NATIONAL LIBRARY OF FINLAND

Collection plan for online materials 2021–2024

Jari Heikkinen, Kaisa Kaunonen, Erik Lindholm, Mikko Merioksa, Matti Pitkälä, Aija Vahtola & Petteri Veikkolainen

National Library of Finland 2021

ISBN 978-951-51-7281-5

Collection plan for online materials 2021–2024

Plan on the scope of the collecting of online materials and on the related deposit practices, pursuant to the Act on Collecting and Preserving Cultural Materials (1433/2007, 9 §).

Development targets and focus areas

- constantly developing the National Library's technical competence and storage media as online publishing becomes more diverse
- further redesigning of the deposit infrastructure to include all materials covered by legal deposit
- increasing cooperation with researchers and experts to survey materials to be stored in the web archive
- making available and promoting research use of the archived materials.

1. Introduction

The duties of the National Library of Finland include archiving Finnish publications according to the Act on Collecting and Preserving Cultural Materials (1433/2007). According to the Act, the National Library must present a plan on the scope of the collecting of online materials and on the related deposit practices, to be approved by the Ministry of Education and Culture (9§).

This collection plan for online materials covers the period 2017–2020. The plan takes into account the needs of pertinent research and cultural-historical archiving as well as the equal treatment of online publishers as specified in the Act (9 §). The plan may be reviewed during its period of validity if the Finnish publishing industry or the technical or financial resources available to the National Library change significantly.

In addition to the Act on Collecting and Preserving Cultural Materials, archiving of cultural heritage is regulated by Unesco (https://unesdoc.unesco.org/ark:/48223/pf0000244280 - link checked on March 9, 2020). International cooperation with e.g. IIPC (International Internet Preservation Consortium) and other organizations archiving online materials will continue to gain importance. As online publishing constantly changes and becomes increasingly international, cooperation is needed on both selecting the materials to be preserved and developing and implementing collection and preservation technologies.

A continuing key issue on the strategic period 2021–2024 is the development of legislation guiding this work, both in terms of archiving material and of the research use of the archived materials. As the amount of digital material continues to rapidly increase, the National Library must increasingly focus on which material to collect and how to develop collection technology. Promotion of research use of the materials will be given even more attention.

2. Materials to be collected and collection methods

Collection to the Finnish Web Archive of the National Library mainly includes websites available online to the public, but also materials protected by a paywall, especially news contents. Social media contents are also collected (e.g. Twitter, Facebook, YouTube, Instagram), both in themed web harvests and as a continuing collection.

Other online publications, such as e-books, digital music, and e-magazines are collected in several different methods:

a) as regular mass deposits directly from a web publisher or aggregator to a National Library server
b) the publisher deposits material through a web service made for e-legal deposit, an online form (at https://luovutuslomake.kansalliskirjasto.fi/?lang=en),
c) automatic collection of materials, when an interface enables this (e.g. government or higher education institutional repositories)
d) exceptionally, material may be submitted on another suitable manner, such as on a storage device, if other collection methods are not possible.

2.1 Collecting websites

Materials openly available to the public through information networks are collected

1. through the yearly Finnish domain web harvests
2. through themed web harvests focusing on a specific topic or event
3. through web harvests of continuously updating contents of newspapers, magazines, and news sites
4. through regular collection of social media contents (e.g. Twitter).

The yearly Finnish domain web harvests collect an overview of Finnish online publishing. In this harvest, conducted at least once a year, the National Library collects websites on the .fi or .ax domains, as well as Finnish webpages from other domains, using language recognition tools.

Themed web harvests cover a broader and more comprehensive sample of online material focusing on a specific topic or material type, also from social media. Themed web harvests are made e.g. on the following topic areas, for example:

1. important national and international events, e.g. elections and state visits
2. other important events, e.g. theme years, cultural events, sports events
3. acute international situations in, e.g. politics, societies, or ecology.

The National Library surveys the online materials to be included in themed web harvests also in cooperation with researchers, other experts, and the general public. Themed web harvests on international topics and phenomena may be conducted in cooperation with, e.g., the IIPC or organizations responsible for online archiving in different countries.

Web harvests of continuously updating contents collect media contents published online. Collecting is made daily, weekly, or monthly, depending on the updating frequency of the website.

Regular collection of social media contents collect publications openly available online by central figures in society, politics, culture, and economic life as well as other figures widely visible in social media.

Web harvesting and collecting methods

The technology for automatic collecting of online material has changed, due to the increased diversity of material to be collected, from a single established method (the Heritrix web crawler) to a selection of different methods which require constant effort to develop and to make compatible with each other. For example, the Heritrix web crawler may not be able to harvest material from behind paywalls or from social media platforms. To collect such material, the National Library has already adopted and continues to develop open-source applications which are used to log on to the websites and to collect material either directly or through the interfaces provided by the platforms. New types of material require new processing solutions in order to ensure their long-term preservation. The National Library is involved in international cooperation aiming to improve collecting tools.

2.2 Collecting other digital publications

Online materials which cannot be collected automatically are requested as deposits to the National Library (§ 8, Act on Collecting and Preserving Cultural Materials; Government proposal 68/2007). Such publications typically include:

1. e-books and e-magazines, as well as newspaper archives
2. online music and games
3. online publications not available through open interfaces, issued by universities, government, NGOs, and other organizations.

New forms of publications emerging with the continuing development of online publishing are also collected through deposit requests if they cannot be collected automatically.

For publications requested to be deposited, the National Library also requests related metadata to be deposited (§ 20, Act on Collecting and Preserving Cultural Materials), primarily in the metadata formats used in the publishing industry (e.g., Dublin Core, ONIX, NewsML, MARC 21). Deposited metadata is then converted – when necessary – to metadata formats used by the National Library. On publications deposited through web form, the depositor enters metadata into the form, which can subsequently be used in bibliographic description and subject indexing of the materials.

Deposit copies of publications included in the institutional repositories maintained by the National Library will be collected directly from the repositories. The institutional repositories include publications from government, universities, and universities of applied sciences. Other institutional repositories as well as other materials available through open interfaces, such as academic online journals, are collected in the same manner. These publications will, then, be incorporated into the National Collection, and their metadata can be used in describing the materials in the national bibliography and in other library databases, through conversions. Thus, apart from archiving and long-term preservation of online publications, this also promotes access to open science.

Deposit requests are sent for material covered by the legal deposit requirement when collecting through requests of making harvesting possible have not been successful with the technology available. Deposit processes are made in cooperation between the National Library and online publishers, and they are carried out so that the effort for both parties is as reasonable as possible. Deposit practices and infrastructures are being harmonized and improved, to better cover all materials under the legal deposit requirement. The National Library gives guidance and advice for depositors of online materials.

Legal deposit of online material and their metadata are primarily made in cooperation with online bookstores, online music stores, or aggregators – they shall regularly deposit online publications of publishers available through their services. The publisher will, however, deposit the material if it is not available through other channels. Government bodies not using the institutional repository services will deposit their material upon request by the National Library.

All material to be deposited must primarily be delivered in the material type specific file formats that are either recommended or acceptable for transfer, as specified by the Digital preservation service of the National Digital Library [http://digitalpreservation.fi/en/specifications - URL corrected September 3, 2020]. Material type is determined by the primary content of the material.

In terms of collecting online material that is particularly problematic for archiving, such as databases, online teaching materials, online games, and maps, the National Library keeps abreast of international development and participates in national and international cooperation projects on the field, whenever possible.

3. Use and long-term preservation of online materials

The National Library is responsible for collecting and submitting the materials for long-term preservation. Collected materials are available at legal deposit workstations maintained by the National Library, pursuant to the Copyright Act (404/1961, § 16 b).

Submission of collected online materials to the National Digital Preservation Service (http://digitalpreservation.fi/en) has been started for the Finnish Web Archive, comprehensive long-term preservation of other materials will be started on this collection plan period. Transfer of materials to long-term preservation will be automatized as much as possible. Apart from long-term preservation, the National Library also has the duty to provide user copies of all materials collected.

4. Conclusion

Online publishing diversifies and grows constantly, which constantly brings new challenges to collecting and preserving online publications as specified in the legislation. Publishing methods, available formats and technologies keep developing and changing. Some technologies remain short-lived, while others became an established practice. Cooperation with the publishing industry and with other domestic and international web archiving organizations is an absolute necessity for collecting and preserving online publications, along with constant development of staff competencies.

Cooperation with researchers and other experts is becoming increasingly important in collecting online materials, both in terms of surveying the material to be collected and in promoting research use of online materials.

The current legislation on collecting and preserving online materials was enacted on 1 January 2008. The entire publishing industry has since been through a powerful change towards the digital, and the current legislation does not fully respond to the needs of collecting cultural heritage, when it comes to digital publishing. Emerging digital research methods have brought forward the need to renew the legislation on use of the collected online materials as well; in part in connection with the European Union on directive on copyright and related rights in the Digital Single Market (the DSM-directive, (EU) 2019/790), and its implementation into national legislation.