

## Web Archive Analytics\*

Michael Völske,<sup>1</sup> Janek Bevendorff,<sup>1</sup> Johannes Kiesel,<sup>1</sup> Benno Stein,<sup>1</sup>  
Maik Fröbe,<sup>2</sup> Matthias Hagen,<sup>2</sup> Martin Potthast<sup>3</sup>

**Abstract:** Web archive analytics is the exploitation of publicly accessible web pages and their evolution for research purposes—to the extent organizationally possible for researchers. In order to better understand the complexity of this task, the first part of this paper puts the entirety of the world’s captured, created, and replicated data (the “Global Datasphere”) in relation to other important data sets such as the public internet and its web pages, or what is preserved thereof by the Internet Archive. Recently, the Webis research group, a network of university chairs to which the authors belong, concluded an agreement with the Internet Archive to download a substantial part of its web archive for research purposes. The second part of the paper in hand describes our infrastructure for processing this data treasure: We will eventually host around 8 PB of web archive data from the Internet Archive and Common Crawl, with the goal of supplementing existing large scale web corpora and forming a non-biased subset of the 30 PB web archive at the Internet Archive.

**Keywords:** Big Data Analytics; Web Archive; Internet Archive; Webis Infrastructure

### 1 The Global Datasphere

In the few decades since its beginnings, the World Wide Web has embedded itself into the fabric of human society at a speed only matched by few technologies that came before it, and in the process, it has given an enormous boost to scientific endeavors across all disciplines. Particularly the discipline of web science—in which the web itself, as a socio-technical system, becomes the subject of inquiry—has proven a productive field of research [KG03, Be06]. It is well understood that the vast size and sheer chaotic structure of the web pose a challenge to those that seek to study it [He08]; yet its precise dimensions are not at all easily quantified.

In what follows, we derive an informed estimate of the web’s size and conclude that it represents, in fact, only a small fraction of all existing data. To this end, we start by considering the broader question of how much data there is overall. This question can be asked in two ways: First, how much data is generated continuously as a result of human activity; and secondly, how much data is persistently stored? To answer the first question is to determine (according to the International Data Corporation) the size of

---

\* An abstract of this paper has been published in the proceedings of the Open Search Symposium 2020.

<sup>1</sup> Bauhaus-Universität Weimar, Germany. <first>.<last>@uni-weimar.de

<sup>2</sup> Martin-Luther-Universität Halle-Wittenberg, Germany. <first>.<last>@informatik.uni-halle.de

<sup>3</sup> Leipzig University, Germany. martin.pothast@uni-leipzig.de

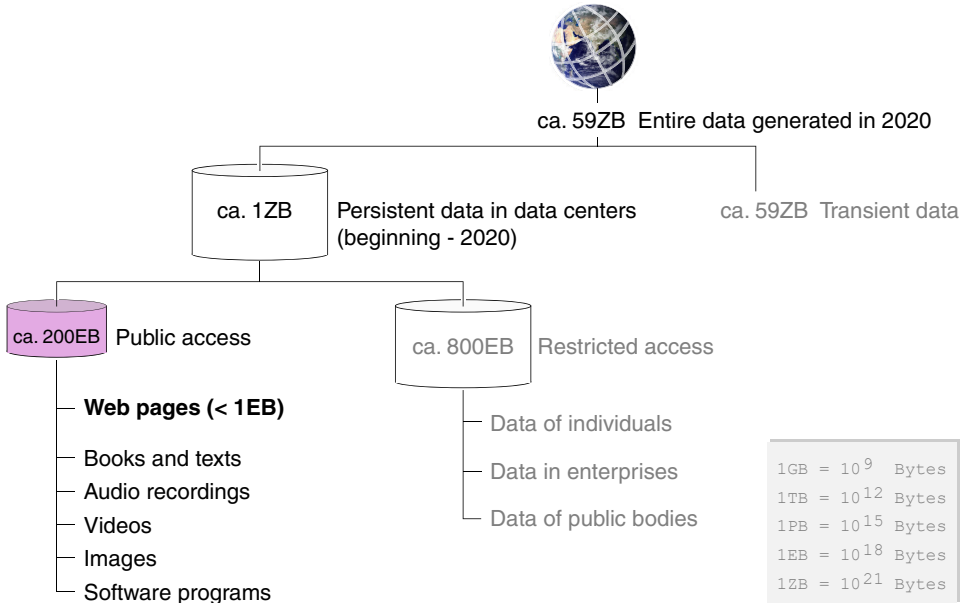


Fig. 1: Overview of the Global DataSphere in 2020. The shown numbers are estimates based on analyses and market forecasts by Cisco Systems, the International Data Corporation, and Statista Inc., among others.

the Global Datasphere, a “measure of all new data captured, created, and replicated in a single year.” [RGR18]. Thanks to efforts at tracking the global market for computer storage devices, spearheaded by storage manufacturers and market researchers alike, we can make an educated guess at how much data that might be (see Figure 1 for an illustration).

Analyses like the “Data Never Sleeps” series [DI19, DI20] by business intelligence company Domo, Inc., provide pieces of the puzzle: While as recently as 2012, only a third of the world’s population had internet access [DI17], that figure has risen to 59%, or 4.5 billion people, today. Throughout 2018, the combined activity of internet users produced almost two petabytes per minute on average [DI19]. This year, in one minute of any given day, users upload 500 hours worth of video to YouTube, and more than a billion of them initiate voice or video calls [DI20]. This combined effort of internet users, however, is nowhere near the total volume of the Global Datasphere. In the same period, as automatic cameras record security footage and ATMs and stock markets transmit banking transactions, the ATLAS particle detector at CERN’s Large Hadron Collider records a staggering amount of nearly four petabytes worth of particle collisions [Br11].

Past and current projections by the aforementioned IDC complete our picture of the Global Datasphere’s size. Already in 2014, more than four zettabytes of data were created yearly,

and this amount has since increased at a rate of more than 150 % per year [Tu14]. The total figure had grown to 33 zettabytes by 2018 [RGR18] and was previously projected to reach 44 zettabytes by 2020, and between 163 and 175 zettabytes by 2025 [Tu14, RGR18, RGR20]. However, in the wake of the COVID-19 pandemic, the growth of the global datasphere has sharply accelerated and updated projections expect it to exceed 59 zettabytes (or  $5.9 \cdot 10^{22}$  bytes) already by the end of 2020 [RRG20].

To put this number into context, consider the following thought experiment: The largest commercially available spinning hard disk drives at the time of writing hold 18 terabytes in a 3.5-inch form factor. To store the entirety of the Global Datasphere would require more than 3.2 billion such disks, which—densely packed together—would more than fill the volume of the Empire State Building and almost cost the equivalent of Apple’s two-trillion-dollar market capitalization in retail price.<sup>4</sup> The densest SSD storage available today would reduce the volume requirements six-fold, while increasing the purchase cost by a factor of eight. However, even by the year 2025, four out of every five stored bytes are still expected to reside on rotating-platter hard drives [RGR20, RR20].

While all of the above makes the vastness of the Global Datasphere quite palpable, it leaves out one key point: Most of the 59 ZB is transient data that is never actually persistently stored. Once again, CERN’s particle detectors quite impressively illustrate this: Almost all recorded particle collisions are discarded as uninteresting already in the detector, such that out of those four petabytes per minute, ATLAS outputs merely 19 gigabytes for further processing [Br11]. Analogously, most data generated by and about internet users does not become part of the permanent web, which in and of itself is only a fraction of what resides on the world’s storage media [Tu14].

## 2 The Archived Web

Statista and Cisco Systems estimate that in the year 2020, about two zettabytes of installed storage capacity exist in datacenters across the globe [Ci20b]. Worldwide storage media shipments support this claim. About 6.5 ZB have been shipped cumulatively since 2010. Accounting for a yearly 3 % loss rate, as well as the fact that only a portion of all storage resides in enterprise facilities, leads to the same estimate of about 2–2.5 ZB installed capacity in 2020 [RGR20]. About half this capacity appears to be effectively utilized in datacenters, which means that the industry stores one zettabyte of data [Ci20a]. Adding to this the remaining 30 % share of data from consumer devices, we can estimate that the world keeps watch over a data treasure of roughly 1.43 zettabytes in total at the time of writing. Although the majority of these sextillion stored bytes is probably quite recent [Pe20], the figure does include all data stored (and retained) since the beginning of recorded history. The majority of persistent data is not publicly accessible, but hides in closed repositories belonging to individuals and private or statutory organizations. But even of what we call the “public

---

<sup>4</sup> We disregard issues like power, cooling, and redundancy, but also what should be a sizeable bulk discount.

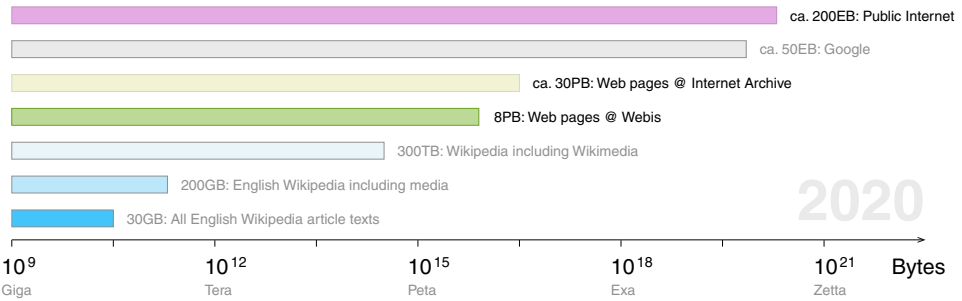


Fig. 2: The 2020 sizes of large, persistent data sets in comparison, illustrated on a logarithmic scale. The English Wikipedia is often compared to the Encyclopædia Britannica, but largely exceeds it by a factor of more than 80 at a combined size of only 30 GB. This is just over a one hundred millionth part of all textual web pages, whose total volume we estimate significantly smaller than 1 EB (cf. Fig. 1).

internet,” only the small “surface web” portion—openly accessible and indexed web pages, books, audiovisual media, and software—is actually *publicly* accessible. What remains is called the “deep web,” which includes email communications, instant messaging, restricted social media groups, video conferencing, online gaming, paid music and video-on-demand services, document cloud storage, productivity tools, and many more, but also the so-called “dark web.” The size of the deep web is much harder to assess than that of both the surface web and the world’s global storage capacity. Folklore estimates range from 20 to 550 times the size of the surface web, but, to the best of our knowledge, no reliable sources exist and previous attempts at measuring its size typically relied on estimating the population of publicly retrievable text document IDs [LL10]. In absence of reliable data, we assume its actual size to be on the larger end and take an educated guess of up to 200 exabytes. That is a fifth of all persistent data and less than half a percent of the Global Datasphere [RGR20].

The indexed surface web is substantially easier to quantify. In 2008, Google reported to have discovered one trillion distinct URLs [Go08], though only a portion of that was actually unique content worth indexing. Based on estimates of the current index sizes of the two largest web search engines [vBd16], there are approximately 60 billion pages (lower bound: 5.6 billion) in the indexed portion of the World Wide Web as of 2020.<sup>5</sup> The HTTP Archive,<sup>6</sup> which aims to track the evolution of the web from a metadata perspective, records (among other metrics) statistics about page weights, i.e., the total number of bytes transferred during page load. Based on the page weight percentiles given in their 2019 report [EH19], we estimate that the combined size of all indexed web pages is at least 1.8 petabytes counting their HTML payload alone. As a rather generous upper bound, we assert that the World Wide Web including all directly embedded style, script, and image assets is (not accounting for redundancy) most likely less than 200 petabytes and—considering the English-language

<sup>5</sup> <https://www.worldwidewebsize.com>

<sup>6</sup> <https://httparchive.org>

bias [vBd16] of the employed method—certainly less than one exabyte in size (Figure 1). As such, the surface web accounts for less than half a percent of the presumed 200 EB constituting the public internet. It is important to note that this figure only captures the *textual* part of the publicly indexable surface web. Publicly accessible on-demand video streaming, accounting for 15% of all downstream internet traffic in 2020 [Sa20] (60% including paid subscriptions) and necessarily taking a sizable portion of enterprises' mass storage, is not included.

Figure 2 compares the sizes of large, persistent data sets on the public internet, starting with an estimate for the total amount of data stored by Google. While Google does not publish any official figures on their data centers' storage capacity, they have maintained a public list of datacenter locations for at least the past eight years [Go13]. In 2013, thirteen datacenters were listed worldwide and contemporary estimates put the total amount of data stored by Google at 15 EB [Mu13]. By mid-2020, Google's list of datacenters had grown to 21 entries [Go20]. At the same time, the areal density of hard disk storage has approximately doubled since 2013 [Co20]. For the total amount of data stored by Google in 2020, we hence propose an updated estimate of 50 EB—this includes both public and non-public data.

The Internet Archive<sup>7</sup> is a large-scale, nonprofit digital library covering a wide range of media, including websites, books, audio, video, and software. In 2009, the Internet Archive's entire inventory comprised only one petabyte [JK09], but according to current statistics,<sup>8</sup> it has grown by a factor of almost 100 over the past decade. The Wayback Machine, an archive of the World Wide Web, forms a significant portion of the Internet Archive's collections. As of this writing, it comprises nearly 500 billion web page captures, which take up approximately 30 petabytes of storage space. In the Webis research group, we aim to store up to 8 PB of web archive data on our own premises, much of it originating from the Internet Archive, but also from other sources, such as the Common Crawl.

The Wikipedia is commonly used in large-scale text analytics tasks and thus we include it in the comparison in three ways: The combined size of all Wikipedias, including the files in the Wikimedia Commons project, comes at approximately 300 TB,<sup>9</sup> of which approximately 3 TB are actual wikitext.<sup>10</sup> The English Wikipedia including all media files is three orders of magnitude smaller, at about 200 GB.<sup>11</sup> Lastly, the plain wikitext content of the English Wikipedia is yet another order of magnitude smaller, adding up to merely 30 GB in 2020.<sup>12</sup>

Figure 3 is based on an analysis by IDC and Seagate [RGR18, RGR20] inquiring *where* the world's persistent data has been, and is going to be stored in the future: A decade ago, most of the data created across the world remained on largely disconnected consumer endpoint

<sup>7</sup> <https://archive.org>

<sup>8</sup> <https://archive.org/~tracey/mrtg/du.html>

<sup>9</sup> <https://commons.wikimedia.org/wiki/Special:MediaStatistics>

<sup>10</sup> <https://stats.wikimedia.org/#/all-projects/content/net-bytes-difference/normal|bar|all|~total|monthly>

<sup>11</sup> <https://en.wikipedia.org/wiki/Special:MediaStatistics>

<sup>12</sup> [https://en.wikipedia.org/w/index.php?title=Wikipedia:Size\\_in\\_volumes&oldid=966973828](https://en.wikipedia.org/w/index.php?title=Wikipedia:Size_in_volumes&oldid=966973828)

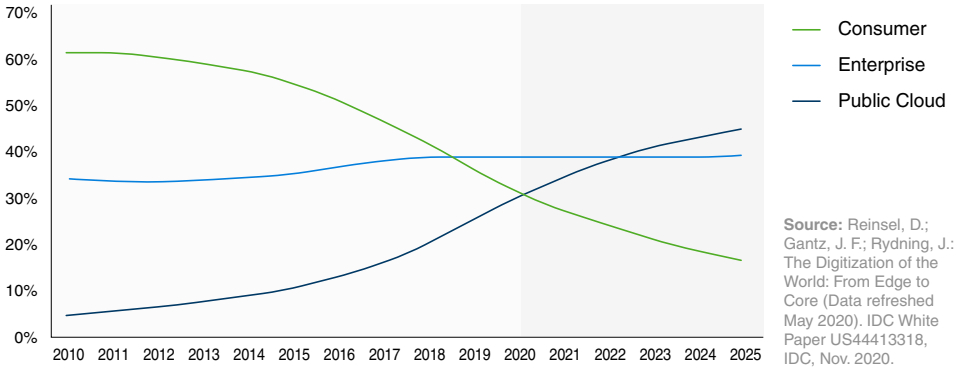


Fig. 3: As of 2020, IDC believes, more bytes are stored in the public cloud than on consumer devices. By 2022, more data is going to be stored in the cloud than in traditional datacenters.

devices, but that has steadily shifted towards public clouds. As of 2020, the share of data stored in public clouds has exceeded that stored on consumer devices, and within the next three years, the majority of persistent data is predicted to reside in public clouds.

The rapid growth in both size and publicness of web-based data sources drives interest in web science globally, not only in the Webis group. We have shown the persistent World Wide Web to be quite a bit smaller than one might expect in contrast with the totality of data created and stored; nevertheless, true “web archive analytics” tasks, which we still consider to involve datasets exceeding hundreds of terabytes, certainly do require a large and specialized hardware infrastructure to be feasible.

### 3 Web Archive Analytics at Webis

Figures 4 and 5 illustrate the software and infrastructure relevant to this effort: Two different computing clusters contribute to our web archive analytics efforts, out of five total currently operated by our research group. We have maintained cluster hardware in some form or another for more than a decade. Figure 4 outlines their specifications. In the following, we give a short overview of typical workloads and research areas each individual cluster is tasked with:  $\alpha$ -web is made up of repurposed desktop hardware and enabled our earliest inroads into web mining and analytics; nowadays it still functions as a teaching and staging environment.  $\beta$ -web is currently our primary environment for web mining, batch and stream processing, data indexing, virtualization, and the provision of public web services.  $\gamma$ -web’s primary purpose is the training of machine learning models, text synthesis, and language modeling.  $\delta$ -web is focused on storage and serves as a scalable and redundant persistence backend for web archiving and other tasks.  $\epsilon$ -web is used for indexing and argument search.





















	$\alpha$ -web [2009]	$\beta$ -web [2015]	$\gamma$ -web [2016+2020]	$\delta$ -web [2018]	$\epsilon$ -web [2020]
Nodes	44 	135 	9 	78 	55 
Disk [PB]	0.2 	4.3 	0.01 	12 	0.1 
Cores	176  $\approx 3.2$ TFLOPs	1,740  $\approx 67.4$ TFLOPs	384 + 227,328  $\approx 690.0$ TFLOPs	1,248  $\approx 119.8$ TFLOPs	1,100  $\approx 44$ TFLOPs
RAM [TB]	0.8 	28 	7.5 	10 	7 

Fig. 4: The cluster hardware owned by the Webis research group, organized from left to right by acquisition date (shown in square brackets).  $\beta$ -web and  $\delta$ -web have been specifically designed to handle large-scale web archive analytics tasks.

The  $\beta$ -web and  $\delta$ -web clusters are the primary workhorses in our ongoing web archiving and analytics efforts. The clusters comprise 135 and 78 Dell PowerEdge servers, respectively, spread across two datacenters joined via a 400 GB/s interconnect. Each individual node is attached to one of two middle-of-row Cisco Nexus switches via a 10 GB/s link. The storage cluster maintains more than 12 PB of raw storage capacity across 1 248 physical spinning disks. The compute cluster possesses an additional 4.3 PB for serving large indexes and storing ephemeral intermediate results across 1 080 physical disks.

Our analytics stack is shown conceptually in Figure 5, beginning with the bottom-most data acquisition layer: Primary sources for data ingestion include web crawls and web archives, such as the aforementioned Internet Archive, the Common Crawl,<sup>13</sup> the older ClueWeb<sup>14</sup> datasets, as well as our own specialized crawls focused on individual research topics, such as argumentation and authorship analytics. Crowdsourcing services, such as Amazon Mechanical Turk, support and complement (distantly) supervised learning tasks.

Both the  $\beta$ -web and  $\delta$ -web cluster are provisioned and orchestrated using the SaltStack configuration management and IT automation software, which manages the deployment and configuration of fundamental file-system- and infrastructure-level services on top of a network-booted Ubuntu Linux base image. We supervise the proper functioning of all system components with the help of a Consul cluster (for service discovery) and a redundant Prometheus deployment (for event monitoring and alerting).

The primary purpose of the  $\delta$ -web cluster is to serve a distributed Ceph [We07] storage system with one object storage daemon (OSD) per physical disk, as well as five redundant

<sup>13</sup> <https://commoncrawl.org>

<sup>14</sup> <https://lemurproject.org>

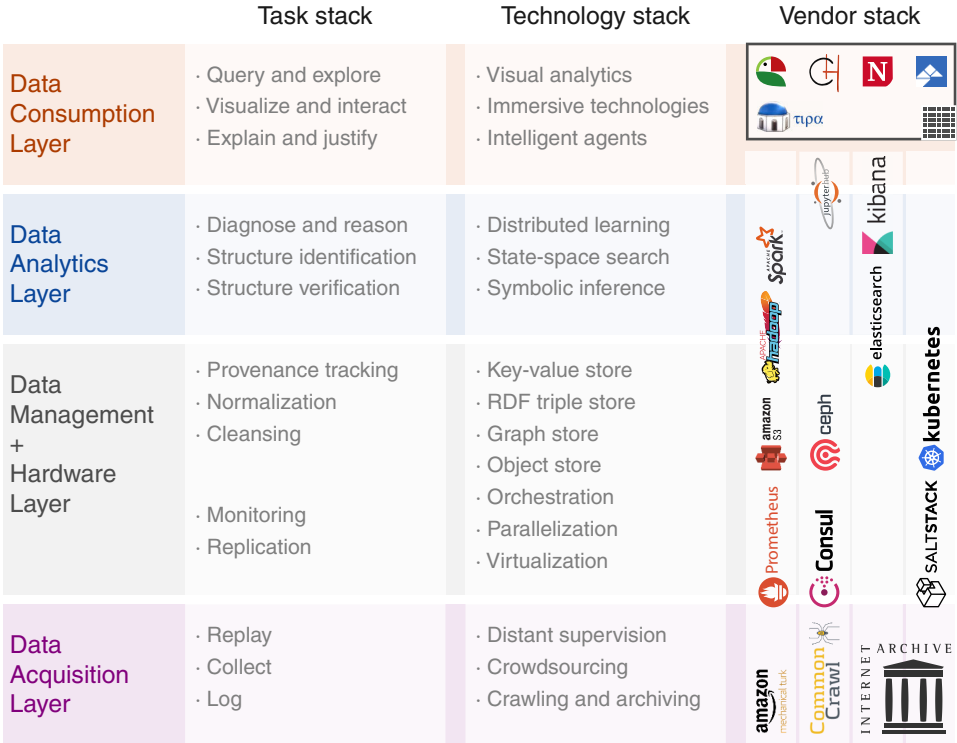


Fig. 5: The Webis web archive analytics infrastructure stack. The y-axis organizes (bottom to top) the processing pipeline from data acquisition to data consumption. The x-axis organizes (right to left) the employed tools (vendor stack), algorithms and methods (technology stack), and tackled problem classes (task stack). The box at the top of the vendor stack shows the resulting frontend services maintained by the Webis research group: args.me, ChatNoir, Netspeak, Picapica, and Tira.

Monitor (MON) / Manager (MGR) daemons. The bulk of the data payload is accessed via the RadosGW S3 API gateway service, which is provided by seven redundant RadosGW daemons and backed by an erasure-coded storage pool. Smaller and more ephemeral datasets live in a CephFS distributed file system with three active and four standby Metadata Servers (MDS).

On  $\beta$ -web, we maintain a Saltstack-provisioned Kubernetes cluster on top of which most of our internal and public-facing services are deployed. Kubernetes services are provided with persistent storage through a RADOS Block Device (RBD) pool in the Ceph cluster. A Hadoop Distributed File System (HDFS) holds temporary intermediate data.



Services operated on top of Kubernetes include a 130-node Elasticsearch deployment powering our search engines, as well as a suite of data analytics tools such as Hadoop Yarn, Spark, and JupyterHub. Our public-facing services are, among others, the search engines *args.me* [Wa17, Po19a] and *ChatNoir* [Be18], the *Netspeak* [SPT10, Po14] writing assistant, the *Picapica* [Ha17] text reuse detection system, and the *Tira* [GSB12, Po19b] evaluation-as-a-service platform. The web-archive-related services at *archive.webis.de* are still largely under construction. Ultimately, we envision also allowing external collaborators to process and analyze web archives on our infrastructure. As of October 2020, almost 2.3 PB of data—of which 560 TB stem from the Internet Archive and the rest from the Common Crawl—have been downloaded and are stored on our infrastructure.

## 4 Conclusion

The web and its ever-growing archives represent an unprecedented opportunity to study humanity’s cultural output at scale. While in 2020 alone, a staggering 59 zettabytes of data will have been created throughout the year, most of that is transient, and the World Wide Web is nearly five orders of magnitude smaller. Nevertheless, the scale of the web is large enough to be challenging, and so web analytics requires potent and robust infrastructure.

The Webis cluster infrastructure has already enabled many research projects on a scale which typically only big industry players can afford due to the massive hardware requirements. This allows us to develop and maintain services such as the *ChatNoir* web search engine [Be18], the argument search engine *args.me* [Wa17], but also facilitates individual research projects which rely on utilizing the bundled computational power and storage capacity of the capable Hadoop, Kubernetes, and Ceph clusters. A few recent examples are the analysis of massive query logs [Bo20], a detailed inquiry into near-duplicates in web search [Fr20b, Fr20a], the preparation, parsing, and assembly of large NLP corpora [Be20, Ke20], and the exploration of large transformer models for argument retrieval [AP20].

With significant portions of the Internet Archive available and further hardware acquisitions in progress, many more of these research projects at a very large and truly representative scale will become feasible, including not only the evaluation, but also the efficient training of ever larger machine learning models.

## Bibliography

- [AP20] Akiki, Christopher; Potthast, Martin: Exploring Argument Retrieval with Transformers. In: Working Notes Papers of the CLEF 2020 Evaluation Labs. September 2020.
- [Be06] Berners-Lee, Tim; Hall, Wendy; Hendler, James A.; O’Hara, Kieron; Shadbolt, Nigel; Weitzner, Daniel J.: A Framework for Web Science. *Found. Trends Web Sci.*, 1(1):1–130, 2006.

- [Be18] Bevendorff, Janek; Stein, Benno; Hagen, Matthias; Potthast, Martin: Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In: 40th European Conference on IR Research (ECIR 2018). LNCS, Springer, Berlin Heidelberg New York, March 2018.
- [Be20] Bevendorff, Janek; Al-Khatib, Khalid; Potthast, Martin; Stein, Benno: Crawling and Preprocessing Mailing Lists At Scale for Dialog Analysis. In: 58th Annual Meeting of the Association for Computational Linguistics. July 2020.
- [Bo20] Bondarenko, Alexander; Braslavski, Pavel; Völske, Michael; Aly, Rami; Fröbe, Maik; Panchenko, Alexander; Biemann, Chris; Stein, Benno; Hagen, Matthias: Comparative Web Search Questions. In: 13th ACM International Conference on Web Search and Data Mining (WSDM 2020). February 2020.
- [Br11] Brumfiel, Geoff: High-Energy Physics: Down the Petabyte Highway. *Nature*, 469(7330):282–283, January 2011.
- [Ci20a] Cisco Systems: Amount of data actually stored in data centers worldwide from 2015 to 2021. Statistics, Statista Inc., New York, March 2020.
- [Ci20b] Cisco Systems: Data center storage capacity worldwide from 2016 to 2021, by segment. Statistics, Statista Inc., New York, April 2020.
- [Co20] Coughlin, Tom: HDD Market History and Projections. Report, May 2020. <https://www.forbes.com/sites/tomcoughlin/2020/05/29/hdd-market-history-and-projections/>.
- [DI17] Domo Inc., American Fork, Utah: Data Never Sleeps 5.0. Leaflet, November 2017. <https://www.domo.com/learn/data-never-sleeps-5>.
- [DI19] Domo Inc., American Fork, Utah: Data Never Sleeps 7.0. Leaflet, August 2019. <https://www.domo.com/learn/data-never-sleeps-7>.
- [DI20] Domo Inc., American Fork, Utah: Data Never Sleeps 8.0. Leaflet, August 2020. <https://www.domo.com/learn/data-never-sleeps-8>.
- [EH19] Everts, Tammy; Hempenius, Katie: The 2019 Web Almanac, Chapter 18: Page Weight. Technical report, The HTTP Archive, 2019. <https://almanac.httparchive.org/en/2019/page-weight>.
- [Fr20a] Fröbe, Maik; Bevendorff, Janek; Reimer, Jan Heinrich; Potthast, Martin; Hagen, Matthias: Sampling Bias Due to Near-Duplicates in Learning to Rank. In: 43rd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2020). July 2020.
- [Fr20b] Fröbe, Maik; Bittner, Jan Philipp; Potthast, Martin; Hagen, Matthias: The Effect of Content-Equivalent Near-Duplicates on the Evaluation of Search Engines. In: Advances in Information Retrieval. 42nd European Conference on IR Research (ECIR 2020), volume 12036 of Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, April 2020.
- [Go08] Google, LLC: We knew the web was big... Web page, August 2008. <https://web.archive.org/web/20200918111648/https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>.
- [Go13] Google, LLC: Data Center Locations. Web page, August 2013. <https://web.archive.org/web/20130801052830/http://www.google.com/about/datacenters/inside/locations/index.html>.

- [Go20] Google, LLC: Discover our Data Center Locations. Web page, July 2020. <https://web.archive.org/web/20200718223350/https://www.google.com/about/datacenters/locations/index.html>.
- [GSB12] Gollub, Tim; Stein, Benno; Burrows, Steven: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2012). August 2012.
- [Ha17] Hagen, Matthias; Potthast, Martin; Adineh, Payam; Fatehifar, Ehsan; Stein, Benno: Source Retrieval for Web-Scale Text Reuse Detection. In: 26th ACM International Conference on Information and Knowledge Management (CIKM 2017). November 2017.
- [He08] Hendler, James A.; Shadbolt, Nigel; Hall, Wendy; Berners-Lee, Tim; Weitzner, Daniel J.: Web science: an interdisciplinary approach to understanding the web. *Commun. ACM*, 51(7):60–69, 2008.
- [JK09] Jaffe, Elliot; Kirkpatrick, Scott: Architecture of the Internet Archive. In: Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference. SYSTOR '09, Association for Computing Machinery, New York, NY, USA, 2009.
- [Ke20] Kestemont, Mike; Manjavacas, Enrique; Markov, Ilija; Bevendorff, Janek; Wiegmann, Matti; Stamatatos, Efstathios; Potthast, Martin; Stein, Benno: Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In: Working Notes Papers of the CLEF 2020 Evaluation Labs. September 2020.
- [KG03] Kilgarriff, Adam; Grefenstette, Gregory: Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3):333–348, 2003.
- [LL10] Lu, Jianguo; Li, Dingding: Estimating deep web data source size by capture-recapture method. *Inf. Retr.*, 13(1):70–95, 2010.
- [Mu13] Munroe, Randall: Google's Datacenters on Punch Cards. Web page, September 2013. <https://web.archive.org/web/20130920004219/http://what-if.xkcd.com:80/63/>.
- [Pe20] Petrov, Christo: 25+ Impressive Big Data Statistics for 2020. Internet blog, techjury.net., September 2020. <https://techjury.net/blog/big-data-statistics/>.
- [Po14] Potthast, Martin; Hagen, Matthias; Beyer, Anna; Stein, Benno: Improving Cloze Test Performance of Language Learners Using Web N-Grams. In: 25th International Conference on Computational Linguistics (COLING 2014). August 2014.
- [Po19a] Potthast, Martin; Gienapp, Lukas; Euchner, Florian; Heilenkötter, Nick; Weidmann, Nico; Wachsmuth, Henning; Stein, Benno; Hagen, Matthias: Argument Search: Assessing Argument Relevance. In: 42nd International ACM Conference on Research and Development in Information Retrieval (SIGIR 2019). July 2019.
- [Po19b] Potthast, Martin; Gollub, Tim; Wiegmann, Matti; Stein, Benno: TIRA Integrated Research Architecture. In: *Information Retrieval Evaluation in a Changing World, The Information Retrieval Series*. Springer, Berlin Heidelberg New York, September 2019.
- [RGR18] Reinsel, David; Gantz, John F.; Rydning, John: The Digitization of the World: From Edge to Core. IDC White Paper US44413318, International Data Corporation (IDC), Framingham, Massachusetts, November 2018.

- [RGR20] Reinsel, David; Gantz, John F.; Rydning, John: The Digitization of the World: From Edge to Core (Data refreshed May 2020). IDC White Paper US44413318, International Data Corporation (IDC), Framingham, Massachusetts, May 2020.
- [RR20] Reinsel, David; Rydning, John: Worldwide Global StorageSphere Forecast, 2020–2024: Continuing to Store More in the Core. IDC Market Forecast US46224920, International Data Corporation (IDC), Framingham, Massachusetts, May 2020.
- [RRG20] Reinsel, David; Rydning, John; Gantz, John F.: Worldwide Global StorageSphere Forecast, 2020–2024: The COVID-19 Data Bump and the Future of Data Growth. IDC Market Forecast US44797920, International Data Corporation (IDC), Framingham, Massachusetts, April 2020.
- [Sa20] Sandvine: The Global Internet Phenomena Report 2020. COVID-19 Spotlight. Technical report, Sandvine Incorporated, May 2020.
- [SPT10] Stein, Benno; Potthast, Martin; Trenkmann, Martin: Retrieving Customary Web Language to Assist Writers. In: *Advances in Information Retrieval*. 32nd European Conference on Information Retrieval (ECIR 2010). Springer, Berlin Heidelberg New York, March 2010.
- [Tu14] Turner, Vernon; Gantz, John F.; Reinsel, David; Minton, Stephen: The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. IDC White Paper 1672, International Data Corporation (IDC), Framingham, Massachusetts, April 2014.
- [vBd16] van den Bosch, Antal; Bogers, Toine; de Kunder, Maurice: Estimating Search Engine Index Size Variability: A 9-Year Longitudinal Study. *Scientometrics*, 107(2):839–856, May 2016.
- [Wa17] Wachsmuth, Henning; Potthast, Martin; Al-Khatib, Khalid; Ajjour, Yamen; Puschmann, Jana; Qu, Jiani; Dorsch, Jonas; Morari, Viorel; Bevendorff, Janek; Stein, Benno: Building an Argument Search Engine for the Web. In: 4th Workshop on Argument Mining (ArgMining 2017) at EMNLP. Association for Computational Linguistics, September 2017.
- [We07] Weil, Sage A.; Leung, Andrew W.; Brandt, Scott A.; Maltzahn, Carlos: RADOS: a scalable, reliable storage service for petabyte-scale storage clusters. In: *Proceedings of the 2nd International Petascale Data Storage Workshop (PDSW '07)*, November 11, 2007, Reno, Nevada, USA. pp. 35–44, 2007.