On the eve of web-harvesting and web-archiving for libraries in Greece in the Openness era

Paper background:

In the post-paper world libraries implement web harvesting and web archiving methods in several countries facing technological and legal issues which are intensified due to the idea of "openness" regarding access to information or openness to partnerships. The core theme of this 2018 ICIL participation could be seen under the prism of "*intellectual property and contemporary issues of openness*" thematic. Web-harvesting and web-archiving as a technological option is usually leveraged upon in the context of legal deposit systems which are set in the legal and technical frameworks of operation of major and/or national libraries, and aim at the collection, download, and archiving of works which are found available on the Internet through an automated process of tracking and pulling of works found online. To collect everything from the web is impractical or simply utopian idea. Most libraries which enact web-harvesting and web-archiving operations leverage on Heritrix[1] and regularly crawl usually once or twice a year on a large scale (legal deposit scope) and manage several supplementary focused crawls depending on special events or topics and/or special collection interests, sometimes common among different countries and libraries.[2] The whole harvested materials may be archived without any selection or cataloguing process.

The access to the web information and works available online is subject to restrictions by regulation, especially laws pertaining to Copyright, Industrial Property Rights, Data Privacy etc. on the same model as the current legal deposit material on hard copies in most libraries empowered to do the legal deposit. It is common for libraries that deploy web-harvesting and web-archiving operations to restrict materials collected and archived through these processes for research to readers registered by the library, only.[3]

Research objective:

There are important issues that pertain to the web-harvesting and web-archiving operations which are not set in the same way and processes from system to

---

[1] Heritrix is a free, open-source, extensible, archiving quality web crawler. It was developed, and is used, by the Internet Archive and is freely available for download and use in web preservation projects under the terms of the GNU GPL. It is implemented in Java, and can therefore run on any system that supports Java (Windows, Apple, Linux/Unix). See more about Heritrix at URL: https://webarchive.jira.com/wiki/display/Heritrix [last check, Sept.1, 2018].

[2]For example, the National Library of France harvests broad-spectrum material from French websites once a year but more often select material from a few thousands websites based on specific topics (literature, sustainability) or events (elections, 2012 Olympic Games). See more at Bibliotheque nationale de France (BnF) through the URL:
http://www.bnf.fr/en/collections_and_services/book_press_media/a.internet_archives.html#SHDC__Attribute_BlocArticle0BnF [last check, Sept. 1, 2018].

[3] The National Library of Sweden and others allows the public to study collected websites only at library place because of copyright matters. See more at URL: http://www.kb.se/english/find/internet/ [last check, Sept 1, 2018]

system, though an attempt to compose a digital roadmap for relevant processes was undertaken in 2013.[4] Web-harvesting and web-archiving has triggered the interest of UNESCO[5] as of the PERSIST project (2013).[6] Common preservation issues that web-harvesting and web-archiving systems may have to deal with are the frequency with which resources online change, the quantity and the range of resources potentially needing preservation, the versioning of resources across a site, the integrity of resources available online, the ownership of resources, the capturing of resources in databases and through deep-linking, the capturing and collection of multimedia resources in terms of quality and quantity of data or range of formats, the technical means for appraisal and selection of resources.[7] In addition, once preserved it has to be considered how access will be provided to the web resources and how to deal with issues of intellectual property rights of the resources. Both personnel and technical resources issues also have to be considered.

The objective of this research is to delve into web-harvesting and web-archiving issues such as the above hereto described and analyze them in consideration of attempts and experimentation to set web-harvesting and web-archiving systems for libraries leveraging on the international experience of relevant systems as well as on the acquis communautaire per subject matter[8] and new legislation in Greece, too. Regarding web harvesting, it is worth looking at the policies that other EU libraries have adopted too.

Methodology:

The methodology for this research on web-harvesting and web-archiving issues involves studying relevant literature and web material as well as using questionnaires and interviews.

---

[4] See Marcel Ras, (2013), **A global Digital Roadmap**, position paper, available at URL: https://www.unesco.nl/sites/default/files/uploads/Comm_Info/position_paper_-_digital_roadmap_meeting_the_hague.pdf [last check, Sept.1, 2018]· see, also, (2013), **A Digital Roadmap for Long-Term Access to Digital Heritage**, Conference organized by UNESCO, ICA, IFLA, report available at URL: https://www.unesco.nl/sites/default/files/uploads/Comm_Info/digital_roadmap_-_report.pdf [last check, Sept.1, 2018].

[5] At the World Library and Information Congress (WLIC) in Lyon the summer of 2014, the UNESCO session was spent on the problem of selection in digital heritage. It is one of the challenging subjects within the PERSIST project, an initiative of UNESCO, ICA, IFLA and other partners, that tries to enhance the sustainability of digital heritage.

[6] In December of 2013 an international meeting took place in The Hague in which the ICT industry, governments and heritage institutions discussed collaboration in the field of digital preservation. The meeting was convened by UNESCO, and was possible thanks to a subvention form the Netherlands Ministry of Education, Culture and Science. Since then UNESCO has been cooperating with ICA, IFLA, LIBER and other partners to continue and intensify the discussion that started in The Hague. Under the new name UNESCO-PERSIST (Platform to Enhance the Sustainability of the Information Society Transglobally) UNESCO addresses globally pressing questions on selection, responsibility and division of labor.

[7] See Sarah CC Choy et al, (2016), **The UNESCO/PERIST Guidelines for the selection of digital heritage for long-term preservation**, available at URL: https://www.unesco.nl/sites/default/files/dossier/persistcontentguidelinesfinal1march2016.pdf?download=1, [last check, Sept.1, 2018], The impact of the Legal environment on selection, p.5.

[8] See **Recommendation 2006/585/EC** of August 24, 2006, EE L 236, of 31/8/2006, available at URL: http://www.opi.gr/images/library/nomothesia/evrwpaiki/systaseis/24_08_2006.pdf [last check, Sept.1, 2018].