INTERNATIONAL JOURNAL OF NATURAL AND APPLIED SCIENCES (IJNAS), VOL. 12, Special Edition (2019); P. 107 – 113, 2 TABLES, 3 FIGS.

Relevance of the Deep Web to Academic Research

E. E. Essien

ABSTRACT

The volume of information on the web is already vast and is increasing at a very fast rate. Most people access Web contents with Surface Web Search Engines; that is, only capable of accessing information in the web surface. The Deep Web in contrast, is a vast repository of web pages, usually generated by database-driven websites, that are available to web users; yet hidden from traditional search engines and consist of 95% of the entire Web. Making meaningful research on the Web must therefore include accessing the surface as well as the deep web. This paper seeks to expose the mysteries of the deep web by exploring various technology tools that will enable researchers to get maximum results with minimal search efforts.

INTRODUCTION

The Internet is the global interconnection of computers, which is supported by many major computer systems called web servers. The Web is the interlinking of electronic pages that run on these servers. Since the servers run the Web on the internet; the Web is called World Wide Web (WWW). The Web is broadly categorized in two; the Surface Web and the Deep Web. The deep web (*Hamilton, 2003) or* invisible web (*Devine, Egger-Sider, 2004)* or hidden web (*Raghavan et al, 2001)* is part of the World Wide Web whose contents are not indexed by standard web search engines. The opposite term to the deep web is the surface web, which is accessible to anyone using the Internet. A Computer scientist Michael K. Bergman is credited with coining the term deep web in 2001 as a search indexing term (*Wright, 2009).*

According to a study by Bright Planet [2005], the deep web is estimated to be up to 550 times larger than the surface web; accessible through traditional search engines and over 200,000 database-driven websites are affected. Sherman & Price [2001], estimates the amount of quality pages on the deep web to be 3 to 4 times more than those pages accessible through search engines like Google, About, Yahoo, etc. While the actual figures are debatable, they made it clear that the deep web is far bigger than the surface web, and is growing at a much faster pace. In a simplified description, the web consists of these two parts: the surface Web and the deep Web (invisible Web or hidden Web) but the deep Web came into public awareness with the publication of the landmark book by Sherman & Price [2001]. Since then, many books, papers and websites have emerged to help further explore this vast landscape and these needs to be brought to your notice too.

The internet and the visible web

The primary focus of this work is on the Web and more specific, the parts of the Web that search engines cannot see (known as the invisible Web) but in order to fully understand the phenomenon called the Invisible Web, it is important to first understand the fundamental differences between the Internet and the Web. Most people tend to use the words “Internet” and “Web” interchangeably, but they are not synonyms. The Internet is an interconnection of computers and networking protocol that allows computers of all types to connect to and communicate with other computers on the Internet. The Internet’s origin traced back to a project sponsored by the U.S. Defense Advanced Research Agency (DARPA) in 1969 as a means for researchers and defense contractors to share information. The Web, on the other hand, is a software protocol that allows users to easily access files stored on the Internet computers. The Web was created in 1990 by Tim Berners-Lee, a computer programmer working for the European Organization for Nuclear Research (CERN). Prior to the Web, accessing files on the Internet was a challenging task, requiring specialized knowledge and skills. The Web made it easy to retrieve a wide variety of files, including text, images, audio, and video by the simple mechanism of clicking a hypertext link.

Corresponding author**.** Email: essieneyo@unical.edu.ng,essieneyo@gmail.com
Department of Computer Science, University of Calabar, Nigeria.

Hypertext is a system that allows computerized objects (text, images, sounds, etc.) to be linked together, while a Hypertext link is a pointer to a specific object, or a specific place with a text; clicking the link opens the file associated with the object (Sherman et al, 2001). The internet is therefore the hardware while the Web is the Software.

The deep and surface web

Surface Web is made up of static and fixed pages, whereas the Deep Web is made up of dynamic pages. Static pages do not depend on a database for their content. They reside on a server waiting to be retrieved, and are basically HTML files whose content never changes. Any changes are made directly to the html code and the new version of the page is uploaded to the server. In static publishing, html text is pre-generated and stored as flat files on Web servers. These pages are less flexible than dynamic pages. Dynamic pages are created as a result of a database search. They are also called database-driven Web pages, wherein the content and the design are housed separately. The content is put in a database and is provided only when requested by the user. The following are the differences between Surface Web and Deep Web (Bergman, 2001): v Public information on the Deep Web is currently 400-550 times larger than the commonly-defined World Wide Web.

v The Deep Web contains 7,500 terabytes of information, compared to 19 on the Surface Web.

v The Deep Web contains nearly 550 billion individual documents compared to one billion on the Surface Web.

v More than 200,000 Deep Web sites presently exist.

v Total quality of the Deep Web is 1,000 to 2,000 times greater than that of the Surface Web.

v Deep Web content is highly relevant to every information need, market, and domain.

v More than half of Deep Web contents reside in topic-specific database.

v Content of the Deep Web is evaluated by experts Information on Deep Web is more comprehensive than that of the Surface Web.

Concept of deep web

The content of the Deep Web is rarely shown in a search engine result, since the search engine spiders do not crawl into databases and extract the data. These spiders can neither think nor type, but jump from link to link. As such, a spider cannot enter pages that are password protected. Web page creators who do not want their page shown in search results can insert special meta tags to keep the page from being indexed. Spiders are also unable to pages created without the use of html, and also links that include a question mark. But now parts of the Deep Web with non-html pages and databases with a question mark in a stable URL are being indexed by search engines, with non-html pages converted to html. Still, it is estimated that even the best search engines can access only 16 percent of information available on the Web. There are other Web search techniques and technologies that can be used to access databases and extract the content, including Librarian's Index to the Internet, which indexes access points to the content of the Deep Web. (Rabia et al 2010) Searching the Deep Web requires time and patience. With a specific URL, users can access a specific database, Search for the specific database directory; while entering search terms, keywords should be entered along with the word “database.” For example, for information on sports, use “sports database”.

The size of the web

The internet contains at least 4.5 billion websites that have been indexed by search engines, according to one Dutch researcher, Antal (2016). That huge number barely scratches the surface of what is really out there, however. The rest is known as the deep web, which is 400 to 500 times larger than the surface internet, according to some estimates. (Maurice, 2017). Today, the size of the Deep web is estimated at 9.7 Zettabyte (a unit of information equal to 2^{70} Bytes) covering 95% of the entire web (Vishwas, 2017), (Bergman, 2001).

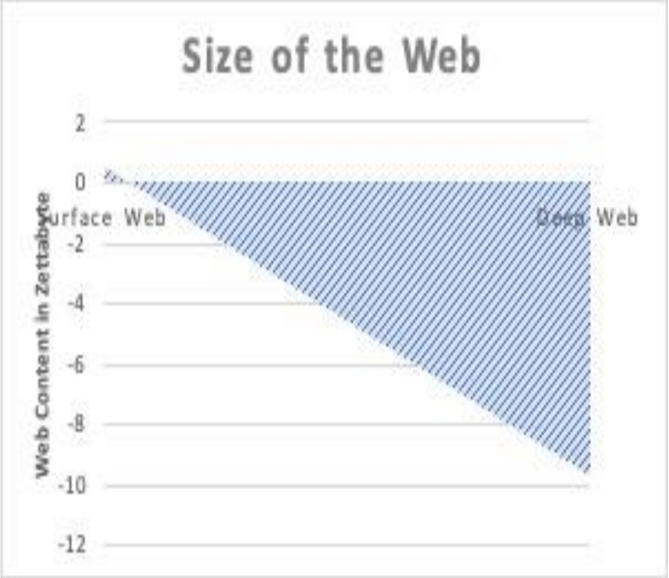


Fig. 1. Size of the Web

AREAS OF THE DEEP WEB

1**. Hidden Wiki**: The Hidden Wiki is the main entryway to the Deep Web. It is a good starting point because, like the Surface Web, Wikipedia provides a wide range of user updated links under a variety of different categories. It includes a section of Introduction Points that are relatively generic and safe to view. As you scroll further down there are links to different databases, Wikileaks, hacking and cracking websites, black marketplaces, peer-to-peer file sharing sites, an unsettling amount of porn, and anonymous forums and blogs.

2**. Darknet (the Dark Web)**: The Darknet is a compilation of networks and technologies that are used to share digital content. Darknet file sharing is a collective process; it is a decentralized database with computers working in harmony to send information from source to destination. Most Internet users have dabbled into the Darknet without realizing it; when downloading a song off a torrent site or a peer-to-peer network. LimeWire, BearShare, and Gnucleus are just a few of the big names of the Darknet. Peer-to-peer file sharing includes the uploading and distributing of things like software programs, songs, movies, and books. The items that are shared are referred to as *objects*, the people who share these objects are *users*, and the computers that are used for these transactions are known as the *hosts*. Millions of users partake in this network of shared information to find ways around restrictions and copyrights.

3**. Illegal Activity**: The Deep Web supplies a gargantuan amount of space for people to anonymously post and view any kind of content they desire. It is no wonder that some consider the Deep Web to be a breeding ground of illicit, unethical, and illegal behavior. While one can argue in favor of the Deep Web by saying that it offers even more valuable knowledge than illegal content, it cannot be ignored that people with bad intentions occupy the dark spaces of the Deep Web for their own interests. Silk Road is an example that comes up often when illegal activity on the Deep Web is explored. It serves as a black marketplace of the Deep Web that allows you to buy virtually any kind of drug in any quantity you desire. Your purchase will be delivered directly to your house disguised in a crate of oranges or some other discrete manner. The identities of the buyers and sellers are completely anonymous during the entire process.

4. **Bitcoin**: On the Surface Web, money transactions are made through websites that are overseen by central authorities. The currency used on the Deep Web is known as Bitcoin and allows for anonymity because it is completely untraceable. Bitcoin took shape in 2008 when a man named Satoshi Nakamoto, which turned out to be a pseudonym, created a currency that libertarian cryptographers had been trying to invent for years. This new crypto-currency was first met with skepticism, but eventually won over the digital world. Bitcoin is known as a crypto-currency because it uses cryptography; the practice of hiding information, to oversee the transfer of money. Transactions are irreversible and verified within anywhere between 10 minutes to an hour. The Bitcoin network is decentralized, making these transactions solely peer-to-peer. As it stands, one Bitcoin equates to about $8.67, although the exchange rate changes from day to day. The biggest challenge in using Bitcoin is what is known as double-spending, a phenomenon that occurs when the digital - currency is simply copied and pasted instead of purchased with real money. Bitcoin avoids double-spending by using a block chain – a record of all transactions and how much currency is left in each account. Bitcoin greatly enforces the anonymity of the deep Web. Unlike traditional Internet commerce, this currency does not pass through a third party. This also arguably encourages illegal behavior because the authorities cannot track where the money goes.

BENEFITS OF DEEP WEB

As the digital world expands, the debate on privacy and data security is on the rise. That is the reason behind the popularity of the deep web and the sudden increase in dark web browsers. Dark web is developing into a community of awesome people who have dedicated their lives towards building a network where no organization, Government, or spy agency can breach anyone's privacy. The deep web community comprises of lawyers, doctors, journalists, activists, CEOs, etc. While the name and the whole concept sound anti-Government, accessing or browsing the deep web is entirely legal with TOR browser. There are tons of illegal activities too all over the deep network as it offers the same privacy and anonymity to the individuals involved in such rackets. Amidst all the negativity, there is a lot of positivity too.

1. Anonymity

Anonymity is the primary reason why many individuals are making their way towards the hidden network and opting for dark web browsers to surf the internet. However, as much as anonymity aids in privacy, it also harbors criminal activities. It is the very reason why the reputation of the dark web is a bit frightening.

2. Freedom of Speech

Freedom of Speech is not universal, and most countries do not consider it as a fundamental right of a citizen. While the United States and a few other nations encourage freedom of speech, it is a distant dream for the citizens of several countries worldwide. Deep web browsing helps the citizens of such nations to hide their identity and allow them to express their opinion freely. Deep web browsing also eliminates any censorship of the country, and everything is accessible in the dark web without any restraint.

3. Knowledge

The Deep web is also an excellent platform for doctors, scientists, researchers, and teachers to share their data without any restriction or censorship. You will find data related to the scientific findings that most Governments want to hide or content redefining some health or social beliefs on the dark web. Most importantly, you will find literature works which are available in their purest form in this hidden web. Knowledge is best delivered when it is not filtered. Dark web harbors all the literature works featuring different thinking or opinions. You can access information on the deep web which is not available even at schools or colleges.

4. Political Activism

The deep web also aids leaders and political activists with their noble causes. In an oppressive Government structure, information is always manipulated and restricted. The administration often keeps a watch on every citizen’s activities on the internet and hinders any spread of revolutionary thoughts. The deep web is the best platform for people belonging to such nations with oppressive Government structure. The dark network offers a private and secure line of communication for the leaders to revolutionize and educate people about their fundamental rights.

The dark web and the deep web

The TOR (The Onion Router) maintains the most popular tool for Dark Web access. On the Tor network, internet traffic is directed through the network of random relays. The browser builds a route of encrypted connections, one-by-one. Each relay knows only the previous and the next relays, but the full connection route stays almost untraceable. The Multiple layers of encryption resemble the structure of an onion. The Deep web coverage and their content proportions are presented in Figure 2.

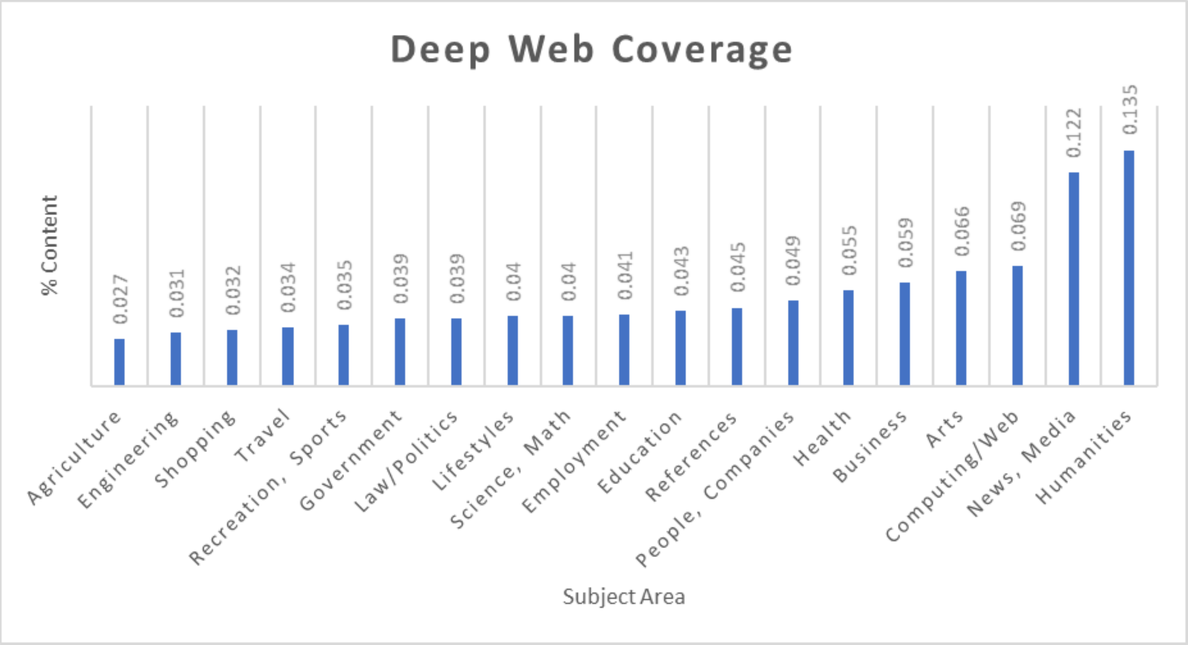
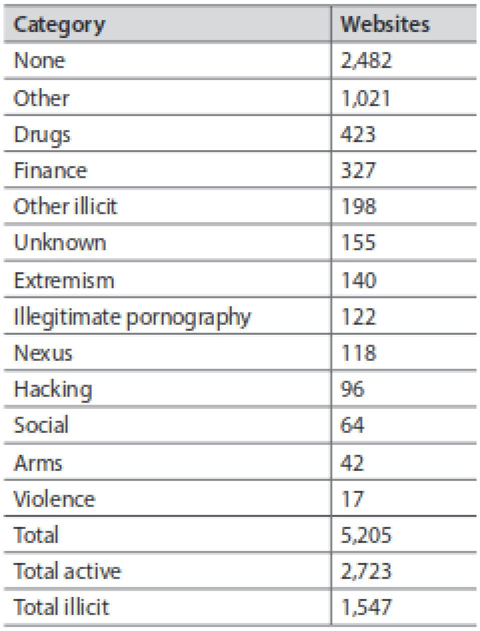


Fig. 2. Distribution of Deep Sites by Subject Area (Source: Statista 2009)

According to Thomas Rid and Daniel Moore (2016); out of 2723 active sites found on the Tor Dark web during several weeks, 1547 or 56.8 percent contained illicit material of some kind. It turns out that a majority of cyber criminals, selling everything from compromised personal and financial data to drugs and hacking tools, constitute over half of Dark Web contents, presented in the table below:

Table 1. Content of the Dark Web



Category	Websites
None	2,482
Other	1,021
Drugs	423
Finance	327
Other illicit	198
Unknown	155
Extremism	140
Illegitimate pornography	122
Nexus	118
Hacking	96
Social	64
Arms	42
Violence	17
Total	5,205
Total active	2,723
Total illicit	1,547

Searching the deep web

Searching and indexing information on the Deep Web is a complex process that differs from the Surface Web. The pages are dynamic changing and not linked to each other, making it impossible for the web “spiders” of traditional search engines to crawl and index data. Although the process is more complicated, it is certainly not impossible to access the Deep Web content. Two main approaches are used to do so. The first approach to find Deep Web content is called virtual integration. It occurs when a user enters a query on a certain topic and that query is directed to relevant databases. Virtual integration brings together vertical search engines in certain domains and allows for a deeper searching experience. The downside of this approach is that it can be difficult to bring together many parallel domains without complicating the keyword that is being searched for because it needs to be translated for each site. The second Deep Web searching approach is known as the surfacing approach. This method pre-computes a set of queries for HTML forms that might be useful. URLs are made for all pre-computed queries and are added to the search engine index. This approach relies on searching HTML forms with sample queries and reviewing how similar the content is on the web pages that are found. The biggest difficulty with this approach is deciding what set of queries to use for certain forms. This approach is more effective than the first because it is automatic and has less human involvement. The Deep Web is so vast that the more automatic the browsing is, the better.

Content and search quality of the deep web

"Quality" is subjective: If you get the results you desire, that is high quality; if you don't, there is no quality at all. In a study carried out by BrightPlanet, five queries were posed across various subject domains using only computational linguistic. Computational linguistic scoring does not introduce systematic bias in comparing deep and surface Web results because the same criteria are used in both. The relative differences between surface and deep Web should maintain, even though the absolute values are preliminary and will overestimate "quality." (Bergman, 2001). A quality search result is not a long list of hits, but the right list. Searchers want answers. Providing those answers has always been a problem for the surface Web, and without appropriate technology will be a problem for the deep Web as well. Effective searches should both identify the relevant information desired and present it in order of potential relevance; quality.

Table 2. "Quality" Document Retrieval, Deep vs. Surface Web

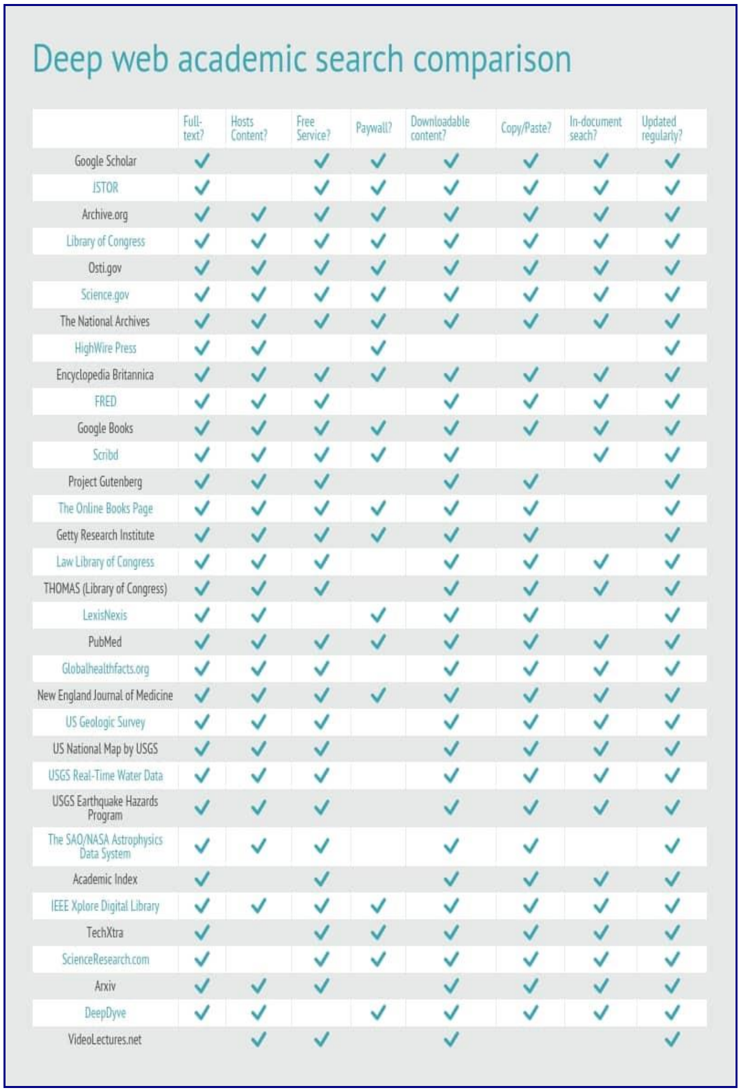
Query	**Surface Web**			**Deep Web**		
	Total	"Quality"	Yield	Total	"Quality"	Yield
Agriculture	400	20	5.0%	300	42	14.0%
Medicine	500	23	4.6%	400	50	12.5%
Finance	350	18	5.1%	600	75	12.5%
Science	700	30	4.3%	700	80	11.4%
Law	260	12	4.6%	320	38	11.9%
TOTAL	2,210	103	4.7%	2,320	285	12.3%

This table shows that there is about a threefold improved likelihood of obtaining quality results from the deep Web as from the surface Web on average for the limited sample set. Also, the absolute number of results shows that deep Web sites tend to return

10% more documents than surface Web sites and nearly triple the number of quality documents.

USING DEEP WEB SEARCH ENGINES FOR ACADEMIC AND SCHOLARLY RESEARCH

The first and most important step in searching the deep web is knowing where to look. While the deep web is almost infinitely vast when it comes to the amount of information you can find, unlike what most people are used to when searching for something in Google, all of that data are not centralized in the same place. This means for as many different subjects you can think of (finance, software, business, economics, academia, etc.), there are an equal number of search engines designed to dive into the deep web archives of those particular subjects. One issue that some researchers run into though is the problem of paywalls. If paywalls are a problem for you, one tool worth checking out is the Google Chrome browser extension Unpaywall. This automatically scours the web for a free version of any content you are trying to access that is behind a paywall. You may not always get back a free result for every paper you search, however, it is still nice to know the option is there. Figure 3 includes a list of some of the services that do the best job of cataloging all the information that may be useful in an academic research.



Deep web academic search comparison

	Full-text?	Hosts Content?	Free Service?	Paywall?	Downloadable content?	Copy/Paste?	In-document seach?	Updated regularly?
Google Scholar	✓		✓	✓	✓	✓	✓	✓
JSTOR	✓		✓	✓	✓	✓	✓	✓
Archive.org	✓	✓	✓	✓	✓	✓	✓	✓
Library of Congress	✓	✓	✓	✓	✓	✓	✓	✓
Osti.gov	✓	✓	✓	✓	✓	✓	✓	✓
Science.gov	✓	✓	✓	✓	✓	✓	✓	✓
The National Archives	✓	✓	✓	✓	✓	✓	✓	✓
HighWire Press	✓	✓		✓				✓
Encyclopedia Britannica	✓	✓	✓	✓	✓	✓	✓	✓
FRED	✓	✓	✓		✓	✓	✓	✓
Google Books	✓	✓	✓	✓	✓	✓	✓	✓
Scribd	✓	✓	✓	✓	✓		✓	✓
Project Gutenberg	✓	✓	✓		✓	✓		✓
The Online Books Page	✓	✓	✓	✓	✓	✓		✓
Getty Research Institute	✓	✓	✓	✓	✓	✓		✓
Law Library of Congress	✓	✓	✓		✓	✓	✓	✓
THOMAS (Library of Congress)	✓	✓	✓		✓	✓	✓	✓
LexisNexis	✓	✓		✓	✓	✓		✓
PubMed	✓	✓	✓	✓	✓	✓	✓	✓
Globalhealthfacts.org	✓	✓	✓		✓	✓	✓	✓
New England Journal of Medicine	✓	✓	✓	✓	✓	✓	✓	✓
US Geologic Survey	✓	✓	✓		✓	✓	✓	✓
US National Map by USGS	✓	✓	✓		✓	✓	✓	✓
USGS Real-Time Water Data	✓	✓	✓		✓	✓	✓	✓
USGS Earthquake Hazards Program	✓	✓	✓		✓	✓	✓	✓
The SAO/NASA Astrophysics Data System	✓	✓	✓		✓	✓		✓
Academic Index	✓		✓		✓	✓	✓	✓
IEEE Xplore Digital Library	✓	✓	✓	✓	✓	✓	✓	✓
TechXtra	✓		✓	✓	✓	✓	✓	✓
ScienceResearch.com	✓		✓	✓	✓	✓	✓	✓
Arxiv	✓	✓	✓		✓	✓	✓	✓
DeepDyve	✓	✓		✓	✓	✓	✓	✓
VideoLectures.net		✓	✓		✓			✓

Fig. 3. Deep Web Academic Search

Source:Chris Stobing, June 2, 2018 (https://www.comparitech.com)

CONCLUSION

The advent of the Internet and access to global information is a great benefit, even though information managers had the difficult task of organizing, retrieving, and providing access to precise information. Despite the vast amount of information, the Deep Web contains, it is still an ambiguous part of the digital world. Many Internet users have not heard of it and believe that what they see on their Google search results is all that the Web has to offer. Others would rather have it abolished, claiming that it is an underground world of crime and unethical behavior. Then there are those who are interested in the possibilities of this unexplored frontier, but simply do not know where to begin. Users depend on the popular search engines and portals, which cannot provide access to the hidden store of valuable information available in the Deep Web. To access the information available on these databases, users will have to become familiar with the structure of the Deep Web. Any information created should be shared and used, since that alone leads to the creation of more information. When a specific database is created, information regarding its existence should be published so that users will be aware and make maximum use of available information.

RECOMMENDATION

The Onion router is the major and cheapest tool used to access the Dark Web. It is beneficial for a researcher to learn how to setup and use this tool for maximal engagement on the web for academic research. It is also useful to study a list of available databases of the Dark Web and their contents using the Hidden wiki. A research subject of search should be crafted carefully so that its category can be clearly understood and applied to the appropriate database(s). This would bring to light the best links for the subject in question in the rank of best to worst as understood by the search engine.

REFERENCES

Antal van den Bosch, T. B and Maurice de K. (2016). Estimating search engine index sizevariability: a 9-year longitudinal study. Scientometrics: An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science. DOI 10.1007/s11192-016-1863-z

Bergman, M. K. (2001). The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing (JEP). 7(1):11-14

Bright, P. (2005). Largest deep-web sites. http://brightplanet.com/infocenter/largest_deepweb_sites.asp

Daniel M and Thomas R. (2016) Cryptopolitik and the Darknet, Survival,58(1):7-38.

Deepweb.com (2008). “Deep web” Retrieved from www.faviki.com/website/www.deepweb.com

Devine, Jane; Egger-Sider, Francine (2004). "Beyond google: the invisible web in the academic library". The Journal of Academic Librarianship. **30** (4): 265–269.

Hamilton, N. (2003). "The Mechanics of a Deep Net Metasearch Engine". In Isaías, Pedro; Palma dos Reis, António (eds.). Proceedings of the IADIS International Conference on e-Society. pp.1034–6.

Lawrence S. and Giles, C. L. (1998). "Searching the World Wide Web," Science 80:98-100.

Lawrence S. and Giles, C.L (1999). "Accessibility of Information on the Web," Nature 400:107-109.

Maurice de Kunder (2007). Geschatte grootte van het geïndexeerde World Wide Web, Universiteit van Tilburg Tilburg, Noord-Brabant maurice dekunder.nl, Versie: 1.4

Prableen B. (2016). Deep Web Vs. Dark Web . ”https://www.investopedia.com/articles/ insights/062416/deep-web-vs-dark-web.asp”

Rabia Iffat, Lalitha K. S. (2010). Understanding the Deep Web, Library Philosophy and Practice. Library Information Science Journal.

Raghavan, S.; Garcia-Molina, H. (2001). "Crawling the Hidden Web". 27th International Conference on Very Large Data Bases.

Sherman, C. and Price, G. (2001). The Invisible Web; Uncovering Information Sources Search Engines Can't See. CyberAge Books, ISBN 0-910965 -51-X.

Vishwas, R. (2017). Know Your Web : Surface Web, Deep Web, and Dark Web. https://csultimates.net/blog/2017/05/what-is-surface-deep-and-dark-web/.