Information Law and Policy Lab

Web harvesting by cultural heritage institutions

Towards adequate facilitation and regulation of web harvesting digital content in order to preserve national cultural heritage

Luna Schumacher, Stefan van Kolfschooten & Daniël Soons

Information Law and Policy Lab

Amsterdam, August 26, 2020

This position paper is written by Luna Schumacher, Stefan van Kolfschooten and Daniël Soons of the Glushko & Samuelson Information Law and Policy Lab (ILP Lab) of the Institute for Information Law (IViR) of the University of Amsterdam. The ILP Lab is a student-run, IViR-led institution which develops and promotes research-based policy solutions that protect fundamental rights and freedoms in the field of European information law.

It has been written in partnership with the Koninklijke Bibliotheek (National Library of the Netherlands) and Nederlands Instituut voor Beeld & Geluid (Netherlands Institute for Sound & Vision). This position paper solely reflects the recommendations and conclusions of the authors of the ILP Lab. We are very grateful for the input we received from Annemarie Beunen, Kees Teszelszky, Jesse de Vos, Maartje Hülsenbeck, Martijn Simon, Peter Scholing, Gaby Wijers, Leontien Bout, Sophie Ham and Stef van Gompel.

Table of contents

Executive summary

As one of few countries, the Netherlands has no legal provision enabling cultural heritage institutions (CHIs) to engage in web harvesting for the purpose of collection building. The need for this type of regulation however has been expressed by several CHIs and has also been acknowledged in several European and international policy documents. It is in the public interest that our digital history and heritage be collected and preserved for future generations, without infringing upon any individual intellectual property rights. This position paper provides an overview of the most important issues that need to be taken into consideration when creating web harvesting legislation for CHIs.

Chapter 2 explores the web harvesting legislation of six selected countries with the purpose of presenting how different countries define web harvesting concepts and how such definitions affect the execution of web harvesting activities by CHIs. This chapter reveals that a legal basis for web harvesting is typically laid down in legal deposit legislation and that legal deposit legislation and legislation enabling web harvesting is very common in countries around the world.

Chapter 3 provides a more in-depth analysis of key issues relevant to web harvesting legislation, focusing on the actors and the type of content which should or may be subject to possible regulation. The authors suggest that the number of CHIs to which web harvesting legislation is addressed be limited to only those with a legal task, while other smaller CHIs can share expertise on what content is important to preserve. The designated CHIs should be allowed to harvest various types of content in light of their collection plans or content strategies, as long as the content is part of the Dutch domain and is publicly accessible.

These conclusions lead to two different suggestions for introducing provisions into Dutch law that provide a legal basis for CHIs to harvest web content without infringing intellectual property rights. As described in Chapter 4, web harvesting can be facilitated by introducing a new copyright exception or by creating a specific provision in legal deposit legislation. There is a high need for regulation enabling web harvesting by CHIs, despite its complexity and the multiple dimensions to be considered. Each day the Netherlands lacks legislation on web harvesting, more aspects and content of our collective digital heritage will be lost. The make-shift solutions that Dutch CHIs are currently utilizing do not suffice. The authors call upon the Dutch legislator to take swift action and help our national institutions to save our society's digital footprint.

1. Introduction

A great part of our lives is spent on the internet. About 93% of the Dutch population owns a smartphone,[1] and the right to access and use the internet is acknowledged universally.[2] As "our history of the present day is written online", future researchers and academics will have to rely on our digital presence to study our age.[3] It is therefore important to collect and preserve a clear and complete digital landscape. This has been recognized in the UNESCO Charter on the Preservation of Digital Heritage[4] and the Recommendation 2011/711/EU on the Digitisation and Online Accessibility of Cultural Material and Digital Preservation.

At present, however, Dutch law does not offer the possibility to archive or preserve this digital part of our society. The Netherlands has no legal deposit legislation, which in many other countries offers a legal basis for web harvesting. Moreover, web harvesting involves several acts which are restricted by copyright, related (neighbouring) rights and/or database rights. Hence, to comply with the law, harvesting entities need to secure individual permission from all the right holders of websites and web content involved. This severely hampers the possibilities of acquiring a complete and insightful web archive.

Cultural heritage institutions (**CHIs**) have a key role in preserving today's memories for future generations. The task assigned to the *Koninklijke Bibliotheek* (National Library of the Netherlands , **KB**) includes taking care of the national library collection and encouraging the development of national facilities in the field of librarianship and information systems.[5] The mission of the *Nederlands Instituut voor Beeld en Geluid* (Netherlands Institute for Sound and Vision, **Sound and Vision**) is to improve people's lives in and with the media by archiving, exploring and contextualising it, to which the freedom of thought and expression in text, image and sound is paramount.[6] Every Dutch citizen has the right to access information and culture, which means everyone must be able to utilize the digital and physical services of the public library.[7]

To be able to effectively fulfil their roles in our digital society, it is important that the responsibilities of CHIs, such as the KB and Sound and Vision, also extend to the online aspect of our society. As shown in the KB's new collection plan, the definition of a library collection has changed over the years, and new digital born content such as websites, podcasts and databases are to be seen as sources of knowledge and culture.[8] The KB and Sound and Vision therefore want to substantially expand their web archives.[9]

---

[1] Smartphone use in the Netherlands, Deloitte report 2019: https://www.consultancy.nl/nieuws/15292/smartphonebezit-gegroeid-naar-93-van-nederlanders-veelvuldig-gebruik-storend (last accessed on 28 August 2020).
[2] See e.g. UN Resolution A/HRC/32/L.20 on the promotion, protection and enjoyment of human rights on the Internet; E. Boyle, 'UN declares online freedom to be a human right that must be protected', *Independent* 5 July 2016; and N. Kivits, 'Is het web een middel of een recht', *Financieele Dagblad* 7 May 2016. See for EU-wide recognition and acknowledgment ECtHR 18 December 2012, *no.* 3111/10, ECLI:CE:ECHR:2012:1218JUD000311110 (Yildirim v. Turkey).
[3] https://www.kb.nl/organisatie/onderzoek-expertise/e-depot-duurzame-opslag/webarchivering
[4] UNESCO Charter on the Preservation of Digital Heritage of 15 October 2003. Accessible via: https://unesdoc.unesco.org/ark:/48223/pf0000133171.page=80" \h.
[5] Art. 1.5 (2) Higher Education and Research Act (Wet op het Hoger Onderwijs en Wetenschappelijk Onderzoek, WHW): "*De Koninklijke Bibliotheek is als de nationale bibliotheek werkzaam op het gebied van het bibliotheekwezen en de informatieverzorging, zowel ten behoeve van het hoger onderwijs en het wetenschappelijk onderzoek als ten behoeve van het openbaar bestuur en de uitoefening van beroep of bedrijf. In dat kader draagt zij in elk geval zorg voor de nationale bibliotheekverzameling, bevordert zij de totstandkoming en instandhouding van nationale voorzieningen op het vorengenoemde gebied en bevordert zij de afstemming met de overige wetenschappelijke bibliotheken*."
[6] See: https://www.beeldengeluid.nl/organisatie/missie-en-visie (last accessed on 19 February 2020).
[7] *Kamerstukken II* 2013/14, 33846, nr. 3, p. 12.
[8] Content Strategy Koninklijke Bibliotheek (not publicly accessible, however the authors were authorised to research the KB's Content Strategy).
[9] See e.g. Response of the Koninklijke Bibliotheek to the Dutch implementation proposal of Directive (EU) 2019/790. Accessible via: https://www.internetconsultatie.nl/auteursrecht/reactie/99554a45-f799-4685-b533-e052ddc3508e.

The Dutch government also recognizes the role of the KB and Sound and Vision as entities for web archiving. The KB and Sound and Vision are mentioned by the Inspectorate for Cultural Heritage of the Ministry of Education, Culture, and Science as possible partners to the Nationaal Archief (National Archive, **NA**) in helping to archive all government websites and social media channels, which the Inspectorate recommends the government to do in the interest of the public.[10] The Minister of Education, Culture and Science, J. (Jet) Bussemaker, has endorsed these recommendations in a letter to Parliament.[11]

Such recognition alone is not enough, however. Without legislation permitting web archiving, the KB and Sound and Vision need to obtain prior authorization from right holders to copy and store websites and web content. They now try to seek permission from right holders of websites and web content on a case-to-case basis, with a possibility for right holders to opt-out and exclude their works from the web archive.[12] This proves to be a time-consuming and labour-intensive strategy. As a result, of the roughly 5.914.650 registered .nl ccTLD's,[13] the KB has only harvested 16.000 URL's in the period between 2007 and 2020. Yearly, about 1500 websites are added to its web archive. The Sound and Vision now has an archive of about 250 websites relating to broadcasting and media. All Dutch web archives together have a collection of 20.000 URL's. This means that only a small percentage (<1%) of .nl ccTLD's are harvested and archived.

By contrast, web harvesting is more common in other EU member states, such as Denmark and Germany, where it is explicitly permitted. Outside of the EU, web harvesting is executed on an even larger scale. The United States, for example, is home to the Internet Archive, which is the oldest and biggest web archive in the world.[14] The Internet Archive is not limited to the U.S. web domain, but also includes many crawls of .nl websites. It is unclear how many webpages it contains exactly, but numbers vary from 300 billion to 411 billion webpages. In no way does this compare to the significantly smaller Dutch archiving projects.

To assure that our own national institutions can crawl and archive websites and web content relevant to our cultural and societal needs and to make sure that future researchers and academics are not dependent on foreign archives, it is important to introduce a legal instrument to facilitate national web harvesting.

In this position paper, the Information Law & Policy Lab of the University of Amsterdam, in partnership with the KB and Sound and Vision, has examined ways to legally facilitate web harvesting by CHIs. This paper provides an overview of issues that need to be taken into consideration when implementing a legal basis for web harvesting and discusses two possible courses for legislative action. The (legislative) suggestions made in this position paper provide a visualization of possible answers to certain questions that arise when considering the implementation of a legal basis for web harvesting. These suggestions are not meant to be read as the sole solution to certain policy dilemmas.

This paper proceeds as follows. Chapter 2 is a legal comparative analysis offering an insight into harvesting legislation in six selected countries. Chapter 3 discusses four main issues regarding the implementation of

[10] Webarchivering bij de centrale overheid: Het archiveren van websites en uitingen op sociale media, Report Erfgoedinspectie, November 2016, p. 4. Accessible via: https://www.inspectie-oe.nl/binaries/inspectie-oe/documenten/rapport/2016/12/8/rapport-webarchivering/Rapport+Erfgoedinspectie+webarchivering+bij+de+centrale+overheid.pdf, p. 11 and 34.
[11] *Kamerstukken II* 2016/17, 29362 nr. 257. Accessible via: https://zoek.officielebekendmakingen.nl/kst-29362-257.html (last accessed on 13 July 2020).
[12] T. Schiphof, M. Ras, E. Cameron, A. Beunen, 'KB kiest voor pragmatische opt-out aanpak', InformatieProfessional 2007, nr. 10.
[13] These numbers have been last checked on 17 February 2020. SIDN stats accessible via: https://stats.sidnlabs.nl/nl/registration.html
[14] https://archive.org.

web harvesting legislation, namely definition questions, the addressees of such a provision, the scope of harvestable content (type and accessibility) and the delineation of the Dutch domain. In Chapter 4, two possible legal options are presented, which can help shape a future legislative initiative.

2. Web harvesting in other countries

In order to find ways in which web harvesting could be made possible in Dutch law, it is useful to examine the legislation of other countries to see how they define different concepts relevant for web harvesting activities. This chapter explores web harvesting legislation in three EU member states and three countries outside the EU: Australia, Denmark, France, Germany, New Zealand and the United Kingdom.

2.1. Australia

The National Library of Australia (**NLA**) states that it has a "mandate and commitment to preservation and has been active in developing infrastructure to collect, manage, preserve and keep the digital collections available into the future".[15] Given its mandate under the 1960 National Library Act to build a comprehensive collection of Australian published materials, collecting online resources has been a necessary extension of the NLA's collecting responsibilities. This led to amendments to the Australian Copyright Act in 2016, extending the national legal deposit requirements to electronic publications, including both offline and online materials. Online material must be provided to the NLA upon request. This includes automated requests made through web harvesting software. Publishers can use the library's eDeposit service to deposit material that is not available on a public website. Website publishers receive notice of the web harvesting process and statutory authority for the request.[16] Harvested websites are publicly available on the electronic archive platform PANDORA.

Section 195CD (1) of the Australian Copyright Act requires a deposit of a copy of the whole work. This includes any illustrations, engravings, photographs, audio-visual elements and, in the case of material available online, in the form in which it was made available online, free from any technological protection measures. In order to be constituted as a complete copy, additional elements like embedded computer scripts, programs and software that are necessary to provide and render the style, presentation and functionality of the work as published could be required. Where literary, dramatic or artistic works include other media elements, such as sound recordings or video, which are an intrinsic part of the work, all media elements should be deposited as part of the work.[17]

The law distinguishes between electronic library material that is 'available online' or 'not available online', to which different deposit requirements apply. Material 'available online' is communicated on (or via) the internet. The NLA refers to 'not online' material as 'offline' material. The NLA's guide for publishers on the deposit of electronic materials explain these terms further:

> *a) Electronic material 'available offline' is distributed on a physical format carrier and supplied to the public (i.e. in a published form), whether for sale or free, by a person in Australia who is the publisher.*
> *b) Electronic material 'available online' is material made available to the public in Australia whether for sale or free, via the internet or some other platform. This includes:*
> *i. Material published on a website within the '.au' top level domain name; or*
> *ii. Material published on a website where the domain name is owned or licensed by an Australian resident;*

[15] Digital Preservation, NLA, https://www.nla.gov.au/content/digital-preservation (last accessed on 12 July 2020).
[16] K. Buchanan, *Digital Legal Deposit in Selected Jurisdictions,* The Law Library of Congress: July 2018, p. 5.
[17] NLA, Deposit of electronic publications with the national library of Australia, June 2016, https://www.nla.gov.au/sites/default/files/deposit-of-electronic-publications.pdf, p. 4-5.

iii. Material accessible online on a website within Australia, where the Director-General of the Library or delegate considers the material should be included in the national collection; or,
iv. Material published on the internet, other than websites, in Australia or by an Australian resident.[18]

The Australian Copyright Act permits the NLA to notify the right holders of a publication using a web harvesting robot. Material by Australians or about Australia can be published anywhere in the world, since online publishing is not constrained by geographic location due to the nature of the internet. Section 195CC of the Act enables the Director-General or a delegate to request any material in which copyright subsists under the Copyright Act 1968 to be deposited with the NLA. This includes works that are supplied via hosting or online publishing services by Australian residents outside Australia as well as works that meet the NLA's national collecting objectives by non-Australian residents.[19]

2.2. Denmark

The Danish web archive *Netarkivet*, established in 2005, aims to collect and preserve the Danish part of the internet. It is a joint venture between the Royal Library and the State and University Library.[20] The legal framework for its activities is the Danish legal deposit law.[21] According to this law, Danish material published in electronic communication networks is subject to legal deposit. The Act states that "the legal deposit obligation is fulfilled by the legal deposit institution having access to, request or produce copies of the material".[22] It further notes that "the institutions are entitled to produce copies of the material with a view to collecting and storing. In so far as access to legally deposited material is not restricted pursuant to other legislation, the institutions are entitled to make it available to the general public within the framework of the Copyright Act".[23]

The Royal Library of Denmark has a long tradition of collecting material that is printed outside the borders of the nation but that is aimed at a Danish audience or treated themes of relevance for a Danish readership. This material is called "Danica".[24] The Netarkivet also collects material published in electronic communication networks when it is published (1) from internet domains that are specifically assigned to Denmark, or (2) from other internet domains if the material is directed at a public in Denmark. The Minister of Culture has laid down detailed rules for the delimitation of the legal deposit obligation according to (1) and (2).[25] Material from other domains aimed at the Danish public is (semi-)manually tracked down based on the criteria whether the material is written in Danish, the person registered as the owner of a domain name is a Danish resident, the material concerns Danish affairs, the author is a Danish citizen or whether the performing artists are Danish.[26]

[18] NLA, Deposit of electronic publications with the national library of Australia, June 2016, https://www.nla.gov.au/sites/default/files/deposit-of-electronic-publications.pdf, p. 5-6.
[19] NLA, Deposit of electronic publications with the national library of Australia, June 2016, https://www.nla.gov.au/sites/default/files/deposit-of-electronic-publications.pdf, p. 6.
[20] N. Brügger, in: *The Routledge Companion to Global Internet Histories,* ed. G. Goggin & M. McLelland, New York: Routledge 2017, p. 66.
[21] Danish Act on Legal Deposit of Published Material 2004.
[22] *Ibid.,* § 18.
[23] *Ibid.*, § 19(3).
[24] N. Brügger, in: *The Routledge Companion to Global Internet Histories,* ed. G. Goggin & M. McLelland, New York: Routledge 2017, p. 64.
[25] Danish Act on Legal Deposit of Published Material 2004, § 8.
[26] N. Brügger, in: *The Routledge Companion to Global Internet Histories,* ed. G. Goggin & M. McLelland, New York: Routledge 2017, p. 64.

With these criteria, the Danish national web domain consists of all material published on the national country code top-level domain name assigned to Denmark (.dk) and material published on other domain names aimed at a Danish audience thereby covering a wide array of information.

2.3. France

On 1 August 2006, a new copyright law was accepted by the French Parliament. Title IV (dépot legal) initiated a change to the Code du Patrimoine, extending the legal deposit to the internet. Internet legal deposit now applies to "all types of publications disseminated on the internet: institutional or personal websites, free or paid-access periodicals, blogs, commercial websites, video platforms or digital books."[27] The CdP explicitly states that it derogates from the French copyright law. The adoption of the new copyright law subjects everything that is published on the internet in France to legal deposit. The legal deposit obligation thus covers websites registered under a ".fr" top-level domain and websites edited by persons or organizations domiciled in France. The National Audiovisual Institute (**INA**) collects websites related to audiovisual productions (mostly radio and TV) and the National Library of France (**BnF**) collects all other websites.[28]

The BnF uses open-source crawler-bot software to conduct bulk automatic harvesting and focused crawls. Bulk harvests collect snapshots of websites belonging to the French domain. Focussed crawls are based on a selection of sites and can be centred on a particular event, like an election, or a specific theme, such as blogs. The BnF may contact the website editor to find technical solutions on a case-by-case basis if, at the moment of capture, content is found to be inaccessible due to technical or commercial reasons (see para. 3.4).[29]

2.4. Germany

The German National Library has a legal mandate to collect, index and preserve every written word or piece of music that is related to Germany since 1913.[30] The mandate was amended in 2006, tasking the German National Library to responding more actively to digital developments. Since then, the mandate includes the collection of online publications such as media works. It is considered that all content disseminated on the publicly accessible part of the internet falls within the collection mandate, including collecting websites.[31] The German National Library initially only focused on digital versions of existing physical publications (monographs (e-books) and university publications (such as online doctoral dissertations), before expanding its activities to other online publications, such as e-papers and e-serials.[32]

The first web crawl took place in 2012 after preparing for two years. The German National Library only collects selected websites whose preservation is in the public interest, which may include news websites, but also forums and blogs. Since websites are subject to constant change, the harvesting is repeated on a regular basis. The addresses of websites, the depth of the collection, and the frequency of runs are determined on a case-by-case basis and entered manually while the harvesting itself is automated.

[27] N. Boring, *Digital Legal Deposit in Selected Jurisdictions,* The Law Library of Congress: July 2018, p. 28.
[28] V. Schafer, *Exploring the "French web" of the 1990s,* New York: Routledge 2017, p. 157.
[29] N. Boring, *Digital Legal Deposit in Selected Jurisdictions,* The Law Library of Congress: July 2018, p. 28.
[30] Collection mandate of the German National Library, accessible via: https://www.dnb.de/SharedDocs/Downloads/EN/Ueber-uns/zumSammelauftragEN.pdf?__blob=publicationFile&v=2
[31] Para. 2(1)(b) jo para. 3(3), Gesetz über die Deutschhe Nationalbibliothek, accessible via: http://www.gesetze-im-internet.de/dnbg/index.html.
[32] J. Gesley, *Digital Legal Deposit in Selected Jurisdictions,* The Law Library of Congress: July 2018, p. 34.

The collection mandate is assumed to include the possibility of web crawling by the German National Library. Periodic harvesting of all ".de" domains, was however prohibited until March 2018. The German Copyright Act only allowed the Library to save online publications on a first and one-time basis only. Repeated retrieval was deemed an extension of existing archival contents and therefore a violation of German copyright law. The legislature therefore proposed amendments to the Copyright Act and National Library Act in 2017. These amendments granted the German National Library the right to automatically and repeatedly harvest works that fall under its collection mandate.[33] Since then, the Library is entitled to crawl, archive and retrieve websites even without requesting permission from the respective right holders.

2.5. New Zealand

Pursuant to the 2003 National Library of New Zealand Act (**NLNZ Act**), the Minister may authorize the National Library "to make a copy, at any time or times and at his or her discretion, of public documents that are internet documents in accordance with any terms and conditions as to format, public access, or other matters that are specified in the notice."[34] Extracted from the NLNZ Act, Buchanan describes that "an *internet document* is a public document that is published on the Internet, whether or not there is any restriction on access to the document; and includes the whole or part of a website.[35] A *public document* is a document of which one or more copies are issued to the public, available to the public on request, or available to the public on the internet, and that is printed or produced in New Zealand, or commissioned to be published in another country by a New Zealand resident or business, and in which copyright exists under the Copyright Act 1994.[36] An *electronic document* is a public document in which information is stored or displayed by means of an electronic recording device, computer, or other electronic medium, and includes an Internet document."[37] The definition of "Internet document" in the NLNZ Act, together with the definition of "public document," essentially means that the National Library can harvest any website produced or hosted in New Zealand without seeking permission from the publisher or website owner. Therefore, a 2006 Notice simply states that "the National Librarian is authorized to copy any Internet document."[38]

The National Library has been selectively harvesting websites since 1999 and allows people to nominate websites for harvesting.[39] Pacific Island websites or websites of New Zealanders that are published outside of the .nz domain are also included in the process. For these overseas websites, a permissions process is followed as they are not covered by the legal deposit legislation.[40] Since 2008, specific harvests of the .nz domain have been undertaken by the National Library on repeated intervals, usually every couple of years.[41]

Since 2015, the National Library has also adopted "whole of domain harvesting" for the .nz domain.[42] The domain harvests are currently unavailable to the public. The NLNZ's "whole of domain harvest" takes a 'snapshot' of the whole .nz domain as it exists on the web during the time of harvesting, thereby

[33] J. Gesley, *Digital Legal Deposit in Selected Jurisdictions,* The Law Library of Congress: July 2018, p. 34.
[34] Article 31(3) NLNZ Act 2003.
[35] Article 29(1) NLNZ Act 2003, K. Buchanan, *Digital Legal Deposit in Selected Jurisdictions*, The Law Library of Congress: July 2018, p. 49-50.
[36] Article 29(1) NLNZ Act 2003.
[37] Article 29(1) NLNZ Act 2003.
[38] K. Buchanan, *Digital Legal Deposit in Selected Jurisdictions*, The Law Library of Congress: July 2018, p. 50 and 53.
[39] *Ibid.*, p. 49.
[40] *Ibid.*, p. 53.
[41] *Ibid.*, p. 53.
[42] *Ibid.*, p. 49.

recognizing the importance of the internet in all areas of New Zealand society and culture. The National Library states that the technical parameters for the harvest were developed after consultation with the public and internet stakeholder groups. The parameters include (1) websites that fall under the .nz country code, (2) websites that fall under .com, .net and .org that can be programmatically determined to be hosted on machines that are physically located in New Zealand and (3) selected websites based overseas that are covered by the provisions of the NLNZ Act.[43]

2.6. The United Kingdom

In 2013, the Legal Deposit (Non-Print Works) Regulations 2013 (**Regulations**) entered into force on the basis of the UK Legal Deposit Libraries Act 2003 framework, under which regulations could be introduced to extend the deposit requirements. The Regulations extended the deposit obligation to non-print materials to enable the legal deposit libraries to build and preserve a "national collection of e-journals, e-books, digitally published news, magazines and other types of content."[44] The Regulations cover offline and online content, including online content that can be obtained through web harvesting, but specifically *exclude* works that contain personal data and that are available only to a restricted group, such as information provided on a social media site with restricted access (e.g. closed groups on Facebook or protected tweets). Publicly available materials on such sites are included within the remit of the Regulations. Also excluded from the scope of the Regulations are works that predominantly consist of film or recorded sound, or material that is incidental to this, and works published prior to the Regulations entering into force, where it concerns electronic materials that cannot be copied but must be requested.[45]

For material to fall *within* the Regulations, a work must be published in the UK. This occurs when (1) it is made available to the public from a website with a domain name which relates to the UK or to a place within the UK, or (2) it is made available to the public by a person and any of that person's activities relating to the creation or the publication of the work take place within the UK. A work published online shall not be treated as published in the UK, if it is only made accessible to persons outside the UK.[46] In practice, this includes all .uk websites, plus websites in potential future geographic top-level domains that relate to the UK such as .scotland, .wales or .london. Websites whose domain name mentions a UK place within a generic top-level domain or another country's geographic top-level domain (e.g., hypothetically, www.oxford.com or www.london.tv) are only treated as being published in the UK if they fit criterion (2).

The Regulations permit deposit libraries to harvest the web to obtain a copy of relevant online materials.[47] Some restrictions apply over how deposit libraries may subsequently handle the materials obtained. To ensure that copyright is not infringed, the Regulations provide that deposit libraries may make copies of non-print materials (1) for preservation purposes, (2) for research purposes and to enable access to visually impaired individuals, and (3) if another copy is not otherwise commercially available. Copies may be preserved in a different form or medium than the original deposit. Deposit libraries may also dispose of deposited materials by destroying them, provided they keep one copy of all relevant material in the most suitable version for the purposes of preservation. The Regulations contain more detailed rules on the subsequent copying of non-print materials by deposit libraries for any of the other purposes, and on the liability for copyright infringement in case of a misuse of collected materials by users.[48]

[43] K. Buchanan, *Digital Legal Deposit in Selected Jurisdictions*, The Law Library of Congress: July 2018, p. 53-54.
[44] C. Feikert-Ahalt, *Digital Legal Deposit in Selected Jurisdictions,* The Law Library of Congress: July 2018, p. 66.
[45] *Ibid.*, p. 67-68.
[46] *Ibid.*, p. 68.
[47] *Ibid.*, p. 69.
[48] *Ibid.*, p. 71.

2.7. Section Summary and conclusion

Legislation enabling web harvesting is common around the world. Several aspects of this legislation can provide useful starting points when considering a Dutch legislative initiative. For example, it is noticeable that most countries enable web harvesting in deposit legislation. Furthermore, a typical definition of the "national domain of a country" includes all websites that are registered at a domain inside the assigned country-code, or that are hosted at an IP-address that belongs to a segment assigned to that country. The Danish definition is helpful because it also includes content that is aimed at the inhabitants of a nation, thus encapsulating more content relevant for demarcating the cultural heritage of a nation. In general, a too restrictive interpretation of the different concepts used in deposit legislation can delimit the use and scope of web harvesting activities while trying to preserve our online cultural heritage.

3. Specific legislative issues

3.1. Which CHIs should be able to harvest?

As observed, CHIs can play a key role in collecting and preserving our digital heritage. To determine which CHIs may be given a mandate to harvest web content, it is necessary to first look at which CHIs currently are legally tasked with the preservation of the Dutch cultural heritage. This might be a solid basis for also attributing to these CHIs the task and possibility to collect and preserve the Dutch digital heritage.

3.1.1. CHIs with a legal task

The public task assigned to the KB includes taking care of the national library collection and encouraging the development of national facilities in the field of librarianship and information systems.[49] It is generally acknowledged that this task has come to include the digital domain as well.

The self-defined mission of the Sound and Vision is to improve people's lives in and with the media by archiving, exploring and contextualising it, to which the freedom of thought and expression in text, image and sound is paramount.[50] The Sound and Vision does not have a clear legal task, although a "media-archive" is referred to in several articles of the Dutch Media Act (*Mediawet*).[51] According to a report by IViR, the tasks of this media-archive, the Sound and Vision, are diverse. It functions as a company archive for the public broadcasters, as a cultural archive with a museum-function, and as an institute for education.[52] The Sound and Vision has an Archival Agreement (*Archiefovereenkomst*) with public broadcasters and several copyright organisations, according to which it can collect and present its collection, thereby also living up to its cultural-historical function to collect, select, access and maintain AV materials.[53] This Archival Agreement underlies the vast majority of Sound and Vision's collection. Besides this, the Sound and Vision has various initiatives through which additional archival materials are collected by asking permission directly from rights holders for the preservation and access to these materials. The Sound and Vision receives subsidies to perform its task, which seems to confirm the government's acknowledgement of the public importance of Sound and Vision's tasks.[54] The Sound and Vision's statutes also set out the goal of collecting, preserving and making accessible audiovisual content. Its assigned role of audiovisual archive in the Media Act shows the Sound and Vision's importance for the Dutch cultural heritage and underscores its public interest mission.[55]

Another CHI specifically designated to a public task is the *Nationaal Archief* (National Archive, **NA**).[56] Its duty is to perform the task attributed to it in the Archival Act 1995 (*Archiefwet* 1995) and to support the government in its administrative and legislative tasks.[57] The task attributed to it in the Archival Act 1995

[49] Art. 1.5 (2) Higher Education and Research Act (WHW), cited above.
[50] See: https://www.beeldengeluid.nl/organisatie/missie-en-visie (last accessed on 19 February 2020).
[51] Articles 2.138a(3)(c), 2.142a(1), 2.146(j), 2.167(1)(c) and 2.180(1) Media Act make reference to a "media-archive".
[52] J.M. Breemen, V.E. Breemen & P.B. Hugenholtz, 'Digitalisering van audiovisueel erfgoed: Naar een wettelijke publieke taak', December 2012, p. 7. Accessible via: https://www.ivir.nl/publicaties/download/Publieke_Taak_Beeld_en_Geluid.pdf (last accessed on 13 May 2020), p. 16.
[53] Collection plan Sound and Vision 2019, p. 35.
[54] J.M. Breemen, V.E. Breemen & P.B. Hugenholtz, 'Digitalisering van audiovisueel erfgoed: Naar een wettelijke publieke taak', December 2012, p. 16.
[55] *Ibid.*, p. 19.
[56] There are more institutions that have the legal task to archive following the Archival Act, but for the purpose of this paper these institutions will be undiscussed.
[57] Art. 4(1) Statuut agentschap nationaal archief, or Regeling van de Staatssecretaris van Onderwijs, Cultuur en Wetenschap, van 7 mei 2006, nr. WJZ/2006/4662 (8175), houdende regels inzake het agentschap Nationaal Archief (Statuut agentschap Nationaal Archief).

entails, inter alia, the management of the state archives, which contain the archived documents of several state bodies and provincial governments.[58] The NA thus has a mostly preservational role. Web harvesting is an important instrument to properly fulfil this role. This has been acknowledged by the Minister of Education, Culture and Science,[59] after a Recommendation by the Dutch Inspectorate for Cultural Heritage to also collect and preserve online government documentation, such as websites and tweets.[60]

While the government has already acknowledged that the NA should be able to engage in web harvesting to perform its task, the KB and Sound and Vision have also indicated that web harvesting should be facilitated, because it is essential for them to be able to adequately perform their public tasks.

3.1.2. Other CHIs

A remaining question is whether other CHIs than the KB, Sound and Vision and NA should be giving similar harvesting capabilities. When introducing a basis for the harvesting of a nation domain, it seems logical to attribute these privileges to national institutes. Furthermore, in practice, not all CHIs express an urgent interest in web harvesting. Dutch film museum EYE, for example, has indicated that web harvesting is not yet part of its collection strategy, nor is it a priority to actively engage in it. EYE does collect, besides its regular collection of movies and short films, "amateur films", but it is not yet interested in, for example, YouTube content. Likewise, the *NIOD Instituut voor Oorlogs-, Holocaust- en Genocide Studies* (NIOD Institute for War, Holocaust and Genocide Studies) has indicated that it is not interested in web harvesting, since it focuses on archiving World War II documentation and has decided to exclude modern websites from its collection. LIMA, a Dutch platform that engages *inter alia* in archiving and preservation of media art, also does not use web harvesting to acquire its collection, but it collects and preserves content in collaboration with institutions and artists and therefore with permission from the right holders concerned.[61]

An initiative like DEN shows the vast interest in digitisation by CHIs.[62] This does not mean that each CHI wants to maintain its own individual web archive. Some CHIs indicate that if they decided to expand their collection to web content in the future, they would search for collaboration with for example the KB. This could possibly take the shape of a request system, in which CHIs can request the KB or the Sound and Vision to harvest certain web content that is important to preserve. This content would then become part of the collection of the harvesting institutions, which assures that it will not be lost. It is also conceivable that the harvested content be consulted by the requesting CHIs or that these CHIs can refer to the harvested content by way of hyperlinks, depending on what the law permits. Such a request system could ensure a diverse web collection that is built on specialized knowledge from various actors. Still, other options remain. The legal basis for web harvesting activities could be arranged in a manner that permits larger CHIs like the KB, Sound and Vision and NA to conduct whole domain crawls, while other smaller CHIs perform selective harvests.

[58] Art. 25 (2(a)) and 26 Archival Act 1995.
[59] *Kamerstukken II* 2016/17, 29362 nr. 257. Accessible via: https://zoek.officielebekendmakingen.nl/kst-29362-257.html (last accessed on 13 July 2020).
[60] Webarchivering bij de centrale overheid: Het archiveren van websites en uitingen op sociale media, Rapport Erfgoedinspectie, November 2016, p. 4. Accessible via: https://www.inspectie-oe.nl/binaries/inspectie-oe/documenten/rapport/2016/12/8/rapport-webarchivering/Rapport+Erfgoedinspectie+webarchivering+bij+de+centrale+overheid.pdf.
[61] Lima offers ArtHost, a paid preservation service, to artists and institutions wanting to preserve their "net art" and other online artworks https://www.li-ma.nl/lima/article/arthost (last accessed on 6 July 2020).
[62] https://www.den.nl (last accessed on 6 July 2020).

3.1.3. Section summary and conclusion

As observed in Chapter 2, in most countries with web harvesting legislation, the legal possibility to harvest is attributed to specific national institutions designated by the law. A solid foundation for such attribution could be to appoint one or more national CHIs that already have a legally attributed task, which in the Netherlands are the KB, Sound and Vision and NA. In light of its legal task and after a Recommendation by the Dutch Inspectorate for Cultural Heritage, the NA can already engage in harvesting government websites and government-related social media content. It is in the public interest that the legislator also enables the KB and the Sound and Vision to harvest the web, as is important for these CHIs to be able to properly fulfil their public task in the digital society and also non-government-related content can be preserved. Extending such web harvesting privileges to other, smaller CHIs may give rise to legal uncertainty. Limiting the applicability of the law to the three aforementioned CHIs does not necessarily hamper the creation of a culturally diverse digital archive, since smaller CHIs do not always want to engage in web harvesting and, if they do, they would prefer to rely on a collaboration with the KB or Sound and Vision. The authors therefore suggest to only create a legal foundation for web harvesting for the aforementioned CHIs with a public task, while considering to permit other CHIs to issue requests to harvest specific web content that is culturally important to preserve.

3.2. What content should be harvestable?

When creating a future-proof legal basis for web harvesting, it is necessary to clearly define what *type* of content web harvesting legislation should apply to. The authors regard content as the actual contents of the objects, including their metadata or properties. This is broader than merely collecting hyperlinks.[63] This section will review the pros and cons of formulating a (broad) legal definition of 'online content'. It will discuss, inter alia, whether a legal definition of harvestable 'content' should cover only websites, or also multimedia content such as web videos; whether it should also include social media content; and what types of content would fit in the collection (strategies) of CHIs. Consequently, this section discusses what content should be harvestable when considering web archiving related legislation and how that can be formulated in a sufficiently neutral and future-proof manner in a legal definition.

3.2.1. The pace of technology

New legislation for online web harvesting should first of all be future proof. The ever-increasing pace of technological innovation prevents us from knowing exactly which type of content will be prevalent in the future. The sudden rise and fall of Dutch social media platform *Hyves*, for example, could never have been anticipated.[64] The legislator cannot reasonably foresee all future developments. To create future-proof legislation, a legal definition of harvestable content should not be limited to content currently available through contemporary technology, but should be broader. This could prevent creating legislation that will be outdated soon after implementation. Restricting the legal definition to specific types of content, such as "web videos" or "websites", seriously limits the scope of web harvesting legislation and its applicability. A broader terminology, like "online content" or "web content", might therefore be preferable.

[63] R. Baeza Yates, C. Castillo & E. N. Efthimiadis, 'Characterization of National Web Domains', *ACM Transactions on Internet Technology* 2007, Vol. 7, No. 2, p. 3.
[64] L. Zandbergen, 'Hoe internetgiganten ten onder gaan', *Het Financieele Dagblad* 14 June 2016, and H. van Lier, 'De opkomst en ondergang van Hyves: hoe heeft het zo ver kunnen komen?', *De Volkskrant* 31 October 2013.

3.2.2. Interweaving types of content

In addition to the development of new types of media, a legal definition must take into account that media types are often interwoven. Content usually not only consists of a website, but also includes embedded and integrated social media applications, web videos and hyperlinks to other websites. When formulating a legal definition of harvestable content, this interdependence must be borne in mind. This also calls for adopting a broad and neutral definition that is not content-specific.

3.2.3. CHIs' expertise in preserving cultural heritage

CHIs have a specific expertise when it comes to deciding which content should be preserved in the public interest. Multiple CHIs have already developed a vision of what content would need to be harvested to preserve a comprehensive image of our digital heritage.[65] Their ideas about what types of content would need to be part of an all-encompassing archive of the Dutch digital heritage are important to take into consideration. The collection policies of the Sound and Vision and the KB perfectly illustrate this.

The Sound and Vision has included a clear web harvesting vision in its collection plan. Its collection includes a large part of the Dutch audiovisual media content from the end of the 19th century until now.[66] Its archives further contain media related photos, objects, memorabilia and publications on media and the media landscape. Over the years, more interactive and online productions have been added such as computer games, web videos and websites.[67] The Sound and Vision has been selecting web videos (videos exclusively distributed via the web) since 2008, and other interactive and web based-media expressions since 2015.[68]

The Sound and Vision's goal is to expand its archive with these types of content in the future. Online productions are specifically mentioned as popular and recent developments that would be additions to its collection.[69] Its priorities lie with filling the gaps in its collection. These gaps consist of web videos, web content of the public broadcasters, podcasts, YouTube and other social media videos and channels, blogs, web based videos, online courses, games, GIFS, vlogs, web-only series, drama content from on demand broadcasters, video clips, websites, memes and other online productions in general.[70] As the Sound and Vision Collection Plan makes clear, harvesting websites exclusively would not be sufficient to fulfil its task of preserving and presenting the Dutch media landscape, since this landscape consists of a wide range of online productions.

The KB has also included online content as a specific focus in the KB Content Strategy. One of the goals mentioned in its Content Strategy is to start harvesting the Dutch web domain.[71] The KB has stressed the importance of a diverse overview of all Dutch online content for its collection, and sees a responsibility for the KB to take a leading role in harvesting and preserving the Dutch online heritage. This includes both websites and other types of content, such as social media content.

[65] See for example Collection plan Sound and Vision 2019 and Content Strategy Koninklijke Bibliotheek (not publicly accessible, however the authors were authorised to research the KB's Content Strategy).
[66] Collection plan Sound and Vision 2019, p. 14. Accessible via: http://files.beeldengeluid.nl/pdf/Collectieplan_BeeldenGeluid_2019.pdf.
[67] *Ibid.*
[68] *Ibid.*, p. 15.
[69] *Ibid.*, p. 18.
[70] *Ibid.*, p. 53-98.
[71] Content Strategy Koninklijke Bibliotheek (not publicly accessible, however the authors were authorised to research the KB's Content Strategy).

3.2.4. The inclusion of social media

Research from the *Centraal Bureau voor de Statistiek* (Statistics Netherlands, **CBS**) shows that 87,4% of Dutch citizens used social media in 2019, most of them every day. 50% of Dutch people used the internet for uploading photos, videos, texts and other types of content.[72] These statistics illustrate how big the role of the internet, and especially social media, is in daily lives of Dutch citizens. Many contemporary societal movements are largely relying on social media platforms like Twitter. One could think of movements like *#metoo* (the movement in which people share their experiences concerning sexual harassment), or more specifically for the Netherlands *#trotsopdeboer* (part of the farmer's protest against new Dutch climate measures). CHIs consider it valuable and key to their public interest mission to collect content produced on these social media, given the important cultural-historical value of such content for future generations. As indicated above, the KB and Sound and Vision have already signalled the need to collect and preserve social media content and have made this type of content part of their collection strategy.

The harvesting of social media content, however, comes with practical issues. First, the platforms' terms and conditions often do not allow archiving and data harvesting. Youtube, for example, prohibits a multitude of actions, like downloading any type of content, without specific or written permission.[73] Facebook also requires written permission "*to modify, create derivative works of, decompile, or otherwise attempt to extract source code from Facebook*".[74] Harvesting social media content can only be done through Application Programming Interfaces (**APIs**). APIs function as middlemen, allowing researcher's computers to engage with a social network in a way that the platforms can control.[75] Agreement to a platform's policies is required to access the API and extract data from the social media platform.[76] The platform thus has control over the API, over its policies and, consequently, over the possibilities to harvest data.[77] These user rules and private policies could only be trumped by a legal provision on national or European level.

Furthermore, technical issues come into play. There are not yet any perfect tools to analyse and harvest Twitter and Facebook data. It cannot be assured that the current tools, offered by the media themselves, give an encompassing overview of the relevant information, nor is it sure that received data is trustworthy. CHIs should therefore ideally be in control of the information which they will receive using their own tools, rather than rely on what the platforms will provide to them. Due to the large quantities of data available, it is also difficult to select what content is relevant and falls under the Dutch domain (see para. 3.3).

In conclusion, even though the harvesting of social media would add significant value to internet archives of CHIs, and their wish to collect this type of content is evident, it does come with practical issues relating to the terms and conditions of the platforms and technical issues due to the underdeveloped harvesting tools and the enormous amounts of content, which require (time-)intensive selection processes. This must be considered when choosing whether to include social media in the definition of 'online content'.

[72] CBS Statline: 'Internet; toegang, gebruik en faciliteiten', October 2019. Accessible via: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/83429NED/table?dl=29015 (last accessed on 18 March 2020).
[73] Rights and limitations, Terms & conditions Youtube, accessible via: https://www.youtube.com/static?template=terms (last accessed on 12 August 2020).
[74] § 3.4 Facebook Legal Terms, accessible via: https://www.facebook.com/legal/terms (last accessed on 8 July 2020).
[75] Gráinne Maedhbh Nic Lochlainn, 'Facebook data harvesting: what you need to know', *The Conversation,* 3 April 2018, accessible via: https://theconversation.com/facebook-data-harvesting-what-you-need-to-know-93959 (last accessed on 1 April 2020).
[76] https://developers.facebook.com/policy (last accessed on 1 April 2020).
[77] Gráinne Maedhbh Nic Lochlainn, 'Facebook data harvesting: what you need to know', *The Conversation,* 3 April 2018. Accessible via: https://theconversation.com/facebook-data-harvesting-what-you-need-to-know-93959 (last accessed on 1 April 2020).

3.2.5. Other issues that may arise when formulating a definition of 'content'

When exploring what content should be harvestable by CHIs many other issues arise. To continue with the earlier example of *#metoo*, harvesting all tweets including this hashtag could be accompanied with ethical issues. Serious privacy concerns may arise, especially in this case where sexual abuse is concerned. CHIs should process this kind of data in accordance with the General Data Protection Regulation (**GDPR**).[78] An option could be to restrict specific access or blacken out any kind of personal data that an archived webpage contains, either in the archived or the accessible version. However, this would require an extensive and active screening of every webpage that is to be harvested, which could prove to be a costly and tedious task. In addition to that, in cases comparable to the *#metoo*-movement, the personal data might add to the meaning of the tweets. One of the movement's strengths lies in the diversity of people tweeting about it, including celebrities whose identities add extra value to the (impact of the) data. Privacy sensitive content could therefore be valuable to preserve from a cultural heritage point of view, but it should be carefully decided – within the boundaries of what is allowed under the GDPR – which content is important enough to be collected and preserved.

It would not be useful to limit the definition of harvestable 'content' *too much*, if this would remove all editorial freedom of CHIs. At the same time, given the specific character of their collections, a too general definition should also be avoided. The authors believe that the decision of what content is subject to web harvesting should be left to the discretion of CHIs, within the limits of the applicable legal framework.

Often, the range of technical choices available to a CHI already shapes the outcome of a collection.[79] The same can be said about the nature of a CHI, which clearly defines its collecting interests. The definition of 'content' may limit the collecting options certain CHIs have. Indeed, as is a common issue, providing a strict definition of harvestable 'content' has the inherent danger of limiting the development of digital collecting attempts. The authors therefore suggest a dynamic definition, leaving the CHIs certain room for interpretation to achieve efficient and useful web archiving. Such a dynamic definition may encompass CHIs harvesting diverse content, judging on a case-to-case basis within the boundaries of their public task. Sound and Vision, for example, is primarily interested in images and sound, including blogs and social media, whilst the KB wants to preserve different kinds of information, such as texts. The judgement can be seen as offsetting the preservation goals of the CHI concerned to the actual content that was harvested, to ensure that the CHI only harvests content that it actually and necessarily needs. If a CHI exceeds its own public task limitation, a copyright infringement might still occur, therefore (proportionally) protecting the intellectual property rights of right holders.

To summarize, it is not useful to enact legislation that strictly defines beforehand what specific "content" may be harvested. Formulating a specific definition holds the inherent danger of significantly limiting the development of web archiving practices. CHIs should be able to consider, at their own discretion, what content best fits their preservation goals. Such an approach would satisfy the urgent need for preservation of the digital heritage. Therefore, the authors suggest that a dynamic definition of 'content', which allows CHIs to determine which content to harvest on a case-to-case basis within their public task.

[78] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons regarding the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
[79] F. Lasfargues, C. Oury, B. Wendland, *Legal deposit of the French Web: harvesting strategies for a national domain,* International Web Archiving Workshop, Sep 2008, Aarhus, Denmark, p. 12.

3.2.6. Possible definitions of 'content'

A few last considerations are in order. To prevent leaving practitioners empty handed, a definition could refer to 'data' rather than to 'content'. Data is "*any information in binary digital form*" and includes digital objects and databases. Additionally, content can be defined as simple ("such as textual files, images or sound files, along with their related identifiers and metadata") or complex ("made by combining a number of other digital objects, such as websites").[80] It is debatable whether 'content' should encompass merely digital objects or more broadly 'data', as the latter also includes databases. Content defined as "all data dynamically covering the scope of the public task of the CHI concerned" is probably the broadest and most relevant definition to uphold.

3.2.7. Section summary and conclusion

To create a legal basis for web harvesting, it is necessary to define the type of harvestable content. What do we want to preserve? What is the goal of harvesting content from the web? The general idea of Dutch CHIs is to preserve the Dutch digital heritage. The scope of the preservation should therefore, naturally, cover content relating to Dutch digital heritage, but Dutch digital heritage exists anywhere. Content could therefore encompass websites providing information on the Netherlands, but also e-books, (web)video's and other digital materials. The authors suggest that the definition of 'content' should be broad, to ensure that it is future-proof and considers the interdependence of types of content on the internet. CHIs would like to include almost all types of content in a definition, not only digital-born, but also uploaded analogue content. Social media content is also of great interest to CHIs, but because of its specific characteristics, the inclusion of social media comes with several consequences that would need serious consideration.

Ideally, a legal definition of harvestable 'content' allows for a dynamic interpretation on a case-to-case basis, within the applicable legal framework. CHIs have different public goals and limiting them by creating unnecessary and unfounded boundaries is a development which must be avoided at all costs. A legal basis for web archiving should support the CHIs' public interest activities. Leaving the decision of what content to include in their web archiving activities to the CHIs themselves should be encouraged.

A possible definition of harvestable 'content', suggested by the authors, might be:
"Content entails all data to be found online. This definition is to be dynamically interpreted by the CHIs concerned, within the limits of their public task and the applicable legal framework."

3.3. What is part of the Dutch 'domain' of the internet

"The historian who sets out to identify a nation's web sphere has to acknowledge that this web sphere did not exist as such beforehand. On the contrary, it has to be constructed, as is the case with any web sphere."[81]

The internet transcends national boundaries. A national web does not comply with sovereignty or international borders. Determining what is part of the national domain can be difficult to ascertain when overlap with other nationalities ought to be avoided. The common interest of Dutch CHIs is to preserve

[80] D.R. Harvey, Preserving Digital Materials (2nd edition), München: De Gruyter Saur 2012, p. 14-23. See for more info Guidelines for the Preservation of Digital Heritage, UNESCO Charter on the Preservation of Digital Heritage of 15 October 2003. Accessible via: https://unesdoc.unesco.org/ark:/48223/pf0000133171.page=80.
[81] N. Brügger, in: *The Routledge Companion to Global Internet Histories,* ed. G. Goggin & M. McLelland, New York: Routledge 2017, p. 63.

Dutch cultural heritage by harvesting the Dutch domain.[82] However, it is unclear what is precisely meant with the term 'Dutch domain'. This section will explain what common definitions of the Dutch domain of the internet are, in order to consider what the scope of the harvesting of the content related to the Dutch (digital) cultural heritage is. This will eventually lead to the conclusion of what should be the scope of a crawl of the Dutch domain.

3.3.1. Definitions provided by CHIs

There are multiple CHIs in the Netherlands that are currently actively trying to harvest online content for the sake of preserving the Dutch digital heritage. Before defining what the Dutch part of the internet is, it is interesting to see how CHIs currently demarcate the scope of their preservation and archiving activities. In this context, the different content and collection strategies of actively harvesting CHIs describe what their interpretation of the Dutch domain of the internet covers. The KB, Sound and Vision and the *Netwerk Digitaal Erfgoed* (Network Digital Heritage, **NDE**)[83] are all heavily interested in the idea of harvesting online cultural heritage, and all three have thought about appropriate definitions of 'the Dutch domain'.[84] In defining the Dutch national domain, it is the Sound and Vision that provides the most complete and relevant definition.

The Sound and Vision collects a representative selection of the Dutch media-heritage.[85] Yearly, about eight thousand hours of Dutch tv-productions and fifty thousand hours of Dutch radiobroadcasts are automatically added to the Sound and Vision's collection. Next to its passive collection, the Sound and Vision also collects and preserves materials actively.[86] In order to successfully carry out its task of preserving the Dutch digital cultural heritage relating to sound and vision, it adheres to the Sound and Vision Collection Plan. Its archiving activities cover Dutch media primarily. The Sound and Vision has narrowed the scope of its activities down to the following:

- Sound and Vision collects media which is produced by Dutch citizens in the Netherlands or abroad;
- Sound and Vision collects media which is produced or created in the Netherlands; and
- Sound and Vision collects media which is produced or created abroad but contributes to a solid illustration of the history of the Dutch society.

What stands out is that the Sound and Vision provides a clear and substantiated scope for their preservation activities. The boundaries of the harvesting mainly extend to content produced or created by Dutchmen, regardless of the country where the content was produced or created, or in the Netherlands. Additionally, it collects content related to the Dutch heritage. In the authors' view, the Sound and Vision Collection Plan holds a very suitable and clear-cut definition of the Dutch domain. It does not

[82] For an extensive history of the Dutch web, see K. Teszelszky, 'The historic context of web archiving and the web archive: reconstructing and saving the Dutch national web using historical methods'. In: The Historical web and digital humanities: the case of national web domains, New York: Routledge 2019.
[83] The Netwerk Digitaal Erfgoed (Network Digital Heritage, NDE) is an organization consisting of CHIs in the Netherlands. Their goal is to provide the CHIs with resources and services that help to improve the exposure and usefulness of the digital heritage, giving it added value. The NDE does not archive and preserve, but rather assists the existing CHIs in fulfilling their respective goals.
[84] See for the KB's considerations Sierman, B., & Teszelszky K. (2017). How can we improve our web collection? An evaluation of web archiving at the KB National Library of the Netherlands (2007–2017). Alexandria, 27(2), p. 99. See for the NDE's considerations Netwerk Digitaal Erfgoed, Nationale Strategie, https://www.netwerkdigitaalerfgoed.nl/wp-content/uploads/2018/02/Nationale_Strategie_Digitaal_Erfgoed_MinOCW.pdf, p. 11.
[85] http://files.beeldengeluid.nl/pdf/Collectieplan_BeeldenGeluid_2019.pdf, p. 11.
[86] *Ibid.*, p. 14-15 and p. 20.

necessarily cover the entire Dutch domain of the internet, but the scope the Sound and Vision has formulated can be applied in a useful manner.

3.3.2. Alternative Dutch definitions

To demarcate the Dutch web domain, inspiration can be taken from other definitions used in the national context. A first definition is the one used by the *Stichting Internet Domeinregistratie Nederland* (SIDN). The SIDN's main task is the registration of .nl domain names, but it also shares its expertise on topics such as internet governance, internet security and the functional stability of .nl.[87] Its task is based on a working group memo on the Domain Name System Structure and Delegation, which states: "The country code domains (for example, FR, NL, KR, US) are each organized by an administrator for that country. (...) These administrators are performing a public service on behalf of the Internet community."[88]

The SIDN however does not have a clear definition of the Dutch domain or the Dutch internet community. According to Martijn Simon, Legal & Policy Manager of the SIDN, this community entails at least the most relevant stakeholders of the .nl domain: the domain name registrars, the domain name owners, and the internet users using the .nl domain name. These parties do not all have the Dutch nationality or domicile in the Netherlands. Any EU citizen can easily apply for a .nl domain name. Parties based outside of the EU have to be assessed by the SIDN first. The functioning of the SIDN shows the international operation of the internet, and the difficulty of defining the Dutch part of it. When asked what should be part of the legal definition of the Dutch domain, Simon first emphasized that limiting it to only .nl is too simple: there are many Dutch websites with a .com TLD (e.g. bol.com), and there are other Dutch TLDs like .amsterdam, .frl. According to Simon, the Dutch web domain is the information available on the web that is relevant for Dutch users (either on the continent or in other parts of the Kingdom) at any moment in time. One could limit this to information in the Dutch language, but that would come with the problem that it would include part of the Belgian web as well and, moreover, excludes websites directed at the Dutch public not written in the Dutch language.[89] The SIDN can help create insights in what this content might entail, but does not offer a clear definition, other than the relevance of the content for Dutch users.

The *Analistennetwerk Nationale Veiligheid* (National Network of Safety and Security Analysts, **ANV**) uses a definition of the Dutch domain in the context of public safety. One of their impact criteria is the violation of the integrity of the 'digital space', which is defined as "the conglomerate of ICT-tools and -services, which contains all entities that (can) be digitally connected. This domain contains both permanent and temporary or local connections, as well as all information (i.a. Data, code, information) situated within this domain, with no geographical restrictions set."[90] Even when essential elements of these services are located physically abroad, these can be part of the risk assessment. This is a very broad definition of the Dutch domain, since there are no geographical restrictions set. Almost anything that influences the use of the digital domain in the Netherlands is part of the Dutch 'digital space', the integrity of which can be attacked. This broad definition also shows the international dimension of the internet, which leads to the same conclusion that the Dutch domain is not easily delineated and very international, which means that a definition would need to be flexible and not limited to the Dutch territory or the .nl-domain.

[87] https://www.sidn.nl/en/about-sidn/what-we-do (last accessed on 15 April 2020).
[88] Network Working Group, Domain Name System Structure and Delegation. Accessible via: https://tools.ietf.org/html/rfc1591 (last accessed on 22 April 2020).
[89] E-mail Martijn Simon 21 April 2020, on file with the authors.
[90] Analistennetwerk Nationale Veiligheid, Leidraad risicobeoordeling 2019: Geïntegreerde risicoanalyse Nationale Veiligheid, 2019, p. 14 ("...het conglomeraat *van ICT-middelen en -diensten en bevat alle entiteiten die digitaal verbonden (kunnen) zijn*". Het domein omvat zowel permanente als tijdelijke of plaatselijke verbindingen, evenals de gegevens (o.a. data, programmacode, informatie) die zich in dit domein bevinden, waarbij geen geografische beperkingen zijn gesteld.").

3.3.3. Demarcating national domains in academic literature

The fact that creating a 'national domain' comes with difficulties is also recognized in academic literature. The biggest problem encountered when demarcating a national web domain is the fact that a nation's web borders are not equivalent to the nation's state borders.[91] Several countries have nevertheless tried to demarcate a national web domain. In its attempt to define the Danish domain, for example, the Library of Denmark considers whether material has been published (1) from internet domains that are specifically assigned to Denmark, like .dk, or (2) from other internet domains but directed at a public in Denmark, such as material written in Danish, owned by a Danish resident or relating to Danish affairs.[92] This is akin to the Sound and Vision-approach mentioned before, which defines material as 'Dutch' if it is (1) in the Dutch language and registered in the Netherlands, (2) in any language and registered in the Netherlands, (3) in the Dutch language and registered outside the Netherlands, or (4) in any language and registered outside the Netherlands, and with a subject matter related to the Netherlands.[93] The fourth criterion has an editorial aspect. Rogers, professor on new media and digital culture at the University of Amsterdam, calls this the "editorial approach", which according to him does not give a complete overview of a national domain. In contrast to selecting .nl websites, which is fairly easy and straightforward, selecting websites related to Dutch subject matters and websites in Dutch but registered outside the Netherlands (outside of .nl) poses challenges to automation, and to working at scale.[94]

Rogers' preferred approach would be to make use of the collected data of web devices, which collect and serve web content territorially or to a particular language group, to be able to find country-specific and/or language-specific webs.[95] A search engine like Google, for example, serves different content to a user in Germany than it does to a user in the Netherlands. Making use of the technology on which this is based could help with demarcating a national web, according to Rogers. Their algorithms and logic decide what content is relevant for a specific country and/or language, which saves librarians and other web harvesters much time, compared to the "editorial approach" used by the Sound and Vision.[96] An automated approach also gives less room for bias in deciding which content is relevant and which is not.[97] Although this research does not offer a clear legal definition of the Dutch domain, it does show that the content that is relevant for citizens of a nation might be way outside of the ccTLD or just the national language, and that only relying on the editorial freedom of CHIs may encounter problems as well.[98]

This is confirmed by other academics. Brügger and Laursen state that only limiting a national domain to a national ccTLD is not enough, because in some cases much of a nation's web activity takes place on Generic Top-Level Domains (**gTLD**) like *.com*, *.org*, or *.net*, or on transnational gTLDs such as .eu, or .africa. Other possibilities are regional or urban TLDs (.amsterdam/.frl) or national ccTLDs that are not relevant to a specific country, but used for other purposes (.tv/.nu).[99] Kahn, at the Humboldt Institute for Internet

[91] N. Brügger, in: *The Routledge Companion to Global Internet Histories,* ed. G. Goggin & M. McLelland, New York: Routledge 2017, p. 64.
[92] *Ibid.*, p. 64-65.
[93] R. Rogers, *Digital Methods*, Cambridge Massachusetts: The MIT Press 2013, p. 129.
[94] *Ibid.*, p. 129.
[95] R. Rogers, *Digital Methods*, Cambridge Massachusetts: The MIT Press 2013, p. 126.
[96] *Ibid.*, p. 132.
[97] Bias in creating a web archive is problematized by several authors, including: I. Milligan & T.J. Smith, in: *The Routledge Companion to Global Internet Histories,* ed. G. Goggin & M. McLelland, New York: Routledge 2017, p. 49, N. Brügger, in: *The Routledge Companion to Global Internet Histories,* ed. G. Goggin & M. McLelland, New York: Routledge 2017, p. 71 & S. A. Hale, G. Blank & V. D. Alexander, 'Live *versus* archive: Comparing a web archive to a population of web pages', in: N. Brügger & R. Schroeder (ed.), *The Web as History: Using Web Archives to Understand the Past and the Present*, London: UCL Press 2017, p. 45-61.
[98] R. Rogers, *Digital Methods*, Cambridge Massachusetts: The MIT Press 2013, p. 150.
[99] N. Brügger & D. Laursen (ed.), The Historical web and digital humanities: the case of national web domains, New York: Routledge 2019, p. 3-4.

and Society, agrees that "the immediately visible indications of national web space may be deceptive."[100] To base a national domain on the ccTLD alone is only useful if this TLD is systematically used.[101]

Language as the only indicator does not suffice either. Only for countries whose language is hardly ever used elsewhere this would be limiting enough.[102] Moreover, material that contributes to a national web may be created outside a country's physical borders.[103] Another option is selecting content belonging to a certain culture, or taking into account the national history, heritage or cultural influences, even from outside the borders.[104] A problem with this type of selection is that it is hard to predict (without bias) what type of content will be relevant and interesting for future researchers. Teszelszky, who reflects on web archiving done by the KB, concludes that when looking back critically, "not all interesting sites of future historical importance have been selected during the last ten years."[105] Only a broad definition of what is part of the national domain, might cover such issues.

3.3.4. Section summary and conclusion

While a clear, all-encompassing definition of a 'national web domain' is not available yet, both academia and practical experts point to specific issues that need to be addressed when creating a legal definition of this domain. First, limiting the national domain to only the .nl ccTLD would most likely make a definition too narrow. It is not representative of how average Dutch citizens use the internet, seeing for example that bol.com is a big stakeholder in the Netherlands. Moreover, it would discount all other specific Dutch TLDs like .amsterdam or .frl. Second, defining the Dutch domain as entailing only websites in the Dutch language would also be inadequate, since Dutch is also spoken in Belgium and research shows that citizens also use websites in other languages on which Dutch content appears. Third, using *a priori* substantive criteria like "content directed at the Netherlands" or "content on Dutch issues" comes with practical issues, since this "editorial approach" is not easily automated. Moreover, it could come with a certain bias on what to harvest and what not: what is relevant for future research might be open for interpretation. This editorial approach however tends to the needs of CHIs, seeing their collection policies. A definition of the Dutch domain should thus seek the right balance between editorial freedom and a broad enough demarcation of the national domain, to indicate the importance of a broad harvesting practice.

One definition suggested by the authors might be as follows:
The appointed CHIs have the possibility to collect:

a) *Web content which is hosted on websites with TLDs hosted by the the national registrar;*
b) *Web content which is produced by Dutch citizens in the Netherlands or abroad;*
c) *Web content which is produced or created in the Netherlands;*
d) *Web content which is produced in the Dutch language; and*
e) *Web content which is produced or created abroad but its subject matter relates to the Netherlands and is valuable as a source of cultural-historic information.*

[100] R. Kahn, 'The nation is in the network: locating a national museum online'. In: *The Historical web and digital humanities: the case of national web domains*, New York: Routledge 2019, p. 164.
[101] *Ibid.*, p. 164.
[102] N. Brügger & D. Laursen (ed.), *The Historical web and digital humanities: the case of national web domains*, New York: Routledge 2019, p. 4.
[103] R. Kahn, 'The nation is in the network: locating a national museum online'. In: *The Historical web and digital humanities: the case of national web domains*, New York: Routledge 2019, p. 164.
[104] K. Teszelszky, 'The historic context of web archiving and the web archive: reconstructing and saving the Dutch national web using historical methods'. In: *The Historical web and digital humanities: the case of national web domains*, New York: Routledge 2019, p. 16.
[105] *Ibid.*, p. 23.

This definition covers all Dutch TLDs, since SIDN hosts inter alia .nl, .aw, .frl, and .amsterdam. Furthermore, it adopts a very broad notion of 'nationality', including content produced abroad but concerning the Dutch cultural heritage, similar to the Danish approach. Finally, it does not have vague criteria like "directed at" the Netherlands, while criterion (e) gives CHIs some editorial freedom in selecting content which would contribute to their web collection.

3.4. Should harvesting be limited to only the publicly accessible part of the internet?

The preceding sections discussed what should be harvestable content and what should be considered part of the Dutch domain of the internet. This section will consider how far web archiving activities may reach in terms of the harvesting of (un)accessible parts of the internet.

3.4.1. Publicly accessible

To determine whether online content that is not publicly accessible should be subject to a web crawl, it is important to first establish what can be considered as publicly accessible. For the purpose of this research, publicly accessible will encompass all content, which is not part of the deep web that is to be accessed online. The deep web is the part of the internet that is supposedly not indexed by web crawlers and search engines, as it is shielded by login pages, paywalls, hidden in databases or protected by certain codes (think of private social media accounts or newspaper articles that require logins and subscriptions).[106]

The dark web, where lots of illegal activities take place, is not further considered in this report. The dark web contains web pages that are not accessible via 'ordinary' web browsers, such as Chrome, Edge or Firefox.[107] For the purpose of this report, it is assumed that the dark web contains no relevant content or information relating to Dutch cultural heritage. The authors therefore believe that the dark web should not be included in web crawls as part of harvesting activities by CHIs. Documenting the dark web is more relevant for investigations by the authorities competent in areas such as criminal law.

3.4.2. Why should the deep web be archived?

One argument in favour of archiving content posted on the deep web is that this content might be relevant for the preservation of Dutch cultural heritage. Therefore, various CHIs might be interested in harvesting parts of the deep web. Granting CHIs a legal mandate to archive content that is hidden in or protected by the deep web could help to achieve the objectives of a web harvesting policy, provided that added access safeguards for CHIs wanting to engage in harvesting parts of the deep web will apply.

It is also debatable how important deep web protection really is, and which interests should prevail: the public interest of preserving digital heritage or the private interest of maintaining the 'fragile' protection of the deep web. Deep web protection measures have limited strength. Creating a fake account to access content that is protected by a login is not very difficult. Hotmail, Gmail and other email providers make it possible to create multiple email accounts. These fake email accounts can subsequently be used to create

[106] K. Teszelszky, 'The historic context of web archiving and the web archive: reconstructing and saving the Dutch national web using historical methods'. In: *The Historical web and digital humanities: the case of national web domains*, New York: Routledge 2019, p. 14; I. Hernández, C.R. Rivero & D. Ruiz, 'Deep Web crawling: a survey', *World Wide Web* 22, (2019), p. 1577-1581; and V. Ciancaglini, M. Balduzzi, R. McArdle, and M. Rösler, 'The Deep Web', *Trend Micro* (2015), p. 4.
[107] K. Teszelszky, 'The historic context of web archiving and the web archive: reconstructing and saving the Dutch national web using historical methods'. In: *The Historical web and digital humanities: the case of national web domains*, New York: Routledge 2019, p. 14.

fake profiles to access login-protected content. Similar workarounds exist for other protection measures. Paywalls, for example, can be circumvented by typing a '.' or a ',' after .com (i.e. the TLD concerned).

3.4.3. Why should the deep web not be archived?

One argument against archiving content posted on the deep web is that it can be assumed that a person who posts content online behind a paywall, login, code or database does not have the intention for this content to be accessed by everyone at any time. The person might have reasons for this, such as the desire to know who is interested in his or her work, or the desire to commercially benefit from the content. Conducting a web crawl and archiving all (relevant) content posted online, only for it to be subsequently accessed (later) in libraries or on location of CHIs could diminish the purpose of using and upholding deep web measures. It is debatable whether the extra step of visiting or contacting a CHI to see the content changes the fact that certain protective measures can be circumvented.

The authors are of the opinion that CHIs should be able to conduct web crawls of all publicly accessible information and content online relating to the Dutch cultural heritage. All said content is put online and is not protected by paywalls, logins, codes or databases. Therefore, it is considered public. Indeed, when posting content online without protection measures, it can be assumed that the author of said content has intended for his work to be accessed by everyone and at any time. Therefore, this content should definitely be eligible to be subject to the web archiving activities by CHIs, taking into account the GDPR and other applicable legislation. Regarding content posted in the so-called deep web, a possible solution would be to allow CHIs to harvest non-publicly accessible content only when direct permission is granted by the person who posted the content, or when authors and CHIs enter into cooperation, in conformity with Dutch copyright law.

3.4.4. Section summary and conclusion

The authors suggest that CHIs should be allowed to harvest all publicly available content (excluding the dark web). When it comes to the deep web, i.e. content that is not publicly accessible, this content should only be subject to harvesting activities with permission of the relevant parties involved. The unimpeded harvesting of such content would not respect individual's rights as they should. It is imaginable that deals or licenses be closed between CHIs and authors of content posted on the deep web to add said content to the web archives of CHIs. Content which is posted without any protection measures, such as logins or paywalls, can be assumed to be intended to be freely accessible by anyone at any time and thus subject to web harvesting activities within the confines of a legal instrument, as discussed in Chapter 4.

The authors therefore suggest adopting a provision of legally allowed web harvesting activities, which reads:
"If content is found to be inaccessible at the moment of capture – whether for technical reasons (such as password-protected contents) or commercial reasons (such as paid-access or subscription-based content) – the CHI may contact the website editor to find solutions on a case-by-case basis."

4. Legislative proposals

This chapter will introduce two possible legislative options. These options should be read as a visualization of possible solutions, instead of being conclusive solutions to the legal issues that arise when discussing web harvesting. The issue of CHI liability will not be addressed in this matter. When considering the introduction of new legislation for web harvesting activities, it is however recommended to include some form of exemption from liability for CHIs. This is necessary to ensure that CHIs cannot be held liable for the accidental harvest of illegal content or the harvest of content that is qualified as being illegal in a later stage. The possibility of being held liable for harvested content could severely hamper harvesting activities. The design of this exemption from liability could take various forms and include various conditions – such as an obligation to restrict access to content concerning privacy matters or an alleviation of duty for the CHI to remove illegal content once it has been archived – but detailed discussion goes beyond the limits of this research.

4.1. Copyright legislation

One possibility to legally enable CHIs to practice web harvesting is to explicitly permit this in the existing Dutch Copyright Act (*Auteurswet*, or **Aw**). The Dutch Copyright Act will soon be adapted to implement the DSM-directive (Directive (EU) 2019/790). CHIs had placed their hope in the implementation proposal to include provisions facilitating web harvesting. This chapter will examine the possibility of including a provision on web harvesting in the Dutch Copyright Act. This chapter will not specifically discuss database rights, but an equivalent provision should be introduced in the Dutch Database Act to avoid CHIs performing web harvesting activities to infringe upon possible database rights.

4.1.1. The Dutch Copyright Act

The Dutch Copyright Act includes various exceptions to the author's exclusive rights but has no provision enabling web harvesting by CHIs. The Dutch Copyright Act must keep within the limits of EU copyright law, including the InfoSoc-directive (Directive 2001/29/EC) that contains an extensive but exhaustive list of admissible exceptions and limitations. Member States have a certain degree of discretion regarding the transposition and interpretation of these exceptions and limitations within the national regime.[108] Art. 5.2(c) InfoSoc-directive, for example, allows Member States to adopt an exception "in respect of specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives, which are not for direct or indirect economic or commercial advantage", which arguably leaves room for web harvesting activities. Such an exception, however, has to pass the three-step test laid down in art. 5.5 InfoSoc-directive, which originates from international conventions. This test states that national exceptions can only be applied in certain special cases which do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holders involved. Any exception allowing web harvesting would have to fit within this legal framework.

Currently, such a provision does not exist in Dutch law. A 'preservation-exception' is included in art. 16n Dutch Copyright Act, but this exception is only applicable to works that are part of the institutions' own collection. The authors believe that the only situation in which this exception might be applicable to web

[108] I. Stamatoudini & P. Torremans (ed.), EU Copyright Law: A Commentary, Cheltenham UK: Edward Elgar Publishing 2014, p. 446.

harvesting, is when the law would state that the whole Dutch web domain would legally be considered to be part of (one of) the aforementioned CHIs' collections. This would require a legislative change.

4.1.2. Comparative legal analysis

Most countries enable web harvesting through legal deposit legislation, while other countries arrange this in their national copyright laws with possible extensions to other laws. A first example is France. A 2006 amendment of the French copyright law and the Code du Partimoine expanded the legal deposit, which has historically been included in copyright law, to web harvesting. Now, everything that is published on the internet in France is subject to legal deposit (see para. 2.3). A similar development took place in Germany. A 2018 amendment of the German Copyright Act and the German National Library Act expanded the National Library's legal collection mandate to web crawling activities. The German National Library now has the right to automatically and repeatedly harvest works that fall under its collection mandate and to archive websites, even without requesting prior permission from the respective right holders (see para. 2.4). These examples show that other EU Member States have found ways to implement web harvesting legislation within the confines of EU copyright law.

4.1.3. Pros and cons

In the wake of the implementation of the DSM-directive, several CHIs – most notably the KB and Sound and Vision – have proposed to include a web harvesting provision in the Dutch Copyright Act.[109] These CHIs stated that in light of the preservation goals of their institutions, an expansion of the preservation-exception that allows for web harvesting is needed. They suggested adding this option to a new art. 16na Dutch Copyright Act, which implements art. 6 of the DSM-directive. This newly proposed art. 16na however did not end up in the final implementation proposal, which has already been presented to Parliament at the time of writing.[110] A new legislative initiative would thus be required to regulate web harvesting.

A valuable parallel can nevertheless be drawn between the implementation of art. 3 DSM-directive and web harvesting. Art. 3 DSM-directive, which will be implemented in art. 15o Dutch Copyright Act, allows research institutions and CHIs to apply the technique of text and datamining (TDM) for scientific research. Even though TDM and web harvesting are two distinct affairs in which to consider different circumstances, the EU legislator at least recognizes that CHIs should be able to use modern techniques to perform their role as preservers of knowledge which in other circumstances would infringe copyright.

One advantage of including a web harvesting exception in the Copyright Act would be that all exceptions to copyright are included in one legislative act. For reasons of legal certainty and consistency, this would be preferred over spreading copyright exceptions between various laws. Another advantage would be the legislative freedom. While bound by international and EU law, the legislator has more freedom to choose the actors to which a copyright exception applies, than if web harvesting were included in a law regulating a specific institution. This could lead to a broader selection of digital archiving initiatives.

A disadvantage is the uncertainty related to the harmonization of copyright law. The EU legislator has not explicitly recognized an exception permitting web harvesting by CHIs. Although various other EU Member

[109] Gezamenlijke reactie van Nederlandse erfgoedinstellingen op het DSM-implementatiewetsvoorstel, 30 August 2019. Accessible via: https://www.den.nl/uploads/5d52c20d493be2b54b2c661aa9d5eebdacf2a8048963b.pdf (last accessed on 8 June 2020).
[110] *Kamerstukken II* 2019/2020, 35 454 nr. 2 (voorstel van wet).

States enable web harvesting by CHIs in their laws and art. 5.2(c) InfoSoc-directive arguably also leaves room for such activities, the legislator may still be reluctant to introduce a copyright exception to this effect. This may be unjustified, given the broad international recognition that digital heritage must be preserved and the existence of similar web harvesting exceptions in other EU Member States.

4.1.4. Possible substance

If the legislator were to permit web harvesting by CHIs in the Dutch Copyright Act, this should take the form of a new copyright exception. Most exceptions in the Dutch Copyright Act have a similar format and are formulated in a similar way. Typically, a copyright exception sets out which act under which conditions cannot be regarded as an infringement of the copyright in a literary, scientific or artistic work. The authors propose a formulation in line with this framework and suggest the following.

First, the act permitted would be web harvesting by certain designated CHIs. The conditions under which web harvesting would be admissible have been discussed in part in the foregoing sections, but they should be tuned to fit the copyright framework. In general, it appears appropriate to provide that web harvesting can only take place (a) in respect of works that have been lawfully disclosed to the public (similar to the citation right of art. 15a Dutch Copyright Act) and (b) insofar as the author's personality rights of art. 25 Dutch Copyright Act are observed (similar to art. 15, 15a & 16 Dutch Copyright Act). In the explanatory memorandum, the legislator could further elaborate on the lawful disclosure criterion for example by proposing the option to contact the website editor to find solutions on a case-by-case basis (para. 3.4).

The law should also elaborate on the CHIs to which the exception applies. It could be provided that the exception should only apply to institutions appointed by the Minister. This would give ample opportunity to only appoint the KB, Sound and Vision and NA as web harvesting institutions, allowing other CHIs to enter into collaborations with them through a request system, as suggested in para. 3.1. It would also be appropriate to include, at least in the explanatory memorandum, a definition stating that harvestable content entails, but is not limited to, all data to be found online, which is to be dynamically interpreted by the designated CHIs, within the limits of their public task and the applicable legal framework (see para. 3.2).

Finally, the law must delineate the national web domain to which the exception pertains. As mentioned in para. 3.3, the authors suggest a domain that covers all Dutch TLDs, that adopts a very broad notion of 'nationality' and that includes content produced abroad but concerning the Dutch cultural heritage.

This would result in the following provision:

***Article # - Web harvesting* (see for Dutch version annex 1)**

1. The collecting and preserving of content through web harvesting by cultural heritage institutions within the Dutch national domain cannot be regarded as an infringement of the copyright in a literary, scientific or artistic work, if:

a. the work harvested has been lawfully disclosed to the public;

b. the provisions of art. 25 are observed.

2. The cultural heritage institutions referred to in the first subsection are appointed by the Minister of Education, Culture and Science.

3. The content referred to in the first subsection should be part of the Dutch national domain, which includes:

- *web content which is hosted on websites with TLDs hosted by the the national registrar;*
- *web content which is produced by Dutch citizens in the Netherlands or abroad;*

- *web content which is produced or created in the Netherlands;*
- *web content which is produced in the Dutch or Frisian language; and*
- *web content which is produced or created abroad but its subject matter relates to the Netherlands and is valuable as a source of cultural-historic information.*

4.2. Deposit legislation

As described in para. 2.7, most countries have created a legal basis for web harvesting by CHIs in deposit legislation. At present, no deposit legislation exists in the Netherlands. Recently, however, advisory committees of the Dutch government have pleaded for the introduction of a mandatory legal deposit requirement in the Netherlands,[111] since digital works are bound to be lost when there is no obligation to deposit its contents.[112]

This section explores the possibility of introducing deposit legislation to enable web harvesting. After a brief description of the Dutch deposit history and a comparison with deposit systems in other jurisdictions, the advantages and disadvantages of deposit legislation as a legal basis for web harvesting are discussed. This section further explains the details of how a legal basis for web harvesting can be created in deposit legislation and how it can be framed to fit the existing practice of voluntary deposit in the Netherlands.

4.2.1. Dutch deposit history

In Europe, the Netherlands stands out for the absence of legislation regarding mandatory legal deposits.[113] Instead, a voluntary deposit system exists in the Netherlands since 1974. This voluntary system generally entails that publishers have freedom to decide whether they will deposit Dutch originated publications to the KB. The KB encourages all publishers to deposit a single copy free of charge. This system is based on the premise that it is also in the interest of publishers that their output be properly preserved.[114]

In the past, however, the Netherlands had mandatory deposit legislation. In 1803, mandatory deposit was introduced in the Netherlands under the influence of the repressive censorship of Lodewijk Napoleon.[115] It was not until 1912 that the mandatory deposit regime was abolished, together with all other copyright formalities. It was then already envisaged that the KB's archive would suffer greatly from this abolition.[116]

Between 1970 and 1972 and again in 1981, the Study Committee on Legal Depot examined the possibility of reintroducing mandatory deposit legislation in the Netherlands. In the Committee's opinion, legal deposit should be a governmental activity and should be laid down in deposit legislation (for which it also made a draft). From the start, the Committee had been in favour of close cooperation with the KB as manager of the statutory deposit. Unwilling to adopt deposit legislation but in order to change the status quo, the government agreed that the KB started collecting works on a voluntary basis from 1974 onwards.[117] Ultimately, the government must decide whether the Netherlands will adopt deposit

[111] Commissie Evaluatie Koninklijke Bibliotheek, accessible via: https://zoek.officielebekendmakingen.nl/blg-932507, *Kamerstukken II* 2019/20, 33846, nr. 58, Bijlage 1, p. 9, en Vaste commissie voor Onderwijs, Cultuur en Wetenschap, 'Inbreng verslag van een schriftelijk overleg', *Kamerstukken II* 2019/20, 2020D21175, p. 7.
[112] Commissie Evaluatie Koninklijke Bibliotheek, 'Evaluatie Koninklijke Bibliotheek', *Kamerstukken II* 2019/20, 33846, nr. 58, Bijlage 1, p. 9.
[113] J. Gesley, *Digital Legal Deposit in Selected Jurisdictions*, The Law Library of Congress: 2018, p. 44.
[114] KB, *Het Nederlands Bibliografisch Centrum*, kb.nl [online].
[115] J. Hallebeek, A.J.B. Sirks, *Nederland in Franse schaduw: recht en bestuur in het Koninkrijk Holland*, Uitgeverij Verloren: Hilversum, 2006, p. 69.
[116] De Groene Amsterdammer, *Historisch Archief 1877–1940*, 7 April 1912, p. 1.
[117] D. van Roekel, 'Hoedster van het geestelijk erfgoed', *Reformatorisch Dagblad* 1980, p. 37.

legislation, enabling CHIs to adequately preserve our cultural heritage, and whether that legislation also extends to web content.

4.2.2. Comparative legal analysis

Some useful insights can be gathered from the overview of web harvesting legislation in Chapter 2, in particular where it relates to deposit laws. Whereas Australia collects online available material through a combination of mandatory deposits and selective web harvesting activities, Denmark's legal deposit law allows CHIs to get access to, request or produce copies of material published in electronic communication networks. Denmark's definition of online material, which includes materials aimed at the Danish public, provides for a great scope of material that can be subject to web harvesting activities, eventually resulting in a more complete archive of Danish online heritage. Supplemented with detailed rules that the Danish Minister of Culture laid down delimiting the legal deposit to the national domain, web harvesting activities can cover a wide array of information while adhering to clear inherent boundaries.

Based on the deposit legislation in France, the French BnF can conduct two types of website collecting. It automatically harvests "snapshots" of websites belonging to the French domain and makes focused crawls of selected websites centred on a particular event or theme. If content is found to be inaccessible at the moment of capture, the BnF may contact the website editor to find technical solutions on a case-by-case basis to preserve the material. The National Library of New Zealand also performs both wide and selective harvests. Regarding the technical parameters for the harvest, the National Library of New Zealand states that these were developed after consultation with the public and internet stakeholder groups.

The UK Legal Deposit (Non-Print Works) Regulations 2013 are interesting, as they exclude some materials from the scope of deposit duties and web harvesting activities, namely works that contain personal data and that are available only to a restricted group, works that predominantly consist of film or recorded sound or material incidental to this, and works published before the regulations came into force. In case of misuse of collected materials, users can also be held liable for copyright infringement in the UK.

4.2.3. Pros and cons

Legal deposit is one of the oldest mechanisms to preserve cultural heritage, exercise freedom of expression and freedom of the press, end censorship and government secrecy, and guarantee public access to information.[118] It allows citizens to access national publications, while cultural development is documented in the process.

Deposit legislation offers flexibility to arrange how web content ought to be preserved. The law can arrange which content is subject to legal deposit and for which content harvesting requires prior consent. Deposit legislation can establish the manner in which collected content and (personal) data is managed, stored and who can access it. The legislator can decide whether exhaustive or selective web harvesting is allowed, for which purposes and by which actors. The law can also provide libraries with a firm basis to conclude agreements with depositors or to set up guidelines for the deposit of materials.[119]

Deposit legislation would provide for a stable basis to perform web harvesting activities. The development of new legislation could be built on the experience and feedback gained from the KB's current opt-out policy, which may create an acceptable legal basis for web harvesting in the eyes of right holders.

[118] J.T. Jasion, *The International Guide to Legal Deposit*, Routledge Revivals, first page of part 1 introduction.
[119] I. Verheul, *Networking for digital preservation*, IFLA publications 119, 2006, p. 25.

Deposit legislation for digital publications has been endorsed in the UNESCO Charter on the Preservation of Digital Heritage,[120] which encourages countries to eventually adopt national deposit legislation which ensures the preservation of and the permanent access to digitally produced materials.

For historical reasons, publishers may resist the reintroduction of mandatory legal deposit, even though it is not entirely clear which arguments hindered the reintroduction of such legislation for physical works in 1974. More recently, the Dutch government and professionals in the audiovisual sector regarded legal deposit of audiovisual heritage as financially unfeasible and potentially damaging to the functioning and quality of the media archives.[121] The preservation of web content, however, was not yet part of these discussions. Given that, to date, the voluntary deposit system for physical works functions properly, and taking into account the flexibility of deposit legislation, the legislator could decide to maintain a voluntary deposit regime for physical works, whilst providing a solid legal basis for harvesting web content, thus allowing the process as a whole to be considered less invasive by right holders.

4.2.4. Possible substance

When introducing a possibility for web harvesting activities in deposit legislation, similar policy decisions have to be made as for the introduction of a new copyright exception (see para. 4.1.4). Such legislation should also include a definition of harvestable content, elaborate on the CHIs that may engage in web harvesting activities, and delineate the national web domain that can be harvested. Regarding the use of the harvested materials, the law could contain a provision obliging a user to declare to the CHI that he or she will only use the materials for a legitimate purpose defined by the law, and that if the materials are used for any other purposes, that user will be liable for copyright infringement.[122] The authors recommend that the relevant provisions of the Copyright Act with regard to the reproduction possibilities and the access regime be declared applicable accordingly in the new deposit legislation, and not be more restrictive than necessary.

Given the current absence of mandatory deposit legislation in the Netherlands and the fact that a well-operating system of voluntary deposit exists for physical works, it seems reasonable if the legislator maintained the voluntary deposit for physical materials and only introduced deposit legislation with a view to enabling designated CHIs to harvest web content. As for physical and offline materials, the KB has good connections with publishers concerning the preservation of these materials and a change to a mandatory deposit regime for all such materials might spark a lot of resistance among publishers and other right holders.[123]

[120] UNESCO Charter on the Preservation of Digital Heritage of 15 October 2003, article 8. Accessible via: https://unesdoc.unesco.org/ark:/48223/pf0000133171.page=80" \h.
[121] *Kamerstukken II* 2011/12, 33000, nr. 148, p. 9.
[122] A similar rule is laid down in Section 27 of the UK Legal Deposit (Non-Print Works) Regulations 2013.
[123] A. de Kemp, E.H. Fredriksson, B. Ortelbach, *Academic Publishing in Europe: The Role of Information in Science and Society*, IOS Press: 2006, p. 132-133.

5. Conclusion

The Netherlands currently has no legal provision enabling web harvesting for the purpose of collection and preservation by CHIs. The need for this type of regulation is real, as the calls by CHIs and several European and international policy documents demonstrate. It is in the public interest that our digital history and heritage is collected and preserved for future generations. This position paper provides an overview of key aspects that need to be taken into consideration when creating web harvesting legislation for CHIs.

Chapter 2, which explores the web harvesting legislation in six selected countries, shows how different jurisdictions define various concepts and approaches, and how this affects the web harvesting activities by CHIs. In addition, it reveals that most countries have created a legal basis for web harvesting in deposit legislation and that such legislation and legislation enabling web harvesting is very common around the globe.

Chapter 3 provides a more in-depth analysis of the addressees of web harvesting legislation and the type of content which should be subjected to it. This paper proposes that the number of CHIs that can benefit from web harvesting legislation be limited to only those with a public task, allowing other smaller CHIs to share expertise on what content is important to harvest. The designated CHIs should be allowed to harvest various types of content in light of their collection plans or content strategies, as long as the content is part of the Dutch web domain and is accessible to the public.

Accordingly, to define the type of content that should be harvestable, it is necessary to ask the question: what do CHIs want to preserve? The common goal of Dutch CHIs is the preservation of the Dutch digital heritage. The scope of the web harvesting legislation should therefore naturally cover content relating to Dutch digital heritage. The definition of harvestable 'content' must be broad to ensure it is future-proof and considers the interdependence of types of content on the internet. To prevent the obstruction of the web harvesting and preservation activities of CHIs and to support their public interest mission, the definition of which content should be harvestable should be interpreted dynamically and on a case-to-case basis by the CHIs concerned, within the limits of their public task and the applicable legal framework.

Furthermore, it is important to agree upon the web domain where harvesting activities may take place. It is obvious that Dutch CHIs are mostly interested in harvesting relevant content in the Dutch web domain. However, a clear and all-encompassing definition of the Dutch domain is not yet available. Academia and practical experts point to several problems that a legal definition of the national web domain encounters. First, limiting the Dutch domain to only the .nl ccTLD would make the definition too narrow. Second, it is not accurate to demarcate the Dutch domain as entailing only websites in the Dutch language, since Dutch is also spoken in other jurisdictions. Third, using *a priori* substantive criteria like "content directed at the Netherlands" or "content on Dutch issues" comes with practical issues, since this "editorial approach" is not easily automated. Moreover, it could come with a certain bias on what to harvest and what not: what is relevant for future research is open to interpretation. This paper suggests that a definition of the Dutch web domain should seek the right balance between editorial freedom and a broad enough delineation of the national domain, to indicate the importance of a broad harvesting practice.

Lastly, the paper recommends that web harvesting legislation only relates to publicly accessible content. If the legislator would want to extend web harvesting activities to the deep web, i.e. content that is not publicly accessible, it would be preferable to make this conditional on the requirement to obtain prior permission of the author and relevant parties involved in conformity with Dutch copyright law. To add

deep web content to their collections, CHIs could conclude deals or licenses with the persons who have posted not publicly available content.

Following these conclusions, Chapter 4 sets out two different suggestions for introducing provisions into Dutch law to provide a legal basis for CHIs to harvest web content without infringing intellectual property rights. First, a new copyright exception permitting web harvesting by CHIs could be adopted in the Dutch Copyright Act (supplemented by an equivalent exception in the Dutch Database Act). This would have the advantage of keeping all exceptions to copyright in one place. Alternatively, web harvesting by CHIs could be facilitated by introducing a provision to this effect in deposit legislation. Given that the Netherlands currently has no mandatory legal deposit, but a well-functioning voluntary deposit system for physical works, it would only be required to create a legal duty to deposit web content or an obligation to tolerate web harvesting by CHIs to adequately preserve the Dutch digital cultural heritage.

In short, the need for regulating web harvesting by CHIs is high. Each day the Netherlands lacks legislation on web harvesting, more aspects and content of our collective digital heritage will be lost. The make-shift solutions and voluntary systems that CHIs currently apply no longer suffice. The Dutch legislator should take swift action to help our national institutions save our society's digital footprint.

Annex 1: Dutch text of a possible new copyright exception

Artikel # Auteurswet:

1. Als inbreuk op het auteursrecht op een werk van letterkunde, wetenschap of kunst wordt niet beschouwd de verveelvoudiging en openbaarmaking van content binnen het Nederlandse nationale domein door middel van web harvesting door cultureel erfgoedinstellingen, mits:
    1. het werk dat onderwerp is van harvesting rechtmatig openbaar is gemaakt;
    2. artikel 25 van deze wet in acht wordt genomen.
2. De cultureel erfgoedinstellingen waarnaar wordt verwezen in het eerste lid worden aangewezen door de Minister van Onderwijs, Cultuur en Wetenschap.
3. De content waarnaar wordt verwezen in het eerste lid moet onderdeel zijn van het Nederlandse nationale domein, wat omvat:
    1. web content die wordt gehost op websites met TLDs die vallen onder het beheer van het nationale domeinregister;
    2. web content die is geproduceerd door Nederlandse staatsburgers in het Koninkrijk of daarbuiten;
    3. web content die geproduceerd of gecreëerd in het Koninkrijk;
    4. web content die is geproduceerd in de Nederlandse of Friese taal; en
    5. web content die is geproduceerd of gecreëerd in het buitenland maar

    waarvan het onderwerp gerelateerd is aan het Koninkrijk en waardevol is als bron van cultureel-historische informatie.