The Web is Lovely, Dark & Deep

A shallow dive beneath the surface

Suzanne Hunt, Scott Davis, Race MoChridhe, Katherine Bauer
Comin’ atcha!

- History
- Brief semantic overview
- Good v. Evil
- Pros and cons to anonymity for public libraries
- Implications for academic libraries
A Brief History of the Deep Dark Web

1960s
Cold-war paranoia runs rampant; US DoD funds creation of ARPANET (Advanced Research Projects Agency Network)
Packet switching = one network, many computers communicate

1970s
Vinton Cerf develops TCP/IP (Transmission Control Protocol and Internet Protocol)
= multiple networks, many computers communicate.
American mathematicians publish the “public-key encryption” = keeps eavesdroppers from deciphering coded online messages.

1980s
(D)ARPANET adopts TCP/IP = “network of networks” assembly begins.
American cryptographer devises a way to keep eavesdroppers from reading email location, destination, and time information.
Bulletin Board Systems (BBS) originate as computers enter many people’s homes.
User groups form as a way for BBS users to communicate.
Outsiders (rebels?) form their own “anti-user” hacker groups and created places to meet online.
Global computer network prepares for ISPs to open the network to the public

1990
Tim Berners-Lee invents the World Wide Web = the Internet; rather the most common means of accessing data online in the form of websites and hyperlinks.
Deep Web contains mostly legal and legit information.

- Contents of databases and other web services that cannot be indexed by standard search engines, e.g.
  - Articles in journal/magazine/newspaper databases
  - Statistical information: weather data, population, and city demographics found in databases
  - Public records on people and businesses
  - Reference materials - dictionaries and thesauri
  - Resources requiring passwords or credentials
  - Online library catalogs
  - Information behind firewalls (company intranets and extranets) and paywalls
  - Subscription databases
  - Sites that purposely exclude some or all of their content
  - Pages purposely excluded by search engine companies to avoid cluttering their results list
  - Technically limited content (CAPTCHA)
  - Text content outside of conventional http:// or https:// protocols
2010: The Farmer’s Market, one of the largest drug markets on the non-anonymized Internet, moves to the Tor network

2011: Iranian government blocks/slows traffic for various anonymizing technologies during protests over a controversial election

2012: Foreign Policy magazine honors Tor developers for protecting the rights of people “who might be silenced online” (such as corporate whistleblowers).

2013: FBI shuts down 1) Freedom Hosting, a large Dark Web hosting service, and charges its operator with dealing in child pornography, and 2) Tor-based marketplace Silk Road, arrests its operator on multiple drug trafficking-related charges (convicted and sentenced to life imprisonment in 2015).

2014: Dark Web illegal drug sales increase as new markets open…British researcher finds that 4 of 5 visits to Tor hidden-service websites were to child-pornography sites

Facebook, longtime advocate of using only real-world identities online, opens a TOR hidden-services site to help people reach Facebook from countries that block it.

2015: In an international survey, 4% of Dark Web drug buyers said they had not consumed illegal drugs before their Dark Web purchase
1995: U.S. Naval Research Laboratory begins developing anonymizing browser software based on an idea called “onion routing” mimicking an onion’s layers.

1999-2000: Irish computer science student Ian Clarke develops a method to keep Web surfers’ locations private; introduces Freenet to the public at no cost.

2001-2003: U.S. NRL enlists civilian computer scientists to help develop onion routing for military and civilian use; Tor, an acronym for the onion router, becomes publicly available for free.

2008: China blocks Tor website to keep residents from using the software to access censored information.

2009: Bitcoin introduced; Drugstore Market becomes the first online marketplace for illegal drug sales, established as a “Tor hidden service.”
The Dark Web is good!

- Anonymous social interaction
- Refuge for those seeking privacy and/or have non-mainstream views
- Keeps cyberspies at bay
- Safe for domestic abuse victims whose abusers are stalking them
- Use of Tor browser thwarts tracking pseudonyms used in online transgender forums to avoid job discrimination if posts are discovered
- Refugees and foreign citizens living in the U.S use Tor to avoid being tracked when they read online political news that’s censored in their home country.
- Handy, free, off-the-shelf tool for intelligence agents.
- Bitcoin* can be used to send money to friends and relatives in hostile countries.

The Dark Web is evil!

Accessing the Dark Net is not in itself illegal. However, one can anonymously buy, sell, trade, or hire illegal products or services (usually using Bitcoin*):

- Light drugs (cannabis is the most traded), hard drugs, and pharmaceuticals such as Ritalin and Xanax, and, most recently opioids.
- Hitmen (and, presumably, hitwomen) with rates based on desired manner of death and status of the target (from now-defunct AlphaBay et al)
- “Doxing” information, personally identifiable information such as birthdays, address, phone numbers, and emails can be attained
- Child exploitation, such as pornography and snuff films that feature children
- Guns, rifles, and other arms
- Hangout for neo-Nazi groups
1990-1996: the number of computers on the Internet grows from 313,000 to ~ 10 million.

1992: Cypherpunks announce a new project to develop technology to allow anyone to navigate the internet without revealing their locations, the websites they visited, or the content of their communications.
Anonymizing the Library: What can libraries do? What should they do?

ALAs public computer privacy checklist recognizes the right to online privacy, but stops short of endorsing TOR for browsing.

Alison Macrina is a librarian, privacy activist, and the founder of the Library Freedom Project, an initiative which aims to make real the promise of intellectual freedom in libraries and to inform librarians about online privacy tools and encourage them to volunteer their computers as Tor relays.

In January 2015 the Library Freedom Project received $244,700 in grant funding from the Knight Foundation and in January 2017 the Library Freedom Project was awarded a $249,504 grant from the Laura Bush 21st Century Librarian Program to facilitate the use of practical privacy tools in libraries using a "training the trainers" model. 40 geographically dispersed Privacy Advocates are expected to be trained in a six month course. New York University (NYU) and the Library Freedom Project have since created a formal collaborative program funded by the Institute of Museum and Library Services called Library Freedom Institute; its inaugural course is expected to begin in June 2018.

Tor Browser in Lebanon Public Library shut down by DHS in August, 2015, resumed one month later.

https://www.propublica.org/article/library-support-anonymous-Internet-browsing-effort-stops-after-dhs-email

Public Libraries: Staff and patrons using anonymous browsers

"Privacy? What's that, human?

Sounds like something bad!"
Anonymous Browsers

most widely used:

**Tor**

Can be downloaded and installed from: [www.torproject.org](http://www.torproject.org)

Other options include:
- Privoxy
- CyberGhost
- TunnelBear
- HotSpot Shield
- SRWare Iron
- Epic
- Comodo Dragon
Pros

Staff:

● Search results unbiased by previous, unrelated searches
● Creates awareness

Public:

● In line with libraries’ history of supporting privacy and individual freedoms
● PC management software may prevent installation for one-time use

Cons

Staff and Public:

● Fear and stigma
● “Nothing to hide” mindset
● Possible abuse - security, crime
● Slower browsing, uploads & downloads
● Video content can become unusable due to multiple layers of encryption
● Potential for strained relationships with law enforcement and government
● Dark web malware: often more of a threat than traditional virus protection software can handle
Best practices for accessing the deep web

1. Buy good Virtual Private Network (VPN) software, and use it **at all times**
2. Install and use the Tor browser (more than just a browser, it’s a network!)
3. Don’t install or enable any browser plugins (Flash, RealPlayer, etc.)
4. Use only HTTPS versions of websites
5. Avoid 3rd party mobile browsers that claim to use Tor. Officially, Tor supports only Windows, Mac and Linux.
How do people find their way around the deep web?

3 Options for navigating the deep web:

1. Know where you’re going (the URL - such as an organization’s intranet page)
2. Check online databases - compilations of Deep Web addresses
3. Use a deep web search engine

*Examples:* Start Page (Tor’s default search engine), DuckDuckGo, Pipl, and Google Scholar (for academic searches)
Possible Implementation Barriers

New or amended internet policies may be needed

Staff education and training

Buy-in from decision makers (IT, library board, public officials, law enforcement, etc.)

Logistics of set up and maintenance

Adequate privacy for patrons using public computers
Looking Ahead

Not on most people’s radar….yet!
- Low awareness
- No policies in place
- Few anecdotes of patron requests

May play an important part in planning for future library technology

Privacy concerns are likely to continue growing, fueled by government action, media, cyber crime

Patrons are becoming more tech savvy - are we ready for their questions about the deep/dark web?
Implications of the Deep Web for Academic Libraries

Race MoChridhe
The Backstory

The hard-coded HTML Web was what the name implied—nodes “threaded” together by hyperlinks.

Crawlers, just like their cutesy spider mascots, followed these “threads” to locate each of the nodes.

It is very easy for the crawler to “click” all the links on this page, but what about the things we will only see after we search?

What does a bot type in the box?
The Backstory (cont’d)

Databases are not webs.

Connections between the items don’t “exist” (for Web purposes) until a query is executed.

A crawler, which is in the business of following links, does not fill out forms or execute queries.

Hence, everything in the database is effectively invisible to our cartoon spider.
In academia, invisibility works both ways...

The deep web can make it **hard** for researchers **to locate** data they need, especially if:

- they work at small institutions (or have no institutional affiliation) and have limited database subscriptions.
- the data they need has been published in smaller, regional or language-specific databases.
- they are millennials who thought everything was on Google Scholar.

The deep web can make it **hard** for researchers **to share** their own work, limiting:

- the number of citations their work receives in the literature.
- the discoverability of their work to non-academic policy makers, social activists, etc. (especially important where research impact is now an important metric for hiring and tenure, as under the UK’s REF).
“We’ve got a really BIG problem!”
“Cat’s right. I can count backwards from five to calm down.”

Possible solutions:

5) Federated searching.

4) Expose data to the shallow Web.

3) Improve the crawlers.

2) Crawl the metadata.

1) Your idea...
5. Federated Searching (aka Meta-Searching)

**Pros:**
- Simultaneously queries multiple databases and integrates results into a single, ranked feed.
- Custom-tailored to the particularities of each database included.
- Consequently, can effectively integrate other systems for search aid (such as machine translation).

**Cons:**
- Generally proprietary, subscription systems.
- Must be custom-tailored to each included database; cannot automatically incorporate new, unknown resources.
- Comes with its own learning curve, and applications requiring maximal thoroughness (legal due diligence, plagiarism investigations, literature search for grant applications, etc.) may still require multiple systems to be queried.

All of which means...
DIALOG is back, baby!

(Laitinen, 2018, 6.)
4. Expose Data to the Shallow Web

Pros:

- Renders data discoverable to crawlers (including the specialized crawlers that power services like Google Scholar).
- Cuts out middlemen.

Cons:

- More difficult to control quality of metadata.
- Lacks the sophisticated querying tools of a database.
- Often conflicts with publishing agreements, and therefore puts research impact and publishing prestige in tension, or limits to pre-prints, etc.
3. Improve the Crawlers

**Pros:**

- Would allow data to remain in place in its databases while also permitting it to appear in shallow web searches.

**Cons:**

- Easier said than done.

“Among the findings of this study is that if the ... crawling process provides search results as script pages, the outcome only collects the first page... [and] it cannot collect deep webpages if the web browser object cannot launch the script, or if the web document contains script errors.” (Oh, 2018, 66).

- Also, fails to address data preservation issues raised by web applications (Xu, Esteva, Beck, & Hsieh, 2017).
2. Crawl the Metadata

mod_oai is a free, GPL Apache module developed at Old Dominion University to allow web crawlers to harvest data through OAI-PMH, in the form of a change log for the server’s contents.

While the initial version only exposed metadata, later releases provided limited options for exposing resources directly.

We (kind of, sort of) solved the problem (by accident)!
1. Your Idea

Though all these approaches have helped, information retrieval in the deep web remains a significant challenge, especially for ECRs seeking to get their work noticed.

New ideas are needed.
Questions?
Thank You

Laitinen, S. (2018). Online services — past and present in a Nordic and international perspective. In E. Mickos, T. Oker-Blom, M. Wallin, L. Klasén, A. Lamvik, & U. Retlev (Eds.), *Online development in the Nordic countries: A history of online information from the 1960s to the ’00s and NORDINFO’s role in its development* (pp. 1–11). Published by the editors.


https://www.slideshare.net/bpichman
https://rlc.radicallibrarianship.org/
www.libraryfreedomproject.org
https://boingboing.net/2014/09/13-radical-librarianship-how-nin.html
https://www.airsassociation.org/airs-articles/11-best-illegal-search-engines-to-browse-the-darknet
https://www.airsassociation.org/search-engines-list/
https://www.airsassociation.org/
https://www.dailydot.com/layer8/best-deep-web-search-engines/
http://www.ala.org/advocacy/privacy/checklists/public-access-computer
https://www.torproject.org/
https://www.atlasobscura.com/articles-radical-reference-collective
https://radicalreference.info/
https://www.airsassociation.org/airs-articles/seven-things-you-didn-t-know-about-the-deep-web
https://darkwebnews.com/tor-guide/
Image Credits

Webcrawler screenshot:

"I am totally freaking out!":
http://www.mamawritesreviews.com/files/images/original/4823.gif

Count backwards:

Surprised Austin Powers:
https://i.kym-cdn.com/photos/images/original/001/038/219/773.gif

Privacy?:
https://cheezburger.com/8557369600

Internetz.:
http://goelastic.com/were-hiring-digital-media-planner-buyer/

Tor network:

Deep Web infographic
http://bit.ly/2y8dQbm

"Banksy"- style librarian

“It’s so simple.”:
https://procreate.art/discussions/3/6/13654

Keep Calm:
http://www.keepcalmandposters.com/poster/2288045_keep_calm_and_be_a_trailblazer

Tor relay http://bit.ly/2C7KInt