Researching the Deep Web

Finding and Using Databases

Albrecht Ude

Albrecht Ude

Journalist | Researcher | Trainer

http://www.ude.de/.

http://eine-woche-ohne.de/.

http://www.netzwerk-recherche.de/newsletter/.
„Deep Web“

Surface Web (4% of WWW content)

Deep Web (96% of WWW content)
„Deep Web“ vs „Surface Web“

“Surface Web”

- everything on WWW you can reach only by clicking a mouse
- researchable through Search Engines (SEs)

“Deep Web”

- on the Web, reachable with a keyboard
- NOT in the range of SEs
„Deep Web“ means „Hidden Web“

The picture is wrong: The deep web is not a monolithic block

But rather a quarry, consisting of
✓ millions of databases
✓ closed web forums
✓ password-protected sites

The contents of “hidden web“ sites is not visible for search engines
„Deep Web“ in Search Engine Results

Search Engine Result Pages

✓ may contain the start pages of databases,
✓ but not their contents

You need totally different search strategies!

✓ don‘t look for your topic,
✓ think about a suitable search tool or database
„Deep Web“ research – an example

An e-mail from August, 2011:

Dear Albrecht,

My good colleague [XXX] said that you might be able to help me: I am looking for a phone number or email for Volker Hauth; a German architect/engineer who studied at the Technical University of Hamburg-Harburg together with Mohamed Atta - the ringleader from the 11. september-attack. I can't find a number or info for Hauth on the internet, so I was hoping that you might know where to look. Hope to hear from you and thanks.
„Deep Web“ research – an example (2)

Googling (or taking any other search engine):

✓ Volker Hauth email
✓ Volker Hauth phone number („Telefonnummer“ in German)
✓ Volker Hauth kontakt

The SEs result pages are full of „hits“...

… but you‘re only wasting your time!
„Deep Web“ research – an example (3)

Don’t look for information on Volker Hauth. Look for information on **architects** in **Hamburg**!

As long as you’re still looking for the database **forget**, what you’re going to look up in the database!

This is really difficult!
Strategies for finding databases (1)

✓ Who should run the database?
  (no Search Engine will help solving this question!)
Strategies for finding databases (2)

✓ Search Engine queries

Look for the category of your topic combined with

database OR directory OR catalogue OR list

in English,
in your language,
in the language that suits the topic
Strategies for finding databases (3)

✓ Weblinks in the Wikipedia

Look for thematic articles in Wikipedia,

check the section „external references“

Do so in English,
in your language,
in the language that suits the topic
Strategies for finding databases (4)

✓ Categories of Wikipedia

Some categories (keywords) in Wikipedia lead to databases:

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

   Category:Databases,
   Category:Digital_libraries,
   Category:Scholarly_databases,
   Category:Scholarly_search_services,
   Category:Bibliographic_databases_and_indexes,
   Category:Scientific_databases..
Strategies for finding databases (5)

✓ Lists in Wikipedia

Finally, you can find lists of databases in the Wikipedia:

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

...List_of_academic_databases_and_search_engines,
...List_of_online_databases..
Strategies for finding databases (6)

✓ Databases of databases

http://www.bibliothek.uni-regensburg.de/dbinfo/

The Database Information System (DBIS) enables the use of 11,554 scientific databases (4,754 of which are free on the Internet).

For Lists in English, check Search Engines looking for "a * z database"
Databases you MUST know (1)

The WayBackMachine

https://archive.org/.

and the other Archive

https://archive.is/.
Databases you MUST know (2)

IANA Root Zone Database

the source of Whois-data

https://www.iana.org/domains/root/db
Databases you MUST know (3)

WorldCat – the biggest library catalog on earth

http://www.worldcat.org/

http://www.worldcat.org/identities/
Databases you MUST know (4)

Common Vulnerabilities and Exposures (CVE)
The One and Only Database on Security Issues

https://cve.mitre.org/,
Databases you should know (1)

TED – Tenders Electronic Daily
where the EU spends her money

Databases you should know (2)

Directory of Open Access Journals (DOAJ)
Scientific Information for free

https://doaj.org/
Databases you should know (3)

Your national library –

pick it up at Wikipedia:

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

List_of_national_and_state_libraries.
Databases you should know (4)

A Database on all Streetnames in Germany
made by Zeit online,
based on the data of Openstreetmap

https://www.zeit.de/interactive/strassennamen/.

Please do it in YOUR country!
Databases you *could* know

... hopefully, you know how to find them!
Thank you for your patience!

Albrecht Ude
albrecht@ude.de
www.ude.de