Open, Extended, Closed or Hidden Data of Cultural Heritage

Tuula Pääkkönen[1[0000-0003-3958-9732]] and Juha Rautiainen[1] and Toni Ryynänen[2] and Eeva Uusitalo[2]

[1] National Library of Finland, Finland
[2] The Ruralia Institute, University of Helsinki, Finland

Abstract. The National Library of Finland (NLF) agreed on an "Open National Library" policy in 2016[1]. In the policy there are eight principles, which are divided into accessibility, openness in actions and collaboration. Accessibility in the NLF means that access to the material needs to exist both for the metadata and the content, while respecting the rights of the rights holders. Openness in operations means that our actions and decision models are transparent and clear, and that the materials are accessible to the researchers and other users. These are one way in which the NLF can implement the findable, accessible, interoperable, re-usable (FAIR) data principles [2] themes in practise.

The purpose of this paper is to view the way in which the policy has impacted our work and how findability and accessibility have been implemented in particular from the aspects of open, extended, closed and hidden data themes. In addition, our aim is to specify the characteristics of existing and potential forms of data produced by the NLF from the research and development perspectives. A continuous challenge is the availability of the digital resources – gaining access to the digitised material for both researchers and the general public, since there are also constant requests for access to newer materials outside the legal deposit libraries' work stations.

Keywords: digital cultural heritage, open data, accessible data, collaboration

1 Introduction

National Library of Finland (NLF) is the oldest and largest scholarly library in Finland [3]. The end-users or clients of the library are researchers, citizen scientists and students covering wide range of material needs. The collection includes books, newspapers, journals, ephemera published in Finland, where part of the collection is digitized and offered either freely online or within the special workstations at the legal deposit libraries in Finland. Recently NLF has also started to offer open data, which it has or it has produced such as the national bibliography [4]. Major achievement was also that a data catalogue was recently created at http://data.nationallibrary.fi to offer information of data and interfaces, which library offers. Open data is important for increasing new kind of availability for citizens, researchers and developers, to allow them to utilize material

by analysing it with the methodological tools at their disposal in their own environment. What then makes the open data "good" for an end-user?

In this paper we will process the following questions. What kind of data categories there exist? What kind of alternatives are there to open digitized material use in Finland? What kind of benefits digitized materials can bring to researchers and other users? These issues have been explored in several development projects of NLF over the years, resulting in few potential alternatives and solutions for the use of the open digitized material.

1.1 Digitised materials and their access rights

NLF has wide and versatile collections. The access to the material depends on the way how the material has become part of the collections of the library and also on the age and the type of the material. The **Fig. 1** lists the many types of the digital materials there is, where different copyright statuses. Material types create one aspect for considering availability, as there might be different rules as for how each can be utilized and even added to the collections of the library.

- Digitised copyright free monographies
- Digitised in-copyright monographies, with due diligence process
- Digitised in-copyright monographies
- Provided e-deposit copies
- Harvested monographies to web archive
- UKK published works
- Digitised newspapers (copyright-free and in-copyright)
- Digitised journals (copyright-free and in-copyright)
- Digitised journals with due diligence process
- Digitised newspapers with due diligence process
- Online news
- Legal deposit e-magazines
- Contractually deposited newspapers (print-pdf, etc)
- Contractually deposited journals (print-pdf,etc)
- Digitised ephemera
- Digitised copyright free music
- Digitised in-copyright music
- Legal deposit copies
- Digitised audio recordings
- Digitised maps
- Manuscripts' collections (letters, etc.)
- Music sheets
- Pergaments
- Collection catalogs
- Digitised board games
- Digitised game cassettes
- Photograps
- Videos, social media content
- Web pages

Fig. 1. The many types of digital material

In Finland the copyright act states that the author holds the copyright to her works during her lifetime and until 70 years after the year of the author's death. This applies both to monographs and articles in newspapers and journals. Published photographs, if they have required level of originality can have 70 years or lower depending on the situation. For example copyright organisation Kopiosto has used a 100 year limit for material being copyright-free [5].

1.2 FAIR principles

The findable, accessible, interoperable and reusable (FAIR) principles state for example that to be findable, metadata or data itself should have identifiers, which can be searched. With regard to accessibility, this requirement is expanded to such data is retrievable via a standardised communication protocol. Interoperability requires that knowledge is in a suitable language and is re-usable, for example, that the data usage license is defined [2]. These require planning in order to get interfaces up and running and to make formats suitable. For reusability and accessibility to the end-user, one solution is to have ongoing negotiations with the copyright organisations, and to negotiate which material could be opened where and with which rules. This enables one to honour the rights of the original rights holder, while still following the original duties and strategy of the National Library [1]. For a more custom solution, one way is to co-operate with the university network, faculties, research groups or even individual researchers to identify what material they would need. Depending on the field or the researcher's research objectives, the need or the viewpoint for the digitisation of materials can differ. The FAIR principles can also be applied together with the incoming copyright directive of the European Union (DSM-directive) [6]. So far, text and data mining (TDM) has been included in the directive; however in 2017, we will see how the final directive might rephrase TDM and who can utilise it. In addition, the general data protection regulation (GDPR) will come into effect in 05/2018 [7], and it will require that data usage is planned when the data contain personally identifiable information – especially when thinking about digital contents and their utilisation within library systems and library workflows; therefore, the planning could relate to FAIR principles.

The findable objective have been met with the new NLF data catalogue[4], actual cataloguing and metadata services of the materials and via the search features of the presentation system. Accessibility has been experienced in pilot projects like [8], ongoing Haka project [9], where education and research use have been in focus, respectively. The requirements for interoperability and reusability are also supported by the data catalogue of the NLF. The data catalogue offers freely downloadable data exports, interfaces to the library systems (bibliography, metadata and presentation systems) and documentation how to utilize data and services.

2 Open, Extended, Closed and Hidden Data

In a broad categorisation, the NLF has four different categories available: open, extended, closed and hidden. Open data are, for example, digitised newspapers in the presentation system at digi.nationallibrary.fi, where old editions of newspapers and journals, up to 1910, have been open for several years. During 2017, this material limit was extended to 1920, so now it is possible to read, for example, newspapers from the days when Finnish independence was declared. By extended materials, we mean materials that are digitised and opened for extended use via contracts. In past pilot projects, for example, this extended the use to a specifically defined workstations for specific

newspapers for a specific time range. This required contracts between publishers, rights holders and the NLF to enable extended usage.

2.1 Data Categories and Possible Transitions

In NLF, the open data means both the metadata and the contents of the digitized materials. Open data is available to all users regardless of time and location. At the minimum there is visual presentation available via library presentation systems, or metadata of the objects themselves. Depending on the needs sometimes the digital presentation via digi.nationallibrary.fi or via other systems, sometimes, e.g. researchers might need the export packages with which it is possible to do text or data mining in their own research infrastructures. The data can also be used via another research data management tools and combined with other data.

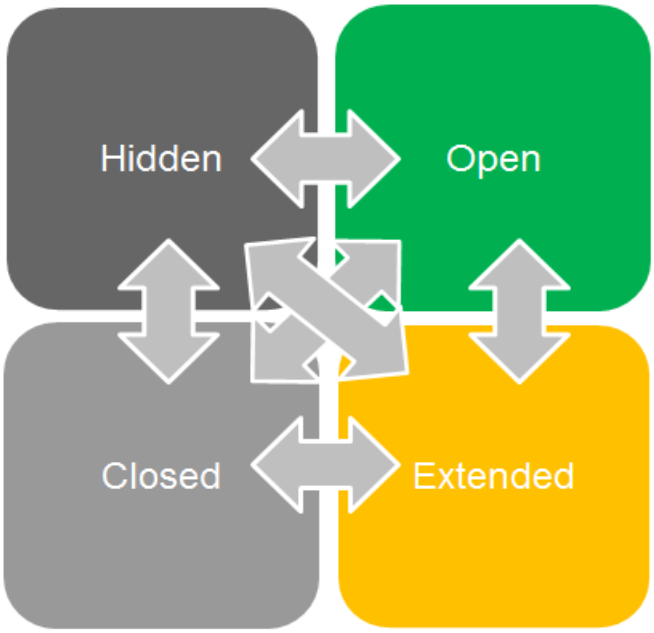

Fig. 2. Different access categories and their potential transitions

The **Fig. 2** illustrates the states in which the data can be accessed and to where data can transition after its production. From the library's point of view, it seems that new requirements in enriching data may highlight the need to hide certain data, while extracting other material could enable extending that particular part of the data onwards.

When looking at the way in which different data states operate, closed materials are the ones that are available in the legal deposit libraries – they are available and only

require a visit to one of the six university libraries in Finland, which have legal deposit rights. Closed materials also include materials that are difficult to digitise or that have complicated copyright structures. This is typically the case with newspapers in which each writer or illustrator of an article has individual rights to the work. Even finding the original authors can be complicated task after decades have passed.

Hidden materials are a versatile selection of various types of materials: works that have not been digitised and might not even be catalogued, or material that has been retracted, i.e., hidden from use at the request of the rights holder. Hidden materials could be quite interesting for research, since they might entail something for which nobody even comes to ask from the library, even though the content could be quite interesting. Even within open, existing data, there can be digitally available materials that require additional enrichment, in order for the user to find specific pieces of information from the large corpus because some phenomena cannot be mastered with well-defined search words.

2.2 Annual and Project-Based Development towards Data Availability

The National Library of Finland (NLF) like other libraries utilize special development projects in order to do improvements to either material availability, to the features of the presentation system or to the creating ways to get born-digital materials to the library in more efficient manner. In one such project, NLF opened its cultural heritage materials to various actors and organisations in the town of Mikkeli and its surrounding regions. We used on-location workstations as a way to extend access to digital, in-copyright newspapers to museums, schools and archives [8]. In this project the lesson-learnt was that the materials are interesting to citizen scientists, genealogists and in education and research [10, 11].

There is increasing aim to offer more materials to all users, who would benefit from specific collection. There are many different access levels to materials: some are copyright free, some in-copyright and in Finland it is also possible to extend usage of the materials via agreements. In addition, some material is so delicate that it can be reserved and viewed only in the reading rooms. In 2017, an agreement was made with the copyright organisation Kopiosto about opening digitized newspapers from years 1917-1920 to general public [12]. In year 2018, it was agreed to open newspapers and journals of the years 1918-1929 for the public [13] for the duration of one year. These agreements widen the access to the general public as one the major user groups.

In our past user surveys, researchers have been identified as one of the distinctive user of the newspapers and journals for research purposes [11]. These surveys [10, 11] have indicated that researchers benefit from having easy access to various digital data. Furthermore, especially with the data mining methodologies, previously hidden viewpoints and research questions can appear from the material. The possibility to new scientific findings and insights is also enhanced by the growing number of users across

the disciplines, whereas earlier the newspaper archives were mainly known by historians. Even more tempting future scenario is the possibility of having the sizeable collections of several memory organisations digitised, which then could be researched as one united material set. As suggested by Monastersky borders between data sources should be seamless [14].

2.3 Technical Implementation for User Access to Available Materials

In the digi.nationallibrary.fi, it is possible to assign rights to different materials based on IP-addresses, individual user accounts or via the HAKA authentication system. The name HAKA is the nickname of the Finnish identity federation of the Finnish universities, polytechnics and research institutes [15], and it is used for the authentication of staff, students and other affiliates of the university.

In an ongoing project, the HAKA authentication was implemented in the digital presentation system of newspapers, journals and technical metadata as a continuance to the earlier work with the aim of extending access [9]. Some of the pilots that have initially shown interested for pilot phase are for example University of Helsinki, and University of Turku. In the pilots the aim is not only to evaluate the technical feasibility but help us to self-assess our capabilities to create functional contracts and collaborations on top of the valuable digital asset of the Finnish cultural heritage. In the future, it will be possible to observe the way in which researchers use digitised materials and how we can improve our collaboration further with researchers. Linked open data could be one way to redefine how cultural heritage materials are used one way create new experiences for new users [16]. This is what the Open National Library policy aims at, to gain more insights into which data is most useful for researchers of different fields; and whether there are some functionalities in the technical environment that might need to be added in the long run to enable users to utilise materials in their own ways.

3 Conclusions

In this paper we have shown what kind of data categories there exist, what kind alternatives we have seen in opening different data for different audiences in the past projects. We have also gone through the potential benefits; because the researcher can access the data and have more versatile access to the research content, which contributes to the validity and reliability of the research.

There are problems to be solved in ensuring equal openness to digitised materials to various user groups; however, they can be overcome with co-operation between researchers and library staff. There are also challenges concerning data-mining methodologies. One solution to overcome them is to form research collectives, including re-

searchers and library staff with skills in data mining. Perhaps, in the long run, one solution is to include basic knowledge about data mining in methodology courses in humanities and social sciences.

As we have learnt to know when working with our research questions via development projects, there are requests both from the general public and the society to open more data, but there are arising needs for extending the access but just for some distinctive user group. The technical capabilities requires work, when new authentication and authorization means are implemented and contract creation requires effort so that new contract models match the technical possibilities. In contract creation, the challenge is two-fold: to find interested parties and to define material sets, which answer to the needs of users. The National Library seeks actively opportunities to negotiate more generic contracts which could act as templates for other similar projects. The recently introduced HAKA-authentication will give a new way to reach more of the users from the research and education fields, so that we can extend the material access towards their needs. It could be taken into account that contribution of the library in the contracts would be the actual task of digitisation, cataloguing and ensuring access as defined in the strategy of the National Library of Finland.

Different data access states enable different solutions also for the researchers. If data is openly available, then the key is how to ensure the easy-to-use and in-depth search functionalities, so that it is possible to get an overview of the data. If data are closed, then solutions range from making the data available by cataloguing and digitising them. Finding possible (collective) contract models and funding sources could create new research solutions, which bring benefits to all parties, namely to original authors, rights holders, library itself and more importantly to the users.

Acknowledgements

Part of this work is funded by the HAKA project (The Ministry of Education and Culture, Kopiosto – copyright organization, Mikkeli University Consortium, The National Library of Finland).

References

1. National Library of Finland: Duties and strategy, https://www.kansalliskirjasto.fi/en/duties-and-strategy.
2. Hagstrom, S.: The FAIR Data Principles, https://www.force11.org/group/fairgroup/fairprinciples.
3. NLF: National Library of Finland (homepage), https://www.kansalliskirjasto.fi/en.
4. Kukkonen, S.: Finnish National Bibliography released as Open Data, https://www.kansalliskirjasto.fi/en/news/finnish-national-bibliography-released-as-open-data.

5. Timonen, J.-P.: Tekijänoikeuksien selvittäminen Kopioston kautta: pilottikauden malli. Digiarkistoista liiketoimintaa. (2016).
6. OKM: DSM-direktiiviehdotus (COM(2016) 593), http://www.minedu.fi/export/sites/default/OPM/Tekijaenoikeus/eu-yhteistyoe/Liitteet/DSM-direktiivi_COMx2016x_593_FINAL.PDF.
7. EUR-Lex: EUR-Lex - 32016L0680 - EN - EUR-Lex, http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016L0680.
8. Pääkkönen, T.: Increasing availability, data privacy and copyrights of digital content via a pilot project of the National Library of Finland. Liber Q. 26, 163–180 (2016).
9. Rautiainen, J.: Digitoitujen sanoma- ja aikakauslehtien käyttö yliopistoissa ja korkeakouluissa helpottuu – Tietolinja, https://tietolinja.kansalliskirjasto.fi/2017-2/1702-digi/, (2017).
10. Matres, I.: Digital historical materials for academics, educators, hobbyists, creatives and browsers: the visitors evaluate digi.kansalliskirjasto.fi, https://journal.fi/inf/article/view/59436.
11. Hölttä, T.: Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat, http://urn.fi/URN:NBN:fi:uta-201603171337.
12. Lehmikoski-Pessa, T.: Kansalliskirjasto avaa itsenäisyyden alun sanomalehdet digitaalisina saataville, https://www.kansalliskirjasto.fi/fi/uutiset/kansalliskirjasto-avaa-itsenaisyyden-alun-sanomalehdet-digitaalisina-saataville.
13. Arpiainen, H.: Kansalliskirjasto avasi lisää digitoituja lehtiaineistoja avoimeen verkkokäyttöön, https://www.kansalliskirjasto.fi/fi/uutiset/kansalliskirjasto-avasi-lisaa-digitoituja-lehtiaineistoja-avoimeen-verkkokayttoon.
14. Monastersky, R.: The library reboot: as scientific publishing moves to embrace open data, libraries and researchers are trying to keep up.(FEATURE: NEWS). Nature. 495, 430 (2013).
15. CSC: Federation - Haka-käyttäjätunnistusjärjestelmä - Eduuni-wiki, https://wiki.eduuni.fi/display/CSCHAKA/Federation.
16. Marden, J., Li-Madeo, C., Whysel, N., Edelstein, J.: Linked Open Data for Cultural Heritage: Evolution of an Information Technology. In: Proceedings of the 31st ACM International Conference on Design of Communication. pp. 107–112. ACM, New York, NY, USA (2013).