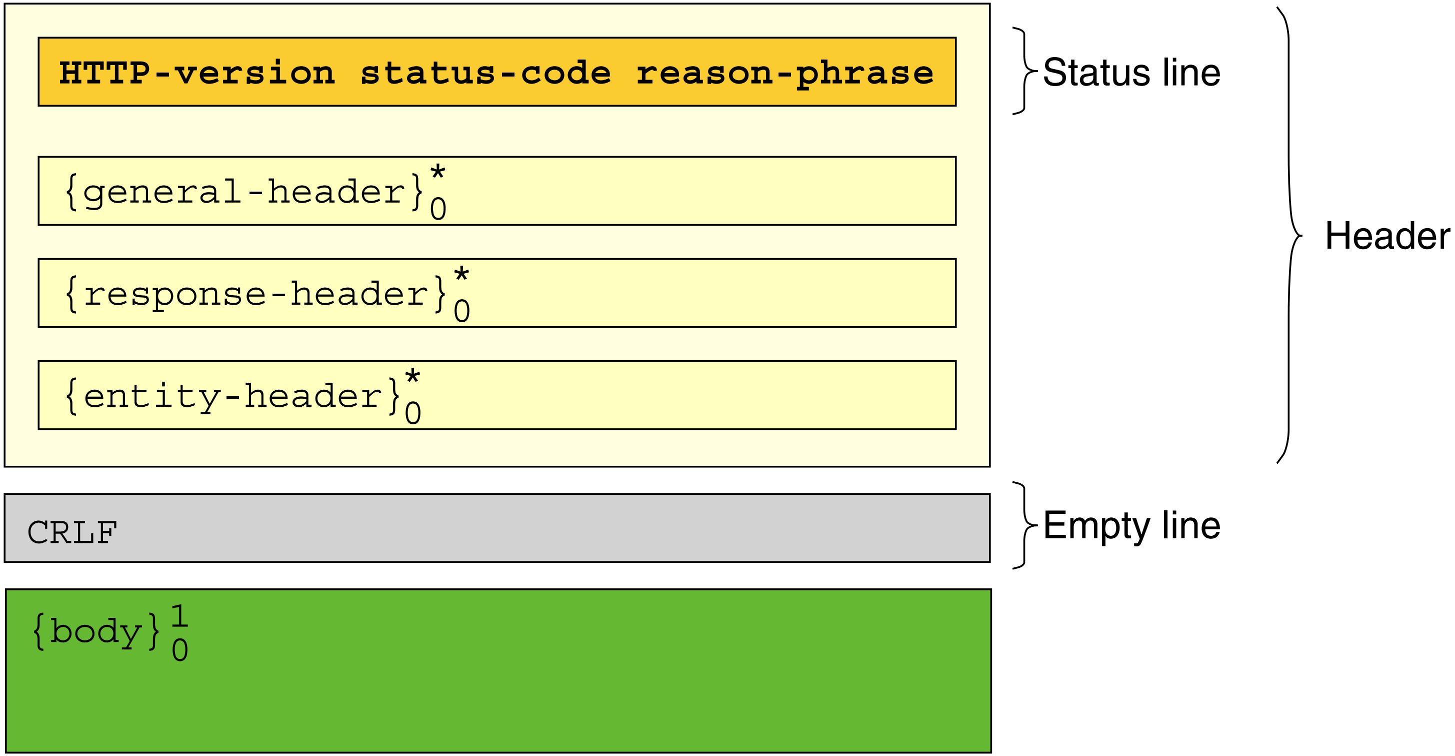
④

Web Archive Processing

Web Archive Data

WARC Standard

- 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:



HTTP-version status-code reason-phrase

{general-header}$_0^*$

{response-header}$_0^*$

{entity-header}$_0^*$

CRLF

{body}$_0^1$

Status line

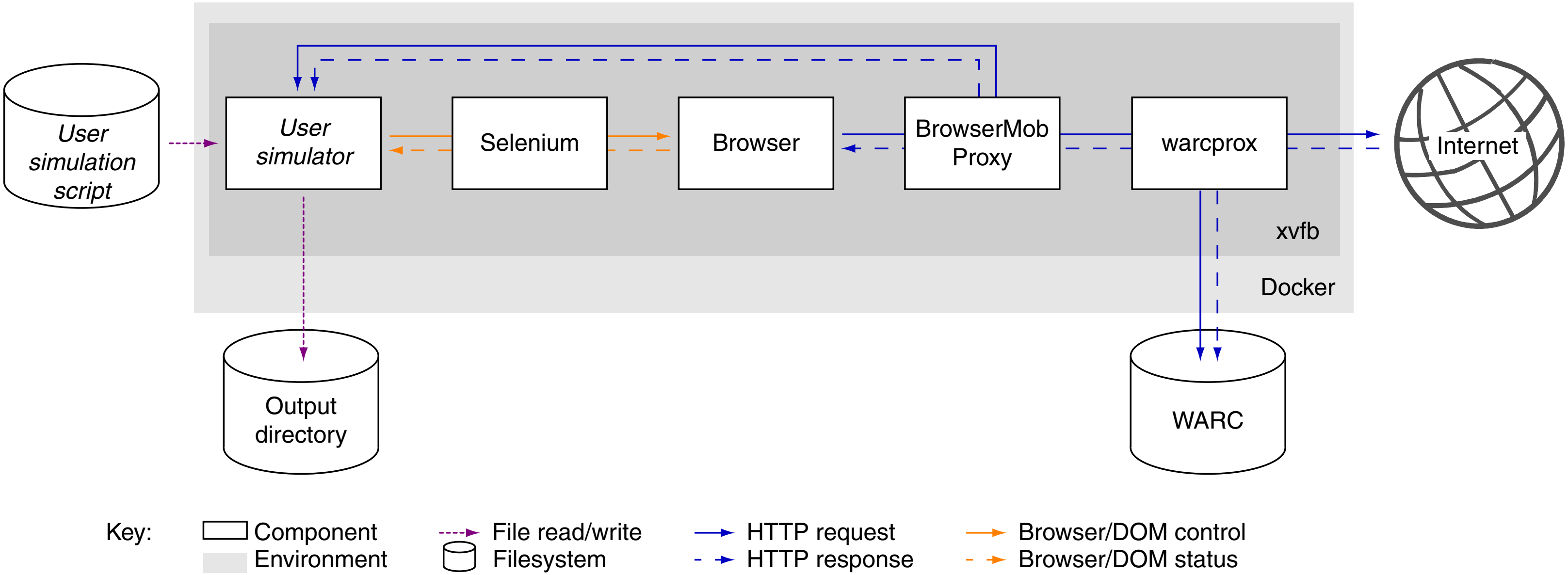
Header

Empty line

Web Archive Data

Web Archiving

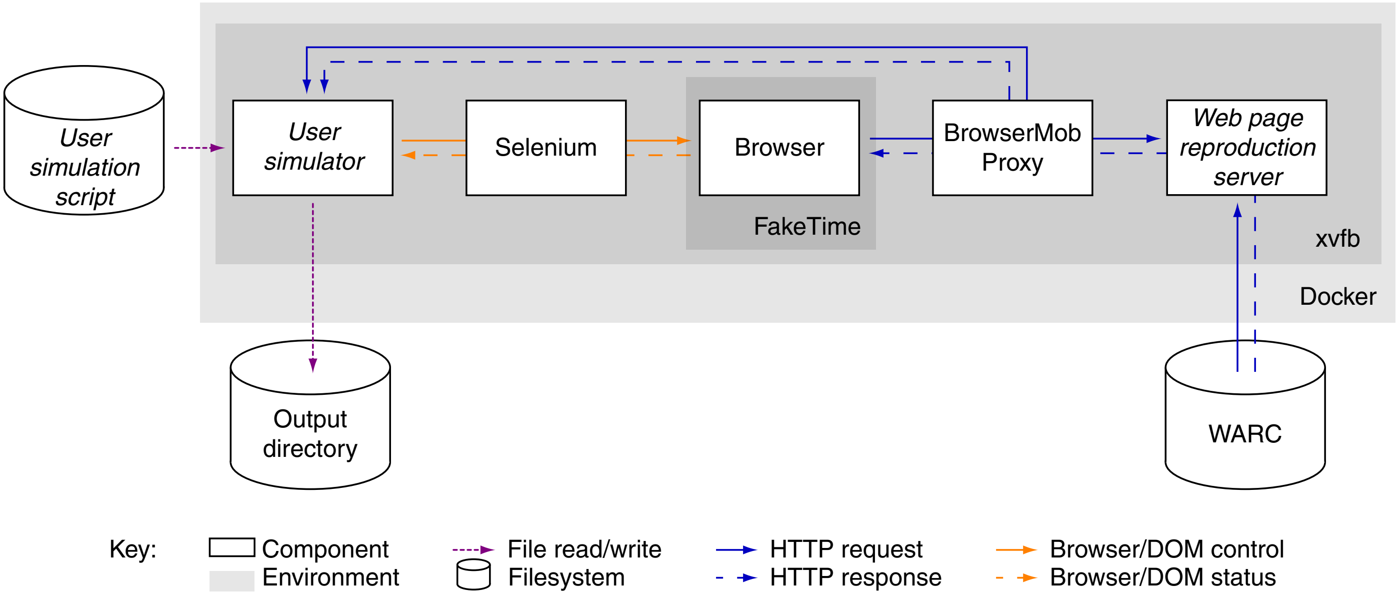
- ❑ A web page: Record all HTTP communication between browser and server.
- ❑ A browser is simulated to ensure the human-readable version is obtained.
- ❑ During web crawling, a web archiver “browses” every crawled page.



Web Archive Data

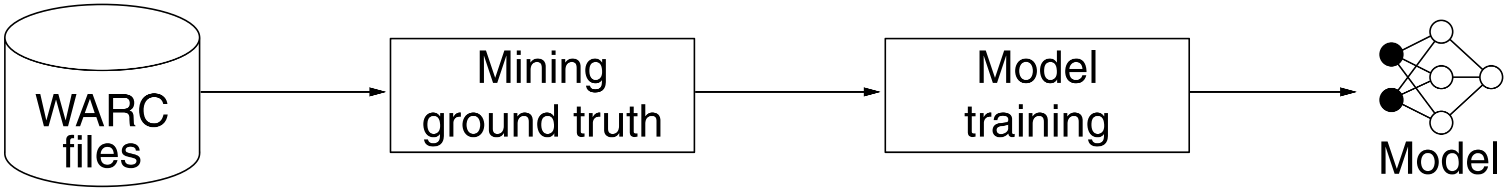
Web Archiving

- ❑ 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.



Web Archive Processing

Streamed Model Training Pipeline

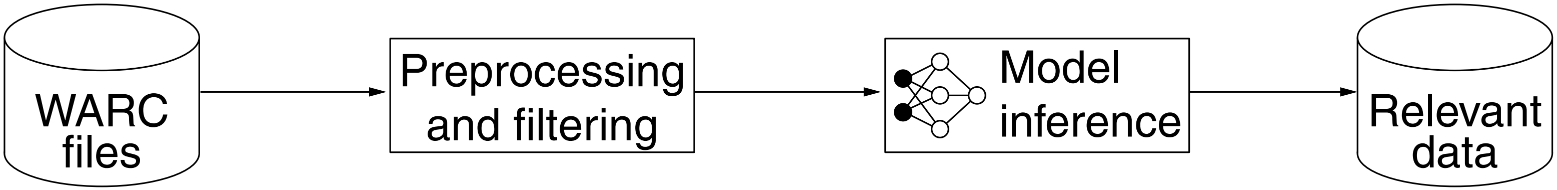


- ❑ Given a learning task and ground truth within WARC files, train a model.
  Only a fraction of the records within the WARC files are ground truth.

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

Web Archive Processing

Streamed Model Training Pipeline



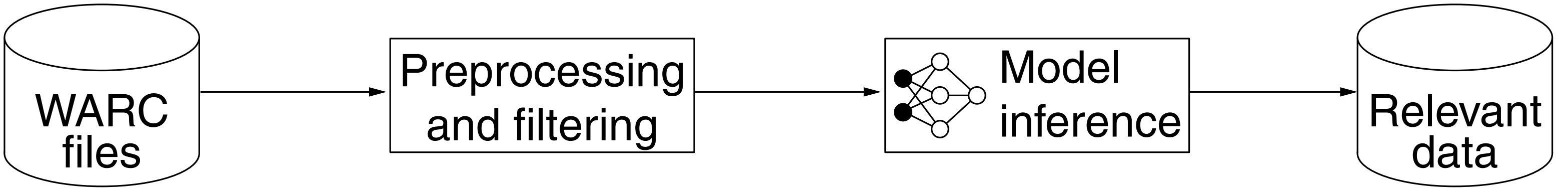
- Given a mining task and a trained (classification) model, collect relevant data.

  Only a fraction of the records within the WARC files are relevant.

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

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.

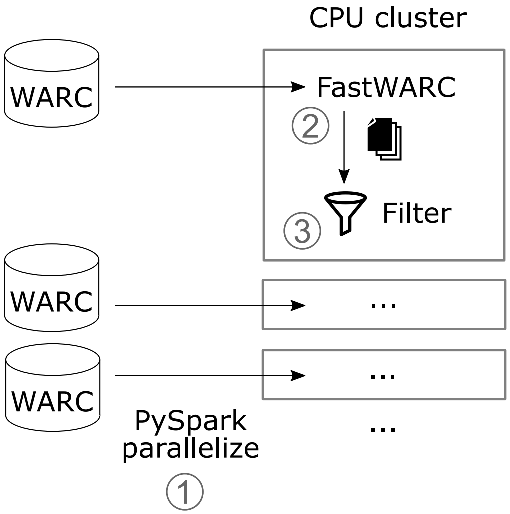
- ❑ Goal: Mining at web scale (billions of WARC files)

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

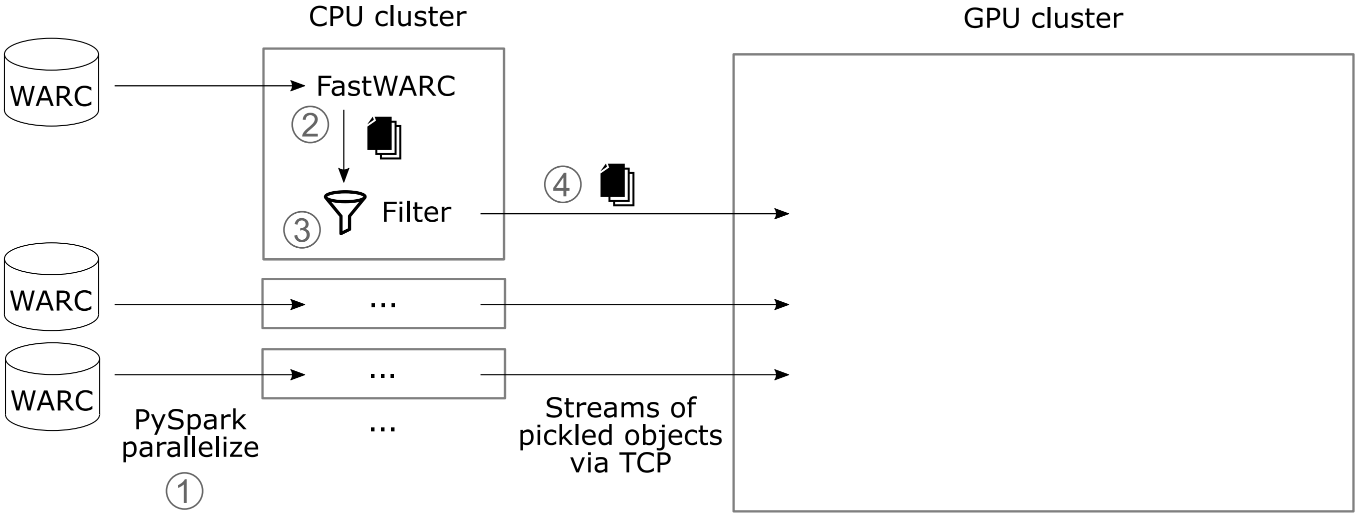
Streamed Model Training Pipeline



1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
3. First filtering step of records

Web Archive Processing

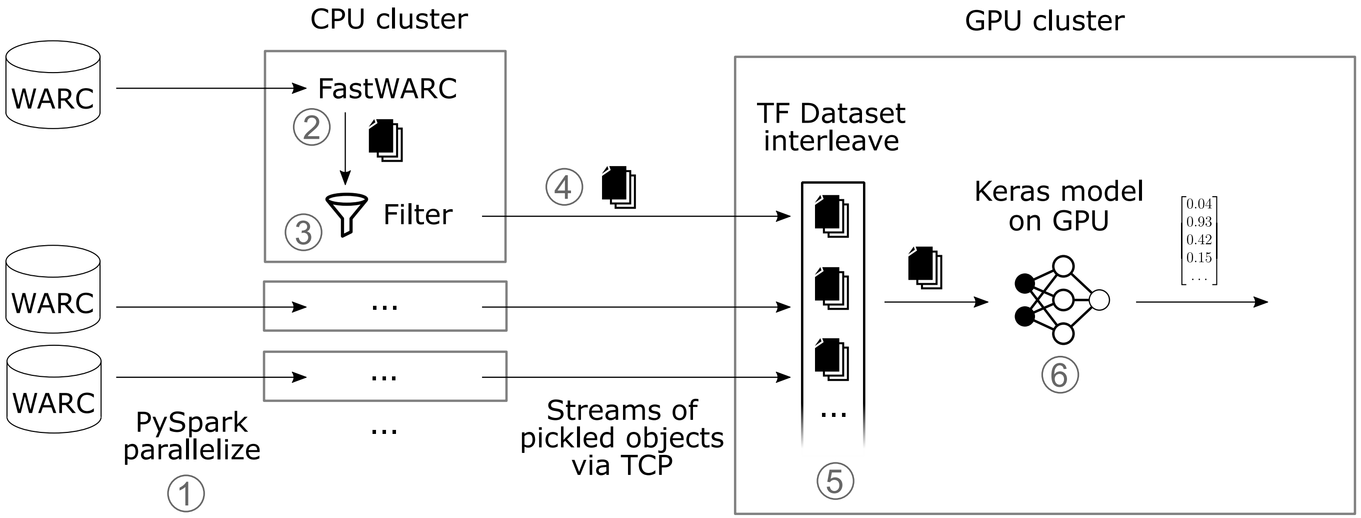
Streamed Model Training Pipeline



1. PySpark distributes WARCs among workers
2. FastWARC decompresses and iterates records
3. First filtering step of records
4. Pickled record streams

Web Archive Processing

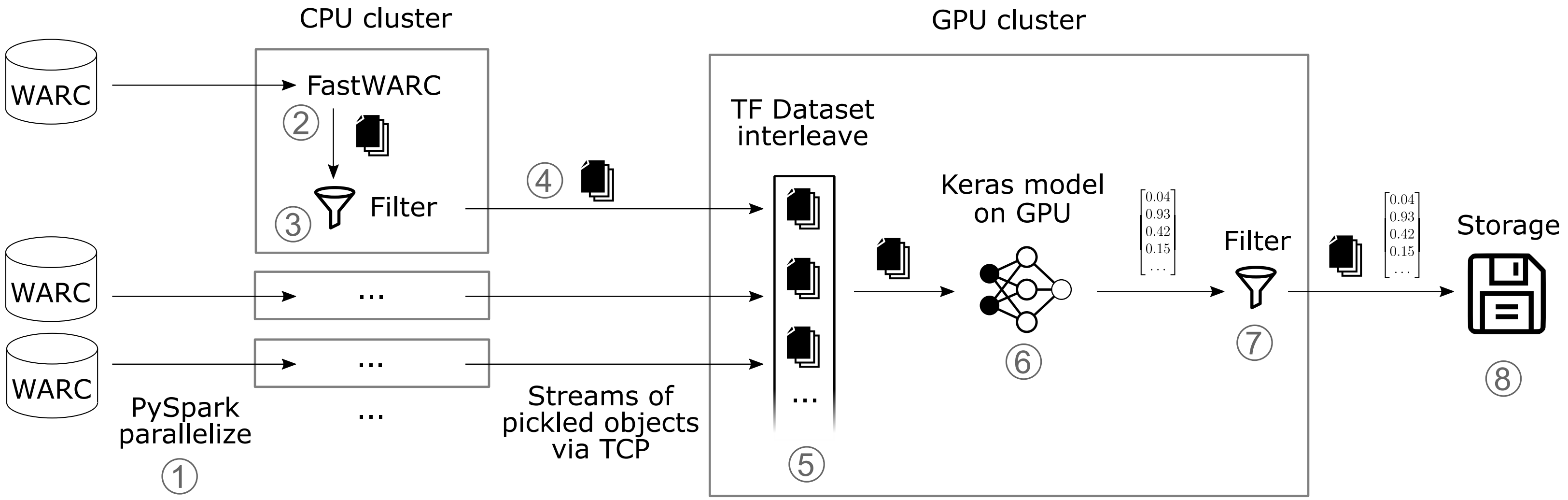
Streamed Model Training Pipeline



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

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
7. Second filtering based on classification results
8. Storage of relevant data

5

Webis Archive Research

Webis Archive Research [publications.webis.de]

Archival support

Argumentation
Language models
Search engines
Social sciences
Text reuse
Text synthesis

Webis Archive Research [publications.webis.de]

Archival support

Argumentation
Language models
Search engines
Social sciences
Text reuse
Text synthesis

❑ 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.

Webis Archive Research [publications.webis.de]

- 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.

Webis Archive Research [publications.webis.de]

- 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?

Webis Archive Research [publications.webis.de]

Archival support

Argumentation
Language models
Search engines
Social sciences
Text reuse
Text synthesis

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.

Webis Archive Research [publications.webis.de]

- 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?

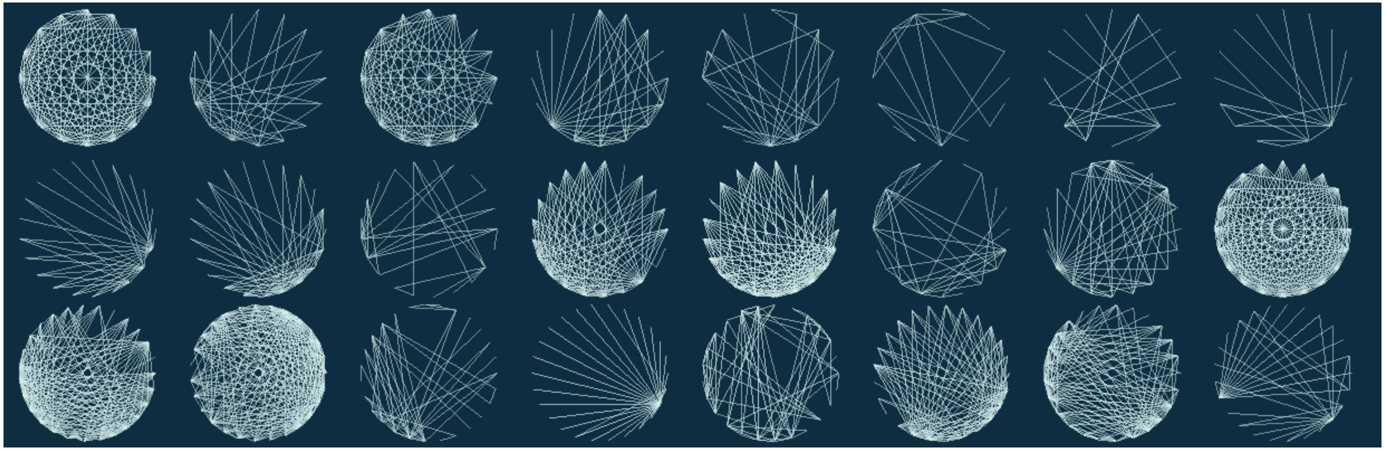
  Applying author identification technology at web-scale.

- 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

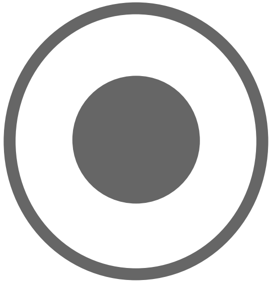
Webis Archive Research [publications.webis.de]

- Abstractive Snippet Generation

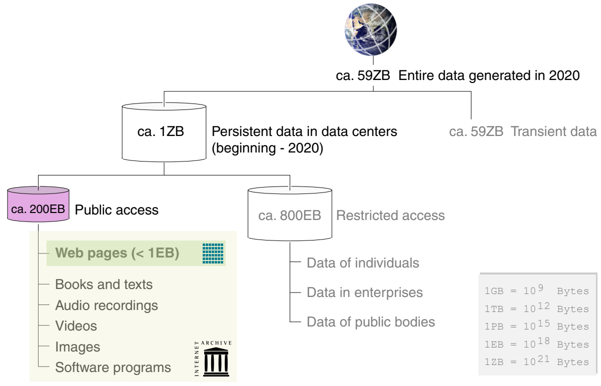
  Approach: Use of anchor contexts to generate abstractive snippets with a pointer-generator network, exploiting ClueWeb09, ClueWeb12, and the DMOZ Open Directory Project.

- Automatic Summarization

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



Summary



ca. 59ZB Entire data generated in 2020

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 59ZB Transient data

ca. 200EB Public access

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

ca. 800EB Restricted access

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

$1GB = 10^{9}$ Bytes

$1TB = 10^{12}$ Bytes

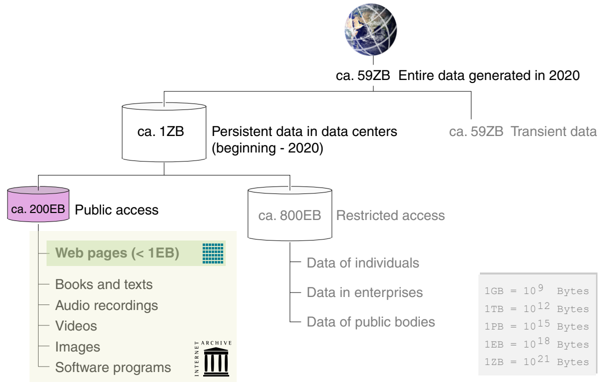
$1PB = 10^{15}$ Bytes

$1EB = 10^{18}$ Bytes

$1ZB = 10^{21}$ Bytes

Global Datasphere

Summary



ca. 59ZB Entire data generated in 2020

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 59ZB Transient data

ca. 200EB Public access

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

ca. 800EB Restricted access

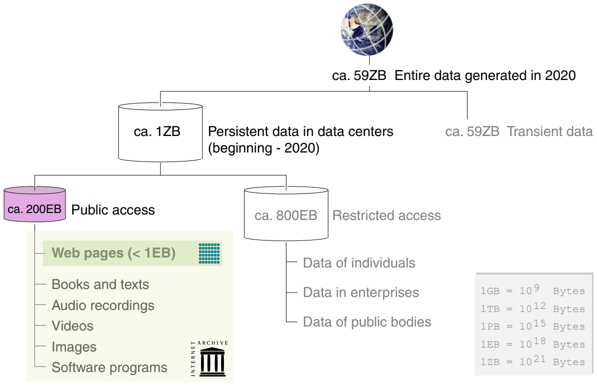
- Data of individuals
- Data in enterprises
- Data of public bodies

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

Global Datasphere

Internet Archive

Summary



ca. 59ZB Entire data generated in 2020

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 59ZB Transient data

ca. 200EB Public access

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

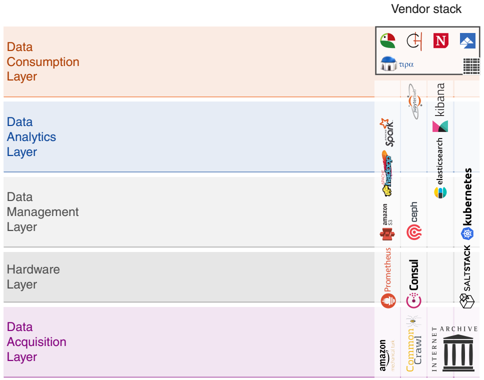
ca. 800EB Restricted access

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

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

Global Datasphere

Internet Archive



Vendor stack

Data Consumption Layer

Data Analytics Layer

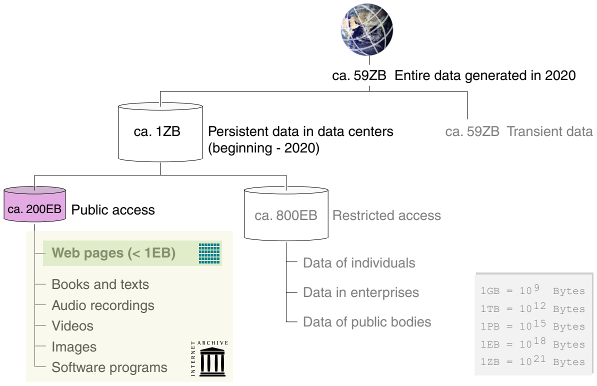
Data Management Layer

Hardware Layer

Data Acquisition Layer

Webis Analytics Stack

Summary



ca. 59ZB Entire data generated in 2020

ca. 1ZB Persistent data in data centers (beginning - 2020)

ca. 59ZB Transient data

ca. 200EB Public access

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

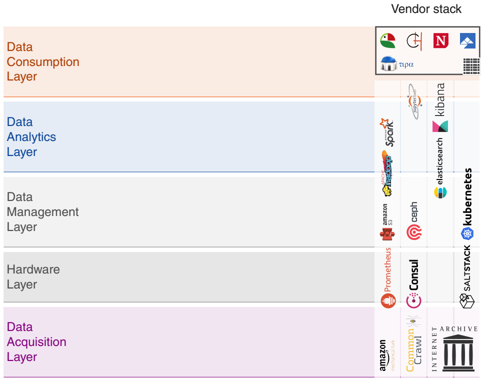
ca. 800EB Restricted access

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

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

Global Datasphere

Internet Archive



Vendor stack

Data Consumption Layer

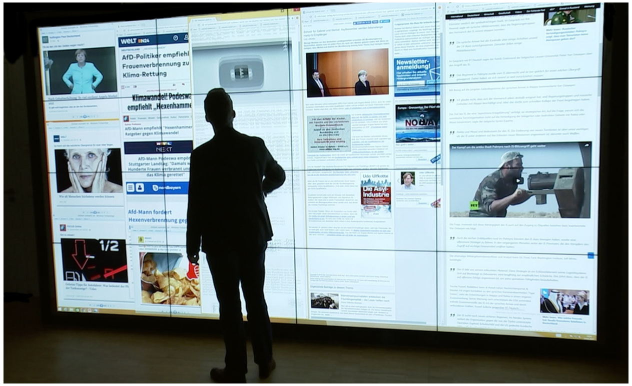
Data Analytics Layer

Data Management Layer

Hardware Layer

Data Acquisition Layer

Webis Analytics Stack



Webis Archive Research

Digital Humanities and the Web Archive

- What questions does digital humanities ask of Web Archives?
- What are classes of research questions for digital humanities?
- How can they be operationalized?
- Can digital humanities scale (be quantifiable, yet representative)?
- How much of recent history and culture does the Web Archive cover?
- What information is missing to fill in gaps?
- ...

Digital Humanities and the Web Archive

- What questions does digital humanities ask of Web Archives?
- What are classes of research questions for digital humanities?
- How can they be operationalized?
- Can digital humanities scale (be quantifiable, yet representative)?
- How much of recent history and culture does the Web Archive cover?
- What information is missing to fill in gaps?
- ...

Thank You!