Benjamin Turkus

Kara Van Malssen

Digital Preservation

1 October 2013

Common Crawl

The Common Crawl Foundation, established in 2007 by software designer and entrepreneur Gilad Elbaz, is a non-profit organization committed to making large-scale web crawl data publically accessible. Currently hosted as a free Public Data Set on Amazon Web Services' S3 (Simple Storage Service, or cloud computing storage), Common Crawl aspires to "change the big data game," democratizing the Internet by collecting, maintaining, and providing access to a corpus of approximately six billion web documents, or crawl information on the order of 40 terabytes of data (Salevan 2011). For Elbaz and the other founders of Common Crawl, the goal is clear: access will beget innovation. As he explained in a 2012 interview:

> "You might ask why it is going to be revolutionary to allow many more engineers and researchers and developers and students access to this data, whereas historically you have to work for one of the big search engines…The question is, the world has the largest-ever corpus of information on the web, and is there more that one can do with it than Google or Microsoft and a handful of other search engines are already doing? The answer is undoubtedly yes" (Zaino "Elbaz Speaks," 2012).

But while Common Crawl is a praiseworthy effort, one largely deserving of the recognition it's received in a variety of publications, is has remained problematically tethered to a vision of treating web information as less a fascinating history of human communication and more as grist for future research and business endeavors.

And despite its trade in the powerful, persuasive language of "the commons," which conjures up images of a grassy English countryside, accessible to all, the question remains: has Common Crawl truly lowered the barriers to entry? What kinds of skills, knowledge, and resources are required to make use of these "public" data sets, and, correspondingly, is it important to highlight rather than glide over the largely unmentioned costs associated with directly extracting Common Crawl data from Amazon's cloud storage? While Common Crawl gives us much to admire, particularly in its dedication to serving as a springboard for human innovation, it is our task, as archivists and historians, to dig even deeper, beginning to investigate the somewhat conspicuous absence of "preservation" from all Common Crawl documentation. In properly assessing Common Crawl, we must proceed forward carefully and critically, insisting that for Common Crawl to be truly considered forward thinking, it must first reconcile some of its inherent contradictions.

The Common Crawl data set, stored on Amazon S3, is subdivided into three major data subsets:

Archived Crawl #1 - s3://aws-publicdatasets/common-crawl/crawl-001/ - crawl data from 2008/2010
Archived Crawl #2 - s3://aws-publicdatasets/common-crawl/crawl-002/ - crawl data from 2009/2010
Current Crawl - s3://aws-publicdatasets/common-crawl/parse-output/ - crawl data from 2012 (Green and Stephens 2012).

Common Crawl harvests web data quarterly, using ccbot, a custom "distributed crawling infrastructure" that is designed to limit the rate at which individual web hosts are crawled (*Common Crawl-FAQ*). The ccbot crawler is a respectful web bot, abiding by any "allow/disallow" protocols found in Robot Exclusion Standards (robots.txt files), and honoring any nofollow attributes (designed to block access to content, nofollow is primarily used to deter search engines such as Google, Bing, Yahoo!, etc.). Yet, as explicitly stated on the Common Crawl website, the organization reserves the right to change its stance in the future. As it describes in painstaking detail: "If we ever did ignore nofollow in the future, we would do so only for the purposes of link discovery and would never create any association between the discovered link and the source document" (*Common Crawl-FAQ*).

After crawling the web via its Tier 1 ICP connection, Common Crawl temporarily stores all downloaded content on an internal Apache Hadoop computer cluster. An open source implementation of the MapReduce framework that originated at Google in 2004, Hadoop has been neatly summarized as, "the driving force behind the growth of the big data industry…[it] brings the ability to cheaply process large amounts of data, regardless of its structure" (Dumbill 2012). Essentially a framework for distributing computation across multiple servers, MapReduce plays a critical role in multiple stages of Common Crawl project: it assists in the processing of content gathered during crawls, the storing of data on Amazon's cloud service, and the provision of end-user access to data.

Common Crawl has also built upon the work of the Internet Archive, a pioneer in the field of web crawling, by compressing its downloaded data into ARC files, a format developed by the IA in 1996 to allow for the combination of "multiple digital resources into an aggregate archival file" (*LOC Sustainability of Digital Formats*, 2013). Alternatively described as format which allows for a "series of concatenated GZIP (compressed) documents," once compressed, Common Crawl's ARC files are uploaded directly to Amazon S3, where they become readily available for either bulk downloading or direct processing through Amazon's EC2 (Elastic Compute Cloud), a service which allows users to take advantage of Amazon's vast computing power by renting pre-configured virtual machines (AMIs, or Amazon Machine Images). For users with access to their own Apache Hadoop cluster or familiar with Java applications, a data processing alternative is available: MapReduce jobs can be built on local computers using Github, then ran in the cloud using Common Crawl data.

For its 2012 corpus, Common Crawl improved accessibility by providing three different types of files: ARC raw content files, Text Only files, and Metadata files (Green and Stephens 2012). As Jennifer Zaino explained in "Common Crawl Corpus Update Makes Web Crawl Data More Efficient, Approachable for Users to Explore," "With metadata files, users don't have to extract the link graph from the raw crawl…Similarly, the full text output that users can now run analysis over is significantly smaller than the .ARC file raw content" (Zaino "Corpus Update," 2012). By adopting the JavaScript Object Notation (JSON) metadata standard, Common Crawl has given users the ability to easily access a variety of information, including "HTML titles, HTML meta tags, RSS/Atom information, and all anchors/hyperlinks from HTML documents" (Stephens and Green 2012[1]).

While the Common Crawl project is still in its early stages, several software developers and researchers have already begun making interesting use of the Common Crawl corpus. One notable example is Tin Eye, a reverse image search engine developed by Idée, Inc. By combining image identification technology with Common Crawl data, Tin Eye allows users to "submit an image to find out where it came from, how it is being used, if modified versions of the image exist, or to find higher resolution versions" (*Tin Eye-About*). In the growing realm of "social search," Seattle-based startup Lucky Oyster has also used Common Crawl data to develop an iPhone app that allows users to share lifestyle recommendations (movies, restaurants, hotels, etc.) with friends and family (*Lucky Oyster-About*). A number of research projects have also incorporated Common Crawl, including the *Web Data Commons*, which has "extracted all Microformat, Microdata and RDFa data from the Common Crawl corpus, packag[ing] the extracted data for each format separately for download;" Matthew Berk's *Study of Web Pages that Reference Facebook*, which found that 22% of ~1.3 billion web pages provided by Common Crawl reference Facebook URLs; and the *Online Sentiment Towards Congressional Bills Study*, which examined individual pieces of Congressional legislation, analyzing "how many times [they] were mentioned across the Internet, what websites talked about [them] most, which sites were most influential, and what language was commonly associated with a bill" (Green and Lester 2012).

As with other organizations that provide access to web crawl data, Common Crawl has been forced to contend with questions related to intellectual property. Common Crawl respects the intellectual property rights of others, and "takes appropriate action" if informed of copyright infringement. Pursuant to Title 17, United States Code, Section 512 (c)(3), Common Crawl insists that written notification be provided to its Copyright Agent. Notice must include the signature of "the person

---

[1] For a complete list of the JSON attributes included in Common Crawl metadata files, see:

https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set

authorized to act on behalf of the owner of the copyright interest," a description of the copyrighted work being infringed, and appropriate contact information (*Common Crawl Terms of Use*).

Common Crawl is a fascinating non-profit big data initiative, one that largely transcends its perhaps too-close-for-comfort relationship to massive multinational Internet-related corporations. Yet, despite its efforts to democratize access to large-scale web crawl data, leveling the playing field for budding American entrepreneurs and researchers, Common Crawl's either intentional or unintentional refusal to take into account questions of preservation is both troubling and difficult to ignore. Common Crawl, and, by extension, all users of its content, would only stand to benefit from a more deliberate, measured approach to preserving this rich, utterly unique resource.

Bibliography

Dumbill, Edd. "What is Apache Hadoop? A Look at the Components and Functions of the Hadoop Ecosystem." *The O'Reilly Media Strata Data Conference.* February 2, 2012.

http://strata.oreilly.com/2012/02/what-is-apache-hadoop.html

Green, Lisa and Chris Stephens. "About the Data Set." *The Common Crawl Wiki*. October 4, 2012.

https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set

Green, Lisa and Dave Lester. "Inspiration and Ideas." *The Common Crawl Wiki*. October 4, 2012.

https://commoncrawl.atlassian.net/wiki/display/CRWL/Inspiration+and+Ideas

Salevan, Steve. "MapReduce for the Masses: Zero to Hadoop in Five Minutes with Common Crawl." *CommonCrawl.org*. December 16, 2011.

http://commoncrawl.org/mapreduce-for-the-masses/

United States. The Library of Congress. *Sustainability of Digital Formats*—ARC_IA, The Internet Archive ARC File Format.

http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml

Zaino, Jennifer. "Common Crawl Founder Gil Elbaz Speaks About New Relationship with Amazon, Semantic Web Projects Using Its Corpus, And Why Open Crawls Matter To Developing Big Data Expertise." *SemanticWeb.com*. January 20, 2012.

http://semanticweb.com/common-crawl-founder-gil-elbaz-speaks-about-new-relationship-with-amazon-semantic-web-projects-using-its-corpus-and-why-open-web-crawls-matter-to-developing-big-data-expertise_b26109

Zaino, Jennifer. "Common Crawl Corpus Update Makes Web Crawl Data More Efficient, Approachable For Users To Explore." *SemanticWeb.com*. July 16, 2012.

http://semanticweb.com/common-crawl-corpus-update-makes-web-crawl-data-more-efficient-approachable-for-users-to-explore_b30771

Websites Consulted

Common Crawl

http://commoncrawl.org/

http://commoncrawl.org/our-work/

http://commoncrawl.org/team/

http://commoncrawl.org/about/media-2/

http://commoncrawl.org/get-started/

The Common Crawl Wiki

https://commoncrawl.atlassian.net/wiki/display/CRWL/Frequently+Asked+Questions

https://commoncrawl.atlassian.net/wiki/display/CRWL/About+the+Data+Set

https://commoncrawl.atlassian.net/wiki/display/CRWL/Quick+Start+-+Amazon+AMI

https://commoncrawl.atlassian.net/wiki/display/CRWL/Quick+Start+-+Build+from+Github

https://commoncrawl.atlassian.net/wiki/display/CRWL/Code+Examples

https://commoncrawl.atlassian.net/wiki/display/CRWL/Additional+Resources

https://commoncrawl.atlassian.net/wiki/display/CRWL/Inspiration+and+Ideas

https://commoncrawl.atlassian.net/wiki/display/CRWL/Helpful+Guides+and+Links

https://commoncrawl.atlassian.net/wiki/pages/viewpage.action?pageId=4292610

Amazon Web Services

http://aws.amazon.com/

http://aws.amazon.com/s3/

http://aws.amazon.com/ec2/

https://aws.amazon.com/amis

Tin Eye

http://www.tineye.com/

http://www.tineye.com/about

Lucky Oyster

http://www.luckyoyster.com/

http://www.luckyoyster.com/about