# **Structured Internet Researches** / by Albrecht Ude

A short résumé

Dataharvest / European Investigative Journalism Conference, 03.06.2023

**The Internet**

The internet is decentralised and open to all types of computers.
There is no central computer.
Hardware and software from all manufacturers can participate as long as they can handle the IP protocol.
The IP protocol regulates how devices and programmes exchange data.
The standards of the internet are open.

The internet was not developed in Silicon Valley, nor by a company.
The World Wide Web was developed in Geneva, Switzerland, in Europe. It was not developed by a company, but at CERN (Conseil européen pour la recherche nucléaire = European Organization for Nuclear Research), a scientific institution.

Google, Amazon, Facebook and many more, and even more shareholders benefit from the performance of these pioneers (this should not be forgotten!).

## Five Areas of Internet Research

| **Area** | **You need to access** |
|---|---|
| Surface Web | Search Engines |
| Deep Web | Databases |
| Social Web | Accounts |
| Dark Web (aka Darknet) | Anonymizing Software |
| Internet ≠ World Wide Web | Expert knowledge |

## What Search Engines really do

Search engines copy a small part of the Internet, more precisely the "surface web".
They are databases that show what is where on the Internet.
The operators of the databases decide what they show us.
Some search engines evaluate who is looking for what.

The Search Engine Result Page (SERP) is a hit list of a search engine influenced by:

- The ranking algorithms of the search engine (SE)
- Information that the SE has about the searcher (personalization)
- SEO (Search Engine Optimization) from third parties

SERPs are battlefields with a lot of money flowing into their contents.
SERPs are never "objective" but always arbitrary!

Google Ranking Factors
**https://backlinko.com/google-ranking-factors**

## Search engines that journalists should know about

The major search engines are Google (USA), Bing (USA), Yandex (Russia) and Baidu (PR China).
All with their own Index, showing the contents of the surface web. All with resonable results, none perfect, all spying on us!

**http://google.com/** ,
**http://bing.com/** ,
**http://yandex.com/** ,
**http://baidu.com/** .

A Google-Clone, not spying, is Startpage:
**http://startpage.com/** .

Bing Clones, not spying, are Duckduckgo, Ecosia, Qwant and Yahoo
**http://duckduckgo.com/** ,
**http://ecosia.org/** ,
**http://qwant.com/** ,
**http://yahoo.com/** .

New Competitors, both working with AI: You and Kagi.
**https://you.com** ,
**https://kagi.com** .

A special search engine for scientific information: WolframAlpha.
**https://www.wolframalpha.com/** .

## Basic Operators: Commands for Search Engines

„Boolean Operators“, correctly interpreted by almost all Search Engines:

| | | |
|---|---|---|
| **+** | AND | and |
| OR | | or (has to be typed in capital letters) |
| **-** | NOT | not |
| ***** | | something missing (single character, word or phrase) |
| **" "** | | exact Phrase |

Be informed, how search engines get their data, how they index them and how they rank results!
Use the Operators AND "+" , "OR", NOT "-" , the Wildcard "*" and the quotation marks " " for phrase search.
Use the Advanced Operators, especially

| | |
|---|---|
| site | returns results from specified Domain (watch for the correct Domain/s) |
| filetype | returns specified filetypes only |
| intitle | searches only in <title>-Tag |
| inurl | saerches only in web address |
| intext | searches only in text of a webpage |
| inanchor | saerches only in links |
| ip | searches specified IP-adress (BING only) |

Use Search Engines parallel - a site one misses, the other one will find. Also, rankings differ.

## What's new in online research?

"Online research is finding, saving, and analyzing files from the Internet."

Files from the internet always have got:

- Content (“internal” Data)
- Hidden contents within the file (Metadata)
- Context (“external” Data, e.g. URL)

### Information about file formats

http://filext.com/ ,
http://mark0.net/soft-trid-deflist.html ,
https://en.wikipedia.org/wiki/List_of_file_formats ,
https://en.wikipedia.org/wiki/List_of_filename_extensions .

### Information about picture files

Image Operations meta tool
A bulging toolbox with links to search engines for reverse image search, metadata analysis, image editing and much more.
http://imgops.com/

## Wikipedia basics

Wikipedia is not a source. PERIOD.

At least three clicks :

1. the Article itself
2. the Talk
3. the History

Wikipedia can *lead* to sources: Look for the external links and the references under the articles.

Wikipedia as a navigation tool:
Categories: https://en.wikipedia.org/wiki/Category:Contents
Portals: https://en.wikipedia.org/wiki/Portal:Contents ,
https://en.wikipedia.org/wiki/Portal:Contents/Portals
Lists: https://en.wikipedia.org/wiki/Category:Lists
Search: https://en.wikipedia.org/w/index.php?title=Special:Search&profile=advanced&search=&fulltext=1

Wikimedia Commons is a rich treasure trove of video, audio and visual material.:
https://commons.wikimedia.org/wiki/Main_Page ,
https://commons.wikimedia.org/wiki/Atlas ,
https://commons.wikimedia.org/wiki/Historical_atlas

Important pages for analysing Wikipedia content
http://en.wiki-watch.de/ ,
http://www.wikigen.org/ ,
https://www.wikishark.com/ ,
https://xtools.wmflabs.org/?uselang=en .

## The „Deep Web“ — Databases

The contents of databases cannot be found in search engines. Therefore:

It‘s two different operations
1. search **for** the database
2. search **in** the database

As long as you search **for** the database, you must hide what you want to search for **in the database**!

Databases can be found by asking the following questions:
- Ask: Who is suitable to run the database (it needs money, work, and time)?
- Run SearchEngine-Queries:
  [search term] + database OR directory OR catalogue OR list
- Check Wikipedia:
  [article on the topic]: see External Links and References for databases mentioned
- Also look for Categories and Lists.
- Check DBIS (only in German, sorry, but remarkable)
  **https://dbis.ur.de** .

## Archives of the World Wide Web

**https://web.archive.org/** ,
**http://archive.is/** ,
**https://en.wikipedia.org/wiki/List of Web archiving initiatives** ,
**https://en.unesco.org/themes/information-preservation** .

### Some databases, you have got to know

Catalogue of Research Databases / Investigative Dashboard
**https://investigativedashboard.org/databases/**

TED – Tenders European Daily – what do public institutions spend money on?
**https://ted.europa.eu/**

IATE - The EU's multilingual term base
**http://iate.europa.eu/**

### Company registers

EU: Business registers in Member States
**https://e-justice.europa.eu/content business registers in member states-106-en.do**

Company Register (Global Open Data Index / Open Knowledge Foundation)
**https://index.okfn.org/dataset/companies/**

Open Corporates - The Open Database Of The Corporate World
**https://opencorporates.com/registers**

**National and international statistical services**

National Statistical Offices / World Bank
http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/SCBEXTERNAL/0,,contentMDK:20445729~menuPK:2740285~pagePK:229544~piPK:229605~theSitePK:239427,00.html

OECD.Stat Statistical Sources
https://stats.oecd.org/

UN Statistics Division: National Statistical Offices Websites
https://unstats.un.org/home/nso_sites/

**Libraries and Archives**

International Federation of Library Associations and Institutions (IFLA)
https://www.ifla.org/

List of national and state libraries / Wikipedia
https://en.wikipedia.org/wiki/List_of_national_and_state_libraries

List of national archives / Wikipedia
https://en.wikipedia.org/wiki/List_of_national_archives

Archives Portal Europe
http://www.archivesportaleurope.net/home

WorldCat
is the world's largest bibliographic database with 1.5 billion records
Inventory records in 450 languages.
http://www.worldcat.org/
Search Engine for persons
https://www.worldcat.org/identities/

Europeana
Europe's largest online collection of art, culture and science. Combines the digital Collections of the institutions mentioned in the source list.
https://www.europeana.eu/portal/
Source List
https://www.europeana.eu/portal/en/explore/sources.html

Directory of Open Access Journals (DOAJ)
https://doaj.org/

The Online Books Page
a service offered by the University of Pennsylvania, gives readers access to more than two million books freely accessible (and readable) on the internet. Users also gain access to significant directories and archives of online texts, as well as special exhibits of particularly interesting classes of online books.
http://digital.library.upenn.edu/books/search.html

## Social Networks

There are always three types of search options for social networks:

- the search options of the network itself (look for the name of the network and “advanced search” in a search engine)
- Search engines with the operator **`site:`**
- external search tools from third party programmers (look for the name of the network and “research tools” in a search engine). Many good analyzing tools are provided by marketers.

Two overviews of social networks in the English Wikipedia:
**https://en.wikipedia.org/wiki/Social_networking_service#External_links** ,
**https://en.wikipedia.org/wiki/Category:Social_networking_services**

World Map of Social Networks
most used and second-most used social networks
**http://vincos.it/world-map-of-social-networks/**

Social Media flower visualised as flowers
**https://www.researchgate.net/profile/Christine_Noonan/publication/275958841/figure/fig13/AS:614090980487189@1523422107916/German-conversations-in-social-media-comScore-2013.png** ,
**https://www.visualcapitalist.com/visual-map-social-media-universe/**

Social Media Image Sizes Cheat Sheet
**https://blog.hootsuite.com/social-media-image-sizes-guide/**

### Links for safety-conscious work

Directory of recommended, free software
**https://prism-break.org/en/all/**

The TOR browser for anonymous surfing and access to the TOR network (Dark Web)
**https://www.torproject.org/**

Information about whether you are affected by Identity theft
**https://haveibeenpwned.com/** ,
**https://sec.hpi.de/ilc/**