Scientific Research Publishing

Open Journal of Philosophy, 2020, 10, 88-112
https://www.scirp.org/journal/ojpp
ISSN Online: 2163-9442
ISSN Print: 2163-9434

Empirical Research on Web Harvesting in the Process of Text and Data Mining in National Libraries of EU Member States

Marinos Papadopoulos[1], Maria Botti[2], M. A. Paraskevi (Vicky) Ganatsiou[3], Christos Zampakolas[4]

[1]Attorney-at-Law, Independent Researcher
[2]Department of Archives, Library Science and Museology, Ionian University
[3]Coordinator of Educational Projects in the Prefecture of Ionian Islands (Education Sustainability)
[4]Archivist-Librarian, Independent Researcher
Email: marinos@marinos.com.gr, botti@otenet.gr, pganatsiou@gmail.com, christoszampakolas@gmail.com

How to cite this paper: Papadopoulos, M., Botti, M., Paraskevi (Vicky) Ganatsiou, M. A., & Zampakolas, C. (2020). Empirical Research on Web Harvesting in the Process of Text and Data Mining in National Libraries of EU Member States. *Open Journal of Philosophy*, *10*, 88-112.
https://doi.org/10.4236/ojpp.2020.101007

Received: November 8, 2019
Accepted: February 4, 2020
Published: February 7, 2020

Abstract

Almost two decades of experience on web harvesting and archiving are counted; the subject of web harvesting and web archiving have been top in the interest of researchers, technologists and librarians-information scientists. Web harvesting projects and pilot programs on archiving content traced on the Web are becoming priorities for national libraries and cultural heritage organizations in the EU. This paper pertains to web harvesting as a process for data mining from web and only through web ("*pull*" function); this paper elaborates upon research implemented in the framework of the funded research project titled "*Web Archiving in Public Libraries and IP Law*" that focused on the processes of web-harvesting and archiving as well as Text and Data Mining (TDM) operations in the national libraries of EU Member States. Web archiving as an official operation in national libraries of EU Member States creates web collections and preserves them for the purpose of being accessible and usable in perpetuity. This paper pertains to research on various components of web harvesting and archiving through an online survey (qualitative research) which targeted the national libraries of EU Member States. The research team of authors posed seventeen questions to EU national libraries. The survey output comes from answers delivered by 22 national libraries of EU Member States. The questionnaire was created through the use of Google forms. The researchers reached the EU national libraries via email and follow up telephone calls seeking libraries' participation in the research. The aim of the research was to delve on participant libraries' Text and Data Mining operation leveraging on Web harvesting and Web archiving technologies and operations. Results analysis reveals that web harvesting is considered among national libraries' top priorities; the relevant projects increase in

number, the web collections become more and more and the technological infrastructures and tools for web harvesting improve. Yet, there are many issues that remain unresolved. A significant number of surveyed libraries consider that legal and technical issues remain the most important to resolve. Access to harvested material is still under legal restrictions. The Directive 2019/790/EU on Copyright in the Digital Single Market (DSM) creates a favorable legal foundation for the deployment of web harvesting operations in national libraries of the EU Member States. TDM technologies make possible new areas of research. Web harvesting that was initially aimed for preservation purposes now expands to unprecedented research of national heritage through state-of-the-art automated TDM processes.

Keywords

TDM, Web Harvesting, Web Archiving, National Libraries, Survey

1. Introduction

From the very beginning of Internet's pioneering appearance in the early 90s, humanity realized that world culture has acquired a new "*vehicle*" for information spreading and dissemination of knowledge, science and research (Masanès 2002); the Internet was also seen as a means for the modification of economy, society and cooperation, and a necessity of new management was derived, consequently. Soon it was realized that a huge volume of web heterogeneous resources that reside online seek for a path to eternity and that the Internet is a very dynamic space in which information is susceptible to loss, though (Miranda, n.d.). Web content is changing at a pace that puts itself at risk of extinction or falsification while humans would probably want to preserve it in the future as part of world cultural heritage. Experts in Portugal report that 80% of the web is disappeared one year after being published excluding any further access[1]. Even printed publications are adversely affected by the ephemeral nature and transience of the Internet as they often refer to websites that have ceased to exist (Gomes, Miranda, & Costa, 2011).

Web harvesting and web archiving have emerged as new official functions of intellectual and cultural heritage preservation organizations leveraged to serve the need for management of content harvested from the web. According to the International Internet Preservation Consortium (IIPC), "*Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use*"[2]. It is

[1]Arquivo.pt is a Portuguese web archive created to preserve their online national heritage and has been providing web pages since 1996. Retrieved April 10, 2019, from https://www.fccn.pt/en/knowledge/arquivo-pt/

[2]The International Internet Preservation Consortium (IIPC) "aims at collecting, preserving and making accessible knowledge from the global web". The definition of web archiving is on IIPC webpage entitled Web Archiving. Why archive the web? (n.d.). Retrieved May 20, 2019, from http://netpreserve.org/web-archiving/

important to keep in mind that the ultimate recipient of the processes chain for cultural heritage aggregation and preservation is the user and therefore the ultimate goal of web harvesting and archiving organizations is to make use and access of archived content that resides on the Web possible. The achievement of this goal will ideally justify national libraries' operation as cultural aggregators and preservation organizations. "*It's the unexpected reuse of information that adds value to the web*," said Sir Tim Berners-Lee (2006), the founder of the World Wide Web talking about linked data[3].

Web harvesting, therefore, is a process that leverages on new technologies and relates to the widespread term of extracting texts and data from the web. The evolution of web harvesting technologies and processes leads to more exciting paths than simple web mining and archiving of online resources in order to become yet another document in the digital "*shelves*" of a library. Text and data mining technology—TDM technology as is simply referred to—used in the process of web harvesting is "*any activity where computer technology is used to index, analyze, evaluate and interpret mass quantities of content and data*" (Caspers at al., 2016; Botti, Papadopoulos, Zampakolas, & Ganatsiou, 2019a, 2019b).

Having passed over twenty years of web harvesting in Europe, this research aims at highlighting the current state of web harvesting and exploitation of TDM technologies in the EU Member States and, in particular, in their national libraries. Qualitative properties and characteristics of this function are researched.

2. Directive 2019/790/EU

The new European Directive 2019/790/EU of 17 April 2019 on copyright in the Digital Signal Market (DSM) introduces the term of text and data mining as compulsory copyright exemption for educational/teaching or scientific research purposes[4]. This new European legislation remains to be implemented Europe-wide through national laws of EU Member States.

For the EU legislator, TDM is just a means to achieve the goal of Digital Single Market (DSM). The goal for a European DSM is a goal for the free movement of goods, persons, services and capital where individuals and businesses can seamlessly access and exercise online activities under conditions of fair competition, and a high level of consumer and personal data protection, irrespective of their nationality or place of residence. "*Everyone has an equal right to access and use a secure and open Internet*" (IRPC, 2014: p. 9)[5].

The TDM exception in the new EU Directive on Copyright in the DSM pertains to the harmonization issue of exceptions and limitations in copyright law of EU Member States, and the creation of legal certainty for cross-border use of

[3]Retrieved April 10, 2018 from https://www.w3.org/DesignIssues/LinkedData.html.

[4]The Directive (EU) 2019/790 of the European Parliament and the Council of European Union published in Official Journal of the European Union (17 May 2019), retrieved May 25, 2019 from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0790.

[5]Internet Rights and Principles Coalition (IRPC) is an open network on human rights promotion especially via Internet. It based at the UN Internet Governance Forum (IGF), retrieved August 2, 2019 from http://internetrightsandprinciples.org/.

content for the purpose of scientific research or other purposes (Papadopoulos & Botti, 2019). The barriers to access, so far, remain because of legislative restrictions on data protection and intellectual property (Directives 96/9/EC and 2001/29/EC), national laws of EU Member States and administrative law (Jacobsen, 2008). At best the legislation itself defines the exact place of access, the scope and the manner of copying the web archived material (digital copies are not permitted) as is the case of National Library of Finland (Keskitalo, 2010).

The new Directive on Copyright in the DSM includes article 3 and article 4 which address the issue of TDM. Article 3 is titled "*Text and data mining for the purpose of scientific research*"; Article 4 is titled "*Exception or limitation for text and data mining*"[6]. The mandatory character in the provision per TDM of Directive 2019/790/EU on Copyright in the DSM prevails (Botti, Papadopoulos, Zampakolas, & Ganatsiou, 2018).

TDM is seen in the broader perspective of Web harvesting. When Web harvesting began in the USA and Europe, two different policies were followed. Web harvesting was done by pre-selecting individual sites which limited its range (USA) or by using "*crawlers*" as an automated process based on good and appropriate technology that allowed for extensive mining as happened in the case of Sweden (Botti, Papadopoulos, Zampakolas, & Ganatsiou, 2018; Masanès, 2002). The first European Web harvesting program was the Swedish Kulturarw3[7].

Various researches have been implemented in the past that have outlined the field of Web harvesting in Europe and in the USA. Empirical researches on TDM technologies and Web harvesting in Europe have been conducted both, by individual researchers and research organizations. IIPC (International Internet Preservation Consortium) is one such international organization, which brings together institutions and organizations from all around the world in the name of preserving the web, developing research and collaborative action between its members and make the web archived collections accessible[8]. In 2004, a new non-profit organization was found named Internet Memory Foundation, as the European Archive to enhance web harvesting and web archiving operation (Toyoda & Kitsuregawa, 2012). A 2010 research[9] deployed by Internet Memory Foundation on Web archiving initiatives indicates that Web archiving has been gaining momentum and is recognized for modern societies around the world af-

[6]See note 4.

[7]National Library of Sweden began harvesting and archiving web resources pertained to Swedish web heritage, from the first time, in 1997. The web content was extracted from Swedish top level domain "se" and other servers identified as Swedish via geolocation. Retrieved May 2, 2019, from, http://dig-hum-nord.eu/projects/kulturarw3-the-web-archive-of-the-national-library-of-sweden/

[8]IIPC was founded in 2003 with twelve institutions as the fist members of the consortium. Today, IIPC members come from more than forty five countries and they have the mission to fund, collaborate and participate in web harvesting projects on web archived data (collections, preservation, usability and accessibility) Retrieved April 2, 2019, fromhttp://netpreserve.org/about-us/

[9]In the framework of the European Research Project "Living Web Archives project" (LiWA), Internet Memory Foundation implemented a survey on web harvesting in Europe. The survey was sent to European and international bodies and the results were released in 2010. Retrieved June 3, 2019 from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6182575

ter 2003 (Internet Memory Foundation, 2011). In 2007 the National Library of the Netherlands conducted a Web archiving user survey (Ras & Bussel, 2007). One of the most recent Web harvesting projects in Europe is the "*Promise*" project for archiving the Belgian Web, in which the experience of other states is explored, too (Chambers, 2018). Also, the initiative to create a Website archiving guide for Dutch government agencies that was presented at IIPC WAC 2018 by Suzi Szabo (2018) concludes that the best practice for Web archiving requires leaving this work to experts for the purpose of "*correcting*" lacking knowledge from website archiving.

In general, the types of Web harvesting are divided into the following categories: broad Web harvesting in top level national domain, selective or thematic web harvesting in selected subject areas, and Web harvesting of events and emergencies (IIPC). Using link crawlers, the whole web page is mined with their interconnections and hyperlinks. The archived webpage preserves the same interface with the original one (Botti, Papadopoulos, Zampakolas, & Ganatsiou, 2018; Nielsen, 2016).

3. Research Method

Authors' research on EU national libraries' TDM through the use of Web harvesting and Web archiving technologies and operations was implemented in the timeframe between March and July 2019; a short questionnaire was prepared in consideration of the assumption that most EU national libraries may not be fully prepared for large scale Web harvesting and Web archiving operations given that the relevant EU legal framework was just set through the new EU Directive on Copyright in the DSM. Besides, authors' legal research on national legal frameworks of EU Member States on TDM and/or Web harvesting or Web archiving had revealed that only a few EU Member States had set provisions in their national legal systems which cater for Web harvesting and/or Web archiving. Therefore, authors did not expect to have a prolonged questionnaire on TDM and/or Web harvesting or Web archiving answered by EU national libraries. The questionnaire focused on various components of web harvesting and archiving through an online survey (qualitative research) which targeted the national libraries of EU Member States. Authors didn't want to limit the posed questions to only one subject area of TDM and/or Web harvesting or Web archiving, but rather opted for spreading the subject matter of the questions to more issues relevant to TDM and/or Web harvesting or Web archiving. The research posed seventeen questions to EU national libraries. The survey output comes from answers delivered by twenty national libraries of EU Member States. The questionnaire was created through the use of Google forms. The researchers reached the EU national libraries via email and follow up telephone calls seeking libraries' participation in the research. The aim of the research was to delve on participant libraries' Text and Data Mining operation leveraging on Web harvesting and Web archiving technologies and operations.

4. Presentation of Empirical Research in 27 EU Member States' National Libraries

Research on TDM technologies, web harvesting and web archiving in the national libraries of the 27 EU Member States of the European Union was carried out in the framework of the funded research project titled "*Web Archiving in Public Libraries and IP Law*" that focused on the processes of Web-harvesting and archiving as well as Text and Data Mining (TDM) operations in the national libraries of EU Member States. In addition to bibliographic/online research, empirical research was conducted via email contained a link to a web-based survey provided by Google-forms. The questionnaire was sent to national libraries of all EU member states (27 sending emails). Eventually, twenty-two (22) responses were collected from twenty (20) different libraries/countries. From these answers, nineteen (19) were received by the online survey and three (3) more answers were received via e-mail (from libraries without any action on web harvesting yet).

Legal deposit legislation and furthermore, digital legal deposit regulation assigned national libraries of EU Member States with the task of Web harvesting and archiving of content that resides on the Internet. Web harvesting isn't just a new challenge for national libraries; it is a new legal and officially assigned obligation and area of their operation. Most national libraries of EU Member States have gained some experience on Web harvesting and archiving of content on the Web, and thus the scope of this empirical research was well served by focusing on the national libraries of EU Member States.

Table 1 below hereto depicts the basic characteristics of the survey implemented through the use of questionnaire furnished to the focus group of EU

Table 1. Survey's identity.

SURVEY'S IDENTITY	
Name	A survey on web archiving in EU Member States' national libraries
Kind	Empirical research via questionnaire
Medium	Internet by Google Forms
Provider	Ionian University
Co-Funded by	Greece and the European Union—European Social Fund
Part of	A research project titled "Web Archiving in Public Libraries and IP Law" within the framework of the Operational Program "Human Resources Development, Education and Lifelong Learning" of NSRF—Partnership Agreement 2014-2020
Duration	March-July 2019
Target group	National Libraries of EU Member States'
Language	English
Basic Fields/components	1) Library's policies on Web-harvesting/Arrangement/Procedures, 2) Technological issues, 3) Legal issues, 4) Access/Utilization, 5) Co-operation & Perspectives 6.Proposals and useful observations
Question's number	17
Main scope	Collecting elements on current web archiving situation
Expected results	Enhancing countries involved in Web Archiving, complications, perspectives, new projects

national libraries (survey's identity). As noted in the table, the survey's interest focuses on collecting elements related to the current situation on Web harvesting operation in the national libraries of the EU Member States (main scope). The ultimate goal is to make this information useful for libraries most of which do not have a long track record with TDM and/or Web harvesting or Web archiving activities. At this initial phase of research on TDM and/or Web harvesting or Web archiving in EU national libraries, the research team tried to identify and shed light upon the main considerations of surveyed national libraries regarding TDM and/or Web harvesting or Web archiving issues that are the most important for the EU national libraries. There's no doubt that certain areas such as TDM and/or Web harvesting and GDPR require more focused research in consideration of the existing European legal framework. The responses in the questionnaire posed to the surveyed national libraries attest to the need for more research focused on certain sub-areas of TDM and/or Web harvesting or Web archiving issues.

The surveyed national libraries responded with their plans on Web harvesting in various question areas, such as "*Library policies on Web harvesting/Arrangement/Procedures*" (question on thematic fields), "*Co-operation & Perspectives*" (question about new plans) and "*Proposals and useful observations.*" The new projects are, directly or indirectly, related to what is deemed important by the national libraries as well as to what is perceived as requirement for fully-developed Web harvesting activity, according to participants' assessment. The surveyed national libraries responded with indicative success or failure factors for TDM and/or Web harvesting; their comments are included at the end of the questionnaire in an optional question which is presented in the relevant section titled "*Proposals and useful observations*". Two (2) introductory questions are included to the survey (about libraries, and the identification of the individual who provided the answers to the posed questions on behalf of the library) and subsequently, seventeen (17) main questions, divided into six (6) fields/components (Table 1). All the questions are referred to in the Appendix of the article. Survey's language was English.

5. The Participants

The EU national libraries which participated in our empirical research and responded to our survey are shown in Table 2 below.

6. Scope and Components of the Research

The purpose of the research was to demonstrate the current situation of Web harvesting in EU members from specific aspects as is mentioned above hereto. The research was mainly qualitative and within this purpose; we collected information from EU Members' national libraries that have gained much experience in Web harvesting and archiving. The type of researched organizations, a.k.a. "*national libraries*" was chosen because they served our research's purpose as these are the Web harvesting and archiving organizations that conduct these operations under a national legal mandate. Our empirical research also aimed at

Table 2. National Libraries which participated in survey.

Country	Institution
Italy	National Central Library of Florence
Luxemburg	National Library of Luxemburg
United Kingdom	The British Library
Estonia	National Library of Estonia
Austria	Austrian National Library
Denmark	The Royal Danish Library
France	National Library of France
Slovenia	National and University Library, Slovenia
Finland	National Library of Finland
Greece	National Library of Greece
Scotland	National Library of Scotland
Cyprus	Cyprus Library
Germany	National Library of Germany
Spain	National Library of Spain
Hungary	National Szechenyi Library
Sweden	National Library of Sweden
Belgium	KBR Royal Library of Belgium
Malta	National Library of Malta
Italy	Biblioteca Nazionale Centrale "Vittorio Emanuele II"—Rome
	Europeana

filling the gaps and/or broadening the research to the most interesting points that emerged from bibliographic/Web research.

These points are linked to different query sections which are defined as follows:

1) Libraries;
2) Responders' professional skills and expertise;
3) Library policies on Web-harvesting/Arrangement/Procedures;
4) Technological issues;
5) Legal issues;
6) Access/Utilization;
7) Co-operation & Perspectives;
8) Proposals and useful observations.

7. Survey Results

Survey results are presented here within relation to the components of our research.

From the responders, 90% of the national libraries which participated in our survey answered that they have Web harvesting/archiving activity. The remaining 10% of surveyed libraries like the National Central Library of Rome and the National Library of Malta expressed interest in both, web harvesting and archiving of content on the Web which is still under development for them. Europeana

has no web harvesting/archiving activity too.

Survey results, via online questionnaire, indicate that Web harvesting is reported to be one of the three most important purpose-specific functions of the surveyed libraries. However, most of the EU national libraries do not employ a full-grown team of experts on their TDM and/or Web harvesting or Web archiving operation; in most cases, operators' number ranges from one person, a librarian with multiple responsibilities and with the help of outsourced collaborators to a well-organized small team of three or four people.

Most surveyed EU national libraries use quality filters for their Web harvesting operations. Thematic Web harvesting is the rule. Surveyed national libraries leverage on a variety of themes for their Web harvesting operations. There are national libraries which consider thematic diversity and peculiar themes in their effort for prominent placement in the niche market of TDM in Europe.

Most surveyed national libraries have an interest in TDM and/or Web harvesting and Web archiving because they want to make certain kind of information and/or works accessible to researchers. The side-effect of storage and preservation of harvested material from the Web comes second in the EU national libraries' goals targeted through their TDM and/or Web harvesting operations.

Most surveyed national libraries prefer leveraging on their own researches for TDM and/or Web harvesting activity; however, it's still too early for all EU national libraries to depend on their own researches for successful large-scale TDM and/or Web harvesting. In most cases of surveyed national libraries, there are still legal issues that remain to be resolved through amendments of national legal frameworks on TDM and/or Web harvesting or Web archiving activities. For example, replies that surveyed national libraries gave to the posed questions indicate that there is not a prior consent mechanism in the Web-harvesting and archiving operations of the libraries. Also, TDM and/or Web harvesting and GDPR is a major issue of concern for surveyed EU national libraries.

Almost all the surveyed libraries do not allow access to harvested material through an online application. They opt for access to such material made possible through their premises and for certain pre-defined scope. The novice of the TDM and/or Web harvesting operations seems to be the cause for the surveyed national libraries limited user-satisfaction inquiries regarding this library's new service.

When in the need of cooperation regarding the implementation of TDM and/or Web harvesting operations the surveyed national libraries replied that they turn to other libraries; a significant pool of EU national libraries' collaborators is the public administration, too. Databases of private entities and publishers of e-books are not connected to surveyed EU national libraries' TDM and/or Web harvesting operations, currently.

The two most significant categories of problems that the surveyed EU national libraries seem to be faced with are related to technical and legal problems.

8. Policies on Web Harvesting, Arrangement and Procedures

National libraries' policies and strategies on Web harvesting have practical and

theoretical perspectives. They are reflected on their priorities, procedures, partnerships and co-operations as well as on their concerns for organization and administration of the department or team responsible for the plan and implementation of each library's TDM and/or Web harvesting or Web archiving operations and service provided to library's constituents and users. Most EU national libraries' policies on TDM and/or Web harvesting are still under development given that national legislation in most EU Member States does not include any or detailed provision on data mining. From this point of view, the first field of the survey ("*Policies/Arrangement/Procedures*") is naturally linked to all subsequent fields up to the section of the participants' perspectives and final observations.

The first section of the survey contains five questions (1) The importance of Web collection, 2) The organization chart, 3) The use of quality filters, 4) Thematic crawl, 5) The purposes and the ways of web archived content use).

The importance of Web harvesting operation as one of the surveyed national libraries' new functions and services is shown in Figure 1. Web harvesting is reported to be one of the three most important purpose-specific functions of the surveyed libraries. The other two are cataloguing/indexation and collecting of digital files.

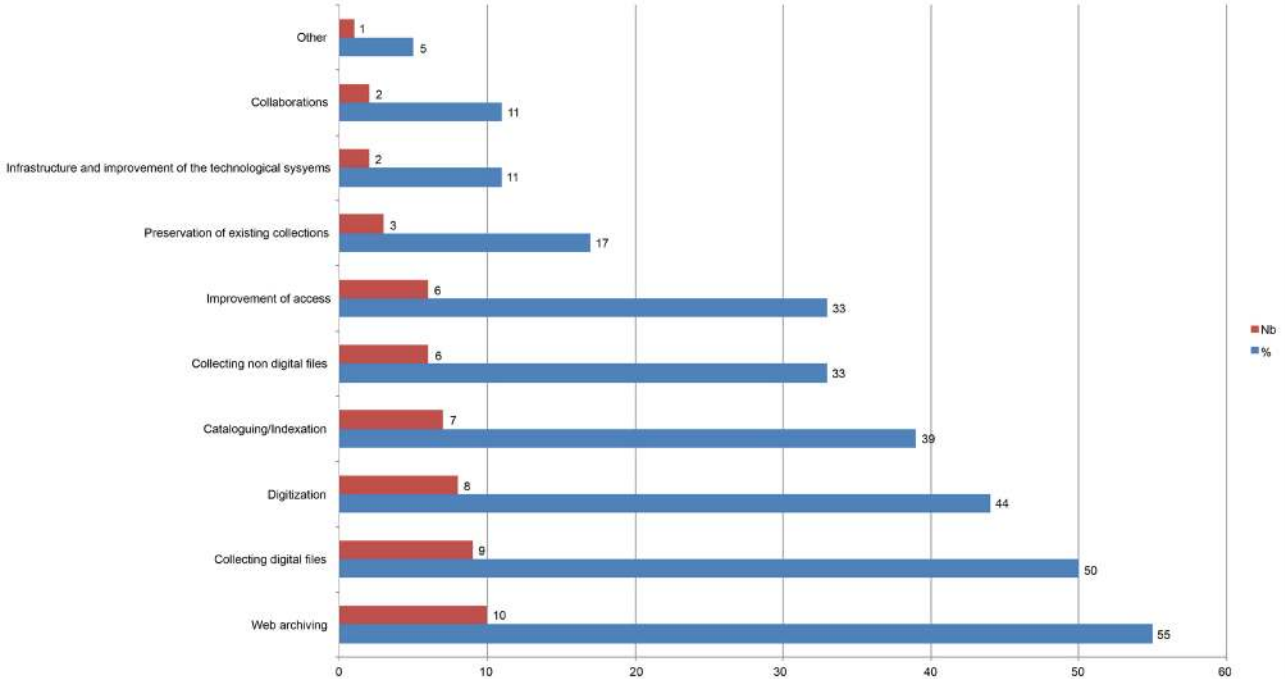

Figure 1. The importance of web harvesting.

The surveyed national libraries indicated the three most important current functions for them which include Web-harvesting as follows, in Table 3.

Regarding the operators in the library's organizational chart who are assigned with planning and running the national libraries' Web harvesting operation, the answers we received in our survey indicate that operators' number ranges from

Table 3. The three most important functions currently in surveyed group.

Answer options	%	Nb
Web archiving	55	10
Collecting digital files	50	9
Digitization	44	8
Cataloguing/Indexation	39	7
Collecting non digital files	33	6
Improvement of access	33	6
Preservation of existing collections	17	3
Infrastructure and improvement of the technological systems	11	2
Collaborations	11	2
Other	5	1
Answered question		18

one person, a librarian with multiple responsibilities and with the help of outsourced collaborators to a well organized working group such as in the case of the national libraries of the UK and Denmark. For example, in Denmark, the national library's organizational chart is described in national legislation (Schostag & Fonss-Jorgensen, 2012). The Royal Danish Library leverages one program manager, a full-time employee assigned with Web harvesting and archiving tasks, one operation manager, two IT specialists and two or three curators; they all report to the head of Digital Cultural Heritage department of the library. The National Library of France has a team of four people, two librarians and two technologists to run the Web harvesting and archiving services of the library. The National and University Library of Slovenia has one IT specialist and one developer working for the Digital Library Development Department, and one librarian working for the Digital Library Management Department, but none of them is a full-time Web-harvesting and archiving employee. The National Library of Greece has a team of three librarians and one IT expert to run the Web-harvesting and archiving operations of the library. The National Library of Spain employs two Web curators full-time and three IT specialists part-time. The National Library of Hungary has one team leader, one web-librarian, one web-curator, and two IT professionals working part-time. The National Library of Sweden has a team of three persons involved in the Web harvesting operation on a part-time basis (librarian, crawling engineer, and systems administrator). The Royal Library of Belgium is still in the research and development phase of its Web-harvesting operation, thus employs one full-time individual working on Web-harvesting and archiving. The National Library of Germany employs a team of four, i.e. three librarians and one IT specialist. The national library of Estonia has a team of four full-time employees, i.e. two librarians and two IT specialists, dealing directly with Web harvesting and archiving. Our research indicates that in most cases the persons involved with the Web harvesting are ei-

ther part-time employees or full-time employees assigned with the Web harvesting and archiving as part-time tasks.

During the web harvesting process, on line survey participants responded that they use quality filters as a percentage 73.7%, as is shown in **Figure 2**. The surveyed national libraries which replied negatively in the question about quality filters are the Austrian National Library, the Bibliothèque Nationale de Luxembourg, the National Szechenyi Library of Hungary, the National Library of Sweden, and the German National Library.

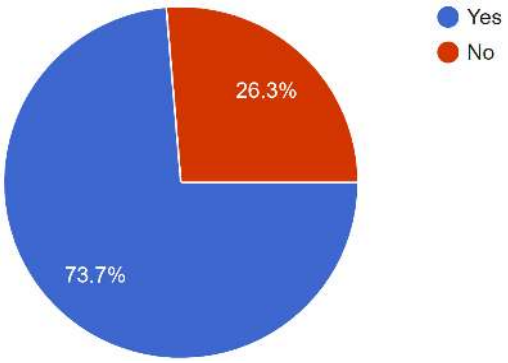

Figure 2. Use of quality filters on web harvesting.

Bibliographic and online research reveals that Web harvesting procedure is directly related to the kind of Web harvesting[10]. For libraries, the initial or following goal is to collect Web content at a broad national domain. Frequently they start with thematic harvesting to control the volume of the content before the implementation of broad mining (IIPC). The workflow can be defined by legislation and indicated by the organizational chart such as the example of Web harvesting in Denmark (Schostag & Fønss-Jørgensen, 2012).

Our survey revealed that thematic fields for Web harvesting may reflect relationships and interests of the public assumed by the national libraries. There are countries such as the United Kingdom in which there are many topics for thematic Web harvesting and the British Library considers the necessity of studying bibliography/Web sources to be intense. In addition to similarities, there are differences in thematic choices for Web harvesting and it would be interesting to explore further prioritization of thematic choices (e.g. what are the priorities and why, how countries' different identities are reflected in national libraries' choices about what they choose to be mined from the Web).

Our observations are due to the participants' responses. Initially, topics are mentioned that are, more or less, common or at least, among the first ones selected by EU national libraries. New thematic fields as surveyed national libraries future plans are presented in section related to "*Co-operation and Perspectives*."

According to our survey, the most common themes specified below:

[10]The different types of web harvesting is the national domain broad crawl, selective crawl which focuses on crawling specific types of web pages, thematic web crawling with crawling specific topic content, events crawl on specific events and/or unexpected (emergencies). Retrieved Aug 10, 2019, from http://palc24.cs.teilar.gr/conference/el/programma.jsp?id=12#a12 (see Papadopoulos et al., 2018)

- Elections and politics, government websites, state agencies, boards and authorities, research and educational institutions, other educational sites and cultural organizations, news websites and big world events such as the Olympics, are topics usually harvested by the national libraries.
- Literature and history include the most famous web harvesting themes, too. E-magazines, e-journals are also selected.
- Other common themes that were found to be top in the interests of three or more EU national libraries are: nature, environment and climate change, sports, religion, minorities, media and digital culture. Moreover, the British Library and the Austrian National Library archive web collections on women/Gender and Women Issues.
- The British Library has the largest variety of themes (almost 200) [11].
- Between the surveyed countries, France, Germany, Spain, Belgium, Finland, Denmark and Slovenia already have a significant range of thematic web harvesting collections.
- Examples of thematic areas that somehow differentiate in the surveyed EU national libraries' interests are: "Family History", "Transgender issues and Third Age", "Health issues (epidemics etc.)" in The British Library. Thematic crawling on Institutions/associations websites relevant to each society like sports associations and religious bodies are reported in Germany as well as harvesting of content in websites on specialized subjects (such as digital long-term preservation, biology, etc). The National Library of Spain harvests works on librarianship and computing science, applied science, popular heritage and on a big variety of themes ranging among art, media, gastronomy, etc. Among other thematic categories, topics about natural sciences, technology tourism, hobbies, traveling sports, etc. are selected in Slovenia; works on song and dance festival sites are harvested in Estonia. Collections about coins and medals, maps and plans are harvested by the Royal Library of Belgium among other subjects

The survey results in **Figure 3** show that the main purpose for harvesting from the web and for archiving the material found on it is to make it accessible to researchers (78.9%) followed by use for storage and preservation (73.7%). In some way, this research indication stands in contrast to the existing situation per allowances provided by law as in most cases access is only possible into the library premises and with its sole equipment. Digital copies are not permitted but only conventional copies are allowed and in many cases, the purpose for requesting access must be stated clearly by the user.

9. Technological Issues

This field consisted of two questions (third party provider, software). National libraries were asked if they leverage on third parties for technological expertise for their Web-harvesting operation. At a percentage of 57.9% of the surveyed

[11]Full list of thematic collection are on line available. Retrieved March 13, 2019 from https://www.webarchive.org.uk/en/ukwa/collection

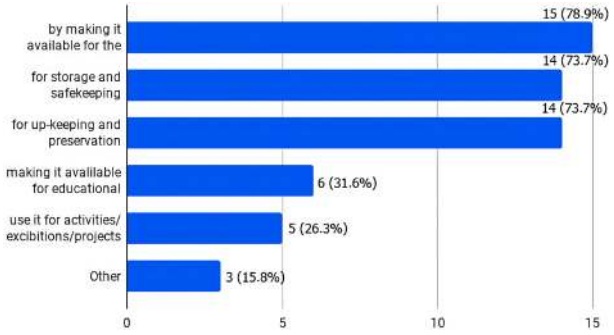

Figure 3. Uses of web archiving collections.

national libraries the answer was that they do not use a third party (technologist) for their Web-harvesting operation, as is shown in Figure 4.

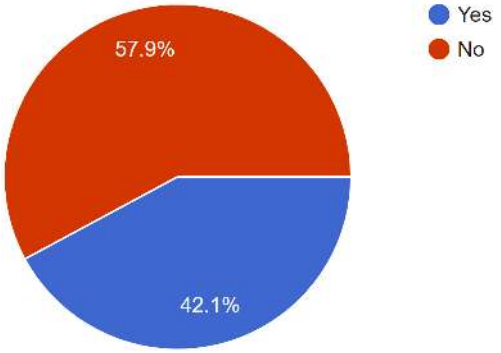

Figure 4. Use of third party (technologist) for web harvesting.

Most of the surveyed national libraries confirmed that they leverage on Heritrix as software program for Web-harvesting, as is shown in Figure 5.

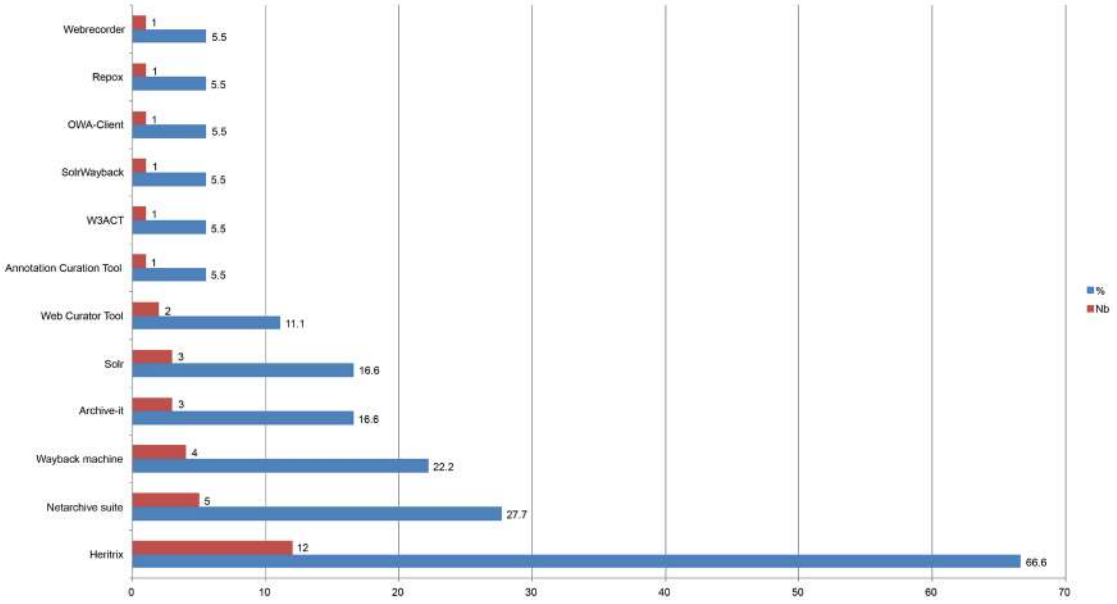

Figure 5. Software programs for web harvesting.

Most surveyed national libraries replied that they've used more than one software programs for their Web-harvesting operation. The answers we received per question on software programs which the surveyed national libraries leverage upon are the following:

- ✓ Archive-it of Internet Archive;
- ✓ Heritrix crawl engine, Annotation Curation Tool (curation software);
- ✓ Heritrix;
- ✓ Heritrix bundled with NetarchiveSuite;
- ✓ Heritrix 3, ArchiveIt, Webrecorder (as an experiment) NetarchiveSuite, Heritrix, Free text search using Solr, and Wayback. Developing search frontend and playback engine SolrWayback;
- ✓ Heritrix, Net Archive Suite, Open Wayback, SolR;
- ✓ Web Curator Tool, Heritrix;
- ✓ Heritrix web harvesting software (Our library is a member of IIPC);
- ✓ Heritrix (harvesting), SOLR (indexing), Wayback (search and representation);
- ✓ W3ACT (please see: https://github.com/ukwa/w3act);
- ✓ Repox Software;
- ✓ Proprietary software of the service provider;
- ✓ Heritrix with Net Archive Suit (NAS);
- ✓ Archive-It;
- ✓ Heritrix (and the Web Curator Tool);
- ✓ NetarchiveSuite and Heritrix;
- ✓ OWA-Client, developed by our service provider;
- ✓ Heritrix.

The decision to leverage on a third party as TDM and/or Web harvesting technology-provider depends in most cases of surveyed national libraries on the fact of library's available personnel for TDM and/or Web harvesting operations. When a single librarian was responsible for this kind of library's operations national libraries turned to a third party as TDM and/or Web harvesting solutions-provider.

Technological issues for TDM and/or Web harvesting are also connected with national libraries' future plans in this area of activities. Most surveyed national libraries described their concerns on technological issues per subject matter of TDM and/or Web harvesting through their interest in updating their own technology systems, in improving them, and in developing new tools for access to and retrieving of information (field "*Operations and Perspectives*"). Technology has a pivotal role in Web harvesting and this was clearly stated in the responses we received through the posed questions. Further study focused on technology issues for TDM and/or Web harvesting such as on improvements to existing software, combined with Web harvesting needs, new software development, complications and restrictions, etc., is an important field for research. National libraries' observations to an "*open*" question in relation to technological issues for TDM and/or Web harvesting indicate that technology is of primary impor-

tance for libraries in order to succeed in Web mining.

10. Consideration of Legal Issues

During the Web-harvesting process, the surveyed national libraries responded that they cater for author's prior consent at a percentage of 26.3%, as is shown in Figure 6. National libraries were asked if there is in place a procedure for securing authors prior consent, i.e. authors' consent before the execution of Web-harvesting and archiving processes by the library. The answers from most libraries indicate that there is not a prior consent mechanism in the Web-harvesting and archiving operations of the libraries.

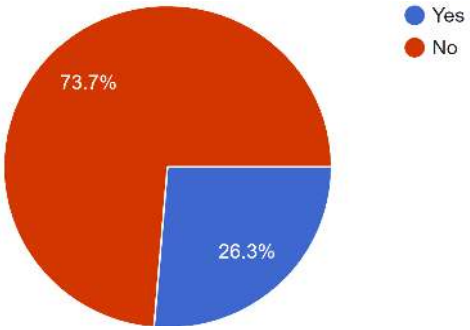

Figure 6. Prior consent of author for web harvesting.

The surveyed national libraries show more concern regarding data protection rights in comparison with intellectual property rights. National libraries were asked if their Web-harvesting systems cater for personal data protection as this issue is delineated in the General Data Protection Regulation (Regulation 2016/679/EU). According to the replies in our survey 57.9% of the national libraries replied that they do take care of means for authors' data protection in relation to their Web-harvesting and archiving operations, as is shown in Figure 7.

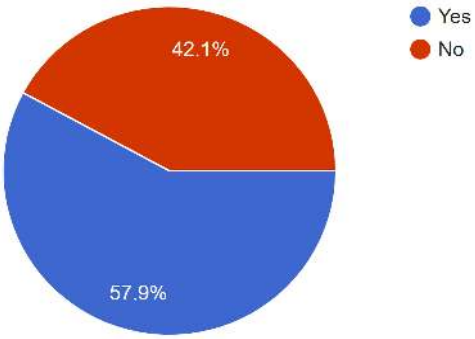

Figure 7. Secure data protection rights for web harvesting.

Regarding the protection of intellectual property rights in Web-harvesting operation, the replies which we received from the surveyed national libraries indicate that 52.6% is still an issue to consider, as is shown in Figure 8. Surveyed national libraries were asked if there's a provision of an application in their

Web-harvesting systems that could prevent the violation of right-holders' intellectual property rights.

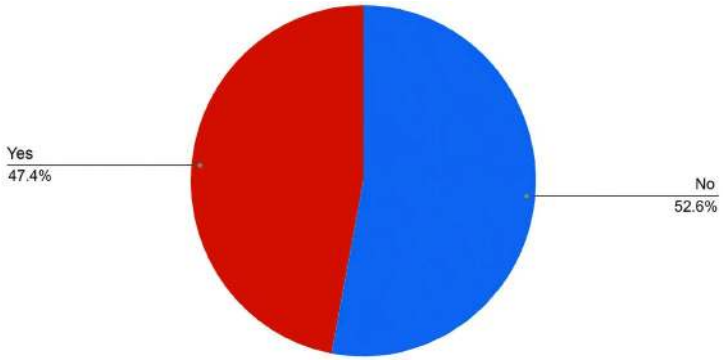

Figure 8. Protection of intellectual property right in web harvesting.

The survey included three (3) questions in the field of "*Legal issues*" regarding the protection of personal data and intellectual property by the parties concerned. The answers illustrate that there is space for development. In section "*Proposals and useful observations*" the survey recorded answers about the balance between the use of Web harvested and archived content and the exercise of legal rights especially in view of the new European Directive 2019/790/EU of 17 April 2019 digital property copyright (DSM) rights.

11. Access/Utilization

Surveyed national libraries were asked about the terms of making harvested works from the Web available to library-users. They replied that due concern is shown regarding access to and use of works harvested from the Web and archived accordingly. Among the replies which the national libraries gave in our inquiry per the terms of access to and use of works harvested from the Web and archived accordingly, are the following (the number in parenthesis represents how many libraries replied with the certain answer):

- ✓ Usually only inside the library in the research reading rooms (7).
- ✓ On legal deposit terminals with firewall (3).
- ✓ Only on Library premises to registered users (6).
- ✓ Available online with the specific permission of the website holder and publishers (5).
- ✓ Available online on the permission of National Library (1)
- ✓ The web archive is publicly available without restrictions. Intellectual property right holders can request their material to be accessible only on library premises (1).
- ✓ The archived websites are available for research purposes only (3).
- ✓ Only printing is permitted and not in all libraries (3).

During the Web-harvesting process, the surveyed national libraries responded that they have inquired library-users for user-satisfaction from their Web-harvesting service at a percentage of 26.3%, as is shown in **Figure 9**.

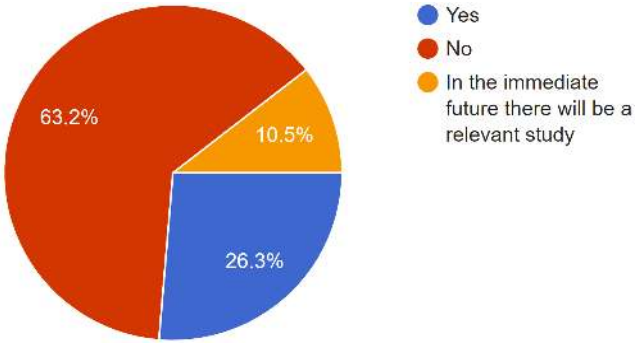

Figure 9. Users satisfaction inquiry for web harvesting.

Two (2) questions are related to this field on user-satisfaction: 1) what are the terms of use (who, where, how, to what extent) for making the Web material collected available to users; and 2) users' satisfaction inquiries by the national libraries). The answers complete the first field of the questionnaire ("*Policies of web harvesting/Arrangement/Procedures*") related to Web harvesting process that reflect the legislation. Another issue affecting access conditions and user satisfaction is technology.

12. Co-Operation & Perspectives

This field included four questions, two on co-operations (1) forms of co-operations, 2) connection with electronic public industry), and two on perspectives (3) immediate plans, 4) important problems). Future plans, which are described in this section, are complementary to the optional field of "*Proposals and useful observations*" but also to field 1 which focused on thematic Web harvesting by the participants in the survey.

Surveyed national libraries were asked about the forms of co-operation which they have developed regarding the utilization of the Web-harvesting results in their attempts. Most of them have sought the cooperation of other libraries—more experienced in web-harvesting libraries—while a significant number of them have turned to public administration for co-operation, as Figure 10 is shown below.

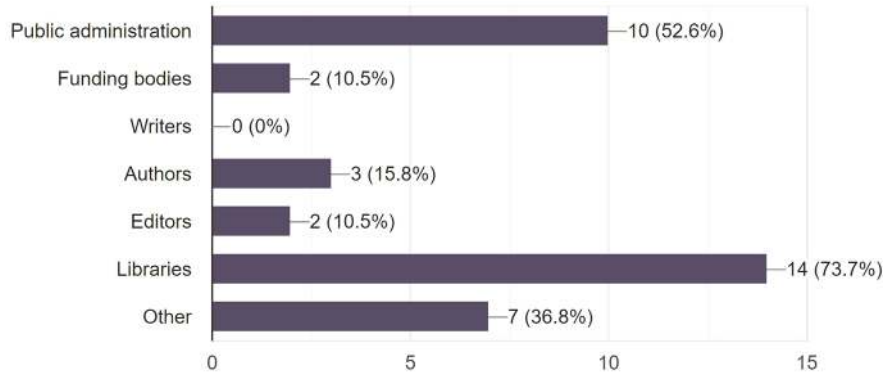

Figure 10. National Libraries cooperation for web harvesting.

We posed the surveyed national libraries with the question if there's a connection between their Web-harvesting system and the publishers of works in electronic format. Almost 95% of the surveyed national libraries replied negatively, as Figure 11 is shown. Further future research could answer questions like why this happens, what is the opinion of the stakeholders (publishers, authors, users, libraries) on possible connection between libraries and publishers of works in electronic format, what could change to make this connection possible, etc.

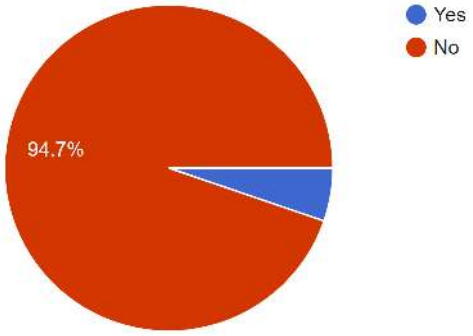

Figure 11. Connecting with publishers for web harvesting.

We asked the surveyed national libraries if they have any immediate plans for a new project on Web-harvesting. Most of them responded that they do have plans for new projects related to web harvesting, which pertain to:

- ✓ Integration of the web documents metadata in the National Library Service Catalog.
- ✓ Exploring using the web recorder tool to archive websites and push the WARCs gathered in this way into library's collection.
- ✓ More stakeholder involvement and projects related to raising awareness on web harvesting.
- ✓ Searching for use of new tools for harvesting content from social and streaming media platforms.
- ✓ Harvesting of press websites with paywall (an automated authentication of the crawler).
- ✓ Cooperation with the Internet Archive, in order to achieve better bulk harvesting.
- ✓ Upgrading library's services with the support of another software (MINT) which will enrich metadata during the harvesting process.
- ✓ Web harvesting of new thematic fields on digital music, climate change etc.
- ✓ Increasing the web harvested collections constantly.
- ✓ Modernizing and expanding the web harvesting environment, including the system used for access to harvested works where library will switch from an in-house system to an Open Wayback system.
- ✓ Social media harvesting depending on whether there will be funding,

We asked the national libraries about their opinion on the most important problem in their Web-harvesting operation. Figure 12 shows their answers.

Technical (42.1%) and legal (31.6%) problems stand out as the most important ones.

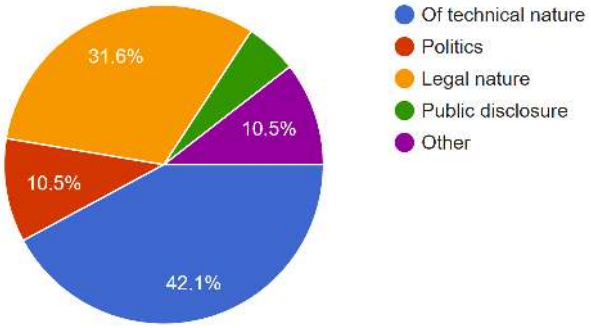

Figure 12. The most important problem for web harvesting.

13. Proposals and Useful Observations by Surveyed EU National Libraries

We asked the national libraries to make proposals and observations regarding the Web-harvesting operation. Of particular interest are the observations and suggestions from the library specialists involved in our research on the issues to be addressed in the near future.

Most national library experts recognize the necessity to continually improve technology in general (e.g. to extract material from large and dynamic web pages that are not yet satisfying or feasible with Heritrix).

Most national library experts consider that legal issues are always at the forefront of interest because the legislation is general and incomplete and allows only for limited access to content harvested from the Web. Library experts also noticed the necessity of protecting and securing their web collections. For example, the replies which we received in our research show that librarians are concerned about the risk of losing items in the library collections in the event of application by data subjects of the right to be forgotten or the right to privacy. There is a widespread belief among library experts that Legal Deposit Law should change in order to give wider access to harvested material from the Web.

With regard to libraries that are now taking their first steps in Web-harvesting their replies in our research shows that they prefer the development of small collections with works harvested from different websites initially (quality and variety is important for them); they consider the development of extensive collections subsequently and at a later stage in their Web-harvesting operation (quantity is not an immediate goal).

Improving technical infrastructures and tools comes at the forefront of upcoming library research projects along with expanding collections, a better description of web archives metadata and extracting pages on new topics and fields such as social media and live streaming. Those national libraries of EU Member States which are most experienced in web harvesting, aim at the extraction of materials from "*difficult*" websites such as complex websites and sites with pay

walls. Less experienced libraries aim at collaboration and co-operation development and awareness raising programs of their Web-harvesting operation.

14. Conclusion

Research on EU Member States national libraries Web-harvesting and archiving operations indicates that most national libraries consider their Web-harvesting and archiving operations to be important. Though they seem to have realized that the new EU legislation through Directive 2019/790/EU on Copyright in the Digital Single Market (DSM) creates a favorable legal foundation for the deployment of Web-harvesting and archiving operations through the national libraries of the EU Member States, they are still not fully and self-capable in executing widespread harvests of works. TDM technologies are the means for national libraries of EU Member States to make possible new areas of research, to enrich qualitatively and quantitatively their collections, and expand their digital services to consumers. However, legal and technological problems still linger and prevent libraries from deploying full resources in their Web-harvesting and archiving operations. Unless there is an amendment to national Copyright legislation and Digital Legal Deposit rules at a national level that meets the core of 2019 amendment of EU legislation through Directive 2019/790/EU national libraries of EU Member States will not be freed from the legal restrains that keep Web-harvesting and archiving limited in scope and implementation. Technological restrains seem less difficult to overcome. National libraries are experienced in outsourcing technological solutions when their own resources could not suffice for state-of-the-art Web-harvesting and TDM. Further research on TDM and GDPR issues is deemed necessary in consideration of EU national libraries significant concerns upon the effects of data protection regulation on Web-harvested and archived content.

Acknowledgements

The authors of this research are sincerely thankful to participants in the survey from the national libraries of EU Member States. We express our thanks to curators, IT specialists and the other librarians and information professionals of the national libraries which eventually replied to our research; knowledge-sharing among experts is the only means to find solutions to new problems and drive innovation in librarianship in the era of taming big data through TDM, Web-harvesting and archiving tools and processes.

Funding

This paper is composed within the framework of a research project titled "Web Archiving in Public Libraries and IP Law" within the framework of the Operational Program "Human Resources Development, Education and Lifelong Learning" of NSRF—Partnership Agreement 2014-2020 and is co-funded by Greece and the European Union—European Social Fund (Law 4314/2014 in ac-

cordance with the requirements of European Regulation (EC) 1303/2013).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Botti, M., Papadopoulos, M., Zambakolas, C., & Ganatsiou, P. (2019a). On the Eve of Web-Harvesting and Web-Archiving for Libraries in Greece. *Erasmus Law Review, 12,* 178-189.

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019b). Text and Data Mining in Directive 2019/790/EU. Enhancing Web-Harvesting and Web-Archiving in Libraries and Archives. *Open Journal of Philosophy (OJPP), 9,* 369-395. https://www.scirp.org/journal/paperinformation.aspx?paperid=94640 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3444204 https://doi.org/10.4236/ojpp.2019.93024

Botti, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2018). Legal and Technical Issues for Text and Data Mining in Greece. In *Computer Ethics—Philosophical Enquiry (CEPE) Proceedings*. https://digitalcommons.odu.edu/cepe_proceedings/vol2019/iss1/11

Caspers, M., Guibault, L., McNeice, K., Piperidis, S., Pouli, K., Eskevich, M., & Gavriilidou, M. (2016). *Barriers and Increasing Uptake of Text and Data Mining for Research Environments Using a Collaborative Knowledge and Open Information Approach* (p. 9). Baseline Report of Policies and Barriers of TDM in Europe (Extended Version of D3.3). https://cordis.europa.eu/project/rcn/197301_en.html

Chambers, S. (2018). *Investigating the Promise of a Belgian Web Archive*. https://www.slideshare.net/schambers3/investigating-the-promise-of-a-belgian-web-archive

Gomes, D., Miranda, J., & Costa, M. (2011). A Survey on Web Archiving Initiatives. In S. Gradmann, F. Borri, C. Meghini, & H. Schuldt (Eds.), *Research and Advanced Technology for Digital Libraries* (Vol. 6966, pp. 408-420). TPDL 2011, Lecture Notes in Computer Science, Berlin: Springer. https://doi.org/10.1007/978-3-642-24469-8_41 https://link.springer.com/chapter/10.1007/978-3-642-24469-8_41#citeas

International Internet Preservation Consortium. http://netpreserve.org

Internet Memory Foundation (2011). *Web Archiving in Europe*. A Survey Provided by the Internet Memory Foundation, 2010.

IRPC Internet Rights and Principles Coalition (2014). *The Charter of Human Rights and Principles for the Internet Educational Resource Guide* (4th ed.). https://www.ohchr.org/Documents/Issues/Opinion/Communications/InternetPrinciplesAndRightsCoalition.pdf

Jacobsen, G. (2008). Web Archiving: Issues and Problems in Collection Building and Access. *LIBER Quarterly, 18,* 366-376. https://doi.org/10.18352/lq.7936 https://www.liberquarterly.eu/articles/10.18352/lq.7936

Keskitalo, E. P. (2010). *Web Archiving in Finland*. https://www.doria.fi/bitstream/handle/10024/67051/webarchivingfinland_cdnl.pdf?sequence=1&isAllowed=y

Lee, B. T. (2006). *Linked Data*. https://www.w3.org/DesignIssues/LinkedData.html

Masanes, J. (2002). Towards Continuous Web Archiving. *D-Lib Magazine, 8*. http://www.dlib.org/dlib/december02/masanes/12masanes.html https://doi.org/10.1045/december2002-masanes

Miranda, J. (n.d.). *Web Harvesting and Archiving*. http://web.ist.utl.pt/joaocarvalhomiranda/docs/other/web_harvesting_and_archiving.pdf

Nielsen, J. (2016). *Using Web Archives in Research: An Introduction*. https://dighumlab.org/wp-content/uploads/2017/06/Nielsen_Using_Web_Archives_in_Research.pdf

Papadopoulos, M., & Botti, M. (2019). Text and Data Mining at the Forefront of the European Copyright Law. Investigating Its Implementation by Libraries in the EU and the Greek National Library. In *World Library and Information Congress*.

Papadopoulos, M., Zambakolas, Ch., Ganatsiou, P., & Botti, M. (2018). *Web-Harvesting Is Ante Portas of Greek Public and Academic Libraries*. http://palc24.cs.teilar.gr/conference/el/programma.jsp?id=12#a12

Ras, M., & Bussel, S. (2007). *Web Archiving Web Survey*. National Library of the Netherlands (Koninklijke Bibliotheek). https://www.kb.nl/sites/default/files/docs/kb_usersurvey_webarchive_en.pdf

Schostag, S., & Fønss-Jørgensen, E. (2012). *Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective*. https://doi.org/10.1515/mir-2012-0018 http://resaw.eu/wp-content/uploads/2013/06/Schostag_F%C3%B8nss-J%C3%B8rgensen_Webarchiving_Legal-Deposit-of-Internet-in-Denmark.pdf

Szabo, S. (2018). Website Archiving Guide Dutch National Archives. In *International Internet Preservation Consortium—Web Archiving Conference*. Wellington: IIPC. http://netpreserve.org/ga2018/wp-content/uploads/2018/11/IIPC_WAC2018-Suzi_Szabo-Webarchiving_Guideline.pdf

Toyoda, M., & Kitsuregawa, M. (2012). The History of Web Archiving. *Proceedings of the IEEE, 100,* 1441-1443. https://doi.org/10.1109/JPROC.2012.2189920 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6182575

Appendix

Web Archiving Survey

A survey on web archiving in the national libraries of the European Union

INITIAL QUESTIONS

About your library (please sent us: your Library's name, Country, Address, Contact details

Respondent identification (Name, Working Department, Status, Years of service, contact details)—Professional capacity of the responders:

MAIN PART

A) POLICIES OF WEB-HARVESTING/ARRANGEMENT/PROCEDURES

1) *Indicate the main three* (3) *current functions of your Library*

(Options: digitization/collecting non digital files/web archiving/collecting digital files/cataloguing-indexation/preservation of existing collections/infrastructure and improvement of the technological systems/improvement of access/collaborations/other)

2) *Indicate the operators who constitute the scientific team involved with the web-harvesting in your Library* (*the organization chart*)

3) *While harvesting the web pages do you raise quality filters or filters of reliability*?

4) *What are the thematic fields that the web-pages you are collecting with selective harvesting are included*? *In which new thematic categories are you planning to extent to in the future*?

5) *Being an institution, how do you make use of the material which is being collected by mining or crawling from the Internet*?

(Options: by making it available for the researchers/for storage and safekeeping/for up-keeping and preservation/making it available for educational purposes/use it for activities of the Library based on the collected material/Other)

B) TECHNOLOGICAL ISSUES

1) *Do you use a third provider of technological expertise for the web-harvesting*?

2) *Which is the basic software program that you use for harvesting purposes*?

C) LEGAL ISSUES

1) *Is there an updating or giving previous consent procedure from the part of the authors/creators as far as their collected works are concerned*?

2) *If your Library uses web harvesting system, does it provide for individuals to exercise citizen rights concerning personal data protection as these rights are described in the New General Data Protection Regulation* (*Regulation* 2016/679/*EU*)?

3) *In the web harvesting system that your Library uses, is there provision of an application which could prevent violation of the creators' and/or beneficiaries' intellectual property rights regarding their works which are on line in the Internet*?

D) UTILIZATION/APPLICATION OF THE WEB HARVESTING MATERIAL

1) *Which are the terms of making the web material collected available to us-*

ers? (*Who is entitled to have access*/*for what purposes*/*the access areas*)

2) *Have you checked the level of satisfaction, from the part of the users, regarding web harvesting material*?

E) CO-OPERATION & PERSPECTIVES

1) *Which forms of co-operation has your Library developed regarding utilization of the web harvesting results*?

2) *Is there a connection established between the web harvesting system and the publishing production and availability of the works in electronic form, which come from the editors' databases*?

3) *Have you any immediate plans regarding a new project on web harvesting*?

4) *Which is the most important problem in your Library as far as the web harvesting is concerned*?

F) OBSERVATIONS

From your experience please make proposals and useful observations regarding web harvesting (*optional*).

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
Επιχειρησιακό Πρόγραμμα
Ανάπτυξη Ανθρώπινου Δυναμικού,
Εκπαίδευση και Διά Βίου Μάθηση
Ειδική Υπηρεσία Διαχείρισης
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη