

# NFDI Web



Harnessing the Web  
for Computer Science,  
the Social Sciences,  
and the Humanities

# NFDI Web: Consortium

## L3S Hanover

Prof. Dr. Wolfgang Nejdl  
Prof. Dr. Avishek Anand

## RWTH Aachen

Prof. Dr. Markus Strohmaier

## MLU Halle-Wittenberg

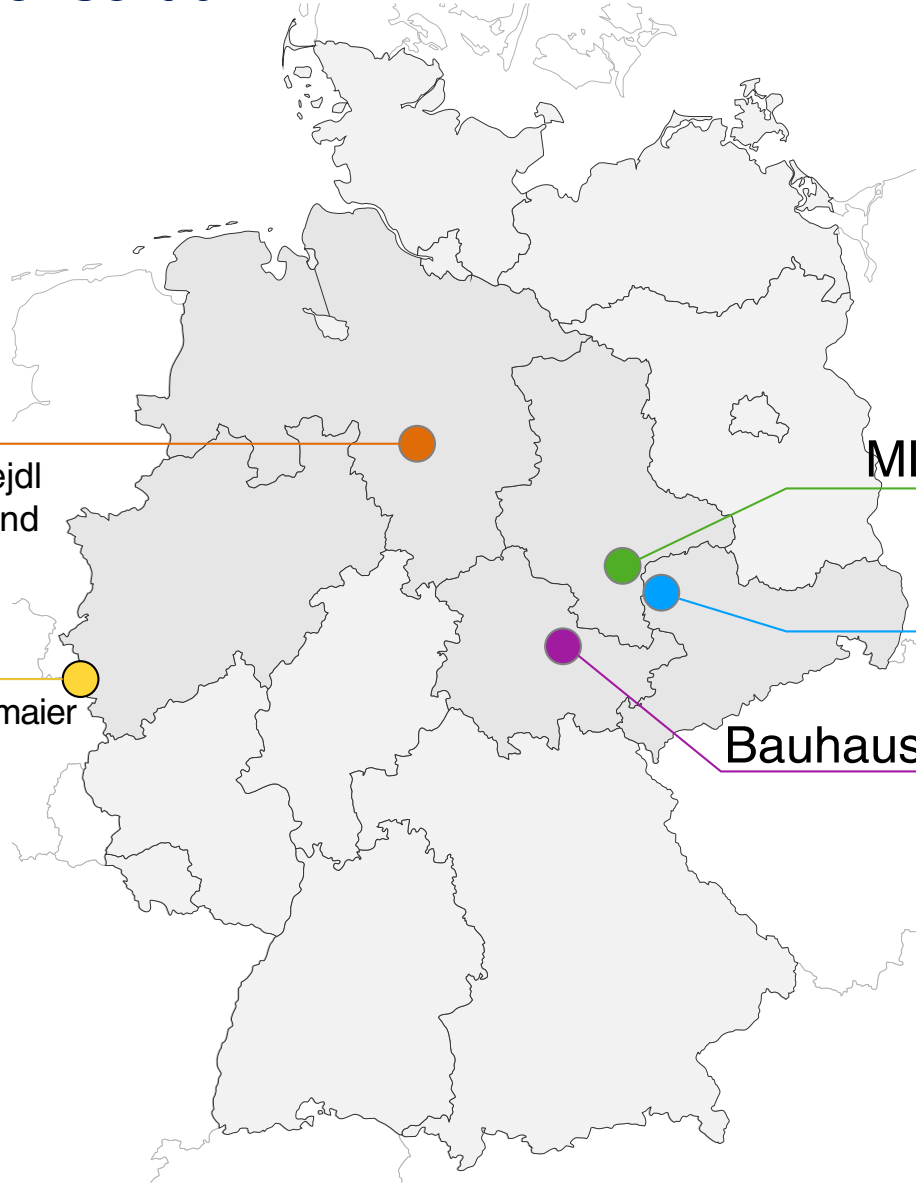
Prof. Dr. Matthias Hagen

## Leipzig University

Prof. Dr. Martin Potthast

## Bauhaus-Universität Weimar

Prof. Dr. Benno Stein



# NFDI Web: Scientific Community

Tier 1

**Web Exploitation and Analytics**



Tier 1

**Infrastructure**



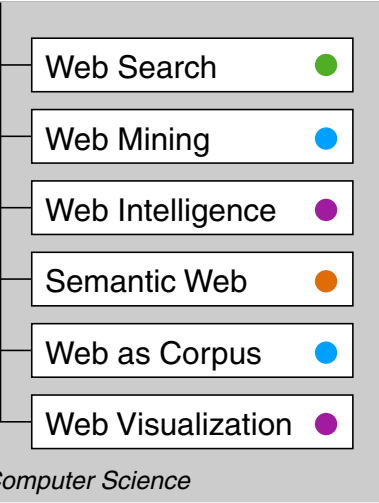
-  Hanover
-  Leipzig
-  Halle
-  Aachen
-  Weimar

# NFDI Web: Scientific Community

Tier 1

**Web Exploitation and Analytics** 

Tier 2



Tier 1

**Infrastructure** 


-  Hanover
-  Leipzig
-  Halle
-  Aachen
-  Weimar


# NFDI Web: Scientific Community

Tier 1

**Web Exploitation and Analytics** 

Tier 2


Web Search 

Web Mining 

Web Intelligence 


Semantic Web 

Web as Corpus 

Web Visualization 

*Computer Science*

Social Web 

Web Economics 

Web Media Studies 

Web Culture 

Digital History 

Digital Libraries 

*Social Sciences, Humanities*

Tier 1

**Infrastructure** 

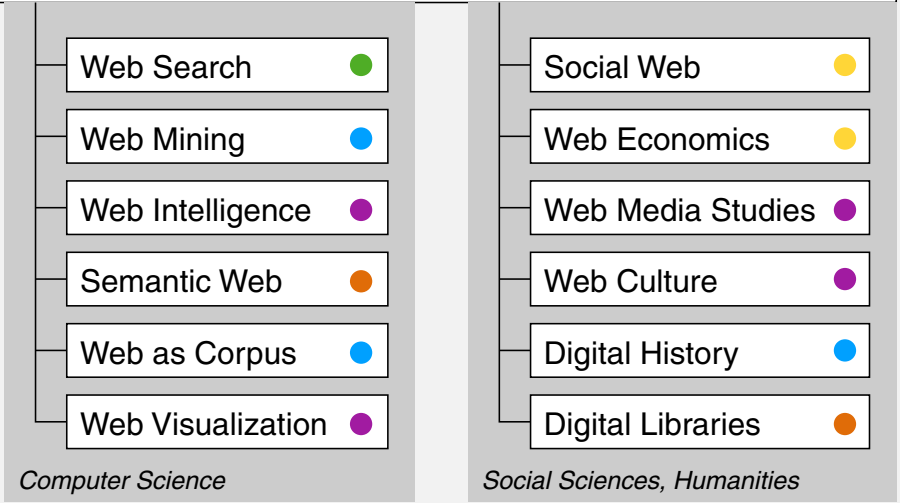
-  Hanover
-  Leipzig
-  Halle
-  Aachen
-  Weimar

# NFDI Web: Scientific Community

Tier 1

## Web Exploitation and Analytics

Tier 2



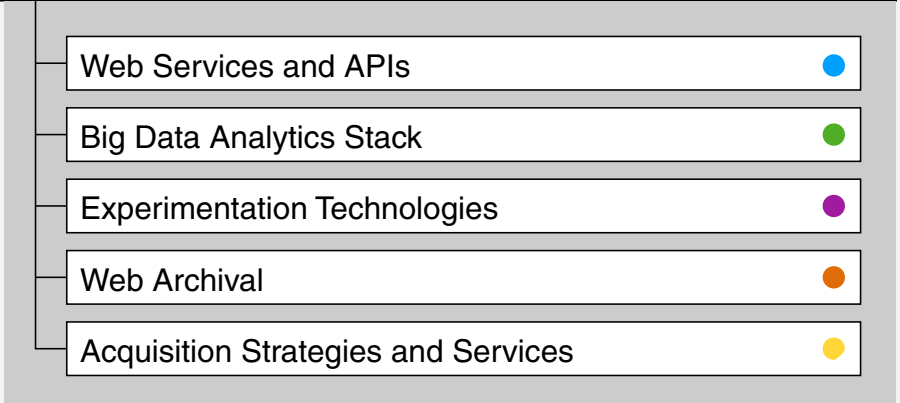
*Computer Science*

*Social Sciences, Humanities*

Tier 1

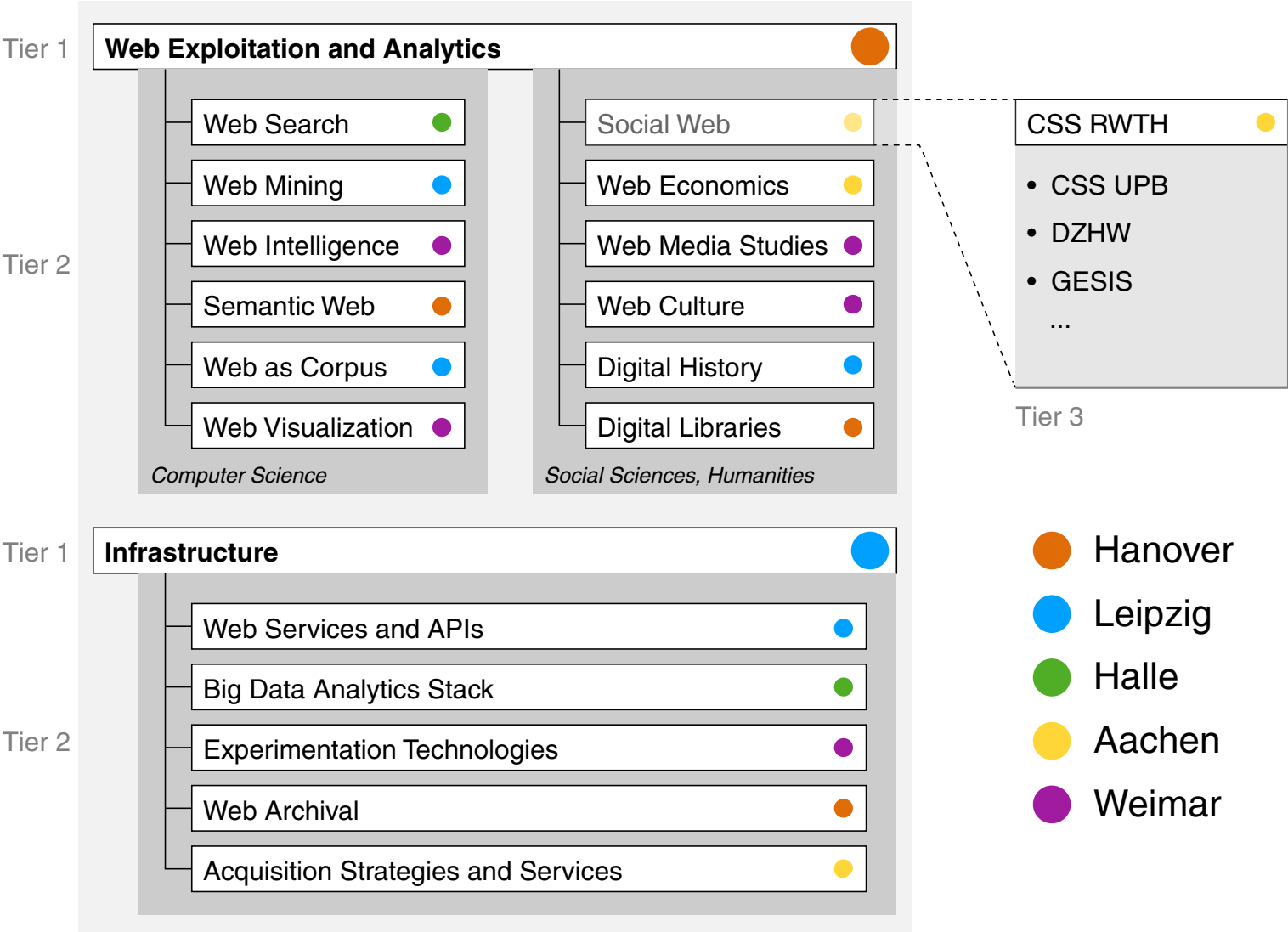
## Infrastructure

Tier 2



-  Hanover
-  Leipzig
-  Halle
-  Aachen
-  Weimar

# NFDI Web: Scientific Community



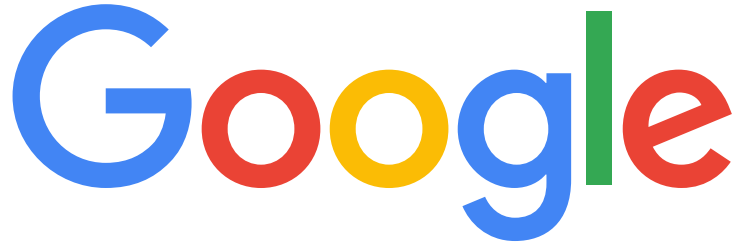
# NFDI Web: Data

One copy of the Web, please.



# NFDI Web: Data

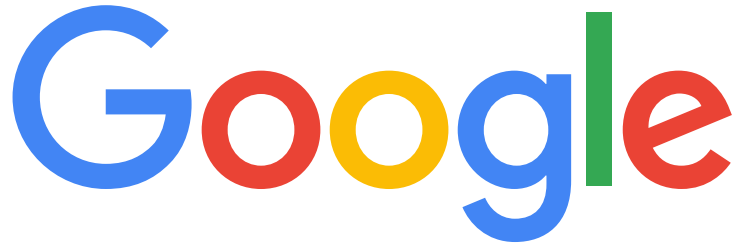
One copy of the Web, please.



- ❑ Started 1996, founded 1998
- ❑ Founders: Larry Page, Sergey Brin  
Then PhD students at Stanford
- ❑ Among the first to crawl the Web
- ❑ Built a billion-dollar business

# NFDI Web: Data

One copy of the Web, please.



- ❑ Started 1996, founded 1998
- ❑ Founders: Larry Page, Sergey Brin  
Then PhD students at Stanford
- ❑ Among the first to crawl the Web
- ❑ Built a billion-dollar business

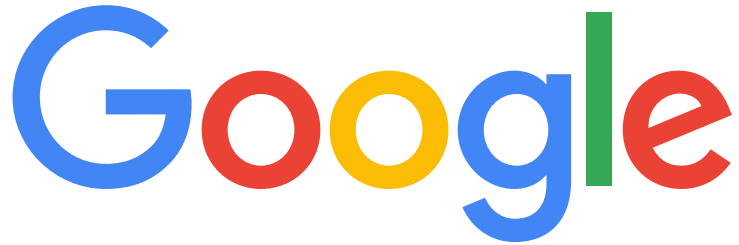


[\[www.archive.org\]](http://www.archive.org)

- ❑ Founded 1996
- ❑ Founder: Brewster Kahle  
Then experienced Internet entrepreneur.
- ❑ Archives the web; all things digital
- ❑ Built the first free digital library

# NFDI Web: Data

One copy of the Web, please.



[\[www.archive.org\]](http://www.archive.org)

- ❑ Started 1996, founded 1998
- ❑ Founders: Larry Page, Sergey Brin  
Then PhD students at Stanford
- ❑ Among the first to crawl the Web
- ❑ Built a billion-dollar business
- ❑ Founded 1996
- ❑ Founder: Brewster Kahle  
Then experienced Internet entrepreneur.
- ❑ Archives the web; all things digital
- ❑ Built the first free digital library

Mission: make the world's information **universally accessible**

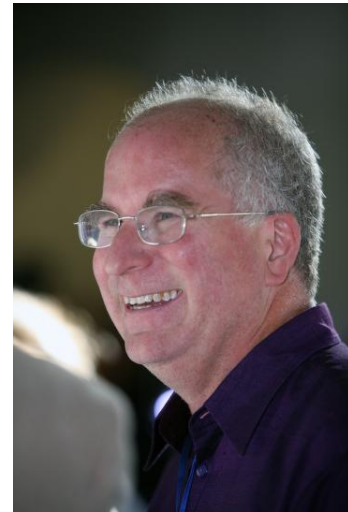
# NFDI Web: Data

One copy of the Web. Coming right up.

- ❑ Web Archive Collection
- ❑ recorded since 1996
- ❑ ~ 17 petabyte
- ❑ ~ 750 billion web pages
- ❑ Accessible at the [Wayback Machine](http://www.archive.org)



[\[www.archive.org\]](http://www.archive.org)



Brewster Kahle, Founder

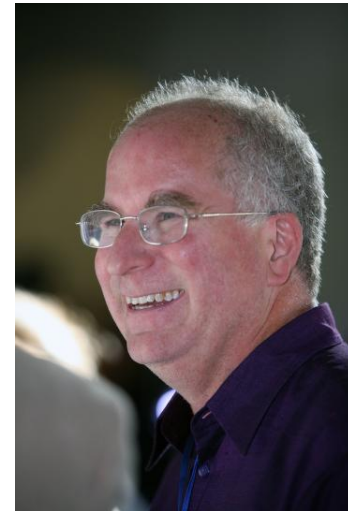
# NFDI Web: Data

One copy of the Web. Coming right up.

- ❑ Web Archive Collection
  - ❑ recorded since 1996
  - ❑ ~ 17 petabyte
  - ❑ ~ 750 billion web pages
  - ❑ Accessible at the [Wayback Machine](http://www.archive.org)
- 
- ❑ One full copy in San Francisco
  - ❑ Part at the new Library of Alexandria
  - ❑ Part in Amsterdam



[\[www.archive.org\]](http://www.archive.org)



Brewster Kahle, Founder

# NFDI Web: Data

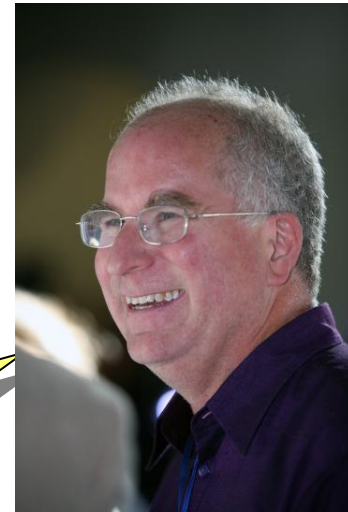
One copy of the Web. Coming right up.

- ❑ Web Archive Collection
  - ❑ recorded since 1996
  - ❑ ~ 17 petabyte
  - ❑ ~ 750 billion web pages
  - ❑ Accessible at the [Wayback Machine](#)
- 
- ❑ One full copy in San Francisco
  - ❑ Part at the new Library of Alexandria
  - ❑ Part in Amsterdam

I won't sleep until there are at least five or six backup sites!



[\[www.archive.org\]](http://www.archive.org)



Brewster Kahle, Founder

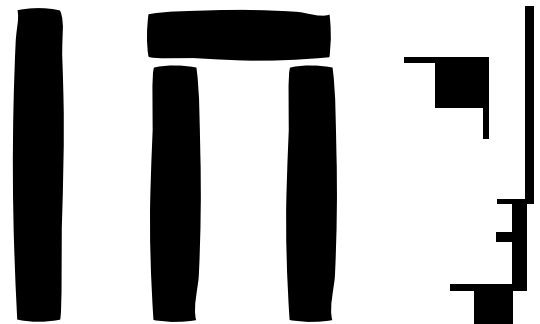
# NFDI Web: Data

One copy of the Web. Coming right up.

- ❑ Web Archive Collection
  - ❑ recorded since 1996
  - ❑ ~ 17 petabyte
  - ❑ ~ 750 billion web pages
  - ❑ Accessible at the [Wayback Machine](https://www.archive.org)
- 
- ❑ Immersive Web Observatory at Bauhaus-Universität Weimar
  - ❑ BMBF-funded infrastructure
  - ❑ ~ 8 petabyte of web pages



[\[www.archive.org\]](http://www.archive.org)



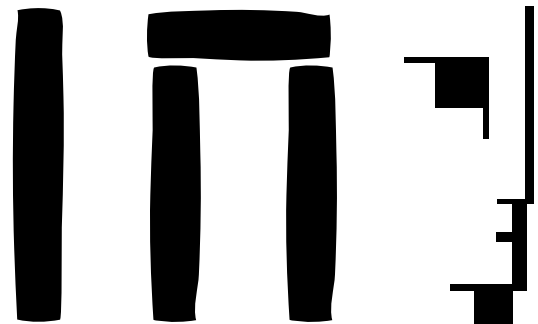
# NFDI Web: Data

One copy of the Web. Coming right up.

- ❑ Web Archive Collection
  - ❑ recorded since 1996
  - ❑ ~ 17 petabyte
  - ❑ ~ 750 billion web pages
  - ❑ **Accessible** at the [Wayback Machine](https://www.archive.org)
- 
- ❑ Immersive Web Observatory at Bauhaus-Universität Weimar
  - ❑ BMBF-funded infrastructure
  - ❑ ~ 8 petabyte of web pages













[\[www.archive.org\]](http://www.archive.org)



Mission: make the world's information **universally processable**



# NFDI Web: Infrastructure

	Task Stack	Technology Stack	Vendor Stack	Roles
Data Consumption Layer	<ul style="list-style-type: none"> <li>- Query and explore</li> <li>- Visualize and interact</li> <li>- Explain and justify</li> </ul>	<ul style="list-style-type: none"> <li>- Visual analytics</li> <li>- Immersive technologies</li> <li>- Intelligent agents</li> </ul>		Data scientist
Data Analytics Layer	<ul style="list-style-type: none"> <li>- Diagnose and reason</li> <li>- Structure identification</li> <li>- Structure verification</li> </ul>	<ul style="list-style-type: none"> <li>- Distributed learning</li> <li>- State-space search</li> <li>- Symbolic inference</li> </ul>		
Data Management Layer	<ul style="list-style-type: none"> <li>- Provenance tracking</li> <li>- Normalize</li> <li>- Cleansing</li> </ul>	<ul style="list-style-type: none"> <li>- Key-value store</li> <li>- RDF triple store</li> <li>- Object store</li> </ul>	  	System architect
Hardware Layer	<ul style="list-style-type: none"> <li>- Virtualization</li> <li>- Orchestration</li> </ul>	<ul style="list-style-type: none"> <li>- Replication</li> <li>- Parallelization</li> </ul>	  	System admin
Data Acquisition Layer	<ul style="list-style-type: none"> <li>- Replay</li> <li>- Log</li> <li>- Collect</li> </ul>	<ul style="list-style-type: none"> <li>- Distant supervision</li> <li>- Crowdsourcing</li> <li>- Crawling and archiving</li> </ul>	 	Data scientist

# NFDI Web: Community Needs, Status and Plans

- ❑ Computer Science
  - Web-based Artificial intelligence: distant and weak supervision
  - Knowledge extraction
  - Next-generation web search: argument search
  - Social network analysis
  - Peta-scale benchmark engineering
  
- ❑ Computational Social Sciences + Digital Humanities
  - Tracing societal / cultural processes on the web
  - Changing biases on the web over time
  - Who wrote the web?
  
- ❑ Status and future plans:
  - We have: Foundation to show web exploitation and analysis at scale.
  - We plan: Scaling the infrastructure to the entire Internet Archive.

# NFDI Web: Benefits, Infrastructure Needs, Data Management

## □ Key contributions

- Lowering the bar for academics to reach **industry scales**
- Infrastructure for web-based learning / AI
- **Provenance and reproducibility** of web analysis results
- **Benchmarks** and reference models
- Everything is **public domain**

## □ Our requirements

- Human resources to support development
- Scaling up hardware to match future growth of the archive
- High-bandwidth network between consortium partners

## □ Research Data Management

- API + tool development to ease external analytics
- Registry for derived data
- Reproducible research through software submission for benchmarks