




Accessible Web Archives: *Rethinking and Designing Usable Infrastructure for Sustainable Research Platforms*

Samantha Fritz, MLIS
Project Manager, Archives Unleashed

25 Years



of logging onto the WWW



*“Meet the demand for automated information-sharing between scientists in universities and institutes **around the world**”*



***fastest growing communications
medium of all time***

The web has shaped how we **connect with one another** and **interact with information**.





We all have a relationship to data

- Organize
- Search
- Provide access



*We also use data to **interpret**
and understand the world
around us*

A photograph of a wall completely covered in old, stacked books. The books are of various colors and sizes, creating a textured, layered appearance. In the center of the wall is a green, double-leaf door with a small red sign on the right leaf. The lighting is bright, highlighting the spines and covers of the books.

The web has provided a new
context for research data

*impacts the way we **produce**, **preserve**
and **interact** with information.*

4.66 BILLION

internet users



1.7MB
Of data
/sec/person



306.4 Billion
Emails
per day



95 Million
Photos & Videos
per day

A laptop is shown from a top-down perspective, slightly angled. The screen is open and displays a colorful, abstract digital interface with various charts and data points. A large, semi-transparent blue circle is centered over the laptop. Inside this circle, the text '800 web pages' is written in a large, white, sans-serif font. Below this, in a smaller, dark blue font, it says 'have been created since the start of this presentation'.

800
web pages

**have been created since the
start of this presentation**

we risk losing
potentially significant
information

404

Hi, the page you were looking for doesn't seem to exist anymore.

[Back to Unsplash](#)

MacBook Pro



Development of Web Archiving

1992

World wide web is launched

.....
1996

Conscious effort to preserve born-digital content

First large-scale preservation projects

.....
1996-2021

Increasing adoption of web archiving mandates among memory institutions around the world

**The web is a critical
source for studying
our digital cultural
heritage**



Opportunities

Expands scope to incorporate a wider and more diverse range of voices and perspectives

Shift in scale from resource scarcity to abundance
(Roy Rosenzweig)



Challenges

Challenges are inevitable when dealing with data

Occur throughout web archiving lifecycle:

- Selection
- Collection
- Organization & Storage
- Description/Metadata
- **Access & Use**



A photograph of a library aisle. The shelves are filled with books, and several warm-toned pendant lights hang from the ceiling, creating a soft, ambient glow. The perspective is from the end of the aisle, looking down its length.

Despite the volume of data captured

web archives have largely remained
inaccessible



Barriers to Access & Use

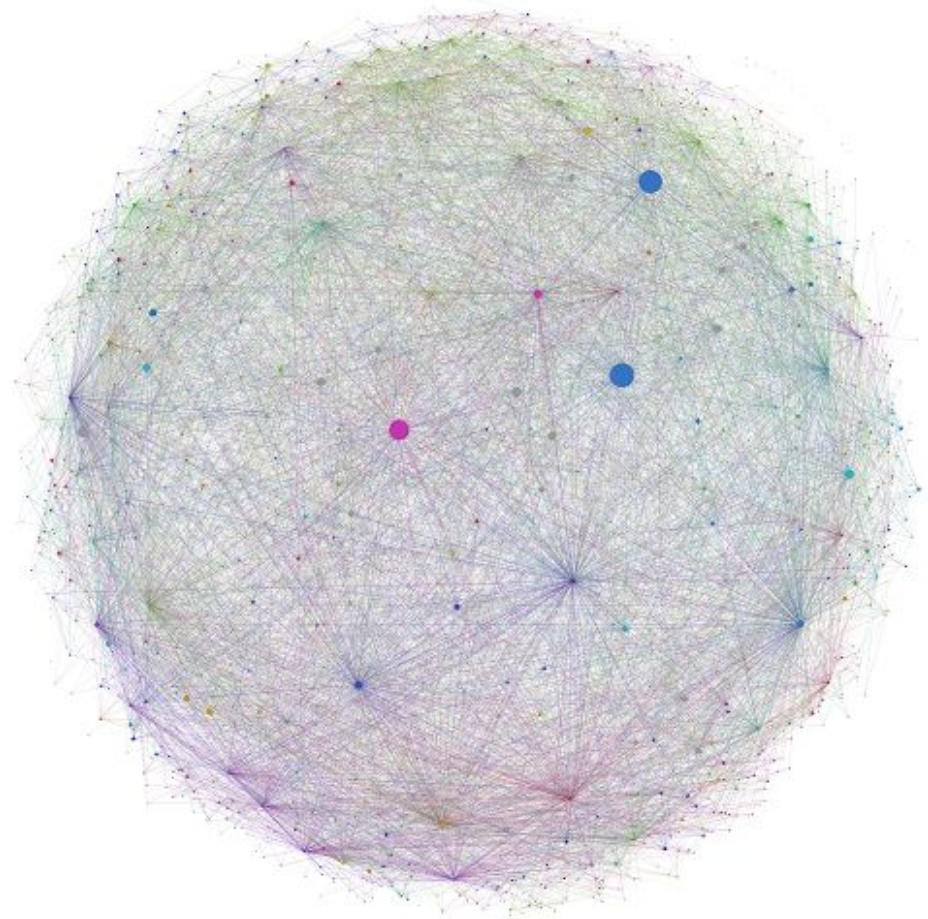
- Required understanding of high-performance computing
- Familiarity with command line
- Lag in analytical tools
- Limitations of time, resources, support

Archives Unleashed



2017-2020

How can we lower
barriers of access and use
to web archives?



Archives Unleashed Project 2017 - 2020

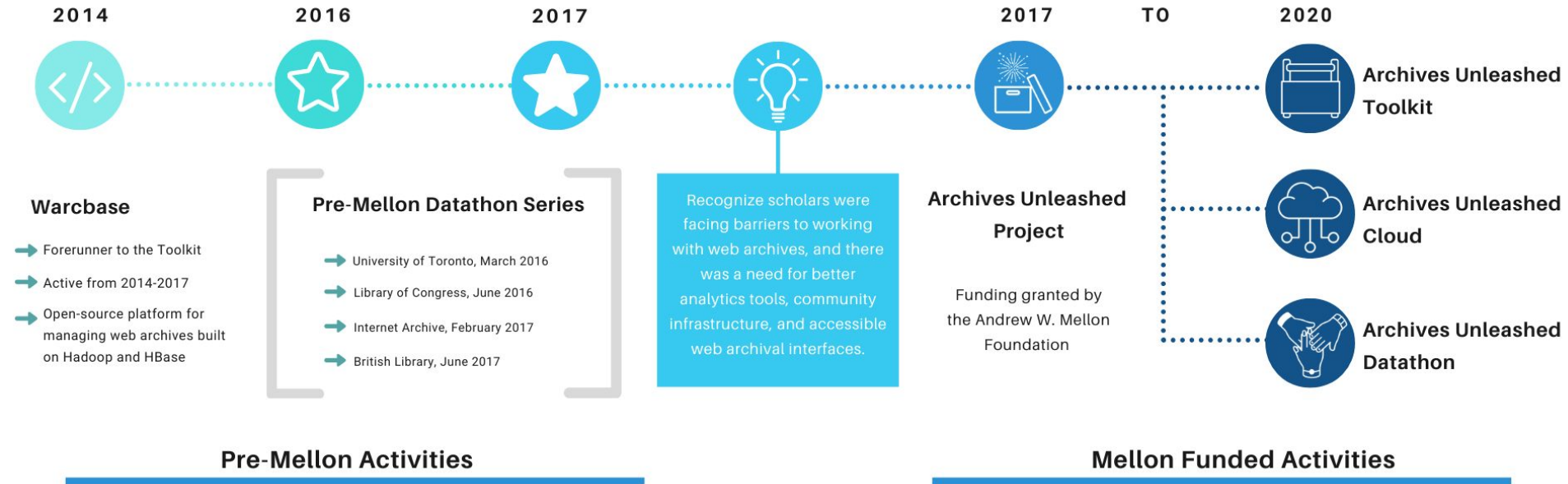
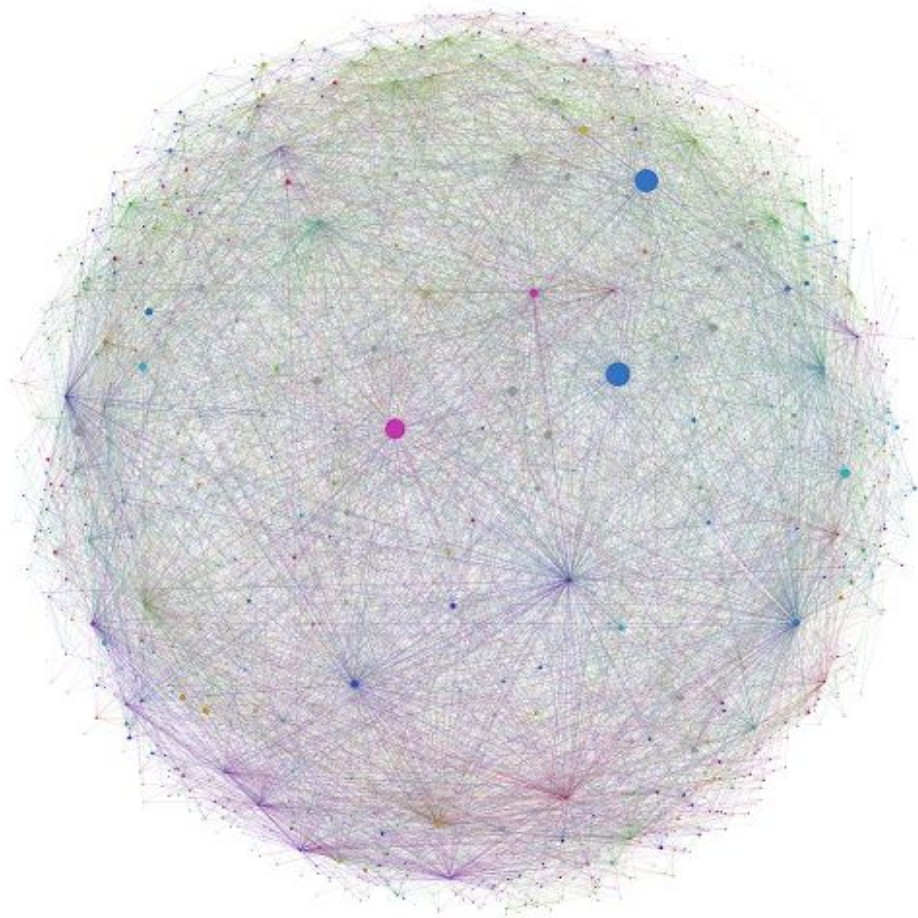


Image: Archives Unleashed Project Timeline

Archives Unleashed



Tools & Platforms



Archives Unleashed Toolkit

→ Documentation provides pre-built scripts

→ Analytic tasks:

- ◆ Collection Analytics
- ◆ Text Analysis
- ◆ Network Analysis
- ◆ Binary Extraction

The screenshot displays the documentation for the Archives Unleashed Toolkit version 0.90.2. The page is titled "Text Analysis" and is part of a navigation menu that includes Home, Getting Started, Generating Results, Filtering Results, Standard Derivatives, and What to do with Results. The main content area is titled "Text Analysis" and contains a sub-section "Extract All Plain Text". This section includes a "Scala RDD" code block with the following code:

```
import io.archivesunleashed._
import io.archivesunleashed.matchbox._

RecordLoader.loadArchives("/path/to/warcs", sc)
  .keepValidPages()
  .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getCo
  .saveAsTextFile("plain-text-rdd/"))
```

Below the code, there is a note: "Note that this will create a new directory to store the output, which cannot already exist." The section also includes a "Python DF" code block with the following code:

```
from aut import *

WebArchive(sc, sqlContext, "/path/to/warcs") \
  .webpages() \
  .select("crawl_date", extract_domain("url").alias("domain"), "url",
  .write.csv("plain-text-df/"))
```

The page also features a sidebar on the right with a list of navigation links, including "Extract All Plain Text", "Scala RDD", "Python DF", "Extract Plain Text Without HTTP Headers", "Extract Plain Text By Domain", "Extract Plain Text by URL Pattern", "Extract Plain Text Minus Boilerplate", "Extract Plain Text Filtered by Date", "Extract Plain Text Filtered by Language", "Extract Plain text Filtered by Keyword", and "Extract Raw HTML".

Archives Unleashed Cloud

- Uses Toolkit code base
- One-stop, web-based portal
- Scholars ingest their Archive-It collections and execute a number of analyses with the click of a mouse
- Generate and explore derivatives and in-browser visualizations

The screenshot displays the Archives Unleashed Cloud interface. On the left, there is a sidebar with the 'Archives Unleashed' logo, an 'AU Cloud Account' section with email 'archivesunleashed@gmail.com' and 'University of Waterloo', and an 'Archive-It Account' section with 'ResearcherDL' and a masked password. Below this are buttons for 'Update', 'Jobs Run' (22), and 'Disk Usage' (529 GB). The main area is titled 'Collections' and contains a table with columns: Title, Status, Date Analyzed, Public, Files, and Size. The table lists various collections such as 'Leonard Cohen Collection', 'University of Toronto Libraries Digital Collections', and 'Toronto 2015 Pan Am & Parapan American Games'. At the bottom, there are logos for Mellon, University of Waterloo, and York University, along with the URL 'archivesunleashed.org/'.

Title	Status	Date Analyzed	Public	Files	Size
Leonard Cohen Collection			No	159	34.3 GB
University of Toronto Libraries Digital Collections			Yes	125	73.2 GB
Toronto 2015 Pan Am & Parapan American Games	Completed	April 16, 2020	Yes	294	50.4 GB
Canadian Political Parties and Political Interest Groups			Yes	6127	691 GB
Federal Election Candidate Sites 2015			Yes	310	206 GB
Toronto Mayoral Election 2014	Completed	April 16, 2020	Yes	292	292 GB
Canadian Government Information			Yes	14358	4.66 TB
Canadian Labour Unions			Yes	7757	1.03 TB
Ontario Provincial Election 2011	Completed	August 5, 2020	No	106	7.91 GB
Snowden Archive	Completed	April 16, 2020	Yes	42	7.16 GB
Canadian Political Interest Groups			Yes	100	8.75 GB
Ontario Provincial Election 2018	Completed	April 16, 2020	Yes	939	113 GB
University of Toronto Archives Web Collection			Yes	10624	1.35 TB
University of Toronto Scarborough	Completed	October 20, 2020	No	3	27.7 MB
Ontario Open Data			No	239	244 GB
Hong Kong Politics			Yes	1106	1.04 TB
Aboriginal Canada Portal	Completed	May 19, 2020	Yes	10	426 MB
Test			No	3	182 MB
Toronto Municipal Election 2018	Completed	April 16, 2020	Yes	1106	34.3 GB
Global Summitry Archive			Yes	660	494 GB

Archives Unleashed Cloud

- The Cloud will be sunsetting at the end of June 2021
- Efforts continue in collaboration with Archive-It to integrate and implement a new Cloud interface



ARS Cloud

Datasets

Learn More: [ARS Cloud Documentation](#)

Collection Analysis +

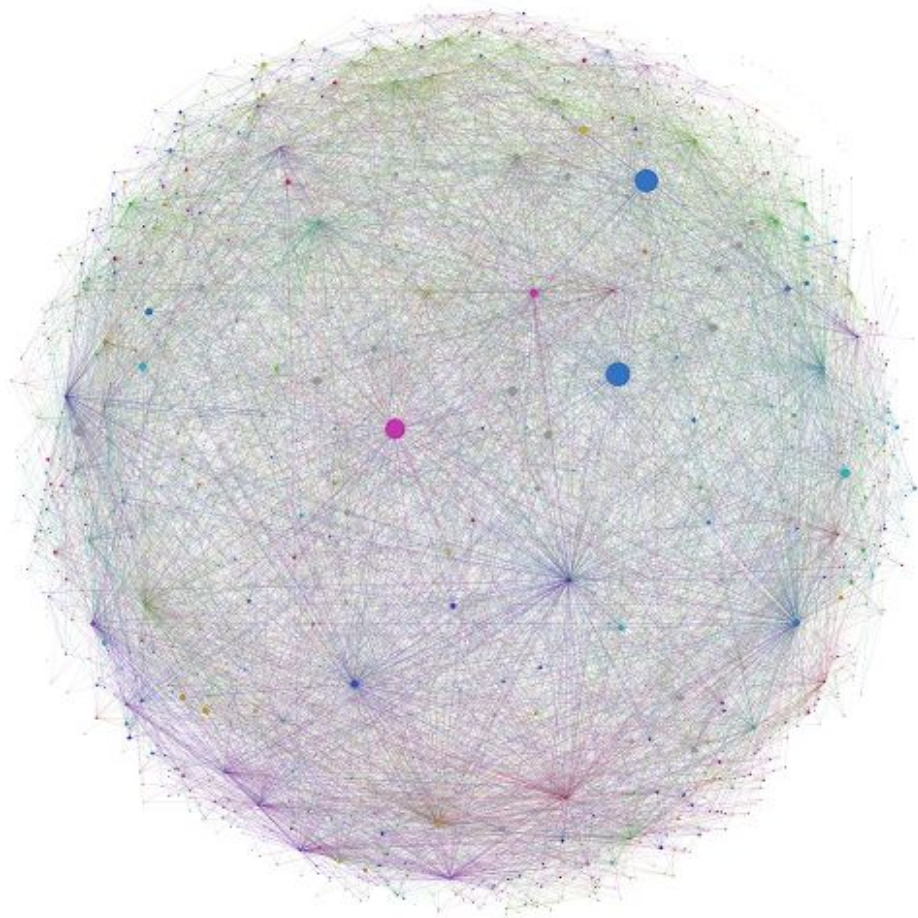
Collection Name	Public Collection	Recently Created	Last Created	Size
AU Team	Yes	Domain Frequency (Sample)	2021-04-08 14:21:05	1.8 GB
Datavis COVID-19	Yes	Extract word processor information	2021-04-01 04:36:23	861.1 MB
Canada U15	Yes	Extract webpages (Sample)	2021-04-01 15:19:27	9.3 GB

© Internet Archive | Archive-It | Help Center

Archives Unleashed



**Accessibility
&
Usability**





Defining Access & Use

Access / Accessibility

the ability to make use of something, or capability of being reached, used, understood or appreciated

Usability

“The quality or state of being usable; ease of use”

We cannot talk about access without acknowledging the vital role usability plays



**Applications of
Access and Use
Concepts**

Code Base

- **Publicly** available, **free**, **open-source**
- **Robust** and **flexible** in conducting **large-scale** web archive analysis
- Integrated widely adopted and **stable programming languages and best practices**

- Transparent tool; there are **no hidden processes** as researcher see analysis requests and output results
- Integrated **feedback** and addressed growing areas of interest

User Interface

- The Cloud provides a **web-base front end** to the Toolkit
- **Addresses hesitancy** of using command-line
- Interface that is **intuitive** and **familiar**
- Individuals tend to be more comfortable with a **click to results** type task process

- Broadens access to web archives using the **WASAPI** Data transfer API.
- Individuals already set up with an **Archive-It account** are able to **ingest and explore their WARC files**
- Examined experimental approaches to **expand access**, use and **interoperability** with other platforms

Supporting Materials

Toolkit Documentation

Cookbook approach, with **pre-built scripts** that users can plug in to address **common analytic tasks**.

Addresses uncertainty of how to use Toolkit

Learning Resources

Learning guides provide **instructions** on how to **use and explore** Cloud **derivatives** with external tools like Gephi and AntConc.

Datasets

Collaboration to process web archive collections and make **derivatives available for all** to use and explore.

Great starting point for scholars who might not have access to a web archive collections

General Takeaways

- ★ **Web archives are critical source** for studying topics post-1990 for many scholars
- ★ Despite the volume of data captured by institutions across the globe, **web archives have largely remained inaccessible and difficult to use.**
- ★ Archives Unleashed has focused on **lowering barriers to access and use** by contributing two substantial tool developments to the community.
- ★ Our team has thoughtfully **integrated** concepts of **access and usability** throughout project development cycles

General Takeaways

- ★ Archives Unleashed incorporates the spirit of access and usability by:
 - Providing **access points for exploring web archives** via development of the Toolkit and Cloud
 - **Tools are created as user-friendly** as possible, which are robust, transparent, flexible, and intuitive
 - Creating **documentation and resources** to support training and learning
- ★ Access and usability need to cooperate in partnership



In designing infrastructure for sustainable research platforms, we need to thoughtfully apply concepts of access and usability

If you build it, they will come



In designing infrastructure for sustainable research platforms, we need to thoughtfully apply concepts of access and usability

If you design it, will it be usable?



CREDITS

This work is primarily supported by the Andrew W. Mellon Foundation. Other financial and in-kind support has come from the Social Sciences and Humanities Research Council, Compute Canada, the Ontario Ministry of Research, Innovation, and Science, York University Libraries, Start Smart Labs, and the Faculty of Arts and David R. Cheriton School of Computer Science at the University of Waterloo.

References

- CERN Accelerating science. (n.d.). *The Birth of the Web*. CERN. <https://home.cern/science/computing/birth-web>.
- Howson, P. 2014. “80 Moments that Shapes the World.” British Council. <https://www.britishcouncil.org/sites/default/files/80-moments-report.pdf>
- Bulao, J. 2021. “how Much Data is Created Every Day in 2021?” <https://techjury.net/blog/how-much-data-is-created-every-day/#gref>
- Huss, N. 2021. How many Websites are there Around the World? <https://siteefy.com/how-many-websites-are-there/>
- Rosenzweig, R. (2003). Scarcity or Abundance? Preserving the Past in a Digital Era, *The American Historical Review*, Volume 108, Issue 3, Pages 735–762, <https://doi.org/10.1086/ahr/108.3.735>
- Milligan, I. (2019). *History in the age of abundance?: how the web is transforming historical research* . McGill-Queen’s University Press.
- Access [Def. 1.b]. (n.d.). In *Merriam Webster Online*, Retrieved September 4, 2019, from <https://www.merriam-webster.com/dictionary/access?src=search-dict-box>
- Accessibility [Def. 1-4]. (n.d.). In *Merriam Webster Online*, Retrieved September 4, 2019, from <https://www.merriam-webster.com/dictionary/accessibility>
- Usability. (n.d.). In *Merriam Webster Online*, Retrieved September 4, 2019, from <https://www.merriam-webster.com/dictionary/usability>

Images Used

In order of appearance; image title provided where possible.

- Photo by [Philipp Katzenberger](https://unsplash.com/photos/S3elo7CMIRA) on Unsplash <https://unsplash.com/photos/S3elo7CMIRA>
- Photo by [Norbert Levajsics](https://unsplash.com/photos/D97n3LR5uN8) on Unsplash. <https://unsplash.com/photos/D97n3LR5uN8>
- Green and Beige Cord by [Brett Sayles](https://www.pexels.com/photo/green-and-beige-cord-1624895/) from Pexels <https://www.pexels.com/photo/green-and-beige-cord-1624895/>
- Photo by [NASA](https://unsplash.com/photos/Q1p7bh3SHj8) on Unsplash <https://unsplash.com/photos/Q1p7bh3SHj8>
- Female software engineer with projected code by [ThisIsEngineering RAEng](https://unsplash.com/photos/8hgmG03spF4) on Unsplash. <https://unsplash.com/photos/8hgmG03spF4>
- Image by [Pete Linforth](https://pixabay.com/photos/earth-internet-globalisation-2254769/) from Pixabay . <https://pixabay.com/photos/earth-internet-globalisation-2254769/>
- Book-covered wall by [Eugenio Mazzone](https://unsplash.com/photos/6ywyo2qtaZ8) on Unsplash. <https://unsplash.com/photos/6ywyo2qtaZ8>
- LED-Keyboard by [Christian Wiediger](https://unsplash.com/photos/WkfDrhxDMC8) on Unsplash <https://unsplash.com/photos/WkfDrhxDMC8>
- Photo by [Tianyi Ma](https://unsplash.com/photos/WIONHd_zYl4) on Unsplash https://unsplash.com/photos/WIONHd_zYl4
- Photo by [Erik Mclean](https://unsplash.com/photos/sxiSod0tyYQ) on Unsplash <https://unsplash.com/photos/sxiSod0tyYQ>
- Grand Rapids lightbulbs by [Kari Shea](https://unsplash.com/photos/OfAX7_xjxm4) on Unsplash https://unsplash.com/photos/OfAX7_xjxm4
- Photo by [Lorenzo Herrera](https://unsplash.com/photos/p0j-mE6mGo4) on Unsplash <https://unsplash.com/photos/p0j-mE6mGo4>
- Reflective perspective by [Nadine Shaabana](https://unsplash.com/photos/VA9xSOekC8c) on Unsplash <https://unsplash.com/photos/VA9xSOekC8c>
- Photo by [x |](https://unsplash.com/photos/N4OTBfNO8Nk) on Unsplash <https://unsplash.com/photos/N4OTBfNO8Nk>
- Library with hanging bulbs by [Jonathan Cooper](https://unsplash.com/photos/sfL_QOomy00) on Unsplash https://unsplash.com/photos/sfL_QOomy00
- Photo by [Jamie Street](https://unsplash.com/photos/dOLgop4tnsc) on Unsplash <https://unsplash.com/photos/dOLgop4tnsc>
- Photo by [Hannah Gullixson](https://unsplash.com/photos/0IEemURq3GM) on Unsplash <https://unsplash.com/photos/0IEemURq3GM>
- Photo by [Patrick Tomasso](https://unsplash.com/photos/Oaqk7qqNh_c) on Unsplash https://unsplash.com/photos/Oaqk7qqNh_c



<https://archivesunleashed.org>