WARCnet

Special Reports

Towards a Glossary for Web Archive Research: Version 1.0

Towards a Glossary for Web Archive Research: Version 1.0

Sharon Healy (Maynooth University), Helena Byrne (British Library), Katharina Schmid (Bavarian State Library), Juan-José Boté-Vericad (Universitat de Barcelona), Lanna Floody (Independent Researcher)

Helena.Byrne@bl.uk

WARCnet Special Report

Aarhus, Denmark 2023

Sharon Healy, Helena Byrne, Katharina Schmid, Juan-José Boté-Vericad, Lanna Floody:
Towards a Glossary for Web Archive Research: Version 1.0

Published by the research network WARCnet, Aarhus, 2023.

Editors of WARCnet Special Reports: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster, Michael Kurzmeier.

Cover design: Julie Brøndum, Kamilla Rosenberg, Emma Lund Nielsen, Thea Laugesen
ISBN: 978-87-94108-18-8

WARCnet
Department of Media and Journalism Studies
School of Communication and Culture
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
warcnet.eu

The WARCnet network is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (May 2020)

Ian Milligan: *You shouldn't Need to be a Web Historian to Use Web Archives* (Aug 2020)

Valérie Schafer and Ben Els: *Exploring special web archive collections related to COVID-19: The case of the BnL* (Sep 2020)

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt: *Exploring special web archive collections related to COVID-19: The case Netarkivet* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the UK Web Archive* (Nov 2020)

Friedel Geeraert and Barbara Signori: *Exploring special web archives collections related to COVID-19: The case of the Swiss National Library* (Nov 2020)

Matthew S. Weber: *Web Archives: A Critical Method for the Future of Digital Research* (Nov 2020)

Niels Brügger: *The WARCnet network: The first year* (Jan 2021)

Susan Aasman, Nicola Bingham, Niels Brügger, Karin de Wild, Sophie Gebeil and Valérie Schafer: *Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections* (Dec 2021)

Olga Holownia, Friedel Geeraert, Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva: *Exploring special web archives collections related to COVID-19: The case of the Library of Congress* (Feb 2022)

Niels Brügger: *The WARCnet network: The second year* (Dec 2022)

Michael Kurzmeier: *Using a national web archive for the study of web defacements? A case-study approach* (Aug 2023)

Helle Strandgaard Jensen: *Any Teletubbies Caught in the Web?* (Aug 2023)

Niels Brügger: *The WARCnet network: The third year* (Aug 2023)

Valérie Schafer, Jérôme Thièvre and Boris Blanckemane: *Exploring special web archives collections related to COVID-19: The case of INA* (Aug 2020)

Niels Brügger, Valérie Schafer, Jane Winters (Eds.): *Perspectives on web archive studies: Taking stock, new ideas, next steps* (Sep 2020)

Friedel Geeraert and Márton Németh: *Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection* (Nov 2020)

Caroline Nyvang and Kristjana Mjöll Jónsdóttir Hjörvar: *Exploring special web archives collections related to COVID-19: The Case of the Icelandic web archive* (Nov 2020)

Sophie Gebeil, Valérie Schafer, David Benoist, Alexandre Faye and Pascal Tanesie: *Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)* (Dec 2020)

Karin de Wild, Ismini Kyritsis, Kees Teszelszky and Peter de Bode: *Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)* (Nov 2021)

Michael Kurzmeier, Joanna Finegan and Maria Ryan: *Exploring special web archives collections related to COVID-19: The National Library of Ireland* (Feb 2022)

Emily Maemura: *Towards an Infrastructural Description of Archived Web Data* (May 2022)

Olga Holownia, Friedel Geeraert and Paul Koerbin: *Exploring special web archives collections related to COVID-19: The case of the National Library of Australia* (Dec 2022)

Helena Byrne, Beatrice Cannelli, Carmen Noguera, Michael Kurzmeier, Karin de Wild: *Looking ahead: after web (archives)?* (Aug 2023)

Friedel Geeraert, Marie Haškovcová, Luboš Svoboda and Markéta Hrdličková: *Exploring special web archives collections related to COVID-19: The case of the National Library of the Czech Republic* (Aug 2023)

WARCnet Special Reports

Sharon Healy, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, Robert Jansma: *Skills, Tools, and Knowledge Ecologies in Web Archive Research* (Aug 2022)

Sharon Healy, Helena Byrne, Katharina Schmid, Juan-José Boté-Vericad, Lanna Floody: *Towards a Glossary for Web Archive Research: Version 1.0* (Aug 2023)

Sharon Healy, Helena Byrne: *Scholarly Use of Web Archives Across Ireland: The Past, Present & Future(s)* (Aug 2023)

All WARCnet Papers and WARCnet Special Reports can be downloaded for free from the project website warcnet.eu.

Author Information

Sharon Healy is an independent researcher and archivist with strong interests in internet/web history, RDM workflows for digital data, and the preservation of digitised, born digital, and reborn digital heritage. Sharon holds a BA (Hons) in Cultural Studies, an MA in Digital Humanities, a PG Diploma in Historical Archives, and a PhD in Digital Humanities & Archives.

Helena Byrne is the Curator of Web Archives at the British Library. She was the Lead Curator on the IIPC Content Development Group 2022, 2018 and 2016 Olympic and Paralympic collections. Helena completed a Master's in Library and Information Studies at University College Dublin, Ireland in 2015. Previously she worked as an English language teacher in Turkey, South Korea, and Ireland. Helena is also an independent researcher that focuses on the history of women's football in Ireland. Her previous publications cover both web archives and sports history.

Katharina Schmid is an IT developer at the Bavarian State Library. She holds an MA in European Literatures and Cultures and an MSc in Computer Science for Graduates in the Humanities or Social Sciences. Previously she contributed to a research project on applying methods from the digital humanities to web archives.

Juan-José Boté-Vericad is a Lecturer professor at the Faculty of Information and Media Studies at the University of Barcelona. He holds a BA in Computer Science, MA in Digital Content Management and PhD in digital preservation. His main research interests are in web archives and digital preservation focusing on the research data lifecycle.

Lanna Floody is an independent researcher based in Ireland with interests in humanities, cultural studies, and information studies.

Acknowledgements

We would like to thank the WARCnet Steering Group for organising network meetings and activities, which provided the capacity for members to develop projects such as this one. We would also like to thank Maynooth University, the British Library, the Bavarian State Library, and the University of Barcelona for providing back bone support to the authors.

Accessibility

As part of our commitment to accessibility, we have tried to ensure that the URLs provided in this document are (i) captured in a web archive close to the time of access on the live web or (ii) saved in a web archive close to the time of access on the live web. In case of future link rot, we have documented which archive the URL may be found in the Bibliography, e.g. [URL Memento: Wayback Machine]. To further assist with accessibility, we utilise the Arial/Calibri fonts, and apply [alt text] for all images contained in this document. Should a reader need to access this document in some other form which would provide better accessibility, please contact the authors.

Abstract

This study offers a novel approach for developing a glossary of terms and concepts for web archive research. In doing so, we offer a selection of glossary entries, which could be used as a starting point for beginners in web archive research or by web archives to communicate their holdings to users and stakeholders.

List of Contents

INTRODUCTION 1

BACKGROUND & METHODOLOGY 2

USING THE ZOTER0 GLOSSARY LIBRARY 6

Glossary Entries 8

REFERENCES 66

INTRODUCTION

In this study we discuss the development of a glossary of terms and concepts for web archive research, using a novel approach. The study may also be regarded as a companion glossary for the WARCnet Special Report, *Skills, Tools, and Ecologies in Web Archive Research* compiled by the project team for Web Archives – Researcher Skills & Tools Survey (WARST). WARST was a collaborative project by researchers from Maynooth University, the British Library, the International Internet Preservation Consortium, the Bavarian State Library, and the University of Siegen. The study sought to identify, and document skills, tools and knowledge required to achieve a range of different research goals within the web archiving lifecycle and explored the challenges for participation in web archive research, and the intersections of such challenges across communities of practice.

The WARST research team were all members of Web ARChive studies network researching web domains and events (WARCnet, warcnet.eu), and the core of the team participated in WARCnet WG3: Digital research methods and tools.[1] WARCnet (Web ARChive studies network researching web domains and events) was a transnational interdisciplinary network, primarily based in Europe. It provided network meetings and activities for web archivists, IT developers and researchers who study the archived web, with the involvement of some leading European web archives, and the International Internet Preservation Consortium (IIPC) (Brügger, 2020; WARCnet, n.d., About WARCnet). WARCnet was funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

In the WARST Report, the project team described web archive research to be inclusive of web archiving, curation, and the use of web archives and archived web content for research or other purposes (Healy et al., 2022), as well as the processes and activities described in the Archive-It's web archiving lifecycle model from appraisal and acquisition to storage and preservation, to replay, access, use, and reuse (Bragg & Hanna, 2013). It is from this point of departure that we started to explore the development of a glossary for web archive research, using a novel approach.

[1] WARCnet WG3: Digital research methods and tools, https://cc.au.dk/en/warcnet/working-groups

BACKGROUND & METHODOLOGY

In compiling the WARST Report, some members of the research team began to compile a glossary of terms and concepts which were used in the writing of the report, but also of terms which might be useful for novices starting out in web archive research. While the terms glossary and dictionary are sometimes used interchangeably, there are some differences. For example, Pediaa (2021) offers an account below of the difference between a glossary and a dictionary:

> The main difference between glossary and dictionary is that a glossary is a reference source that includes terms specific to a particular subject, while a dictionary is a reference source that gives you information about words, their meanings, pronunciation, and usage (Pediaa, 2021).

The description above offers some form of boundary for the terms dictionary and glossary. Therefore, to avoid spending too much time on semantics, for the purpose of this project we consider this glossary to be "a reference source" that includes a list of terms, concepts, protocols, standards, and tools that are specific or related to web archive research, and the full web archiving lifecycle.

From there, we developed the glossary further with descriptions and definitions drawn from a wide range of web resources such as the glossaries by the Archive-It Help Center, the Library of Congress, and the Digital Preservation Coalition (DPC). We also incorporated descriptions/definitions from other online resources such as Wikipedia, Wikidata, GitHub, and the PREMIS events vocabulary. In some cases, we draw directly from web pages for descriptions of tools and software. We further used Zotero open-source software, to collect and organise the web resources, and we populated the Zotero metadata field elements with relevant information inclusive of a term description in the [abstract] field. Other interested colleagues also joined us to complete this undertaking.

Simultaneously, WARCnet WG5 were also examining existing glossaries, dictionaries, and data vocabularies which would contribute to their aims of compiling a codebook of web archive data formats.[2] Thus, WG3 & WG5 proposed to join forces to do a collaborative workshop at the WARCnet London Meeting in June 2022, in the form of a Glossary Sprint. More terms and concepts were added to the glossary resource as a result of the sprint.

[2] WARCnet W5: The WARCnet Code Book of web archive data formats, https://cc.au.dk/en/warcnet/working-groups

Moreover, a meaningful discussion ensued amongst the participants on how useful the resource might be for individuals with different levels of experience in web archive research, and how useful the resource might be as a teaching aid. It was also suggested that the resource might serve as a tool for web archives in communicating their holdings to users, e.g., by appending the glossary to web archive reports, and sharing the resource with researchers and other stakeholders to better understand the dynamics of their web archive collections. The participants discussed further whether it would be worth expanding the resource through crowdsourcing, to build a more inclusive glossary which would not only incorporate more granular processes and activities within the web archiving lifecycle, but also terms and concepts that represent a multitude of disciplines which cross over into web archive research. Most certainly, this type of investigation is worth pursuing in the future and version 1.0 could be adapted to incorporate new terms.

As mentioned, we developed the glossary with descriptions and definitions drawn from a wide range of web resources and used Zotero for collection and organisation. Described as a free, open-source research tool, Zotero assists researchers in the collection, organisation, analysis and sharing of research (Zotero, n.d., About). During the glossary workshop, participants who were first-time users found Zotero relatively easy to use, and several participants mentioned that they were already familiar with the software. However, we acknowledge that we would need to conduct a user study with a larger, more diverse, participant sample, to be able to draw more substantive conclusions, nonetheless, it provided us with some confidence in the choice of software.

For this project, we use the current version, Zotero 6 for Windows, but it is also available for Mac and Linux. While individual items may be added manually to the application, Zotero also comes with a browser plug-in connector for Chrome, Firefox, Edge, or Safari. Through a click of the connector icon in the browser, Zotero can automatically create an item of the appropriate type, populate the metadata fields, and download a PDF attachment if available (Figure 1). It may also attach useful links (e.g., a link to a PubMed entry) or supplemental data files or notes (Zotero, n.d., Quick Start Guide). Items may also be added by entering a DOI, ISBN, PMID, arXIV IDs etc. (Figure 2).

Figure 1: Adding items through the Zotero browser connector

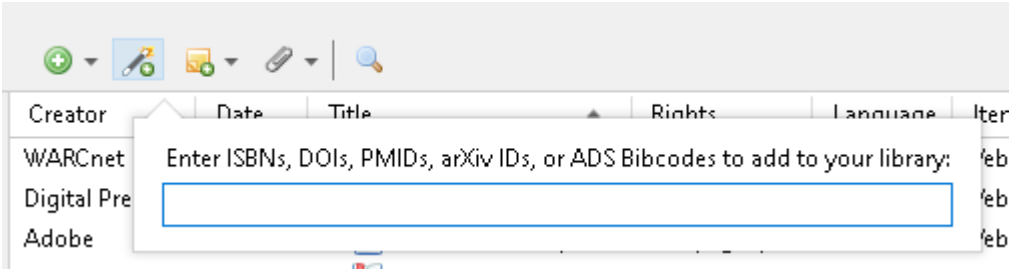

Figure 2: Adding items to Zotero through an identifier such as a DOI or PMID

While the Zotero browser connector can extract embedded metadata from a web resource, and automatically integrate it into the application metadata fields, it is quite often the case that there is not enough metadata provided in the HTML source code, thus, one may need to enter a lot of metadata information manually. On the other hand, some web resources provide a great deal of metadata, such as Wikipedia and GitHub, inclusive of an abstract or description of the contents of the page, which we could easily modify to use as a term description. This is further explained below.

> The quality of the data Zotero imports is determined by the information supplied on the webpage. Some websites provide very high-quality data using a standard way to provide Zotero with data (via embedded metadata). Other websites provide only limited metadata (e.g., only the title of a blog post) or no metadata at all. For many sites, Zotero has website-specific "translators" to obtain the best quality metadata. Zotero recognizes almost all library catalogs, most news sites, research databases and scientific publishers (Zotero, n.d., Quick Start Guide).

When applying metadata manually, Zotero offers the user multiple 'Item Types' to choose from covering a vast range of media types as outlined below. Zotero suggests that "Item types [...] should be regarded as flexible, broad categories. Item types are generally determined based on how items should be cited" (Zotero, n.d., Item Types and Fields).

Zotero Item Types

Artwork	Letter
Audio Recording	Magazine Article
Bill	Manuscript
Blog Post	Map
Book	Newspaper Article
Book Section	Patent
Case	Podcast
Conference Paper	Presentation
Dictionary Entry	Radio Broadcast
Document	Report
Email	Software
Encyclopedia Article	Statute
Film	Thesis
Forum Post	TV Broadcast
Hearing	Video Recording
Instant Message	Webpage
Interview	Attachment
Journal Article	Note

The glossary (version 1.0) is publicly accessible through a Zotero Group web library, so there is no need to download the application, or register for an account. For those who do wish to download, or are already using a Zotero desktop application, one can easily add the web library to their application by logging in/or registering online, and then by joining the group. This will automatically add the web library to the individual's desktop application in the folder window on the left. Note, it will need a few minutes to synchronise. For the most part, the Zotero web library interface replicates the interface of the desktop application, although there are some functional differences. However, this does not affect the basic user functions which we describe in the next section. The current version of the project web library is available as, Zotero Groups - Towards a Glossary for Web Archive Research.[3]

[3] Healy, S., Byrne, H., Schmid, K., Boté-Vericad, J.-J., Floody, L. (2021+) Zotero Groups - Towards a Glossary for Web Archive Research. Zotero, https://www.zotero.org/groups/4380600/towards_a_glossary_for_web_archive_research.

In the interface (web library or desktop), users may browse the glossary item entries through an A-Z or Z-A scroll of any column, such as the title, creator, or rights. Users can add or remove columns by clicking on the column icon at the top right of the item list window. This will open a dialogue box with more options, and users can add or remove columns as required (Figure 3). The entries may also be searched in the top search bar for a title, creator, and year, or through a free-text search against the full-text content in the metadata. Additionally, users may navigate the resource by scrolling through the tags in the bottom left window, or by entering a keyword in the tag search bar (Figure 4).

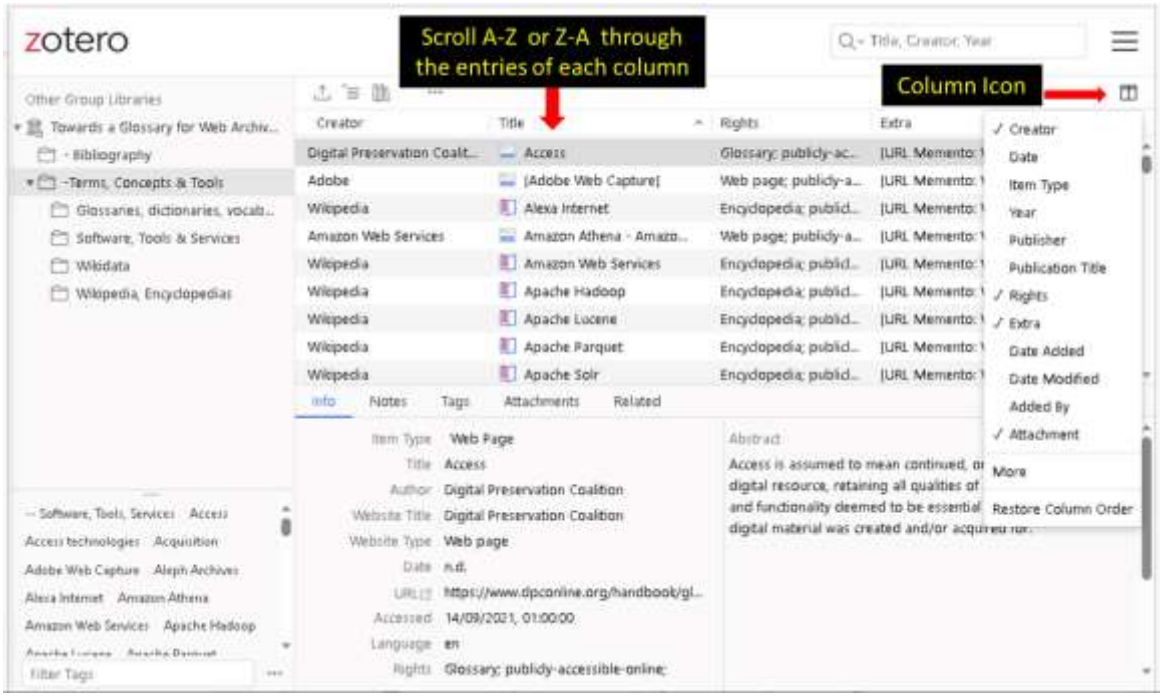

Figure 3: Screenshot of Zotero group glossary interface with options of scrolling A-Z through columns and adding or removing a column

The tags were contributed by the research team. In some instances, the tags are derived from a taxonomy for web archive research that is being developed by one of the authors, which includes instances from controlled vocabularies for subject headings and name authority files (e.g., Library of Congress subject headings), as well as other concepts such as Wikidata and the UNESCO Thesaurus. Tags were also compiled from the item itself such as the titles of organisations, services, or software, as well as being derived from a non-controlled vocabulary. The non-controlled vocabulary was compiled through subjects suggested by the

researchers themselves and includes subjects or keywords for tools and concepts which may not, as yet, have entered the lexicon of more formal controlled vocabularies. This is not surprising due to the rate at which technologies change within the domain of web archive research (Healy et al. 2022), and thus new software, methodologies, standards, and terminologies are bound to keep emerging.

To view the metadata of an individual item, simply click on the item once to open the metadata dialogue box or double-click on an item to be directed to a new window to view the source origin for an entry (Figure 5). Users may also wish to browse the items by resource type, in the left window with folders.

To conclude, at the end of this document we provide a written compilation of the Glossary Entries from Version 1.0, and hope that it may be of some assistance as a glossary companion for the WARCnet Special Report, *Skills Tools, and Knowledge Ecologies in Web Archive Research*. But also, we hope that the glossary may be helpful for novices starting out in web archive research, whether this is in web archiving or curation, IT services, legal services, or the use of the archived web for research or any other purposes.

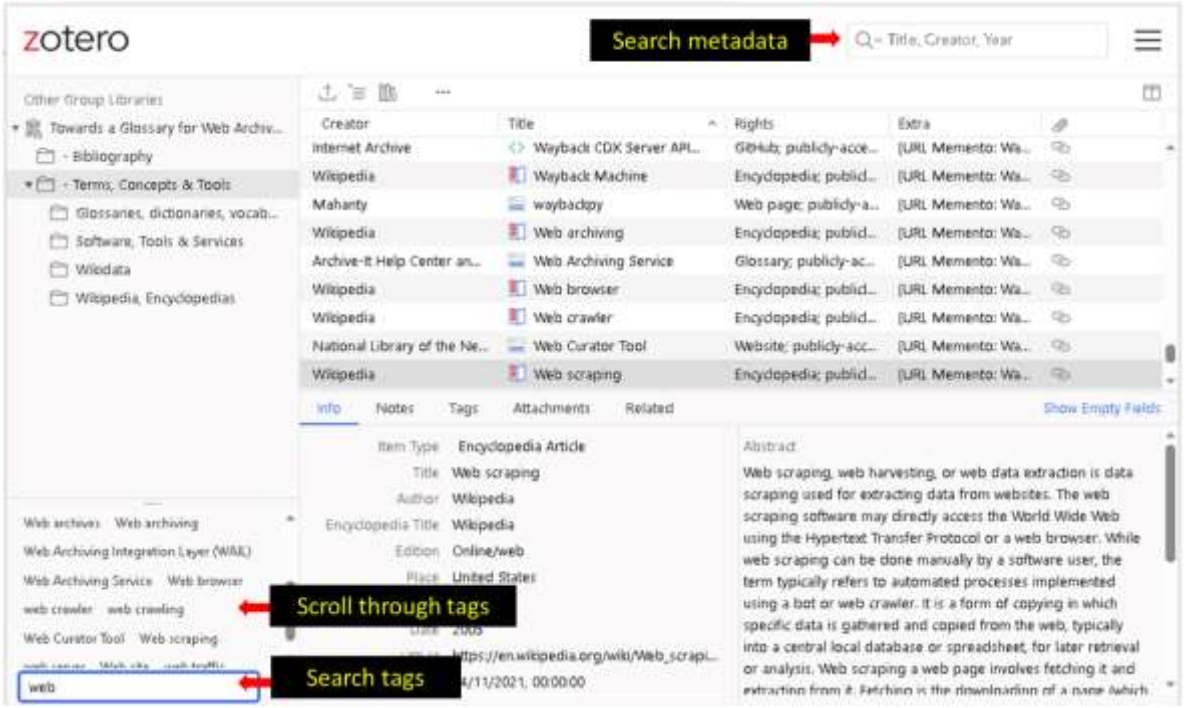

Figure 4: Screenshot of Zotero group glossary interface with search options through metadata or tags

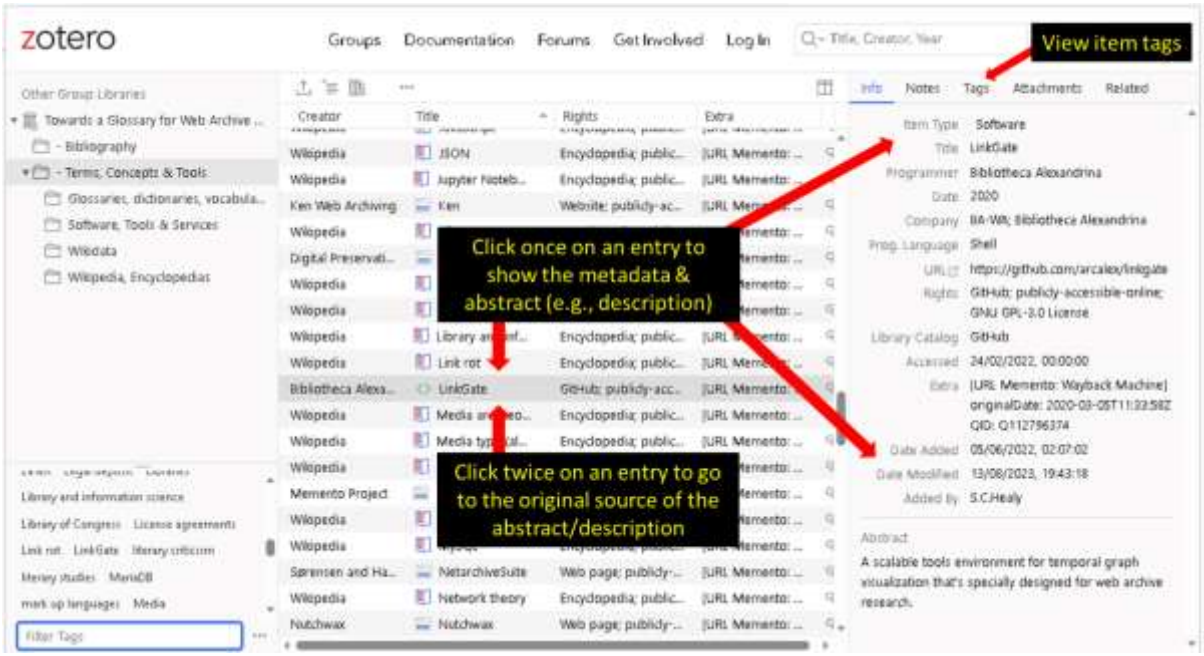

Figure 5: Screenshot of Zotero group glossary interface with search options through metadata or tags

Glossary Entries

Please note, the text in the description fields is taken, for the most part, directly from a web resource, thus, there may be errors in the spelling and/or grammar in the originals which are reflected in this glossary. The entries are taken from a wide range of resources which may use en-US, or en-GB.

Access

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Access is assumed to mean continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential for the purposes the digital material was created and/or acquired for.
URL	https://www.dpconline.org/handbook/glossary/#A

Adobe web capture

Type	Web Page
Author	Adobe
Description	Captures and archives a website as a PDF using Adobe Acrobat 9, by converting HTML to PDF.

URL	https://acrobatusers.com/tutorials/capture-and-archive-website-using-adobe-acrobat-9/

Alexa Internet

Type	Encyclopedia Article
Author	Wikipedia
Description	Alexa Internet, Inc. is an American web traffic analysis company based in San Francisco. It is a wholly owned subsidiary of Amazon. In December 2021, Amazon announced that it was closing down Alexa Internet, with service to discontinue as of May 1, 2022. Alexa was founded in April 1996 by Brewster Kahle and Bruce Gilliat as an independent company in 1996 and acquired by Amazon in 1999 for $250 million in stock. Alexa provides web traffic data, global rankings, and other information on over 30 million websites. Alexa estimates website traffic based on a sample of millions of Internet users using browser extensions, as well as from sites that have chosen to install an Alexa script.
URL	https://en.wikipedia.org/wiki/Alexa_Internet

Amazon Athena - Amazon Web Services

Type	Web Page
Author	Amazon Web Services
Description	Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run. Athena is easy to use. Simply point to your data in Amazon S3, define the schema, and start querying using standard SQL. Most results are delivered within seconds. Athena is out-of-the-box integrated with AWS Glue Data Catalog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions, and maintain schema versioning.
URL	https://aws.amazon.com/athena/

Amazon Web Services

Type	Encyclopedia Article
Author	Wikipedia
Description	Amazon Web Services, Inc. (AWS) is a subsidiary of Amazon providing on-demand cloud computing platforms and APIs to individuals, companies, and governments, on a metered pay-as-you-go basis. These cloud computing web

	services provide a variety of basic abstract technical infrastructure and distributed computing building blocks and tools. AWS's virtual computers emulate most of the attributes of a real computer, including hardware central processing units (CPUs) and graphics processing units (GPUs) for processing; local/RAM memory; hard-disk/SSD storage; a choice of operating systems; networking; and pre-loaded application software such as web servers, databases, and customer relationship management (CRM).
URL	https://en.wikipedia.org/wiki/Amazon_Web_Services

Apache Hadoop

Type	Encyclopedia Article
Author	Wikipedia
Description	Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. Hadoop was originally designed for computer clusters built from commodity hardware, which is still the common use. It has since also found use on clusters of higher-end hardware.
URL	https://en.wikipedia.org/wiki/Apache_Hadoop

Apache Lucene

Type	Encyclopedia Article
Author	Wikipedia
Description	Apache Lucene is a free and open-source search engine software library, originally written in Java by Doug Cutting. It is supported by the Apache Software Foundation and is released under the Apache Software License. Lucene is widely used as a standard foundation for non-research search applications. Lucene has been ported to other programming languages including Object Pascal, Perl, C#, C++, Python, Ruby and PHP.
URL	https://en.wikipedia.org/wiki/Apache_Lucene

Apache Parquet

Type	Encyclopedia Article
Author	Wikipedia
Description	Apache Parquet is a free and open-source column-oriented data storage format of the Apache Hadoop ecosystem. It is similar to the other columnar-storage file formats available in Hadoop namely RCFile and ORC. It is compatible with most of the data processing frameworks in the Hadoop

	environment. It provides efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk.
URL	https://en.wikipedia.org/wiki/Apache_Parquet

Apache Solr

Type	Encyclopedia Article
Author	Wikipedia
Description	Apache Solr is an open-source enterprise-search platform, written in Java. Its major features include full-text search, hit highlighting, faceted search, real-time indexing, dynamic clustering, database integration, NoSQL features and rich document (e.g., Word, PDF) handling. Providing distributed search and index replication, Solr is designed for scalability and fault tolerance. Solr is widely used for enterprise search and analytics use cases and has an active development community and regular releases. Solr runs as a standalone full-text search server. It uses the Lucene Java search library at its core for full-text indexing and search and has REST-like HTTP/XML and JSON APIs that make it usable from most popular programming languages.
URL	https://en.wikipedia.org/wiki/Apache_Solr

Apache Spark

Type	Encyclopedia Article
Author	Wikipedia
Description	Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance. Originally developed at the University of California, Berkeley's AMPLab, the Spark codebase was later donated to the Apache Software Foundation, which has maintained it since.
URL	https://en.wikipedia.org/wiki/API

API

Type	Encyclopedia Article
Author	Wikipedia
Description	An application programming interface (API) is a connection between computers or between computer programs. It is a type of software interface, offering a service to other pieces of software. A document or standard that describes how to build or use such a connection or interface is called an API specification. A computer system that meets this standard is said to implement or expose an API. The term API may refer either to the specification or to the

	implementation. In contrast to a user interface, which connects a computer to a person, an application programming interface connects computers or pieces of software to each other. It is not intended to be used directly by a person (the end user) other than a computer programmer who is incorporating it into software.
URL	https://en.wikipedia.org/wiki/API

Arc File Format

Type	Web Page/Format Reference
Author	Mike Burner
Author	Brewster Kahle
Description	The Internet Archive stores the data it collects in large aggregate files for ease of storage in a conventional file system. It is the Archive's experience that it is difficult to manage hundreds of millions of small files in most existing file systems. This document describes the format of the aggregate files. The file format was designed to meet several requirements: The file must be self-contained: it must permit the aggregated objects to be identified and unpacked without the use of a companion index file. The format must be extensible to accommodate files retrieved via a variety of network protocols, including http, ftp, news, gopher, and mail. The file must be "stream able": it must be possible to concatenate multiple archive files in a data stream. Once written, a record must be viable: the integrity of the file must not depend on subsequent creation of an in-file index of the contents. The reader will quickly recognize, however, that an external index of the contents and object-offsets will greatly enhance the retrievability of objects stored in this format. The Archive maintains such indices but does not seek to standardize their format.
URL	https://archive.org/web/researcher/ArcFileFormat.php

Archival science

Type	Encyclopedia Article
Author	Wikipedia
Description	Archival science, or archival studies, is the study and theory of building and curating archives, which are collections of documents, recordings and data storage devices. To build and curate an archive, one must acquire and evaluate recorded materials, and be able to access them later. To this end, archival science seeks to improve methods for appraising, storing, preserving, and cataloging recorded materials. An archival record preserves data that is

	not intended to change. In order to be of value to society, archives must be trustworthy. Therefore, an archivist has a responsibility to authenticate archival materials, such as historical documents, and to ensure their reliability, integrity, and usability. Archival records must be what they claim to be; accurately represent the activity they were created for; present a coherent picture through an array of content; and be in usable condition in an accessible location.
URL	https://en.wikipedia.org/wiki/Archival_science

Archive Now

Type	Software
Programmer	Old Dominion University Web Science and Digital Libraries Research Group
Description	A tool to push web resources into web archives. Archive Now (archivenow) currently is configured to push resources into four public web archives. You can easily add more archives by writing a new archive handler.
URL	https://github.com/oduwsdl/archivenow

Archives Unleashed Toolkit

Type	Web Page
Author	The Archives Unleashed Project
Description	The Archives Unleashed Toolkit is an open-source platform for analyzing web archives built on Apache Spark, which provides powerful tools for analytics and data processing.
URL	https://archivesunleashed.org/aut/

ArchiveWeb.page

Type	Web Page
Author	Webrecorder
Description	ArchiveWeb.page is a tool from Webrecorder to turn your browser into a full-featured interactive web archiving system. ArchiveWeb.page is available as an extension for any Chrome or Chromium based browsers. To create web archives, the extension (or app) will be needed. Once created, the archives can be viewed in any modern browser using ReplayWeb.page -- no extension required
URL	https://archiveweb.page/

Authenticity

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Authenticity The digital material is what it purports to be. In the case of electronic records, it refers to the trustworthiness of the electronic record as a record. In the case of "born digital" and digitised materials, it refers to the fact that whatever is being cited is the same as it was when it was first created unless the accompanying metadata indicates any changes. Confidence in the authenticity of digital materials over time is particularly crucial owing to the ease with which alterations can be made.
URL	https://www.dpconline.org/handbook/glossary

Automatic indexing

Type	Encyclopedia Article
Author	Wikipedia
Description	Automatic indexing is the computerized process of scanning large volumes of documents against a controlled vocabulary, taxonomy, thesaurus, or ontology and using those controlled terms to quickly and effectively index large electronic document depositories. These keywords or language are applied by training a system on the rules that determine what words to match. There are additional parts to this such as syntax, usage, proximity, and other algorithms based on the system and what is required for indexing. This is taken into account using Boolean statements to gather and capture the indexing information out of the text. As the number of documents exponentially increases with the proliferation of the Internet, automatic indexing will become essential to maintaining the ability to find relevant information in a sea of irrelevant information.
URL	https://www.dpconline.org/handbook/glossary

Bandwidth (computing)

Type	Encyclopedia Article
Author	Wikipedia
Description	In computing, bandwidth is the maximum rate of data transfer across a given path. Bandwidth may be characterized as network bandwidth, data bandwidth, or digital bandwidth. This definition of bandwidth is in contrast to the field of signal processing, wireless communications, modem data transmission, digital communications, and electronics, in which bandwidth is

	used to refer to analog signal bandwidth measured in hertz, meaning the frequency range between lowest and highest attainable frequency while meeting a well-defined impairment level in signal power.
URL	https://en.wikipedia.org/wiki/Bandwidth_(computing)

BibTeX

Type	Encyclopedia Article
Author	Wikipedia
Description	BibTeX is reference management software for formatting lists of references. The BibTeX tool is typically used together with the LaTeX document preparation system. The name is a portmanteau of the word bibliography and the name of the TeX typesetting software. The purpose of BibTeX is to make it easy to cite sources in a consistent manner, by separating bibliographic information from the presentation of this information, similarly to the separation of content and presentation/style supported by LaTeX itself.
URL	https://en.wikipedia.org/wiki/BibTeX

Bit loss

Type	Wiki
Author	Wikidata
Description	bit loss (Q112796335) The corruption of the lowest level of information digital data in transmission or during storage.
URL	https://www.wikidata.org/wiki/Q112796335

Bit preservation

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Bit preservation is a term used to denote a very basic level of preservation of digital resource as it was submitted (literally preservation of the bits forming a digital resource). It may include maintaining onsite and offsite backup copies, virus checking, fixity-checking, and periodic refreshment to new storage media. Bit preservation is not digital preservation, but it does provide one building block for the more complete set of digital preservation practices and processes that ensure the survival of digital content and also its usability, display, context and interpretation over time.
URL	https://www.dpconline.org/handbook/glossary/#B

Bit rot

Type	Wiki
Author	Wikidata
Description	bit rot (Q1390705) accumulation of data corruption on a storage device over time.
URL	https://www.wikidata.org/wiki/Q1390705

BitCurator

Type	Web Page
Author	BitCurator NLP
Description	BitCurator is open source software for collecting institutions to extract, analyze, and produce reports on features of interest in text extracted from born-digital materials contained in collections.
URL	https://bitcurator.net/

Born-digital

Type	Encyclopedia Article
Author	Wikipedia
Description	The term born-digital refers to materials that originate in a digital form. This is in contrast to digital reformatting, through which analog materials become digital, as in the case of files created by scanning physical paper records. It is most often used in relation to digital libraries and the issues that go along with said organizations, such as digital preservation and intellectual property. However, as technologies have advanced and spread, the concept of being born-digital has also been discussed in relation to personal consumer-based sectors, with the rise of e-books and evolving digital music. Other terms that might be encountered as synonymous include "natively digital", "digital-first", and "digital-exclusive".
URL	https://en.wikipedia.org/wiki/Born-digital

Browsertrix Cloud

Type	Software
Programmer	Webrecorder
Description	Browsertrix Cloud is an open-source cloud-native high-fidelity browser-based crawling service designed to make web archiving easier and more accessible for everyone. The service provides an API and UI for scheduling crawls and viewing results and managing all aspects of crawling process. This system

	provides the orchestration and management around crawling, while the actual crawling is performed using Browsertrix Crawler containers, which are launched for each crawl. The system is designed to run in both Kubernetes and Docker Swarm, as well as locally under Podman.
URL	https://github.com/webrecorder/browsertrix

Browsertrix Crawler

Type	Software
Programmer	Webrecorder
Description	Browsertrix Crawler is a simplified (Chrome) browser-based high-fidelity crawling system, designed to run a complex, customizable browser-based crawl in a single Docker container. Browsertrix Crawler uses puppeteer-cluster and puppeteer to control one or more browsers in parallel.
URL	https://github.com/webrecorder/browsertrix

Brozzler

Type	Web Page/Glossary
Author	Archive-It Help Center
Description	Brozzler is a distributed web crawler that uses a real browser (chrome or chromium) to fetch pages and embedded urls and to extract links. It also uses youtube-dl to enhance media capture capabilities.
URL	https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms

CDX file format

Type	Web Page/ Format Reference
Author	Internet Archive
Description	A CDX file consists of individual lines of text, each of which summarizes a single web document. The first line in the file is a legend for interpreting the data, and the other lines contain the data for referencing the corresponding pages within the host. The first character of the file is the field delimiter used in the rest of the file. This is followed by the literal "CDX" and then individual field markers.
URL	https://archive.org/web/researcher/cdx_file_format.php

Close reading

Type	Encyclopedia Article

Author Wikipedia

Description In literary criticism, close reading is the careful, sustained interpretation of a brief passage of a text. A close reading emphasizes the single and the particular over the general, effected by close attention to individual words, the syntax, the order in which the sentences unfold ideas, as well as formal structures. A truly attentive close reading means thinking about both what is being said in a passage (the content), and how it is being said (the form, i.e., the way the content is presented) and leading it to possibilities for observation and insight.

URL https://en.wikipedia.org/wiki/Close_reading

Collection development

Type Encyclopedia Article

Author Wikipedia

Description Library collection development is the process of systematically building the collection of a particular library to meet the information needs of the library users (a service population) in a timely and economical manner using information resources locally held as well as resources from other organizations. According to the International Federation of Library Associations and Institutions (IFLA), acquisition and collection development focuses on methodological and topical themes pertaining to acquisition of print and other analogue library materials (by purchase, exchange, gift, legal deposit), and the licensing and purchase of electronic information resources. Collection development involves activities that need a librarian or information professional who is specialized in improving the library's collection. The process includes the selection of information materials that respond to the users or patrons need as well as de-selection of unwanted information materials, called weeding. It also involves the planning strategies for continuing acquisition, evaluation of new information materials and the existing collection in order to determine how well a particular library serves its users.

URL https://en.wikipedia.org/wiki/Close_reading

Comma-separated values (CSV)

Type Encyclopedia Article

Author Wikipedia

Description A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field

	separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields. The CSV file format is not fully standardized. Separating fields with commas is the foundation, but commas in the data or embedded line breaks must be handled specially. Some implementations disallow such content while others surround the field with quotation marks, which yet again creates the need for escaping if quotation marks are present in the data. The term "CSV" also denotes several closely-related delimiter-separated formats that use other field delimiters such as semicolons. These include tab-separated values and space-separated values. A delimiter guaranteed not to be part of the data greatly simplifies parsing.
URL	https://en.wikipedia.org/wiki/Comma-separated_values

Computer-assisted qualitative data analysis software (CAQDAS)

Type	Encyclopedia Article
Author	Wikipedia
Description	Computer-assisted (or aided) qualitative data analysis software (CAQDAS) offers tools that assist with qualitative research such as transcription analysis, coding and text interpretation, recursive abstraction, content analysis, discourse analysis, grounded theory methodology, etc.
URL	https://en.wikipedia.org/wiki/Computer-assisted_qualitative_data_analysis_software

Conifer (previously Webrecorder)

Type	Web Page
Author	Rhizome
Description	Conifer is an open-source web archiving service that creates an interactive copy of any web page that you browse, including content revealed by your interactions such as playing video and audio, scrolling, clicking buttons, and so forth. Conifer is both a tool to create high-fidelity, interactive captures of any web site you browse and a platform to make those captured websites accessible.
URL	https://conifer.rhizome.org

Content analysis

Type	Encyclopedia Article
Author	Wikipedia

Description: Content analysis is the study of documents and communication artifacts, which might be texts of various formats, pictures, audio, or video. Social scientists use content analysis to examine patterns in communication in a replicable and systematic manner. Practices and philosophies of content analysis vary between academic disciplines. They all involve systematic reading or observation of texts or artifacts which are assigned labels (sometimes called codes) to indicate the presence of interesting, meaningful pieces of content. By systematically labeling the content of a set of texts, researchers can analyse patterns of content quantitatively using statistical methods or use qualitative methods to analyse meanings of content within texts. Computers are increasingly used in content analysis to automate the labeling (or coding) of documents. Simple computational techniques can provide descriptive data such as word frequencies and document lengths.

URL: https://en.wikipedia.org/w/index.php?title=Content_analysis&oldid=1065940520

Copyright

Type: Encyclopedia Article

Author: Wikipedia

Description: A copyright is a type of intellectual property that gives its owner the exclusive right to copy and distribute a creative work, usually for a limited time. The creative work may be in a literary, artistic, educational, or musical form. Copyright is intended to protect the original expression of an idea in the form of a creative work, but not the idea itself. A copyright is subject to limitations based on public interest considerations, such as the fair use doctrine in the United States. Some jurisdictions require "fixing" copyrighted works in a tangible form. It is often shared among multiple authors, each of whom holds a set of rights to use or license the work, and who are commonly referred to as rights holders. These rights frequently include reproduction, control over derivative works, distribution, public performance, and moral rights such as attribution. Copyrights can be granted by public law and are in that case considered "territorial rights". This means that copyrights granted by the law of a certain state, do not extend beyond the territory of that specific jurisdiction. Copyrights of this type vary by country; many countries, and sometimes a large group of countries, have made agreements with other countries on procedures applicable when works "cross" national borders or national rights are inconsistent. Typically, the public law duration of a copyright expires 50 to 100 years after the creator dies, depending on the jurisdiction. Some countries require certain copyright formalities to establishing copyright, others recognize copyright in any completed work, without a formal registration. When the copyright of a work expires, it enters the public domain.

URL	https://en.wikipedia.org/wiki/Copyright

Country code top-level domain (ccTLD)

Type	Encyclopedia Article
Author	Wikipedia
Description	A country code top-level domain (ccTLD) is an Internet top-level domain generally used or reserved for a country, sovereign state, or dependent territory identified with a country code. All ASCII ccTLD identifiers are two letters long, and all two-letter top-level domains are ccTLDs. Creation and delegation of ccTLDs is described in RFC 1591, corresponding to ISO 3166-1 alpha-2 country codes. ccTLDs are subjected to requirements that are determined by each country's domain name regulation corporation.
URL	https://en.wikipedia.org/wiki/Country_code_top-level_domain

Crawl frequency

Type	Web Page/Glossary
Author	Maria Praetzellis; Archive-It Help Center
Description	Crawl frequency is the rate at which you set your seeds to be crawled. The frequency is on a per seed basis and can be set to one time, twice daily, daily, weekly, monthly, bi-monthly, quarterly, semiannual, or annual.
URL	https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms

Crawl/Capture/Harvest

Type	Web Page/Glossary
Author	Library of Congress
Description	Terms used interchangeably to all mean the process of downloading all code, images, documents, and other files essential to completely reproduce a website, ultimately preserving the original form of the retrieved content. Also involves capturing metadata about the conditions of the crawl.
URL	https://www.loc.gov/programs/web-archiving/about-this-program/glossary/

Creative Commons license (CC)

Type	Encyclopedia Article
Author	Wikipedia
Description	A Creative Commons (CC) license is one of several public copyright licenses that enable the free distribution of an otherwise copyrighted "work". A CC

	license is used when an author wants to give other people the right to share, use, and build upon a work that the author has created. There are several types of Creative Commons license. Each license differs by several combinations that condition the terms of distribution. They were initially released on December 16, 2002, by Creative Commons, a U.S. non-profit corporation founded in 2001.
URL	https://en.wikipedia.org/wiki/Creative_Commons_license

CSS

Type	Encyclopedia Article
Author	Wikipedia
Description	Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language such as HTML or XML (including XML dialects such as SVG, MathML or XHTML). CSS is a cornerstone technology of the World Wide Web, alongside HTML and JavaScript. CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts. This separation can improve content accessibility; provide more flexibility and control in the specification of presentation characteristics; enable multiple web pages to share formatting by specifying the relevant CSS in a separate .css file, which reduces complexity and repetition in the structural content; and enable the .css file to be cached to improve the page load speed between the pages that share the file and its formatting. CSS also has rules for alternate formatting if the content is accessed on a mobile device.
URL	https://en.wikipedia.org/wiki/CSS

Curator

Type	Encyclopedia Article
Author	Wikipedia
Description	A curator (from Latin: cura, meaning "to take care") is a manager or overseer. When working with cultural organizations, a curator is typically a "collections curator" or an "exhibitions curator" and has multifaceted tasks dependent on the particular institution and its mission. In recent years the role of curator has evolved alongside the changing role of museums, and the term "curator" may designate the head of any given division. More recently, new kinds of curators have started to emerge: "community curators", "literary curators", "digital curators" and "biocurators".
URL	https://en.wikipedia.org/wiki/Curator

Dark and Stormy Archives

Type	Web Page
Author	Shawn M. Jones
Description	The Dark and Stormy Archives (DSA) project exists to provide storytelling solutions to improve the understanding of web archive collections. With search engines, collection users must first have a query in mind. What if the user does not know enough about the collection to form a query? Our goal is to provide a "summary of summaries" in the form of social media storytelling that describes a collection sufficiently for a user to decide if that collection merits further time.
URL	https://oduwsdl.github.io/dsa/

DAT file format

Type	Web Page/Format Reference
Author	Internet Archive
Description	A DAT file contains meta-data about the documents stored in ARC files. The header line for a document in a DAT file always has mime type alexa/dat. The data that follows is separated into individual lines of the form <tag><space><value> where <tag> is defined in the cdx/dat legend, and value is text that does not contain a newline, perhaps further constrained by the definition of the tag.
URL	https://archive.org/web/researcher/dat_file_format.php

Data compression

Type	Wiki
Author	Wikidata
Description	data compression (Q2493) is a process of encoding information using fewer bits than the original representation
URL	https://www.wikidata.org/wiki/Q2493

Data degradation

Type	Encyclopedia Article
Author	Wikipedia
Description	Data degradation is the gradual corruption of computer data due to an accumulation of non-critical failures in a data storage device. The phenomenon is also known as data decay, data rot or bit rot.
URL	https://en.wikipedia.org/wiki/Data_degradation

Data management

Type	Encyclopedia Article
Author	Wikipedia
Description	Data management comprises all disciplines related to managing data as a valuable resource.
URL	https://en.wikipedia.org/wiki/Data_management

Data mining

Type	Encyclopedia Article
Author	Wikipedia
Description	Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of extracting information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use.
URL	https://en.wikipedia.org/wiki/Data_mining

Data set

Type	Encyclopedia Article
Author	Wikipedia
Description	A data set (or dataset) is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as for example height and weight of an object, for each member of the data set. Data sets can also consist of a collection of documents or files. In the open data discipline, data set is the unit to measure the information released in a public open data repository.
URL	https://en.wikipedia.org/wiki/Data_mining

Data transmission

Type	Wiki
Author	Wikidata
Description	data transmission (Q389772) physical transfer of data; transfer of data (a digital bit stream or a digitized analog

	signal) over a point-to-point or point-to-multipoint communication channel.
URL	https://www.wikidata.org/wiki/Q389772

Digiboard

Type	Web Page/Glossary
Author	Library of Congress
Description	A custom-built tool used by the Library of Congress to manage many aspects of the Library's web archiving processes.
URL	https://www.loc.gov/programs/web-archiving/about-this-program/glossary/

Digital curation

Type	Encyclopedia Article
Author	Wikipedia
Description	Digital curation is the selection, preservation, maintenance, collection, and archiving of digital assets. Digital curation establishes, maintains, and adds value to repositories of digital data for present and future use. This is often accomplished by archivists, librarians, scientists, historians, and scholars. Enterprises are starting to use digital curation to improve the quality of information and data within their operational and strategic processes. Successful digital curation will mitigate digital obsolescence, keeping the information accessible to users indefinitely. Digital curation includes digital asset management, data curation, digital preservation, and electronic records management.
URL	https://en.wikipedia.org/wiki/Digital_curation

Digital dark age

Type	Encyclopedia Article
Author	Wikipedia
Description	The digital dark age is a lack of historical information in the digital age as a direct result of outdated file formats, software, or hardware that becomes corrupt, scarce, or inaccessible as technologies evolve and data decay. Future generations may find it difficult or impossible to retrieve electronic documents and multimedia, because they have been recorded in an obsolete and obscure file format, or on an obsolete physical medium, for example, floppy disks. The name derives from the term Dark Ages in the sense that there could be a relative lack of records in the digital age, as documents are transferred to digital formats and original copies are lost. An early mention of the term came

	from Terry Kuny at a conference of the International Federation of Library Associations and Institutions (IFLA) in 1997.
URL	https://en.wikipedia.org/wiki/Digital_dark_age

Digital history

Type	Encyclopedia Article
Author	Wikipedia
Description	Digital history is the use of digital media to further historical analysis, presentation, and research. It is a branch of the digital humanities and an extension of quantitative history, cliometrics, and computing. Digital history is commonly digital public history, concerned primarily with engaging online audiences with historical content, or, digital research methods, that further academic research. Digital history outputs include: digital archives, online presentations, data visualizations, interactive maps, time-lines, audio files, and virtual worlds to make history more accessible to the user. Recent digital history projects focus on creativity, collaboration, and technical innovation, text mining, corpus linguistics, network analysis, 3D modeling, and big data analysis. By utilizing these resources, the user can rapidly develop new analyses that can link to, extend, and bring to life existing histories.
URL	https://en.wikipedia.org/wiki/Digital_history

Digital humanities

Type	Encyclopedia Article
Author	Wikipedia
Description	Digital humanities (DH) is an area of scholarly activity at the intersection of computing or digital technologies and the disciplines of the humanities. It includes the systematic use of digital resources in the humanities, as well as the analysis of their application. DH can be defined as new ways of doing scholarship that involve collaborative, transdisciplinary, and computationally engaged research, teaching, and publishing. It brings digital tools and methods to the study of the humanities with the recognition that the printed word is no longer the main medium for knowledge production and distribution. By producing and using new applications and techniques, DH makes new kinds of teaching possible, while at the same time studying and critiquing how these impact cultural heritage and digital culture. DH is also applied in research. Thus, a distinctive feature of DH is its cultivation of a two-way relationship between the humanities and the digital: the field both employs technology in the pursuit of humanities research and subjects technology to humanistic questioning and interrogation, often simultaneously.

URL https://en.wikipedia.org/wiki/Digital_humanities

Digital media

Type Encyclopedia Article

Author Wikipedia

Description Digital media means any communication media that operate with the use of any of various encoded machine-readable data formats. Digital media can be created, viewed, distributed, modified, listened to, and preserved on a digital electronics device. Digital can be defined as any data represented by a series of digits, while media refers to methods of broadcasting or communicating this information. Together, digital media refers to mediums of digitized information broadcast to us through a screen and/or a speaker. This also includes text, audio, video, and graphics that are transmitted over the internet for viewing or listening to on the internet.

URL https://en.wikipedia.org/wiki/Digital_media

Digital preservation

Type Encyclopedia Article

Author Wikipedia

Description In library and archival science, digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies, and actions to ensure access to reformatted and "born-digital" content, regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time. The need for digital preservation mainly arises because of the relatively short lifespan of digital media.

URL https://en.wikipedia.org/wiki/Digital_preservation

Digital record

Type Wiki

Author Wikidata

Description digital record (Q111664863) compilation of recorded information created, modified, stored, retrieved, and distributed by digital means.

URL https://www.wikidata.org/wiki/Q111664863

Digital stewardship

Type	Wiki
Author	Wikidata
Description	digital stewardship (Q112796603) management of digital objects with an eye to long-term preservation and accessibility.
URL	https://www.wikidata.org/wiki/Q112796603

Distant reading

Type	Encyclopedia Article
Author	Wikipedia
Description	Distant reading is an approach in literary studies that applies computational methods to literary data, usually derived from large digital libraries, for the purposes of literary history and theory. While the term is collective and is used to refer to a range of different computational methods of analysing literary data, similar approaches also include macroanalysis, cultural analytics, computational formalism, computational literary studies, quantitative literary studies, and algorithmic literary criticism.
URL	https://en.wikipedia.org/wiki/Distant_reading

Distributed computing

Type	Encyclopedia Article
Author	Wikipedia
Description	Distributed computing is a field of computer science that studies distributed systems. A distributed system is a system whose components are located on different networked computers, which communicate and coordinate their actions by passing messages to one another from any system. The components interact with one another to achieve a common goal. Three significant characteristics of distributed systems are: concurrency of components, lack of a global clock, and independent failure of components. It deals with a central challenge that, when components of a system fails, it doesn't imply the entire system fails. A computer program that runs within a distributed system is called a distributed program (and distributed programming is the process of writing such programs). Distributed computing also refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers, which communicate with each other via message passing.

URL	https://en.wikipedia.org/wiki/Distributed_computing

Documenting The Now

Type	Wiki
Author	Wikidata
Description	Documenting The Now (Q75971350) A set of tools and an online community that support the ethical collection, use, and preservation of social media content.
URL	https://www.wikidata.org/wiki/Q75971350

DROID: file format identification tool

Type	Web Page
Author	The National Archives
Description	DROID is a software tool developed by The National Archives to perform automated batch identification of file formats. DROID is designed to meet the fundamental requirement of any digital repository to be able to identify the precise format of all stored digital objects, and to link that identification to a central registry of technical information about that format and its dependencies.
URL	https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

Dynamic web page

Type	Encyclopedia Article
Author	Wikipedia
Description	A server-side dynamic web page is a web page whose construction is controlled by an application server processing server-side scripts. In server-side scripting, parameters determine how the assembly of every new web page proceeds and including the setting up of more client-side processing. A client-side dynamic web page processes the web page using JavaScript running in the browser as it loads. JavaScript can interact with the page via Document Object Model, or DOM, to query page state and modify it. Even though a web page can be dynamic on the client-side, it can still be hosted on a static hosting service such as GitHub Pages or Amazon S3 as long as there isn't any server-side code included. Client-side-scripting, server-side scripting, or a combination of these make for the dynamic web experience in a browser.
URL	https://en.wikipedia.org/wiki/Dynamic_web_page

Elastic Stack: Elasticsearch, Kibana, Beats & Logstash

Type	Web Page
Author	Elasticsearch
Description	Elastic Stack is comprised of Elasticsearch, Kibana, Beats, and Logstash (also known as the ELK Stack) and more. Reliably and securely take data from any source, in any format, then search, analyze, and visualize.
URL	https://www.elastic.co/elastic-stack

Elasticsearch

Type	Encyclopedia Article
Author	Wikipedia
Description	Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents. Elasticsearch is developed in Java and is dual-licensed under the source-available Server-Side Public License and the Elastic license, while other parts fall under the proprietary (source-available) Elastic License. Official clients are available in Java, .NET (C#), PHP, Python, Apache Groovy, Ruby and many other languages.
URL	https://en.wikipedia.org/wiki/Elasticsearch

Electrolyte

Type	Web Page
Author	MirrorWeb
Description	The firm [MirrorWeb] currently deploys two archival webcrawlers. The first is Heritrix, a tried-and-true open-source tool in widespread usage. The second is Electrolyte, MirrorWeb's proprietary tool which, as Harriet Christie, COO at MirrorWeb explains, is a more sophisticated, agile, capable, and thorough tool for exploration and capture in the most rigorous and dynamic digital domains
URL	https://www.mirrorweb.com/solutions/sec-17a-4

File format

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	A file format is a standard way that information is encoded for storage in a computer file. It tells the computer how to display, print, and process, and save the information. It is dictated by the application program which created the file, and the operating system under which it was created and stored. Some

	file formats are designed for very particular types of data, others can act as a container for different types. A particular file format is often indicated by a file name extension containing three or four letters that identify the format.
URL	https://www.dpconline.org/handbook/glossary

Fixity check

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Fixity check is a method for ensuring the integrity of a file and verifying it has not been altered or corrupted. During transfer, an archive may run a fixity check to ensure a transmitted file has not been altered en route. Within the archive, fixity checking is used to ensure that digital files have not been altered or corrupted. It is most often accomplished by computing checksums such as MD5, SHA1 or SHA256 for a file and comparing them to a stored value.
URL	https://www.dpconline.org/handbook/glossary

Full-text search

Type	Encyclopedia Article
Author	Wikipedia
Description	In text retrieval, full-text search refers to techniques for searching a single computer-stored document or a collection in a full-text database. Full-text search is distinguished from searches based on metadata or on parts of the original texts represented in databases (such as titles, abstracts, selected sections, or bibliographical references). In a full-text search, a search engine examines all of the words in every stored document as it tries to match search criteria (for example, text specified by a user). Full-text-searching techniques appeared in the 1960s, for example IBM STAIRS from 1969, and became common in online bibliographic databases in the 1990s.
URL	https://en.wikipedia.org/wiki/Full-text_search

Gephi

Type	Encyclopedia Article
Author	Wikipedia
Description	Gephi is an open-source network analysis and visualization software package written in Java on the NetBeans platform.
URL	https://en.wikipedia.org/wiki/Gephi

GLAM Workbench - Web Archives

Type	Web Page
Author	Tim Sherratt; Andrew Jackson
Description	A collection of tools and examples to help individuals work with data from galleries, libraries, archives, and museums. We tend to think of a web archive as a site we go to when links are broken – a useful fallback, rather than a source of new research data. But web archives don't just store old web pages, they capture multiple versions of web resources over time. Using web archives we can observe change – we can ask historical questions. This collection of notebooks is intended to help historians, and other researchers, frame those questions by revealing what sort of data is available, how to get it, and what you can do with it. These notebooks focus on four particular web archives: the UK Web Archive, the Australian Web Archive (National Library of Australia), the New Zealand Web Archive (National Library of New Zealand), and the Internet Archive. These notebooks focus on data that is readily accessible and able to be used without the need for special equipment. They use existing APIs to get data in manageable chunks.
URL	https://glam-workbench.github.io/web-archives

HeidiSQL - MariaDB, MySQL, MSSQL, PostgreSQL and SQLite made easy

Type	Web Page
Author	HeidiSQL
Description	HeidiSQL is free software and has the aim to be easy to learn. "Heidi" lets you see and edit data and structures from computers running one of the database systems MariaDB, MySQL, Microsoft SQL, PostgreSQL and SQLite. Invented in 2002 by Ansgar, HeidiSQL belongs to the most popular tools for MariaDB and MySQL worldwide.
URL	https://www.heidisql.com

Heritrix

Type	Encyclopedia Article
Author	Wikipedia
Description	Heritrix is a web crawler designed for web archiving. It was written by the Internet Archive. It is available under a free software license and written in Java. The main interface is accessible using a web browser, and there is a command-line tool that can optionally be used to initiate crawls. Heritrix was developed jointly by the Internet Archive and the Nordic national libraries on specifications written in early 2003. The first official release was in January

	2004, and it has been continually improved by employees of the Internet Archive and other interested parties.
URL	https://en.wikipedia.org/wiki/Heritrix

HTML

Type	Encyclopedia Article
Author	Wikipedia
Description	The HyperText Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript. Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document. HTML provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists, links, quotes and other items. HTML elements are delineated by tags, written using angle brackets. Browsers do not display the HTML tags but use them to interpret the content of the page. HTML can embed programs written in a scripting language such as JavaScript, which affects the behavior and content of web pages. Inclusion of CSS defines the look and layout of content. The World Wide Web Consortium (W3C), former maintainer of the HTML and current maintainer of the CSS standards, has encouraged the use of CSS over explicit presentational HTML since 1997.
URL	https://en.wikipedia.org/wiki/HTML

HTTrack

Type	Encyclopedia Article
Author	Wikipedia
Description	HTTrack is a free and open-source Web crawler and offline browser, developed by Xavier Roche and licensed under the GNU General Public License Version 3. HTTrack allows users to download World Wide Web sites from the Internet to a local computer. By default, HTTrack arranges the downloaded site by the original site's relative link-structure. The downloaded (or "mirrored") website can be browsed by opening a page of the site in a browser.
URL	https://en.wikipedia.org/wiki/HTTrack

Human-readable

Type	Wiki

Author	Wikidata
Description	human-readable (Q16716513) is a representation of information readable directly by humans
URL	https://www.wikidata.org/wiki/Q16716513

Hypermedia

Type	Encyclopedia Article
Author	Wikipedia
Description	Hypermedia, an extension of the term hypertext, is a nonlinear medium of information that includes graphics, audio, video, plain text, and hyperlinks. This designation contrasts with the broader term multimedia, which may include non-interactive linear presentations as well as hypermedia. It is also related to the field of electronic literature. The World Wide Web is a classic example of hypermedia to access web content, whereas a non-interactive cinema presentation is an example of standard multimedia due to the absence of hyperlinks. Most modern hypermedia is delivered via electronic pages from a variety of systems including media players, web browsers, and stand-alone applications (i.e., software that does not require network access). Audio hypermedia is emerging with voice command devices and voice browsing.
URL	https://en.wikipedia.org/wiki/Hypermedia

Hypertext

Type	Encyclopedia Article
Author	Wikipedia
Description	Hypertext is text displayed on a computer display or other electronic devices with references (hyperlinks) to other text that the reader can immediately access. Hypertext documents are interconnected by hyperlinks, which are typically activated by a mouse click, keypress set, or screen touch. Apart from text, the term "hypertext" is also sometimes used to describe tables, images, and other presentational content formats with integrated hyperlinks. Hypertext is one of the key underlying concepts of the World Wide Web, where Web pages are often written in the Hypertext Markup Language (HTML). As implemented on the Web, hypertext enables the easy-to-use publication of information over the Internet. The term was coined by Ted H. Nelson in 1965.
URL	https://en.wikipedia.org/wiki/Hypertext

Information retrieval (IR)

Type	Encyclopedia Article
Author	Wikipedia
Description	Information retrieval (IR) in computing and information science is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing. Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds. Automated information retrieval systems are used to reduce what has been called information overload. An IR system is a software system that provides access to books, journals, and other documents; stores and manages those documents.
URL	https://en.wikipedia.org/wiki/Information_retrieval

Information science (also known as information studies)

Type	Encyclopedia Article
Author	Wikipedia
Description	Information science (also known as information studies) is an academic field which is primarily concerned with analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information. Practitioners within and outside the field study the application and the usage of knowledge in organizations in addition to the interaction between people, organizations, and any existing information systems with the aim of creating, replacing, improving, or understanding information systems. Historically, information science is associated with computer science, data science, psychology, technology, and intelligence agencies. However, information science also incorporates aspects of diverse fields such as archival science, cognitive science, commerce, law, linguistics, museology, management, mathematics, philosophy, public policy, and social sciences.
URL	https://en.wikipedia.org/wiki/Information_science

Information seeking

Type	Encyclopedia Article
Author	Wikipedia
Description	Information seeking is the process or activity of attempting to obtain information in both human and technological contexts. Information seeking is related to, but different from, information retrieval (IR).

URL	https://en.wikipedia.org/wiki/Information_seeking

Information technology (IT)

Type	Encyclopedia Article
Author	Wikipedia
Description	Information technology (IT) is the use of computers to create, process, store, and exchange all kinds of electronic data and information. IT is typically used within the context of business operations as opposed to personal or entertainment technologies. IT is considered to be a subset of information and communications technology (ICT). An information technology system (IT system) is generally an information system, a communications system, or, more specifically speaking, a computer system – including all hardware, software, and peripheral equipment – operated by a limited group of IT users. The term is commonly used as a synonym for computers and computer networks, but it also encompasses other information distribution technologies such as television and telephones. Several products or services within an economy are associated with information technology, including computer hardware, software, electronics, semiconductors, internet, telecom equipment, and e-commerce.
URL	https://en.wikipedia.org/wiki/Information_technology

Instaloader

Type	Web Page
Author	Alexander Graf
Description	Instaloader is a tool to download pictures (or videos) along with their captions and other metadata from Instagram.
URL	https://instaloader.github.io/

interoperability

Type	Wiki
Author	Wikidata
Description	interoperability (Q749647) ability of products or systems to work with each other via compatible interface.
URL	https://www.wikidata.org/wiki/Q749647

Iramuteq — IRaMuTeQ

Type	Web Page
Author	Pierre Ratinaud
Description	Iramuteq - R interface for Multidimensional Analyses of Texts and Questionnaires. Free software built with free software.
URL	http://www.iramuteq.org/

JavaScript

Type	Encyclopedia Article
Author	Wikipedia
Description	JavaScript, often abbreviated JS, is a programming language that is one of the core technologies of the World Wide Web, alongside HTML and CSS. Over 97% of websites use JavaScript on the client side for web page behavior, often incorporating third-party libraries. All major web browsers have a dedicated JavaScript engine to execute the code on users' devices. Although Java and JavaScript are similar in name, syntax, and respective standard libraries, the two languages are distinct and differ greatly in design.
URL	https://en.wikipedia.org/wiki/JavaScript

JSON

Type	Encyclopedia Article
Author	Wikipedia
Description	JSON (JavaScript Object Notation) is an open standard file format and data interchange format that uses human-readable text to store and transmit data objects consisting of attribute–value pairs and arrays (or other serializable values). It is a common data format with diverse uses in electronic data interchange, including that of web applications with servers. JSON is a language-independent data format. It was derived from JavaScript, but many modern programming languages include code to generate and parse JSON-format data. JSON filenames use the extension .json.
URL	https://en.wikipedia.org/wiki/JSON

Jupyter Notebook

Type	Encyclopedia Article
Author	Wikipedia
Description	Jupyter Notebook (formerly IPython Notebooks) is an open source web-based interactive computational environment for creating notebook

	documents. A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension. Jupyter notebooks are built upon a number of popular open-source libraries
URL	https://en.wikipedia.org/wiki/Project_Jupyter#Jupyter_Notebook

Ken

Type	Web Page
Author	Ken Web Archiving
Description	Ken is an e-discovery and archiving software suite that helps organizations gain control of the data from collaboration apps and dynamic websites.
URL	https://ken-webarchiving.com/

Kibana

Type	Encyclopedia Article
Author	Wikipedia
Description	Kibana is a source-available data visualization dashboard software for Elasticsearch, whose free and open source successor in OpenSearch is OpenSearch Dashboards. The combination of Elasticsearch, Logstash, and Kibana, referred to as the "Elastic Stack" (formerly the "ELK stack"), is available as a product or service.
URL	https://en.wikipedia.org/wiki/Kibana

Kilobyte (KB)

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Kilobyte (KB) A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 Bytes
URL	https://www.dpconline.org/handbook/glossary

LaTeX

Type	Encyclopedia Article
Author	Wikipedia
Description	LaTeX is a software system for document preparation. When writing, the writer uses plain text as opposed to the formatted text found in "What You

	See Is What You Get" word processors like Microsoft Word, LibreOffice Writer and Apple Pages. The writer uses markup tagging conventions to define the general structure of a document to stylise text throughout a document (such as bold and italics), and to add citations and cross-references. A TeX distribution such as TeX Live or MiKTeX is used to produce an output file (such as PDF or DVI) suitable for printing or digital distribution. LaTeX is widely used in academia for the communication and publication of scientific documents in many fields, including mathematics, computer science, engineering, physics, chemistry, economics, linguistics, quantitative psychology, philosophy, and political science. It also has a prominent role in the preparation and publication of books and articles that contain complex multilingual materials, such as Sanskrit and Greek. LaTeX uses the TeX typesetting program for formatting its output, and is itself written in the TeX macro language. LaTeX can be used as a standalone document preparation system, or as an intermediate format.
URL	https://en.wikipedia.org/wiki/LaTeX

Legal deposit

Type	Encyclopedia Article
Author	Wikipedia
Description	Legal deposit is a legal requirement that a person or group submit copies of their publications to a repository, usually a library. The number of copies required varies from country to country. Typically, the national library is the primary repository of these copies. In some countries there is also a legal deposit requirement placed on the government, and it is required to send copies of documents to publicly accessible libraries. The legislation covering the requirement varies from country to country, but is often enshrined in copyright law. Until the late 20th century, legal deposit covered only printed and sometimes audio-visual materials, but in the 21st century, most countries have had to extend their legislation to cover digital documents as well. In 2000, UNESCO published a new and enlarged edition of Jean Lunn's 1981 Guidelines for Legal Deposit Legislation, which addresses the issue of electronic formats in its recommendations for the construction of legal deposit legislation.
URL	https://en.wikipedia.org/wiki/Legal_deposit

Library and information science

Type	Encyclopedia Article
Author	Wikipedia

Description	Library and information science (LIS) (sometimes given as the plural library and information sciences) is a branch of academic disciplines that deals generally with organization, access, collection, and protection/regulation of information, whether in physical (e.g. art, legal proceedings) or digital forms. By the late 1960s, mainly due to the meteoric rise of human computing power and the new academic disciplines formed therefrom, academic institutions began to add the term "information science" to their names. The first school to do this was at the University of Pittsburgh in 1964. More schools followed during the 1970s and 1980s, and by the 1990s almost all library schools in the USA had added information science to their names. Although there are exceptions, similar developments have taken place in other parts of the world. In Denmark, for example, the 'Royal School of Librarianship' changed its English name to The Royal School of Library and Information Science in 1997. In spite of various trends to merge the two fields, some consider the two original disciplines, library science and information science, to be separate. However, it is common today is to use the terms as synonyms or to drop the term "library" and to speak about information departments or I-schools. There have also been attempts to revive the concept of documentation and to speak of Library, information and documentation studies (or science).
URL	https://en.wikipedia.org/wiki/Library_and_information_science

Link rot

Type	Encyclopedia Article
Author	Wikipedia
Description	Link rot (also called link death, link breaking, or reference rot) is the phenomenon of hyperlinks tending over time to cease to point to their originally targeted file, web page, or server due to that resource being relocated to a new address or becoming permanently unavailable. A link that no longer points to its target, often called a broken or dead link (or sometimes orphan link), is a specific form of dangling pointer. The rate of link rot is a subject of study and research due to its significance to the internet's ability to preserve information. Estimates of that rate vary dramatically between studies.
URL	https://en.wikipedia.org/wiki/Link_rot

LinkGate

Type	Software
Programmer	Bibliotheca Alexandrina

Description	A scalable tools environment for temporal graph visualization that's specially designed for web archive research.
URL	https://github.com/arcalex/linkgate

Media archaeology

Type	Encyclopedia Article
Author	Wikipedia
Description	Media archaeology or media archeology is a field that attempts to understand new and emerging media through close examination of the past, and especially through critical scrutiny of dominant progressivist narratives of popular commercial media such as film and television. Media archaeologists often evince strong interest in so-called dead media, noting that new media often revive and recirculate material and techniques of communication that had been lost, neglected, or obscured. Some media archaeologists are also concerned with the relationship between media fantasies and technological development, especially the ways in which ideas about imaginary or speculative media affect the media that actually emerge.
URL	https://en.wikipedia.org/wiki/Media_archaeology

Media type (also known as MIME type)

Type	Encyclopedia Article
Author	Wikipedia
Description	A media type (also known as MIME type) is a two-part identifier for file formats and format contents transmitted on the Internet. The Internet Assigned Numbers Authority (IANA) is the official authority for the standardization and publication of these classifications. Media types were originally defined in Request for Comments RFC 2045 (MIME) Part One: Format of Internet Message Bodies (Nov 1996) in November 1996 as a part of MIME (Multipurpose Internet Mail Extensions) specification, for denoting type of email message content and attachments; hence the original name, MIME type. Media types are also used by other internet protocols such as HTTP and document file formats such as HTML, for similar purposes.
URL	https://en.wikipedia.org/wiki/Media_type

Memento Project

Type	Encyclopedia Article
Author	Wikipedia

Description	Memento is a United States National Digital Information Infrastructure and Preservation Program (NDIIPP)–funded project aimed at making Web-archived content more readily discoverable. The project is being led by the Los Alamos National Laboratory and Old Dominion University. Rather than expecting people to know about the growing number of Web archives, and to guess which archive might hold an older version of the resource they're looking for, Memento proposes to make archived content discoverable via the original URL that the searcher already knew about.
URL	https://en.wikipedia.org/wiki/Memento_Project

Memento Time Travel

Type	Web Page
Author	Memento Project
Description	Time Travel helps you find and view versions of web pages that existed at some time in the past. These prior versions of web pages are named Mementos. Mementos can be found in web archives or in systems that support versioning such as wikis and revision control systems. The Time Travel portal offers two distinct services: Find and Reconstruct.
URL	http://timetravel.mementoweb.org/about/

Metadata

Type	Encyclopedia Article
Author	Wikipedia
Description	Metadata is "data that provides information about other data", but not the content of the data, such as the text of a message or the image itself. There are many distinct types of metadata, including: Descriptive metadata — the descriptive information about a resource. It is used for discovery and identification. It includes elements such as title, abstract, author, and keywords. Structural metadata — metadata about containers of data and indicates how compound objects are put together, for example, how pages are ordered to form chapters. It describes the types, versions, relationships and other characteristics of digital materials. Administrative metadata — the information to help manage a resource, like resource type, permissions, and when and how it was created. Reference metadata — the information about the contents and quality of statistical data. Statistical metadata, also called process data, may describe processes that collect, process, or produce statistical data. Legal metadata — provides information about the creator, copyright holder, and public licensing, if provided. Metadata is not strictly

	bounded to one of these categories, as it can describe a piece of data in many other ways.
URL	https://en.wikipedia.org/wiki/Metadata

MySQL

Type	Encyclopedia Article
Author	Wikipedia
Description	MySQL is an open-source relational database management system (RDBMS). Its name is a combination of "My", the name of co-founder Michael Widenius's daughter, and "SQL", the abbreviation for Structured Query Language. A relational database organizes data into one or more data tables in which data types may be related to each other; these relations help structure the data. SQL is a language programmers use to create, modify, and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a relational database in a computer's storage system, manages users, allows for network access, and facilitates testing database integrity and creation of backups. MySQL is free and open-source software under the terms of the GNU General Public License and is also available under a variety of proprietary licenses.
URL	https://en.wikipedia.org/wiki/MySQL

NetarchiveSuite

Type	Web Page
Author	Mikis Seth Sørensen
Author	Ulrich Karstoft Have
Description	The NetarchiveSuite is a complete web archiving software package, developed from 2004 and onwards, and used in production to harvest the Danish web since 2005. The primary function of the NetarchiveSuite is to plan, schedule and run web harvests of parts of the Internet. It scales to a wide range of tasks, from small, thematic harvests (e.g. related to special events, or special domains) to harvesting and archiving the content of an entire national domain. The software has built-in bit preservation functionality. The systems architecture allows for the software to be distributed among several machines, possibly on more than one geographical location. The NetarchiveSuite is built around the Heritrix web crawler, which it uses to harvest the web. The NetarchiveSuite software is developed and maintained by The Royal Danish Library (previously two separate institutions - The Royal Library Copenhagen and the State and University Library Aarhus)

	together with the National Library of France (BnF) and the National Library of Austria (ONB). The National Library of Spain (BNE) and National Library of Sweden (KB-SE) are also active members of the development project.
URL	https://sbforge.org/display/NAS

Network theory

Type	Encyclopedia Article
Author	Wikipedia
Description	Network theory is the study of graphs as a representation of either symmetric relations or asymmetric relations between discrete objects. In computer science and network science, network theory is a part of graph theory: a network can be defined as a graph in which nodes and/or edges have attributes (e.g. names). Network theory has applications in many disciplines including statistical physics, particle physics, computer science, electrical engineering, biology, economics, finance, operations research, climatology, ecology, public health, sociology, and neuroscience. Applications of network theory include logistical networks, the World Wide Web, Internet, gene regulatory networks, metabolic networks, social networks, epistemological networks, etc.; see List of network theory topics for more examples.
URL	https://en.wikipedia.org/wiki/Network_theory

Nutchwax

Type	Web Page
Author	Nutchwax
Description	NutchWAX ("Nutch + Web Archive eXtensions") searches web archive collections. The Web Archive eXtensions (WAX) include adaptation of the Nutch fetcher step to go against web archives rather than crawl the open net -- adaptation currently does Internet Archive ARC files only -- and plugins to add extra fields to the index that return an Archive Records' location in the repository, its collection name, etc.
URL	http://archive-access.sourceforge.net/projects/nutchwax/

OldWeb.today

Type	Software
Programmer	Ilya Kreymer
Description	oldweb.today (OWT) is a system that connects emulated web browsers to web archives, allowing users to browse the old web, today, as it was!

URL	https://github.com/oldweb-today/oldweb-today

OpenRefine

Type	Web Page
Author	OpenRefine
Description	OpenRefine (previously Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data. A free, open source, powerful tool for working with messy data
URL	https://openrefine.org

OpenWayback

Type	Software
Programmer	OpenWayback Development
Contributor	International Internet Preservation Consortium (IIPC)
Description	OpenWayback is the open source project aimed to develop Wayback Machine, the key software used by web archives worldwide to play back archived websites in the user's browser. OpenWayback is supported by the members of the International Internet Preservation Consortium (IIPC). Note: OpenWayback is no longer under active development.
URL	https://github.com/iipc/openwayback

Pandas (software)

Type	Encyclopedia Article
Author	Wikipedia
Description	Pandas is a software library written for the Python programming language for data manipulation and analysis. It offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010.
URL	https://en.wikipedia.org/wiki/Pandas_(software)

Petabyte (PB)

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Petabyte (PB) A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 Terabytes (TB).
URL	https://www.dpconline.org/handbook/glossary

Provenance

Type	Encyclopedia Article
Author	Wikipedia
Description	Provenance – also known as custodial history – is a core concept within archival science and archival processing. The term refers to the individuals, groups, or organizations that originally created or received the items in an accumulation of records, and to the items' subsequent chain of custody. The principle of provenance (also termed the principle of "archival integrity", and a major strand in the broader principle of respect des fonds) stipulates that records originating from a common source (or fonds) should be kept together – where practicable, physically; but in all cases intellectually, in the way in which they are catalogued and arranged in finding aids. Conversely, records of different provenance should be preserved and documented separately. In archival practice, proof of provenance is provided by the operation of control systems that document the history of records kept in archives, including details of amendments made to them. The authority of an archival document or set of documents of which the provenance is uncertain (because of gaps in the recorded chain of custody) will be considered to be severely compromised.
URL	https://en.wikipedia.org/wiki/Provenance

Public domain

Type	Encyclopedia Article
Author	Wikipedia
Description	The public domain consists of all the creative work to which no exclusive intellectual property rights apply. Those rights may have expired, been forfeited, expressly waived, or may be inapplicable. As examples, the works of William Shakespeare, Ludwig van Beethoven, Leonardo da Vinci, and Georges Méliès are in the public domain either by virtue of their having been created before copyright existed, or by their copyright term having expired. Some works are not covered by a country's copyright laws and are therefore in the

	public domain; for example, in the United States, items excluded from copyright include the formulae of Newtonian physics, cooking recipes, and all computer software created before 1974. Other works are actively dedicated by their authors to the public domain (see waiver); examples include reference implementations of cryptographic algorithms, and the image-processing software ImageJ (created by the National Institutes of Health). The term public domain is not normally applied to situations where the creator of a work retains residual rights, in which case use of the work is referred to as "under license" or "with permission". As rights vary by country and jurisdiction, a work may be subject to rights in one country and be in the public domain in another. Some rights depend on registrations on a country-by-country basis, and the absence of registration in a particular country, if required, gives rise to public-domain status for a work in that country. The term public domain may also be interchangeably used with other imprecise or undefined terms such as the public sphere or commons, including concepts such as the "commons of the mind", the "intellectual commons", and the "information commons".
URL	https://en.wikipedia.org/wiki/Public_domain

Public records

Type	Encyclopedia Article
Author	Wikipedia
Description	Public records are documents or pieces of information that are not considered confidential and generally pertain to the conduct of government. Depending on jurisdiction, examples of public records include information pertaining to births, deaths, marriages, and documented transactions with government agencies.
URL	https://en.wikipedia.org/wiki/Public_records

Python (programming language)

Type	Encyclopedia Article
Author	Wikipedia
Description	Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.

	Guido van Rossum began working on Python in the late 1980s, as a successor to the ABC programming language, and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features, such as list comprehensions and a cycle-detecting garbage collection system (in addition to reference counting). Python 3.0 was released in 2008 and was a major revision of the language that is not completely backward-compatible.
URL	https://en.wikipedia.org/wiki/Python_(programming_language)

pywb

Type	Web Page
Author	Ilya Kreymer
Description	The Webrecorder (pywb) toolkit is a full-featured, advanced web archiving capture and replay framework for python. It provides command-line tools and an extensible framework for high-fidelity web archive access and creation. A subset of features provides the basic functionality of a “Wayback Machine”.
URL	https://pywb.readthedocs.io/en/latest/

R (programming language)

Type	Encyclopedia Article
Author	Wikipedia
Description	R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners and statisticians for data analysis and developing statistical software. Users have created packages to augment the functions of the R language. According to surveys like Rexer's Annual Data Miner Survey and studies of scholarly literature databases, R is one of the most commonly used programming language used in data mining. As of February 2022, R ranks 13th in the TIOBE index, a measure of programming language popularity. The official R software environment is an open-source free software environment within the GNU package, available under the GNU General Public License. It is written primarily in C, Fortran, and R itself (partially self-hosting). Precompiled executables are provided for various operating systems. R has a command line interface. Multiple third-party graphical user interfaces are also available, such as RStudio, an integrated development environment, and Jupyter, a notebook interface.
URL	https://en.wikipedia.org/wiki/R_(programming_language)

Reborn digital

Type	Wiki
Author	Wikidata
Description	reborn digital (Q112795525) Niels Brügger (2016) describes reborn digital media, as media that has been collected and preserved and has undergone a change due to this process, such as emulations of computer games or materials in a web archive. http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html
URL	https://www.wikidata.org/wiki/Q112795525

Redaction

Type	Document
Author	Library of Congress
Description	redaction The process of modifying the content of a digital object to remove or mask information considered to be sensitive in nature (that is, the information cannot be viewed by non-authorized users of the repository).
URL	https://www.loc.gov/standards/premis/v3/preservation-events.pdf

Regular expression

Type	Encyclopedia Article
Author	Wikipedia
Description	A regular expression (shortened as regex or regexp; also referred to as rational expression) is a sequence of characters that specifies a search pattern. Usually such patterns are used by string-searching algorithms for "find" or "find and replace" operations on strings, or for input validation. It is a technique developed in theoretical computer science and formal language theory. The concept of regular expressions began in the 1950s, when the American mathematician Stephen Cole Kleene formalized the description of a regular language. They came into common use with Unix text-processing utilities. Different syntaxes for writing regular expressions have existed since the 1980s, one being the POSIX standard and another, widely used, being the Perl syntax. Regular expressions are used in search engines, search and replace dialogs of word processors and text editors, in text processing utilities such as sed and AWK and in lexical analysis. Many programming languages provide regex capabilities either built-in or via libraries, as it has uses in many situations.
URL	https://en.wikipedia.org/wiki/Regular_expression

Rhizome (organization)

Type	Encyclopedia Article
Author	Wikipedia
Description	Rhizome is an American not-for-profit arts organization that supports and provides a platform for new media art.
URL	https://en.wikipedia.org/wiki/Rhizome_(organization)

Robots exclusion standard

Type	Encyclopedia Article
Author	Wikipedia
Description	The robots exclusion standard, also known as the robots exclusion protocol or simply robots.txt, is a standard used by websites to communicate with web crawlers and other web robots. The standard specifies how to inform the web robot about which areas of the website should not be processed or scanned. Robots are often used by search engines to categorize websites. Not all robots cooperate with the standard; email harvesters, spambots, malware and robots that scan for security vulnerabilities may even start with the portions of the website where they have been told to stay out. The standard can be used in conjunction with Sitemaps, a robot inclusion standard for websites.
URL	https://en.wikipedia.org/wiki/Robots_exclusion_standard

RStudio

Type	Encyclopedia Article
Author	Wikipedia
Description	RStudio is an integrated development environment for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser.
URL	https://en.wikipedia.org/wiki/RStudio

Sandbox containment

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Sandbox Containment A secure computing environment for running novel, unattested or experimental code or changes in code, including potentially malicious code.

	The environment is self-contained with tightly controlled resources and is characteristically virtual.
URL	https://www.dpconline.org/handbook/glossary

Save Page Now (Internet Archive)

Type	Web Page
Author	NetLab
Description	The Internet Archive has a “Save page now” function that is highly convenient for two purposes: – It provides a way to instantly save content that you have found, and – It provides you with a stable reference to said content.
URL	https://www.netlab.dk/services/tools-and-tutorials/save-page-now/

Screencast

Type	Encyclopedia Article
Author	Wikipedia
Description	A screencast is a digital recording of computer screen output, also known as a video screen capture or a screen recording, often containing audio narration. The term screencast compares with the related term screenshot; whereas screenshot generates a single picture of a computer screen, a screencast is essentially a movie of the changes over time that a user sees on a computer screen, that can be enhanced with audio narration and captions.
URL	https://en.wikipedia.org/wiki/Screencast

Screenshot

Type	Encyclopedia Article
Author	Wikipedia
Description	A screenshot, also known as screen capture or screen grab, is a digital image that shows the contents of a computer display. A screenshot is created by the operating system or software running on the device powering the display. Additionally, screenshots can be captured by an external camera, using photography to capture contents on the screen.
URL	https://en.wikipedia.org/wiki/Screenshot

Seed URL

Type	Web Page/Glossary
Author	Library of Congress

Description	The crawler's starting or entry point and the access point within the archive. The seed URL is typically the URL selected for archiving by Library staff. The crawler follows links from the seed URL pages to subsequent pages.
URL	https://www.loc.gov/programs/web-archiving/about-this-program/glossary/

Share-alike

Type	Web Page
Author	Wikidata
Description	share-alike (Q2672228) condition for works or licences that require copies or adaptations of the work to be released under the same or similar licence as the original
URL	https://www.wikidata.org/wiki/Q2672228

SHINE (a UK Web Archive service)

Type	Web Page
Author	UK Web Archive
Description	SHINE, UK Web Archive prototype historical search engine. This tool has been developed as part of the Big UK Data Arts and Humanities project funded by the AHRC. The data was acquired by JISC from the Internet Archive (IA) and includes all .uk websites in the IA web collection crawled between around 1996 until April 2013. The collection comprises over 3.5 billion items (urls, images and other documents) and has been full-text indexed by the UK Web Archive. Also useful for trend analysis.
URL	https://www.webarchive.org.uk/shine

Snagit

Type	Web Page
Author	NetLab
Description	Snagit is a tool for screen capture in the form of images or video. But it can also be used to capture an individual webpage.
URL	https://www.netlab.dk/services/tools-and-tutorials/snagit/

Social Feed Manager

Type	Web Page
Author	GW Libraries and Academic Innovation
Description	Social Feed Manager is open source software that harvests social media data and related content from Twitter, Tumblr, Flickr, and Sina Weibo.

URL https://gwu-libraries.github.io/sfm-ui/

Social media

Type Encyclopedia Article

Author Wikipedia

Description Social media are interactive digital channels that facilitate the creation and sharing of information, ideas, interests, and other forms of expression through virtual communities and networks. While challenges to the definition of social media arise due to the variety of stand-alone and built-in social media services currently available, there are some common features: social media are interactive Web 2.0 Internet-based applications. User-generated content—such as text posts or comments, digital photos or videos, and data generated through all online interactions—is the lifeblood of social media. Users create service-specific profiles for the website or app that are designed and maintained by the social media organization. The term "social" in regard to media suggests that platforms are user-centric and enable communal activity. Users usually access social media services through web-based apps on desktops or download services that offer social media functionality to their mobile devices (e.g., smartphones and tablets). As users engage with these electronic services, they create highly interactive platforms which individuals, communities, and organizations can share, co-create, discuss, participate, and modify user-generated or self-curated content posted online. Some of the most popular social media websites, with more than 100 million registered users, include Facebook (and its associated Facebook Messenger), TikTok, WeChat, Instagram, QZone, Weibo, Twitter, Tumblr, Baidu Tieba, and LinkedIn. Depending on interpretation, other popular platforms that are sometimes referred to as social media services include YouTube, QQ, Quora, Telegram, WhatsApp, Signal, LINE, Snapchat, Pinterest, Viber, Reddit, Discord, VK, Microsoft Teams, and more. Wikis are examples of collaborative content creation. Observers have noted a wide range of positive and negative impacts when it comes to the use of social media.

URL https://en.wikipedia.org/wiki/Social_media

SolrWayback

Type Software

Programmer Thomas Egense

Description SolrWayback
A search interface and wayback machine for the UKWA Solr based warc-indexer framework.

URL	https://github.com/netarchivesuite/solrwayback

SQL

Type	Encyclopedia Article
Author	Wikipedia
Description	SQL (S-Q-L, "sequel"; Structured Query Language) is a domain-specific language used in programming and designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS). It is particularly useful in handling structured data, i.e. data incorporating relations among entities and variables. The scope of SQL includes data query, data manipulation (insert, update and delete), data definition (schema creation and modification), and data access control. Although SQL is essentially a declarative language (4GL), it also includes procedural elements. SQL was one of the first commercial languages to use Edgar F. Codd's relational model. The model was described in his influential 1970 paper, "A Relational Model of Data for Large Shared Data Banks". Despite not entirely adhering to the relational model as described by Codd, it became the most widely used database language. SQL became a standard of the American National Standards Institute (ANSI) in 1986, and of the International Organization for Standardization (ISO) in 1987. Since then, the standard has been revised to include a larger set of features. Despite the existence of standards, most SQL code requires at least some changes before being ported to different database systems.
URL	https://en.wikipedia.org/wiki/SQL

Static website

Type	Wiki
Author	Wikidata
Description	static website (Q52720701) website composed of static web pages.
URL	https://www.wikidata.org/wiki/Q52720701

Style sheet (web development)

Type	Encyclopedia Article
Author	Wikipedia
Description	A web style sheet is a form of separation of presentation and content for web design in which the markup (i.e., HTML or XHTML) of a webpage contains the page's semantic content and structure but does not define its visual layout (style). Instead, the style is defined in an external style sheet file using a style

	sheet language such as CSS or XSLT. This design approach is identified as a "separation" because it largely supersedes the antecedent methodology in which a page's markup defined both style and structure. The philosophy underlying this methodology is a specific case of separation of concerns.
URL	https://en.wikipedia.org/wiki/Style_sheet_(web_development)

Sub-domain

Type	Web Page
Author	Archive-It Help Center
Description	Sub-domain A directory named before the root web address, for example crawler.archive.org, in which crawler is the sub-domain.
URL	https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms

Tableau Software

Type	Encyclopedia Article
Author	Wikipedia
Description	Tableau Software is an American interactive data visualization software company focused on business intelligence. It was founded in 2003 in Mountain View, California, and is currently headquartered in Seattle, Washington. In 2019 the company was acquired by Salesforce for $15.7 billion. At the time, this was the largest acquisition by Salesforce (a leader in the CRM field) since its foundation. It was later surpassed by Salesforce's acquisition of Slack. The company's founders, Christian Chabot, Pat Hanrahan and Chris Stolte, were researchers at the Department of Computer Science at Stanford University. They specialized in visualization techniques for exploring and analyzing relational databases and data cubes and started the company as a commercial outlet for research at Stanford from 1999 to 2002. Tableau products query relational databases, online analytical processing cubes, cloud databases, and spreadsheets to generate graph-type data visualizations. The software can also extract, store, and retrieve data from an in-memory data engine.
URL	https://en.wikipedia.org/wiki/Tableau_Software

Terabyte (TB)

Type	Web Page/Glossary
Author	Digital Preservation Coalition
Description	Terabyte (TB)

	A unit of digital information often used to describe data or data storage size, equates to approximately 1,000 Gigabytes (GB).
URL	https://www.dpconline.org/handbook/glossary

Timestamp

Type	Web Page
Author	Wikidata
Description	timestamp (Q186885) sequence of characters or encoded information identifying when a certain event occurred.
URL	https://www.wikidata.org/wiki/Q186885

Topic model

Type	Encyclopedia Article
Author	Wikipedia
Description	In statistics and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear approximately equally in both. The "topics" produced by topic modeling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.
URL	https://en.wikipedia.org/wiki/Topic_model

twarc

Type	Software
Programmer	Documenting the Now
Description	twarc twarc is a command line tool and Python library for archiving Twitter JSON data. It has separate commands (twarc and twarc2) for working with the older v1.1 API and the newer v2 API and Academic Access (respectively). It also has an ecosystem of plugins for doing things with the collected data.

URL https://github.com/DocNow/twarc

Umbra

Type Software

Programmer Internet Archive

Description
Umbra
A queue-controlled browser automation tool for improving web crawl quality. Umbra is a browser automation tool, developed for the web archiving service https://archive-it.org/. Umbra receives urls via AMQP. It opens them in the chrome or chromium browser, with which it communicates using the chrome remote debug protocol. It runs javascript behaviors to simulate user interaction with the page. It publishes information about the the urls requested by the browser back to AMQP. The format of the incoming and outgoing AMQP messages is described in pydoc umbra.controller.

URL https://github.com/internetarchive/umbra

Uniform Resource Identifier

Type Encyclopedia Article

Author Wikipedia

Description
A Uniform Resource Identifier (URI) is a unique sequence of characters that identifies a logical or physical resource used by web technologies. URIs may be used to identify anything, including real-world objects, such as people and places, concepts, or information resources such as web pages and books. Some URIs provide a means of locating and retrieving information resources on a network (either on the Internet or on another private network, such as a computer filesystem or an Intranet); these are Uniform Resource Locators (URLs). A URL provides the location of the resource. A URI identifies the resource by name at the specified location or URL. Other URIs provide only a unique name, without a means of locating or retrieving the resource or information about it, these are Uniform Resource Names (URNs). The web technologies that use URIs are not limited to web browsers. URIs are used to identify anything described using the Resource Description Framework (RDF), for example, concepts that are part of an ontology defined using the Web Ontology Language (OWL), and people who are described using the Friend of a Friend vocabulary would each have an individual URI.

URL https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

Uniform Resource Name

Type	Encyclopedia Article
Author	Wikipedia
Description	A Uniform Resource Name (URN) is a Uniform Resource Identifier (URI) that uses the urn scheme. URNs are globally unique persistent identifiers assigned within defined namespaces so they will be available for a long period of time, even after the resource which they identify ceases to exist or becomes unavailable. URNs cannot be used to directly locate an item and need not be resolvable, as they are simply templates that another parser may use to find an item.
URL	https://en.wikipedia.org/wiki/Uniform_Resource_Name

UXTR: Universal Links Extractor

Type	Web Page
Author	Aleph Archives
Description	UXTR is a cross-platform Erlang based RESTful Web Service that lets you extract links from any webpage. Such a service is mainly used to develop Web Crawlers backed on High-Quality Web Archives.
URL	http://webarchivingbucket.com/uxtr/doc/

Virus check

Type	Document
Author	Library of Congress (PREMIS)
Description	virus check (also malware check) The process of scanning a file for malicious programs.
URL	https://www.loc.gov/standards/premis/v3/preservation-events.pdf

W3Act

Type	Software
Programmer	UK Web Archive
Description	w3act is an annotation and curation tool for building web archive collections
URL	https://github.com/ukwa/w3act

WAIL - Web Archiving Integration Layer

Type	Web Page
Author	Matt Kelly
Description	Web Archiving Integration Layer (WAIL) is a desktop application that provides a graphical user interface (GUI) atop multiple pre-configured web archiving tools. WAIL acts as an easy way for anyone to preserve and replay web pages. WAIL includes Heritrix 3.2.0 for web crawling and OpenWayback 2.4.0 for replaying web archives. Both these tools and others are accessible from an easy-to-use, native system interface. WAIL is written in Python and compiled to a native application using PyInstaller.
URL	http://machawk1.github.io/wail/

WARC (Web ARChive) file format

Type	Web Page
Author	The National Archives/PRONOM
Description	The WARC (Web ARChive) file format offers a convention for concatenating multiple resource records (data objects), each consisting of a set of simple text headers and an arbitrary data block into one long file. The WARC format is an extension of the ARC file format (ARC) that has traditionally been used to store "web crawls" as sequences of content blocks harvested from the World Wide Web (…) Besides the primary content recorded in ARCs, the extended WARC format accommodates related secondary content, such as assigned metadata, abbreviated duplicate detection events, later-date transformations, and segmentation of large resources". WARC format has been written by the members of the IIPC (http://www.netpreserve.org/) grouped within the ISO/TC46/SC4/WG12.
URL	http://www.nationalarchives.gov.uk/pronom/fmt/289

Wayback CDX Server API - BETA

Type	Software
Programmer	Internet Archive
Description	The wayback-cdx-server is a standalone HTTP servlet that serves the index that the wayback machine uses to lookup captures. The index format is known as 'cdx' and contains various fields representing the capture, usually sorted by url and date, http://archive.org/web/researcher/cdx_file_format.php The server responds to GET queries and returns either the plain text CDX data, or optionally a JSON array of the CDX.The CDX server is deployed as part of

	web.archive.org Wayback Machine and the usage below reference this deployment. However, the cdx server is freely available with the rest of the open-source wayback machine software in this repository. Further documentation will focus on configuration and deployment in other environments.
URL	https://github.com/internetarchive/wayback/blob/a331f968c2a8b3ef56b82dc3a9faed11900e84b2/wayback-cdx-server/README.md

Wayback Machine

Type	Encyclopedia Article
Author	Wikipedia
Description	The Wayback Machine is a digital archive of the World Wide Web and other information on the Internet created by the Internet Archive, a non-profit organization, based in San Francisco, California. It was set up by Brewster Kahle and Bruce Gilliat, and is maintained with content from Alexa Internet. The service enables users to see archived versions of web pages across time, which the Archive calls a "three dimensional index." Since 1996, they have been archiving cached pages of web sites onto their large cluster of Linux nodes. They revisit sites every few weeks or months and archive a new version if the content has changed. The intent is to capture and archive content that would otherwise be lost whenever a site is changed or closed down. Their grand vision is to archive the entire Internet. The name Wayback Machine was chosen as a droll reference to a plot device in an animated cartoon series, The Rocky and Bullwinkle Show. In one of that animated cartoon's component segments, Peabody's Improbable History, lead characters Mr. Peabody and Sherman routinely used a time machine called the "WABAC machine" (pronounced wayback) to witness, participate in, and, more often than not, alter famous events in history.
URL	https://en.wikipedia.org/wiki/Wayback_Machine

waybackpy

Type	Web Page
Author	Akash Mahanty
Description	Waybackpy A Python package that interfaces with the Internet Archive's Wayback Machine API. Archive pages and retrieve archived pages easily.
URL	https://akamhy.github.io/waybackpy/

Web archiving

Type	Encyclopedia Article
Author	Wikipedia
Description	Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public. Web archivists typically employ web crawlers for automated capture due to the massive size and amount of information on the Web. The largest web archiving organization based on a bulk crawling approach is the Wayback Machine, which strives to maintain an archive of the entire Web. The growing portion of human culture created and recorded on the web makes it inevitable that more and more libraries and archives will have to face the challenges of web archiving. National libraries, national archives and various consortia of organizations are also involved in archiving culturally important Web content. Commercial web archiving software and services are also available to organizations who need to archive their own web content for corporate heritage, regulatory, or legal purposes.
URL	https://en.wikipedia.org/wiki/Web_archiving

Web Archiving Service

Type	Web Page
Author	Archive-It Help Center/ Maria Praetzellis
Description	Web Archiving Service Enables curators to build collections of web-published materials that are stored in either local and/or remote repositories. The service includes a set of tools for selection, curation, and preservation of the archives. It also includes repositories for storage, preservation services (e.g., replication, emulation, and persistent naming), and administrative services (e.g., templates for collection strategies, content provider agreements, repository provider agreements). Archive-It is a web archiving service.
URL	https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms

Web browser

Type	Encyclopedia Article
Author	Wikipedia
Description	A web browser (also referred to as an Internet browser or simply a browser) is application software for accessing the World Wide Web or a local website. When a user requests a web page from a particular website, the web browser

	retrieves the necessary content from a web server and then displays the page on the user's device. A web browser is not the same thing as a search engine, though the two are often confused. A search engine is a website that provides links to other websites. However, to connect to a website's server and display its web pages, a user must have a web browser installed. Web browsers are used on a range of devices, including desktops, laptops, tablets, and smartphones. In 2020, an estimated 4.9 billion people used a browser. The most used browser is Google Chrome, with a 63% global market share on all devices, followed by Safari with 20%.
URL	https://en.wikipedia.org/wiki/Web_browser

Web crawler

Type	Encyclopedia Article
Author	Wikipedia
Description	A Web crawler, sometimes called a spider or spiderbot and often shortened to crawler, is an Internet bot that systematically browses the World Wide Web, typically operated by search engines for the purpose of Web indexing (web spidering).Web search engines and some other websites use Web crawling or spidering software to update their web content or indices of other sites' web content. Web crawlers copy pages for processing by a search engine, which indexes the downloaded pages so that users can search more efficiently. Crawlers consume resources on visited systems and often visit sites without approval. Issues of schedule, load, and "politeness" come into play when large collections of pages are accessed. Mechanisms exist for public sites not wishing to be crawled to make this known to the crawling agent. For example, including a robots.txt file can request bots to index only parts of a website, or nothing at all. The number of Internet pages is extremely large; even the largest crawlers fall short of making a complete index. For this reason, search engines struggled to give relevant search results in the early years of the World Wide Web, before 2000. Today, relevant results are given almost instantly. Crawlers can validate hyperlinks and HTML code. They can also be used for web scraping and data-driven programming.
URL	https://en.wikipedia.org/wiki/Web_crawler

Web Curator Tool (WCT)

Type	Web Page
Author	National Library of the Netherlands
Author	National Library of New Zealand

Description: The Web Curator Tool (WCT) is a tool for managing the selective web harvesting process, and is designed for use in libraries by non-technical users. It is integrated with v3 of the Heritrix web crawler which is used to download web material (but technical details are handled behind the scenes by system administrators).

URL: https://webcuratortool.org

Web scraping

Type: Encyclopedia Article

Author: Wikipedia

Description: Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis. Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet or loaded into a database. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else.

URL: https://en.wikipedia.org/wiki/Web_scraping

Web server

Type: Encyclopedia Article

Author: Wikipedia

Description: A web server is computer software and underlying hardware that accepts requests via HTTP (the network protocol created to distribute web content) or its secure variant HTTPS. A user agent, commonly a web browser or web crawler, initiates communication by making a request for a web page or other resource using HTTP, and the server responds with the content of that resource or an error message. A web server can also accept and store resources sent from the user agent if configured to do so. The hardware used to run a web server can vary according to the volume of requests that it needs to handle. At the low end of the range are embedded systems, such as a router

	that runs a small web server as its configuration interface. A high-traffic Internet website might handle requests with hundreds of servers that run on racks of high-speed computers. A resource sent from a web server can be a preexisting file (static content) available to the web server, or it can be generated at the time of the request (dynamic content) by another program that communicates with the server software. The former usually can be served faster and can be more easily cached for repeated requests, while the latter supports a broader range of applications.
URL	https://en.wikipedia.org/wiki/Web_server

WebSnapperPro

Type	Web Page
Author	NetLab
Description	Web Snapper is a tool for capturing web pages (including subpages) as they appear in the browser. Pages can be stored in a variety of file formats.
URL	https://www.netlab.dk/services/tools-and-tutorials/websnapperpro/

Wget

Type	Encyclopedia Article
Author	Wikipedia
Description	GNU Wget (or just Wget, formerly Geturl, also written as its package name, wget) is a computer program that retrieves content from web servers. It is part of the GNU Project. Its name derives from "World Wide Web" and "get." It supports downloading via HTTP, HTTPS, and FTP. Its features include recursive download, conversion of links for offline viewing of local HTML, and support for proxies. It appeared in 1996, coinciding with the boom of popularity of the Web, causing its wide use among Unix users and distribution with most major Linux distributions. Written in portable C, Wget can be easily installed on any Unix-like system.
URL	https://en.wikipedia.org/wiki/Wget

XML

Type	Encyclopedia Article
Author	Wikipedia
Description	Extensible Markup Language (XML) is a markup language and file format for storing, transmitting, and reconstructing arbitrary data. It defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The World Wide Web Consortium's XML 1.0 Specification

	of 1998 and several other related specifications—all of them free open standards—define XML. The design goals of XML emphasize simplicity, generality, and usability across the Internet. It is a textual data format with strong support via Unicode for different human languages. Although the design of XML focuses on documents, the language is widely used for the representation of arbitrary data structures such as those used in web services. Several schema systems exist to aid in the definition of XML-based languages, while programmers have developed many application programming interfaces (APIs) to aid the processing of XML data.
URL	https://en.wikipedia.org/wiki/XML

youtube-dl

Type	Encyclopedia Article
Author	Wikipedia
Description	youtube-dl is a free and open source download manager for video and audio from YouTube and over 1,000 other video hosting websites. It is released under the Unlicense software license. As of September 2021, youtube-dl is one of the most starred projects on GitHub, with over 100k stars. According to libraries.io, 308 other packages and 1.43k repositories depend on it. Numerous forks exist of the project, including yt-dlp, with over 24k stars.
URL	https://en.wikipedia.org/wiki/Youtube-dl

Zotero

Type	Encyclopedia Article
Author	Wikipedia
Description	Zotero is a free and open-source reference management software to manage bibliographic data and related research materials (such as PDF files). Notable features include web browser integration, online syncing, generation of in-text citations, footnotes, and bibliographies, as well as integration with the word processors Microsoft Word, LibreOffice Writer, and Google Docs. It is developed as a project by the non-profit Corporation for Digital Scholarship. It was originally created at the Center for History and New Media at George Mason University.
URL	https://en.wikipedia.org/wiki/Zotero

REFERENCES

Archive-It Help Center, & Praetzellis, M. (n.d.). Glossary of Archive-It and Web Archiving Terms [Web page]. Glossary of Archive-It and Web Archiving Terms. Retrieved 2021-05-06, from https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms. [URL Memento: Wayback Machine]

Bragg, M., & Hanna, K. (2013). The Web Archiving Lifecycle Model [White Paper]. The Archive-It Team, Internet Archive. Retrieved 2021-10-07, from http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf. [URL Memento: Wayback Machine]

Brügger, N. (2020). Welcome to WARCnet. *WARCnet Papers*. Aarhus, Denmark: WARCnet. Retrieved 2021-01-27, from https://cc.au.dk/fileadmin/user_upload/WARCnet/1.Bru__gger_Welcome_to_WARCnet.pdf. [URL Memento: Wayback Machine]

Digital Preservation Coalition. (n.d.). Glossary—Digital Preservation Handbook [Web page]. Digital Preservation Coalition. Retrieved 2021-11-2, from https://www.dpconline.org/handbook/glossary. [URL Memento: Wayback Machine]

GitHub. (n.d.). GitHub: Where the world builds software [Website]. GitHub. Retrieved 2021-10-22, from https://github.com/. [URL Memento: Wayback Machine]

Healy, S.; Byrne, H.; Schmid, K.; Bingham, N.; Holownia, O.; Kurzmeier, M.; Jansma, R. (2022) *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Special Reports, September 2022. Aarhus, Denmark: WARCnet. Retrieved 2022-10-02, from https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf.

Healy, S., Byrne, H., Schmid, K., Floody, L., Boté-Vericad, J.-J. (2021+) Zotero Groups - Towards a Glossary for Web Archive Research. Zotero. Retrieved from https://www.zotero.org/groups/4380600/towards_a_glossary_for_web_archive_research.

Library of Congress (n.d). Glossary - Web Archiving. Library of Congress. Retrieved 2021-05-14, from https://www.loc.gov/programs/web-archiving/about-this-program/glossary/. [URL Memento: Wayback Machine]

Library of Congress (2017). PREMIS: Preservation Events Controlled Vocabulary. PREMIS: Preservation Metadata Maintenance Activity. Retrieved 2021-10-09, from https://www.loc.gov/standards/premis/v3/preservation-events.pdf. [URL Memento: Wayback Machine]

Pediaa. (2021, September 23). What is the Difference Between Glossary and Dictionary [Web page]. Pediaa.Com. Retrieved 2022-06-02, from https://pediaa.com/what-is-the-difference-between-glossary-and-dictionary/. [URL Memento: Wayback Machine]

UNESCO Thesaurus. (n.d.). About UNESCO Thesaurus [Web page]. UNESCO. Retrieved 2021-06-21, from http://vocabularies.unesco.org/browser/en/about. [URL Memento: Wayback Machine]

WARCnet. (n.d.). About WARCnet – WARCnet [Web page]. Aarhus University. Retrieved 2021-04-23, from https://cc.au.dk/en/warcnet/about/. [URL Memento: Wayback Machine]

WARCnet. (n.d.). London Spring 2022 [Web page]. Aarhus University. Retrieved 2022-08-04, from https://cc.au.dk/en/warcnet/meetings/london-2022. [URL Memento: archive.today]

WARCnet. (n.d.). Working Groups. [Web page]. Aarhus University. Retrieved 2022-05-07, from https://cc.au.dk/en/warcnet/working-groups. [URL Memento: Wayback Machine]

Wikidata. (n.d.). Welcome to Wikidata [Wiki | MediaWiki]. Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page. [URL Memento: Wayback Machine]

Wikipedia. (2002). Welcome to Wikipedia. In Wikipedia (Online/web). Wikipedia. https://en.wikipedia.org/wiki/Main_Page. [URL Memento: Wayback Machine]

Zotero & Corporation for Digital Scholarship. (n.d.). Zotero – About [Web page]. Zotero. Retrieved 2021-08-15, from https://www.zotero.org/about/. [URL Memento: Wayback Machine]

Zotero & Corporation for Digital Scholarship. (n.d.). Zotero – Downloads [Web page]. Zotero. Retrieved 2021-02-18, from https://www.zotero.org/download/. [URL Memento: Wayback Machine]

Zotero & Corporation for Digital Scholarship. (n.d.). Zotero - Home [Website]. Zotero. Retrieved 2021-12-18, from https://www.zotero.org. [URL Memento: Wayback Machine]

Zotero & Corporation for Digital Scholarship. (n.d.) Zotero – Item Types and Fields [Web page]. Zotero. Retrieved 2021-01-06, from http://zotero.org/support/quick_start_guide. [URL Memento: Wayback Machine]

Zotero & Corporation for Digital Scholarship. (n.d.). Zotero – Quick Start Guide [Web page]. Zotero. Retrieved 2021-01-06, from http://zotero.org/support/quick_start_guide. [URL Memento: Wayback Machine]

WARCnet
web archive studies

WARCnet *Special Reports* is a series of reports related to the activities of the WARCnet network. To ensure the relevance of the publications, WARCnet strives to publish with a rapid turnover. WARCnet Special Reports are edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster and Michael Kurzmeier. In cases where a WARCnet Special Report has gone through a process of single blind review, this is mentioned in the individual publication.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-23, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).