IFLA WLIC 2022 DUBLIN

Submitted on: 20.06.2022

Digitized Finnish Newspapers in Digital Humanities Research Projects: Challenges and Solutions from the Library Perspective

Juha Rautiainen
Research Library, The National Library of Finland, Mikkeli, Finland.
E-mail address: juha.rautiainen@helsinki.fi

Abstract:

Libraries have a direct influence on digital humanities (DH) research projects that utilize digitized newspapers from their collections. However, a library's level of participation in a project will vary and may range from resource provider to active collaborator. Libraries must manage the different expectations of each project; thus, a variety of services are needed.

The National Library of Finland continuously aims to improve its services and provide researchers with broader access to digitized collections. The library's DH research team has extensive experience in research projects, and they actively support the work of DH researchers. In this paper, I present an overview of the experiences of the DH research team. I focus on projects that have used digitized historical newspapers and journals, and I outline a range of potential problems and suggest possible solutions.

Keywords: research, newspapers, digitization, digital humanities.

1 INTRODUCTION

A library can have various roles in digital humanities (DH) projects, such as being an external source of the research data, a project partner, or something in between. The library will always influence the project, regardless of its specific role.

The DH research team of the National Library of Finland (NLF) has many years of experience in research projects that utilize digitized materials. The team includes the head of planning, three information systems specialists, and fixed-term project employees; their key responsibilities include supporting (DH) researchers, participating in research projects, developing the Digi system, and promoting access to digitized materials. The DH research team also works closely with several other units of the NLF, including Digitization and Research Services.

Digital humanities has many definitions[1], and researchers from various fields understand the concept somewhat differently. In general, DH can be said to operate at an intersection between the humanities and computer science, and in this article, I broadly apply this concept.

I use the models by [Oberbichler], which are briefly outlined in section 2, as an aid in this review, as they reflect the expectations that are placed on libraries involved in collaborative research. Section 3 presents several examples of different DH projects, and in section 4, I use the models as a reference to illustrate a range of challenges that can arise during a research project. As the models do not cover all project-related work, I also address several other relevant issues.

The NLF is constantly developing its services to meet the needs of its customers. In section 5, I describe several key solutions that have been introduced by the NLF to resolve the problems that have been encountered in research projects. Some of the solutions are newly implemented, and thus more time is required for a full evaluation; however, I present them here as useful examples for those considering similar questions.

1.1 The digitized newspaper collection and the Digi system

The NLF's digitized newspapers and the new electronically deposited newspapers are both hosted in Digi[2]. Digi contains over 23 million pages of digitized material, including over 14 million newspaper pages. About 2 million pages are added each year, and newspapers represent about half of the newly digitized content.

All material in Digi can be searched with free text, and the full-text of all non-copyrighted and some copyrighted resources can be accessed online either freely or by logging in as a research user. All copyrighted material is only available on the customer workstations in the legal deposit libraries. Some of the oldest non-copyrighted digitized resources, such as the digitized Finnish newspapers from the years 1771 to 1911, are also available as downloadable open data packages and through an application programming interface.

According to the user surveys conducted in 2018, 2019, and 2022, multiple user groups are served by Digi. Although the options for answering the question about intended use have varied between the surveys, genealogists have always represented the largest user group. Research for private purposes or browsing have also been among the most popular options in all the surveys, whereas scientific research has been selected as the fourth or sixth option.

2 THE PIPELINE AND WORKFLOW-BASED MODELS OF INTERDISCIPLINARY COLLABORATION

A mutual understanding between the parties involved in a project emerges through communication and designing a workflow for a project is one tool that can effectively initiate this exchange of ideas. A detailed workflow ensures that the necessary tasks and components are made visible and explicit, which helps identify the practical boundaries of each discipline and facilitates the negotiation of compromises.

I utilize two models presented by [Oberbichler] when reviewing the experiences of the NLF's DH research team; however, it should be noted that these models have not been used directly in any of the projects described in this article.

[1] See https://whatisdigitalhumanities.com/
[2] See https://digi.kansalliskirjasto.fi/

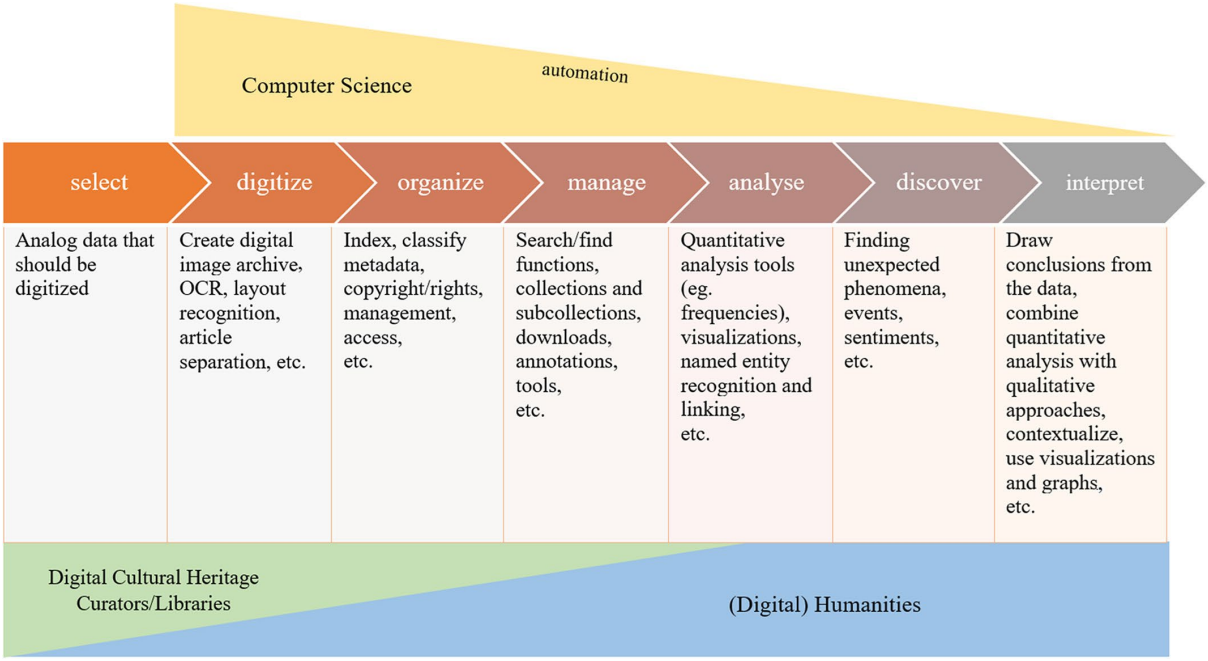

Figure 1: A workflow-oriented view of interdisciplinary collaboration as described in [Oberbichler] (CC BY 4.0).

[Oberbichler] first presented a workflow-oriented pipeline model for interdisciplinary collaboration. Figure 1 shows the involvement of each discipline and the application of the (automated) computational methods. Although the disciplines share common goals for working with digital newspaper collections, they have different objectives and apply different computational methods.

For example, adding metadata to provide context for the material is primarily a technical task for a computer scientist, but it is also an intellectual and labor-intensive task for a library team. In turn, DH researchers move back and forth between data, metadata, and qualitative analysis as they draw conclusions that combine all the research stages.

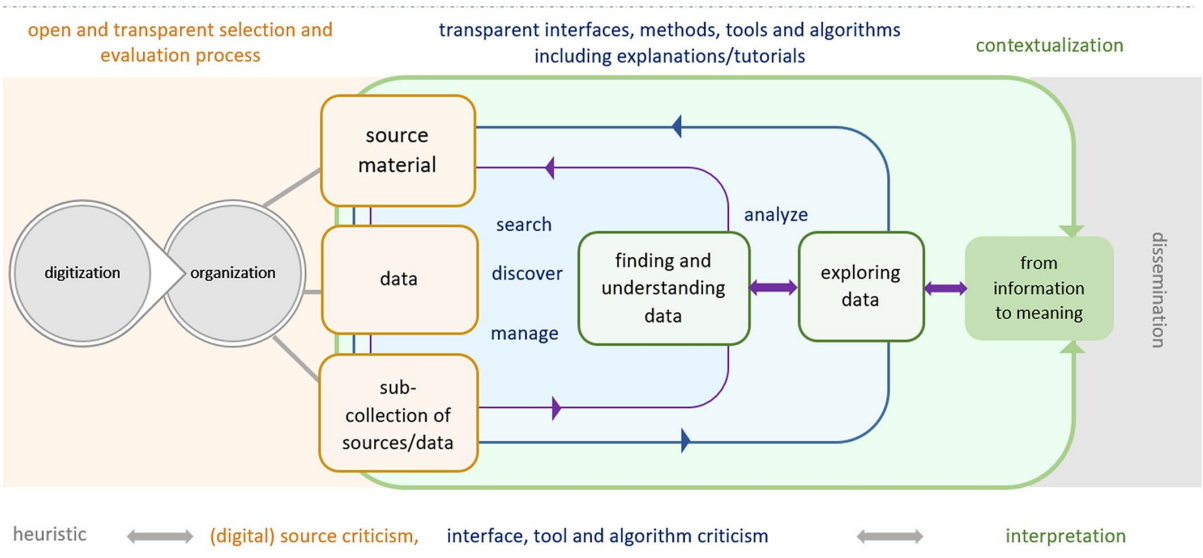

Figure 2: An integrated digital hermeneutics workflow as described in [Oberbichler] (CC BY 4.0).

[Oberbichler] also proposed an integrated digital hermeneutics workflow that combines disciplinary research approaches from computer science, humanities, and library work. This simple model summarizes the experiences and the insights gained from the multidisciplinary NewsEye project discussed in section 3.3. The workflow for the model is presented in Figure 2.

The integrated hermeneutics workflow model differs from the pipeline model in three main aspects: the role of data is emphasized, the iterative (qualitative) analytical steps are prioritized in order to gradually gain deeper insights into the data, and the model includes critical reflections of both the data and tools in the spirit of hermeneutics.

From the point of view of libraries as the curators of cultural heritage, it is important that each organization is aware of why and how they select material for digitization. The sharing of metadata is crucial, and each step of the digitization process must be traceable.

3 RECENT AND ONGOING PROJECTS

The following section provides examples of recent or ongoing projects that have used or are using the NLF's digitized newspapers. For this review, the projects are divided into three categories. First, there are projects in which the researchers work independently and only access the basic services of the library. In the second group of projects, the library's role is primarily to observe and advise, and in the third group, the library is actively and directly involved in the projects.

3.1 Independent projects

It is safe to assume that in most projects, researchers only use the library's standard services. A key issue from the library's perspective is that it can be difficult to obtain reliable information about these projects and their requirements. Therefore, the improvement and development of the related services are frequently based in part on assumptions.

Although the process is not always straightforward, there are several ways to obtain information about independent projects. Surveys can be used to directly ask the researchers about their work, or information can be sourced from research publications and news. The information collected during the use of the services can also be a good source of specific detail, or researchers themselves can inform the library about their own projects. The information gathered via these processes is frequently anecdotal, but it can still be helpful when developing the library's services.

An interesting study that initially came to our attention through a news article is *Historical trends in spring ice breakup for the Aura River in Southwest Finland, AD 1749–2018* by Norrgård and Helama[3]. The study compiled a fully revised and extended ice breakup series (1749–2018) for the Aura River for climate research purposes. The researchers used local newspapers as a complementary source of information when there were shortcomings in the statistics. In this case, all the newspapers were openly available online and the researchers did not appear to require any direct assistance from the NLF.

When an individual logs in to Digi as a research user, they must complete a short questionnaire before they can access restricted materials. The questionnaire does not specifically address research projects; however, many respondents freely provide information about their research. Although the group of respondents is somewhat selective, the answers provide an important perspective on the use of digitized newspapers as research material in Finland.

To date, the Digi questionnaire has received over 4500 answers, and it appears that the digitized material is often used as a traditional library tool to browse newspapers or search with free text.

[3] Norrgård, S., Helama S. (2019) Historical trends in spring ice breakup for the Aura River in Southwest Finland, AD 1749–2018. The Holocene. 2019;29(6):953-963. doi:10.1177/0959683619831429

However, other research approaches have also been listed. For example, several questionnaire answers have mentioned data mining, and one social science research project compared newspaper content with social media material using computational methods.

3.2 Projects in which the NLF is an observer

In between the independent research and the library's active involvement is a group of projects in which the library provides additional services but does not participate directly. Generally, the library will provide insights for these projects or use the opportunity to develop new information or skills.

An ongoing example is a subproject of FinClariah[4], which is led by the National Archives of Finland. FinClariah is a research infrastructure for Social Sciences and Humanities (SSH). The overall aims of the project are to significantly improve access to the research infrastructure and utilize resources across the SSH disciplines.

The subproject will further develop tools from participating universities to recognize named entities (personal, organizational and place names) in the extensive dataset of digitized archive files. The NLF has expertise in named entities recognition following several recent projects and the focused development in this area. It is intended that as an observer, the NLF will provide important insights for the project as well as benefit from the associated work.

3.3 Projects in which the NLF is actively involved

The library has a highly visible role in a project when it is actively and directly involved in the work. The degree of participation may vary from completing a well-defined task in an external project to managing an entire project internally. Both the library's expectations of a project and the partners' expectations of the library shift in relation to the share of the project's resources allocated to the library; NewsEye and Translocalis are two examples of this type of project.

NewsEye[5] was a research project funded by the European Union's Horizon 2020 research and innovation program. The three-and-a-half-year project, which ended in January 2022, involved three national libraries, several humanities and social science research groups, and computer science research groups from five universities. The main goal of the project was to develop methods and tools for effective research and exploitation of the digitized newspapers; the project utilized new technologies and 'big data' approaches and combined the 'close' and 'distant reading' methods of DH.

A particularly notable achievement of NewsEye for the NLF was the significant improvement in the quality of the optical character recognition (OCR) of the early newspapers printed with gothic typefaces. During the project, selected newspaper titles were re-OCRed using Transkribus. As the results were positive, the NLF ordered the re-OCR of all the digitized newspapers published before 1914.

The NLF participates in the Translocalis[6] project, which is managed by the Academy of Finland Centre of Excellence in the History of Experiences (HEX) located in Tampere University. The project collects and conducts research on the readers' letters that were written in the names of local communities and published in 19th century Finnish-language newspapers. The letters have been manually collected by HEX from the digitalized newspaper material with tools provided by the NLF. The project is creating a database containing all the readers' letters published between 1775 and 1885, and the NLF is developing some new features in the Digi system to support the use of the database.

[4] https://www.kielipankki.fi/organization/roadmap/
[5] https://www.newseye.eu/
[6] https://projects.tuni.fi/translocalis-en/

4 PRACTICAL PROBLEMS IN RESEARCH PROJECTS

In this section of the article, I present a range of challenges from the library's point of view that are related to the various DH research projects that have used digitized newspapers. These observations are based on both the experience gained from practical work and the information gathered in three user surveys [Pääkkönen 2018], [Pääkkönen 2019] and [Pääkkönen 2022]. I also highlight additional points for consideration that have been raised following a recent study completed for the Digital Open Memory (DAM) project led by the South-Eastern Finland University of Applied Sciences [Kosonen].

The first step in the pipeline model described in section 2 is the selection of the analog data for digitization. From a library's perspective, there are number of competing requirements that guide the selection and the outcome, including the condition of the material and its cataloging status, and these requirements do not always meet the researchers' expectations.

The digitization of the NLF's newspaper collection is yet to be completed; there is still material from the end of the 1940s through to 2017 that has not been processed. At the current pace of digitization, it is estimated that the library will take at least a decade to fill the existing gaps. Generally, it is not possible to select newspapers to digitize based on the needs of an individual project; however, a selection of recent titles has been digitized in cooperation with publishers or foundations involved in special projects. In principle, such projects help to fill the gaps in the collection. However, if sufficient additional resources are not added over the course of these projects, their implementation will slow down other digitization activities.

The integrated hermeneutics workflow model described in section 2 requires an open selection and evaluation process within libraries; however, this can be difficult to implement in practice. For example, there is limited publicly available information communicating the rationale for the large-scale digitization of the NLF's newspaper collection and the reasons for why certain titles have been digitized for longer periods.

The second step in the pipeline model involves digitization activities such as OCR, layout recognition, and article separation. OCR and, to some extent, layout recognition have been part of the NLF's standard processes for a long time. In contrast, article separation has not been applied because of the lack of advanced automated tools.

The problems in OCR quality tend to cumulate in the workflow, although some error-tolerant tools can hide the issues to a degree. In the 2022 Digi user survey, the most frequently used feature was the free text search. The quality of the OCR in the data varies, and this significantly affects the relevance of the free text search results. Based on the survey answers, however, the users appeared to be satisfied with the search results; therefore, we must consider how well the researchers understand the effects of OCR quality.

Article separation has been addressed in several recent research projects, including NewsEye. The user interfaces implemented in these projects have incorporated some advanced functions that have utilized article separation, and it is expected that the success of these tools will increase interest in article separation among researchers. However, it appears that user interest in this area is still relatively low, as it was not raised in the responses of the latest user survey.

Organize is the third step and involves processes such as copyright management. While the issue of copyright is not a factor for older material, it can limit the use of publications that are less than one hundred and fifty years old. Requests to increase the amount of material that is openly available have frequently appeared in the user surveys and in the customer service feedback.

[Kosonen] has observed that researchers do not always understand every aspect of the copyright legislation and the restrictions it imposes on libraries. This knowledge barrier has occasionally caused unnecessary friction between the library and researchers. From a research perspective, the copyright

protection, which is valid for 70 years from the author's death, can appear extremely long. Even if the material is available for use in the library, this access may not meet the needs of the DH researchers who frequently require electronic tools and computing power that are not available on library premises.

For example, the historians involved in NewsEye wanted to examine the newspapers from 1850–1950 with tools they had developed for the project. However, because of the copyright restrictions, the library could not release the more recent material, and the researchers could only use the newspapers in the NLF collection that were published before 1919.

Both models, the pipeline and the hermeneutic workflow, acknowledge that digitization produces data for the later steps. To maximize the usability of the data, the library responsible for the digitization must have a sufficient understanding of the tools that will be required in future stages. In practice, however, it has been difficult to obtain information about the applicable tools.

The models also state that libraries are responsible for producing adequate descriptions of the data. In the NLF's digitization process, the required metadata that is generated is attached to the digitized data. However, the metadata is often limited and may not meet all user needs; for example, it is difficult to obtain information on missing issues in a machine-readable form.

In this article, I have used the models presented in section 2 as a framework. Although the models provide a good basis for outlining the important discussions between the parties involved in a project, they do not cover all the noteworthy issues. For example, software engineering and performing manual annotations are generally not considered in the research process and are therefore easily overlooked in a project's planning stage. These issues can sometimes be resolved as they arise; however, the tasks that must be completed before and after the project are more problematic.

The NLF has recently given special consideration to the issues related to the project preparation phase, including sourcing sufficient expertise to assess a project's feasibility and ensuring the equal treatment of researchers. One area that has required attention is how researchers can propose collaborations with the library. To date, researchers have generally approached employees of the library who are familiar because of a previous connection; thus, the interest or workload of the individual can have a significant impact on the progress of the initiative. This informal approach should be considered a risk for both the library and the researcher.

Frequently, the researchers do not have a full understanding of the library's roles in the proposed collaborations. This is reflected in the fact that the researchers will often only contact the library at the end of the project planning phase, even when the implementation of the project requires considerable library resources. From a library's perspective, however, it would be more beneficial to be involved early in the planning in order to assess the resources required before applying for funding and during other stages of the project preparation.

At the end of a project, the outputs should be managed sustainably. Some projects, such as NewsEye, produce outputs that are relatively easy to utilize in the library; however, this process is not always straightforward or even possible. One reason for this difficulty is that the traditional methods of humanities studies produce data that can only be understood by the researchers involved in the specific project. This issue is particularly significant if an independent researcher contacts the library at the end of the project and asks the library to take responsibility for preserving the outputs and making them available for future use.

On the other hand, a recent survey [Näpärä, Lilja] revealed that only around 20 per cent of researchers deposit their material in a data repository after their research is complete. This indicates that the outputs offered to libraries are still limited in number, and overall, libraries are not set up to manage extensive data repositories.

5 THE NLF'S ATTEMPTS TO SOLVE THE PROBLEMS

The NLF is continually developing its operations to meet the needs of its customers. Adhering to the cornerstones of its operations[7] – openness, renewal, and bildung – the NLF has frequently sought to address the problems described above. While one solution may benefit a range of projects, some will target specific cases.

Increasing the information available is of particular benefit to independent projects. At the same time, it also supports the work of researchers whose projects require a higher level of library involvement.

The NLF has published the Digitization Program for 2021–2024[8], which provides researchers with an overview of the material that will be digitized in the coming years. The document explains the selection criteria for the digitized resources, the digitization process, and the associated resources. In addition, the strategic objectives are clearly outlined, including digitizing extensive, unified sets of resources, and supporting open science and the development of methods and tools for DH.

When new information is added to the Digi user interface, the NLF aims to inform the users about the decisions behind the development of the existing digitized collection. This process has shown that not all decisions have been documented with sufficient accuracy; for example, there may be information stating that an item was digitized in a certain project, but a clear justification for why the item was selected may be lacking.

The favorable results of the NewsEye project have been encouraging, and the NLF has requested that Transkribus carry out the re-OCR of the collection's oldest newspapers. It is thought that better quality data will support all research projects that use digitized newspapers. However, it is important that users are informed that the changes may lead to differences in both search results and data analyses.

Independent researchers have been able to download digitized newspapers as open data packages for several years. The feedback has indicated that some users have found the pre-made packages difficult to apply in their own projects. To provide more flexibility when downloading large amounts of material, the NLF has introduced a tool that allows the user to download any part of the open access material to their own device. The NLF is also addressing the metadata that has been limited and insufficient for all users; the internal data description group is currently working on improving the metadata, and this project should be completed later in the year.

Article separation has been targeted for future projects, and it has already been considered in the evaluation of Digi's development. Support for component parts has been integrated into the system for a project on digitized sheet music, and the article separation implemented during the digitization of several magazines has been made available via the user interface. Thus, the basis for the functionality exists when the automation of the article separation has been developed to a useful level.

The NLF has sought to resolve the constraints of copyright protection by licensing material for both public online use and limited research use. The Finnish Copyright Act defines an extended collective license through which an organization as the copyright holder can extend a license to cover authors and additional copyright holders not represented by the organization. This has allowed the NLF to reach an agreement with the copyright organization Kopiosto to provide open online access to Finland's oldest digitized newspapers and magazines, including all years up to the end of 1939. The NLF has also negotiated another agreement that provides authenticated researchers with online access to more recent digitized newspapers; this agreement also allows data mining for research purposes.

[7] Strategic plan of the National Library of Finland 2021–2030, https://urn.fi/URN:NBN:fi-fe2020042822711
[8] https://urn.fi/URN:ISBN:978-951-51-7103-0

The amount of copyrighted material available to users online has steadily increased, and this has led to a growing number of questions about the use of the material. Digi's user interface displays clear information about the licenses to prevent users from unknowingly violating the terms and conditions. This information is also aimed at helping independent researchers plan their data management, and it also raises the awareness of copyrighted material that may not be available for use.

As part of the DAM project, the NLF investigated the introduction of a library lab model. Although the lab was not established, there was a clear indication that many of its practices should be adopted. For example, as some digitized materials are not included in the license agreements, such as the most recent newspapers, the NLF is piloting a residency model. During the residency, researchers have wider access to digitized materials inside the library and, unlike standard local users, they can apply a variety of analysis tools.

The NLF has announced a new process to clarify the handling of initiatives involving research collaboration. Individuals who are interested in collaborative projects are now directed to contact the library through a standard form, rather than sending emails to individual library staff. This process provides similar input data for the projects for the evaluation phase. At the same time, the NLF aims to communicate the timeframes for processing the initiatives in the library, which should help reduce the number of last-minute requests.

The Translocalis project, described in section 3.3, is a good example of effective collaboration between a research project and a library. The research group and the NLF had contact during the project preparation phase, and they reached an agreement regarding the integration of the project's database into the NLF's Digi system. Thus, the researchers were provided with information about the type of data that was required, and the NLF was able to start developing the necessary functionality.

In order to follow the developments in the field of research, a wide-ranging dialogue with researchers and other key participants is essential. In addition to attending scientific conferences and other events, the NLF has established several collaboration forums.

The Steering Group of the Digital Humanities is one of the three advisory boards established by the NLF to develop services in cooperation with research communities. The steering group monitors the implementation of the NLF's DH policy and, if necessary, comments on the policy guidelines. The group also assesses the compatibility of the NLF's data services with the methods used by DH, and to avoid overlaps, they discuss the distribution of duties with other member organizations; these duties include, for example, the digital preservation services.

The Newspaper Symposium is a platform for interaction between the NLF and the research community. The symposium provides the NLF with information about on-going research projects, and it presents an opportunity to discuss the NLF's plans with the researchers in a setting that is less formal than a steering group. Unlike a general conference, the focus of the symposium is very specific; thus, the topics surrounding research on newspapers can be addressed in more depth. The symposium also provides up-to-date information on the tools and methods used in the field.

A few years ago, the NLF introduced a data clinic for the research community and graduating student. The clinic enables open discussions and provides information about the NLF's data resources. The aim is to increase the understanding of the library's processes and its potential role in DH research projects. A data clinic session consists of two parts: first, a lecture-style presentation provides background information about the NLF's data resources and processes, and second, the participants get to explore the NLF's digitized data and thus gain a practical understanding of its properties.

6 DISCUSSION

In some cases, the problems encountered in DH projects can be resolved by adding additional information or implementing technical improvements, including better quality digitized data. However, ultimately, effective communication between the library and the researchers is a critical success factor. At the very least, the library services should facilitate the researchers' use of the collection.

[Oberbichler] proposes the workflow concept as a tool to enable better communication and recommends that such a workflow should be implemented at the start of each interdisciplinary project. Even if a specific workflow has not been designed, the model can provide a useful starting point for planning the interdisciplinary collaboration in research projects, including ensuring that implicit assumptions are made explicit, increasing the understanding between disciplines, and identifying and motivating tasks that would otherwise could be overlooked.

Events such as data clinics and the Newspaper Symposium have been a good way to initiate conversations with DH researchers who have not previously been involved in research projects with the NLF. These discussions are important, as they provide both parties with new information and introduce new opportunities for cooperation.

While the library should be open to investigating new directions and opportunities, care must be taken to ensure that researchers fully understand the scope of the library's existing services. Well described and transparent basic processes should help support the success of any collaborative project.

References

Kosonen, M. ed. (2020) Kulttuuriperintöä käyttäjä edellä. South-Eastern Finland University of Applied Sciences. https://urn.fi/URN:ISBN:978-952-344-357-0

Näpärä, L., & Lilja, J. (2021). Kansalliskirjaston aineistojen jatkokäyttö tutkijakyselyssä. Informaatiotutkimus, 40(1), 27–62. https://doi.org/10.23978/inf.99425

Oberbichler S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., & Tolonen, M. (2021). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. Journal of the Association for Information Science and Technology, 73(2), 225– 239. https://doi.org/10.1002/asi.24565

Pääkkönen, T., Lilja J (2018). Hieno palvelu, mutta sisältöä lisää – Kansalliskirjasto kyseli Digi-palvelun käyttökokemuksia. Tietolinja, 2018(2). http://urn.fi/URN:NBN:fi-fe2018092336401

Pääkkönen, T. (2019) Internal report. The National Library of Finland.

Pääkkönen, T. (2022). Syvään päähän: Digi.kansalliskirjasto.fi-palvelun käyttäjäkyselyn tuloksia. Tietolinja, 2022(1). https://urn.fi/URN:NBN:fi-fe2022061546944