Webharvesting

Samenvatting (p. 3) / Summary (p. 11)

Instituut voor Informatierecht
Institute for Information Law

UNIVERSITEIT VAN AMSTERDAM

Martin R.F. Senftleben

Stef J. van Gompel

Anne Helmond

Luna D. Schumacher

Jef Ausloos

Joris V.J. van Hoboken

João Pedro Quintais

20 september 2021

Summary

Desirability from a policy perspective

From a policy perspective, web harvesting can be considered desirable. This follows from an analysis of the European and Dutch constitutional framework. In light of the need to protect freedom of expression, freedom of information and the right to research, it is legitimate and advisable to take measures that make web harvesting possible. With the creation of web archives, web harvesting contributes substantially to the realisation of these fundamental rights and freedoms. This applies not only to specific harvesting activities, such as event harvesting, but also to overarching national domain crawls (3.2.6).

From a fundamental rights perspective, however, it is also necessary to take rights into account that could be affected by web harvesting: copyright and related rights, database rights, the right to privacy and the right to data protection. Policy makers must therefore find a solution that allows web harvesting without unduly compromising intellectual property rights, privacy rights and data protection rights.

Hence, desirability from a policy perspective does not mean carte blanche. In the choice of measures to support web harvesting, the Dutch legislator is not entirely free. Developments in the case law of the Court of Justice of the European Union show that the required balance – between fundamental rights that are supported by web harvesting and fundamental rights that could be affected by web harvesting – should preferably be created *within* the legal framework that exists for these rights at the level of secondary EU legislation. This applies in particular to the system of exclusive rights and limitations in EU copyright law (3.2.6).

Implementation in practice

Dutch cultural heritage institutions that wish to use web crawling and other harvesting techniques to build an archive of Dutch digital heritage face various practical problems. These include problems of an organisational (coordination of activities with other institutions), financial (limited capacity because of limited budget) and technical (lack of knowledge, need to keep pace with rapid technological developments) nature. In addition, conceptual questions arise with regard to the demarcation of the national domain when embarking on a national domain crawl (2.2.2). In the light of these challenges, there is little appetite among Dutch heritage institutions for a legal obligation and responsibility to archive the national domain in its entirety. They prefer a system that offers sufficient room to adjust harvesting activities to the capacity, financial resources and technical expertise available. The Dutch National Library (*Koninklijke Bibliotheek*) is interested in carrying out a national domain crawl. It has already developed extensive knowledge and expertise in this area.

From a legal perspective, it can be added that the opt-out system which is currently used in the Netherlands to request permission from rightholders of online content in the context of web harvesting is very time-consuming (2.2.3.1). Moreover, the rightholder of website content cannot always be traced. A webpage may contain third-party content, in respect of which the website manager is not in a position to give consent for web harvesting purposes. Against this background, it can be concluded that the existing harvesting practice poses particular problems. The focus on individual consent hinders the practical implementation of web harvesting in the Netherlands. This applies in particular to large-scale harvesting projects, such as a national domain crawl. An alternative approach – based on a statutory web harvesting entitlement or measures that

minimise legal risks – could simplify and speed up the existing web harvesting process to a significant degree. Ideally, it would make the current opt-out practice redundant.

An alternative scenario based on the development of specific legislation raises the further question of which institutions should be given the power to harvest the web. One option is a *centralised system* in which the legislator assigns the task of web harvesting to a few central heritage institutions, such as the Dutch National Library, The Netherlands Institute for Sound and Vision and the National Archive. In order to also meet the wishes and support collection strategies of smaller institutions, a system could be developed that allows smaller heritage institutions to submit web archiving requests to one of these central institutions. While the harvesting process would be outsourced to central institutions in this scenario, the selection would remain decentralised and, therefore, in the hands of specialised professionals in different cultural heritage institutions with specific knowledge. The implementation of this combined approach – a centralised system with the possibility for smaller institutions to submit individual harvesting requests – would require a careful further examination of the capacity and financial resources that are necessary to enable the dedicated central institutions to carry out the harvesting task not only with regard to their own collection strategy but also in reaction to requests received from smaller institutions.

It is also possible to envisage the creation of a *decentralised system* whereby a broad web-harvesting competence would be assigned globally to a large(r) group of cultural heritage institutions, such as all archives and publicly accessible libraries. To foster the exchange of knowledge and avoid the duplication of website copying, archiving and making available to the public, this system would require agreement about the scope of harvesting activities of the different institutions and the accessibility of resulting web archives. Existing collaboration initiatives, such as the *Netwerk Digitaal Erfgoed*, the *Nationaal Register Webarchieven* and national web archive consultation meetings, could serve as platforms to coordinate web archiving activities. In this scenario, each individual heritage institution would thus create its own web collection. Mutual agreement on archiving and distribution plans, however, could lead to central access and consultation facilities.

A final option is a *universal system* that offers all legal entities the opportunity to carry out web harvesting activities. In this case, individuals, NGOs and companies that pursue a commercial goal would also be able to invoke the web harvesting rules and create their own web archives.

Legal solutions

The development of any of these policy options depends on the possibilities offered by the legal framework at the national and European level. In that respect, the analysis has provided the following guidelines for legal solutions with respect to copyright, related rights and database rights, privacy and data protection law, and issues arising in other fields of civil, criminal and administrative law.

Copyright, related rights and database rights

The policy options in the areas of copyright, related rights and database rights can be presented on the basis of a matrix that distinguishes between the parameters of central/decentralised web harvesting competence and general national domain crawl/specific thematic selection. In the following overview, “NO” means that it is not advisable to develop legislation that permits web harvesting in this area. “YES” means that policy space exists which the Dutch legislator could use as a basis for the regulation of web harvesting:

	National domaincrawl	Specific selection
Competence: decentralised	NO	YES
Competence: centralised	YES	YES

Decentralised competence - national domain crawl

The analysis shows that the existing legal framework for the protection of copyright, related rights and database rights does not provide sufficient room for the adoption of a limitation of exclusive rights that would allow all cultural heritage institutions to carry out national domain crawls.

The option in the field of copyright and related rights to exempt "specific acts of reproduction" made by publicly accessible cultural heritage institutions offers the broadest basis for allowing acts of reproduction, such as the scraping, storing, indexing and archiving of web pages in the context of a national domain crawl (3.3.1.5). This rule is included in the European list of permitted exceptions and limitations which the national legislator is free to implement (Article 5(2)(c) InfoSoc Directive). The Dutch legislator has not made use of this possibility so far. The neutral condition of "specific" acts could offer room to introduce a clearly defined exception allowing non-commercial archives and publicly accessible libraries to make reproductions of websites in the context of web harvesting.

However, Recital 40 of the InfoSoc Directive raises several questions regarding this solution. Recital 40 excludes the application of the exception where the ultimate purpose is to make the harvested materials available in the context of online delivery. This wish – making archived websites available online, for example via an existing online infrastructure such as CLARIAH – was expressed by Dutch cultural heritage institutions (2.6). Recital 40 also clarifies that the exception should be limited to certain special cases that are covered by the reproduction right. This raises the question whether an exception can be deemed sufficiently specific when web harvesting involves various acts of reproduction and encompasses harvesting activities of all non-commercial archives and publicly accessible libraries in the Netherlands. Even if the three-step test does not seem to present insurmountable obstacles in this respect (3.3.4.2), the notion of "specific" acts could indicate a level of specification that would not allow the creation of a general harvesting competence for all heritage institutions. In this regard, the analysis of foreign legal systems shows that no examined country invoked Article 5(2)(c) InfoSoc Directive as a basis to justify a decentralised approach entitling all cultural heritage institutions to carry out national domain crawls. Furthermore, the CJEU applied Article 5(2)(c) in *TU Darmstadt* only to books that were already part of the collection of a university library (3.3.4.1). Therefore, it cannot readily be inferred from the *TU Darmstadt* decision that Article 5(2)(c) of the Infosoc Directive is applicable in cases where protected material, such as a copyright-protected web page, is not already part of the collection of a cultural heritage institution, but must first be integrated into the collection.

A broader interpretation of Article 5(2)(c) of the InfoSoc Directive in the Netherlands would therefore create legal uncertainty. It cannot be ruled out that a global domain crawl competence would be considered excessive and not "well-defined" when enabling all cultural heritage institutions to make web content part of their collections. Moreover, database law does not contain a corresponding exception (3.3.3.2). Thus, a regime based on "specific acts" of reproduction would not cover database rights.

In sum, it does not seem advisable to establish a decentralised competence entitling all cultural heritage institutions to carry out national domain crawls. This applies even more strongly to the universal approach outlined above, whereby in addition to cultural heritage institutions, individuals, NGOs and companies would have the option of conducting a national domain crawl.

Central competence - national domain crawl

In contrast to the described legal uncertainties arising from a broad decentralised approach, a centralised national domain crawl competence for selected cultural heritage institutions seems defensible when this harvesting possibility is based on a statutory deposit obligation covering online material. Denmark, Germany, France and the UK have chosen this route (Chapter 4). A legal obligation to deposit online material ensures that web content becomes part of the collection of cultural heritage institutions. Once this step has been taken, it becomes possible to rely on limitations of copyright, neighbouring rights and database rights that are linked to the collection status, such as the exemption of preservation copies for the archiving of web pages (3.3.1.2) and the reading terminal exception for providing access to the web archive (3.3.2.1). Furthermore, it is possible to introduce limitations of protection that allow for reproductions and acts of making available for the purpose of scientific research (3.3.1.4). In this respect, it should be mentioned that the database right can be considered as a potential source of legal uncertainty because of the lack of exceptions for acts of making available to researchers and via reading terminals (3.3.3.2). However, in the light of the special infringement criteria that have evolved, inter alia, in the case law of the CJEU, infringement of sui generis database rights does not seem plausible. Historical archival versions of databases that are incorporated into a web archive as a result of web harvesting and accessed according to specific access rules will usually not prevent the rightholder from recovering investments (3.3.3).

In all foreign legal systems under examination, the application of a statutory obligation to deposit online material led to a centralisation of web-harvesting activities in the hands of individual cultural heritage institutions, such as the national library. This ensures a clear institutional demarcation of the web-harvesting competence. Furthermore, it can be said that the bundling of the harvesting authority contributes to a clear and precise definition of the use privilege. Legislators in countries with a legal deposit obligation do not hesitate to refer to web harvesting activities as "specific acts" of reproduction in the sense of Article 5(2)(c) of the InfoSoc Directive (4.2.2.3).

Legislation that centralises the web-harvesting competence in the hands of one or more heritage institutions need not take the form of deposit legislation. Although this is the case in most of the foreign legal systems that have been studied, the starting points for this solution in other countries differ from the situation in the Netherlands. In the other countries under examination, a legal deposit system already existed for analogue material prior to the adoption of web-harvesting legislation. This pre-existing system has simply been extended to online publications, with all associated consequences: a legal obligation arises for the website holder (publisher) and/or the library to deposit the material, based on the rules and guidelines that come along with traditional legal deposit legislation.[2] However, the Netherlands need not follow the same approach. A statutory deposit obligation ensuring that relevant websites become part of the collection of cultural heritage institutions can be included in specific web-harvesting legislation. It is not necessary to create general legal deposit legislation for all publications – analogue and digital – in this context. A place to introduce and accurately describe a specific competence and deposit obligation for online sources could be found, for example, in the laws that define the tasks of selected web-harvesting institution(s). In the Netherlands, for example, this would be the Higher Education and Scientific Research Act (*Wet op het hoger onderwijs en wetenschappelijk*

[2] In 2000, UNESCO issued a renewal of the *Guidelines for Legal Deposit Legislation* developed by J. Lunn in 1981, which addresses the deposit of electronic material in more detail. The guidelines also set out what a legal deposit system should look like. Choosing for deposit legislation means that the legislator will have to take this framework into account. It strongly advises against a voluntary system. See: J. Larivière, *Guidelines for Legal Deposit Legislation,* UNESCO 2000, available at: http://www.unesco.org/new/en/communication-and- information/resources/publications-and-communication-materials/publications/full-list/guidelines-for-legal- deposit-legislation/.

onderzoek) for the Dutch National Library and the Media Act (*Mediawet 2008*) for the Institute for Sound and Vision (2.2.4.3).

An alternative solution for web harvesting can be found in Portugal. The Portuguese system is primarily based on a societal consensus that is also shared by rightholders (4.2.5). In Portugal, a centralised harvesting system has emerged on this basis – without a legal deposit obligation – which, in terms of volume and frequency of the national domain crawl and the degree of accessibility of the web archive, appears to be more extensive than the harvesting activities in many other countries. It is particularly noteworthy that the Portuguese institution Arquivo.pt has managed to make the web archive – with search functionality – available to the entire Internet public. Despite the considerable volume of harvesting activities, Arquivo.pt has not yet been confronted with infringement claims in Portugal. The system appears to be largely the result of implicit consent and/or tacitly granted licences.

This alternative model may offer new impulses for the discussion in the Netherlands as well. In the light of the experiences in Portugal, the question arises whether heritage institutions could expand their harvesting activities – both acts of reproduction (crawling and archiving) as well as acts of making available – based on an analysis and assessment of infringement risks and the development of appropriate measures to reduce those risks. In Portugal, an embargo period and an efficient notice-and-takedown system are central pillars of the harvesting system. The embargo period prevents direct competition with the exploitation of recent versions of source material offered on the Internet. The notice-and-takedown system offers rightholders the possibility to report infringements and ensure the blocking of access to problematic archive material. In the Netherlands, similar building blocks could also allow a switch from the current burdensome system with permissions and opt-out mechanisms, to a system with an embargo period and notice-and-takedown options. The Dutch legislator could support this solution by introducing an indemnity against liability comparable to the safe harbour regime for hosting services (3.5.3). It is also conceivable to give specific cultural heritage institutions an explicit legal mandate (perhaps even reflecting the desirability of giving the entire Internet public access to archived material) to emphasise the legitimacy of web harvesting and dispel potential doubts about the scope and reach of harvesting activities developed by heritage institutions.

A further solution could follow from a system of extended collective licensing (ECL). A basis for this can be found in Article 12 of the DSM Directive and Article 45 of the Dutch Copyright Act (*Auteurswet*). Article 12 allows EU Member States to introduce an ECL scheme for clearly circumscribed purposes of use on their own territory. The legal form of ECL is characterised by a combination of voluntary collective management with a statutory extension of the scope of the collective license to rightholders who are not affiliated with the collecting society in question (3.3.4.5). The ECL system should only be applied in well-defined fields of application, where obtaining permissions from rightholders on an individual basis is, as a rule, cumbersome and impractical to such an extent that the required licensing transaction is unlikely to take place. In the case of web harvesting, this condition appears to be fulfilled. The question is, however, whether sufficiently representative collecting societies exist in the Netherlands with regard to all protected elements of web sites that would require rights clearance.

Various collecting societies exist in the Netherlands, including those for writers, translators and freelance journalists (Lira) and visual artists (Pictoright). However, these collecting societies do not represent all rightholders whose works are used in the context of web harvesting. For example, web designers are not represented as such in any Dutch collecting society. The ECL system, therefore, does not seem capable of providing a complete solution for web harvesting. In addition, it is questionable whether it is desirable from a cultural policy perspective to make the future of web harvesting dependent on collective licenses that have to be agreed upon voluntarily. If no agreement with a collecting society can be reached, no web

harvesting will take place. Even Scandinavian countries, which traditionally have very well-developed ECL systems, have given preference to a solution based on a statutory deposit obligation to enable web harvesting. The analysis of the system in Denmark (4.2.1) attests to this policy choice.

Decentralised competence - specific selection

The fact that centralised systems provide greater legal certainty for a national domain crawl need not lead to the conclusion that heritage institutions outside the centralised system can no longer undertake any harvesting activity. While, as described, it is not advisable to give all cultural heritage institutions general authority to carry out national domain crawls, the legal regimes for the protection of copyright, neighbouring rights and database rights leave room for more specific harvesting projects, covering only pre-selected web sites. The existing legal framework offers room for harvesting activities with this specific focus. The available breathing space goes beyond the option to exempt specific acts of reproduction.

For example, the new rules relating to text and data mining (Articles 3 and 4 of the DSM Directive and Articles 15n and 15o of the Dutch Copyright Act, 3.3.1.3) provide a basis for targeted harvesting activities in the context of scientific research projects, if web content is retrieved for the purpose of analysing it and generating information for research purposes.

A similar conclusion can be drawn with regard to the possibility of permitting acts of reproduction and disclosure "for the sole purpose of […] scientific research" (Article 5(3)(a) of the InfoSoc Directive, 3.3.1.4). As the text and data mining rules, this copyright limitation requires a sufficiently clear connection with scientific research. Web harvesting based on a "reasoned selection" of web content or following from an "event harvesting" initiative, might be justifiable under this provision, when it is carried out in the framework of joint projects between heritage institutions and research organisations. However, there is currently no room for this solution in Dutch law because the optional limitation of exclusive rights for the "sole purpose" of scientific research has not been implemented in the Netherlands.

A legal regulation of selective harvesting activities on the basis of the new legislation on text and data mining and the exemption of use for the purposes of scientific research has the advantage that sui generis database law – at least with regard to acts of reproduction – explicitly offers comparable breathing space (3.3.3.1 and 3.3.3.2). This solution also makes it possible for individuals, NGOs and companies to harvest websites in the context of text and data mining activities. The scope of Article 4 of the DSM Directive and Article 15o of the Dutch Copyright Act is not limited to cultural heritage institutions. Hence, these provisions offer a basis for a broader regulation including other interested parties.

Central competence - specific selection

The described possibilities for justifying web harvesting with a specific focus on pre-selected online content are available to all cultural heritage institutions. This includes those heritage institutions that are specifically entitled to carry out general national domain crawls in a centralised system. They can thus conduct both national domain crawls and specific harvesting projects with a particular focus.

Privacy and data protection law

The analysis also explored the possibilities in the field of privacy and data protection law. In this respect, the Dutch legislator seems to have made a conscious (see the Explanatory Memorandum) albeit implicit (see the actual text of the Dutch implementating legislation of the General Data

Protection Regulation (GDPR), the *Uitvoeringswet Algemene verordening gegevensbescherming* (UAVG)) choice to bring the processing of personal data in the context of web archiving under the umbrella of the exception regime for journalistic, academic, artistic and literary purposes (3.4.2.4). A possible reason for this is that the GDPR provides only limited room to set up a broad exemption framework for "archiving purposes in the public interest". Although an explicit mention of web archiving by heritage institutions under the generous exception regime in Article 43 UAVG ("for journalistic purposes and the purposes of academic, artistic or literary expression") would create more clarity, one can also wonder whether the Dutch legislator is thereby pushing the limit of what is allowed under the GDPR.

However, given the scope offered by Article 85 GDPR, this is a legally defensible choice that also explicitly links the contribution made by heritage institutions to freedom of expression and freedom of information. Consequently, the Dutch legislator has exempted heritage institutions such as the Dutch National Library from many duties and responsibilities under the GDPR and the UAVG. The exemption regime for journalistic, academic, artistic and literary purposes practically exempts web harvesting project from the entire spectrum of obligations in the UAVG, as well as various provisions in the GDPR. As a result of this decision, the Dutch National Library – and other heritage institutions with similar objectives – have significantly less work in interpreting and implementing the GDPR and the UAVG.

An alternative regulatory option emerged from the examination of foreign legal systems. Unlike the Dutch solution on the basis of Article 43 UAVG, Germany, France, Portugal and the UK include web harvesting in the exemption for "archiving purposes in the public interest" (Article 89 GDPR). All these countries also list various rights of data subjects that cannot be invoked under certain conditions (see overview in 4.4).

Other civil, criminal and administrative legal problems and possibilities

Finally, the analysis also offers starting points for a legal solution in the area of unlawful content that may end up in a web archive as "by-catch" in the context of a national domain crawl. For several reasons, it makes sense to use the liability privilege for hosting (Article 14 of the E-Commerce Directive and Article 6:196c of the Dutch Civil Code (*Burgerlijk Wetboek*)) as a yardstick in this respect and assume that cultural heritage institutions can benefit from the same liability privilege with regard to unlawful content that is available for hosting services, unless there are special circumstances.

Firstly, cultural heritage institutions that engage in web harvesting are unintentionally collecting unlawful or illegal content which third parties make available on the Internet. This can at least be assumed as long as a cultural heritage institution has not itself played an active role in the selection of content that finally becomes a part of the web archive and the inclusion of problematic content is the result of an automatic harvesting process. Secondly, it should not be overlooked that web harvesting makes an important contribution to freedom of expression, freedom of information and the right to science (3.2.6). Against this background, it would be problematic to burden cultural heritage institutions with a liability risk that could frustrate harvesting activities. The analogous application of the hosting liability privilege removes this risk. Third, a similar limitation of liability can be deduced from general principles of liability law. Fourth, from a practical point of view, a solution based on the liability regime for hosting – known from the E-Commerce Directive – provides maximum guidance for establishing mechanisms for notifying potentially problematic content and ensuring the takedown (or blocking of access), namely the introduction of notice-and-takedown procedures that could be refined in the light of future developments in the field of the proposed Digital Services Act.