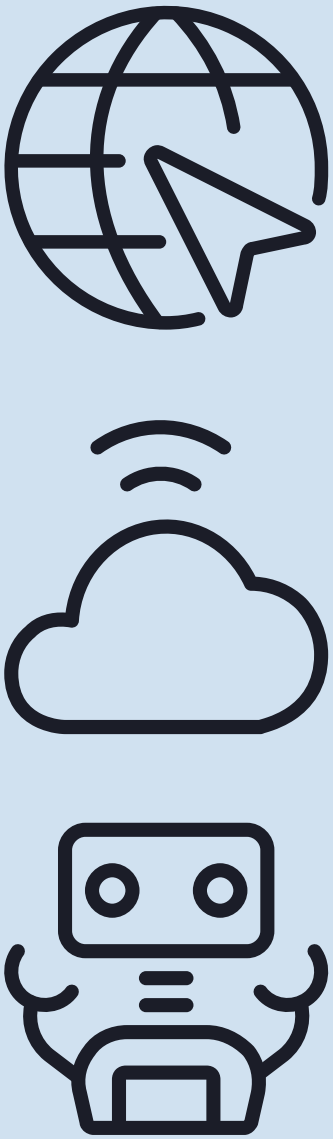
Virtual Lab at the National Library of Estonia



How to develop a virtual lab as a service?

RaRa NATIONAL LIBRARY OF ESTONIA

open digital libraries

Creative Europe MEDIA

Introduction

In recent years, GLAM Labs or virtual labs have gained popularity in cultural heritage institutions as an opportunity to manage their data more efficiently. In most cases, this means datasets derived from the collections of these institutions but also tools which assist users and facilitate access to the datasets. In order to engage with users, share expertise and test new methods, sectoral events are organised and cooperation networks developed.

For several years, the National Library of Estonia (NLE) has operated an open data portal (data.digar.ee), through which you can download limited versions of metadata both from the Estonian National Bibliography (ERB) and NLE's digital archive DIGAR; full texts of open access publications are available in DIGAR as well.

However, NLE's virtual lab should enable users to access the datasets as easily as possible, collect tools and guides, assist in using data and support the community through campaigns, resources and networking. In addition to an online environment, this calls for a comprehensive support service and a strong team.

Creating virtual labs for memory institutions is still in its early stages. Therefore, when setting up and in the future providing a virtual lab at the National Library of Estonia, we keep our minds full of experimental optimism and are not afraid of failures which we could learn from.

Contents

Introduction

1. What does a GLAM Lab or a virtual lab look like today and in the future?

2. Design process

2.1 Mapping of international analogues and examples

2.2 Mapping of user groups

2.3 User survey results

3. Visions of a virtual lab at NLE

The goal of the project is to create simple, modern and versatile opportunities for researchers, students, lecturers and other interested parties to use and work with the digital resources of the National Library of Estonia (NLE), specifically for text and data mining.

The project partnerts – the Royal Library of the Netherlands (Koninklijke Bibliotheek) and the Austrian National Library (Österreichischen Nationalbibliothek) – gave valuable input during the creation of these opportunities through the development of the virtual lab service.

The project was financed through the "Creative Europe" programme of the European Union. The content presented in this report reflects only the views of the authors. The European Education and Culture Executive Agency and the European Commission are not responsible for the use of the information contained therein.

1 What does a GLAM Lab or a virtual lab look like today and in the future?

The abbreviation GLAM stands for galleries, libraries, archives and museums.

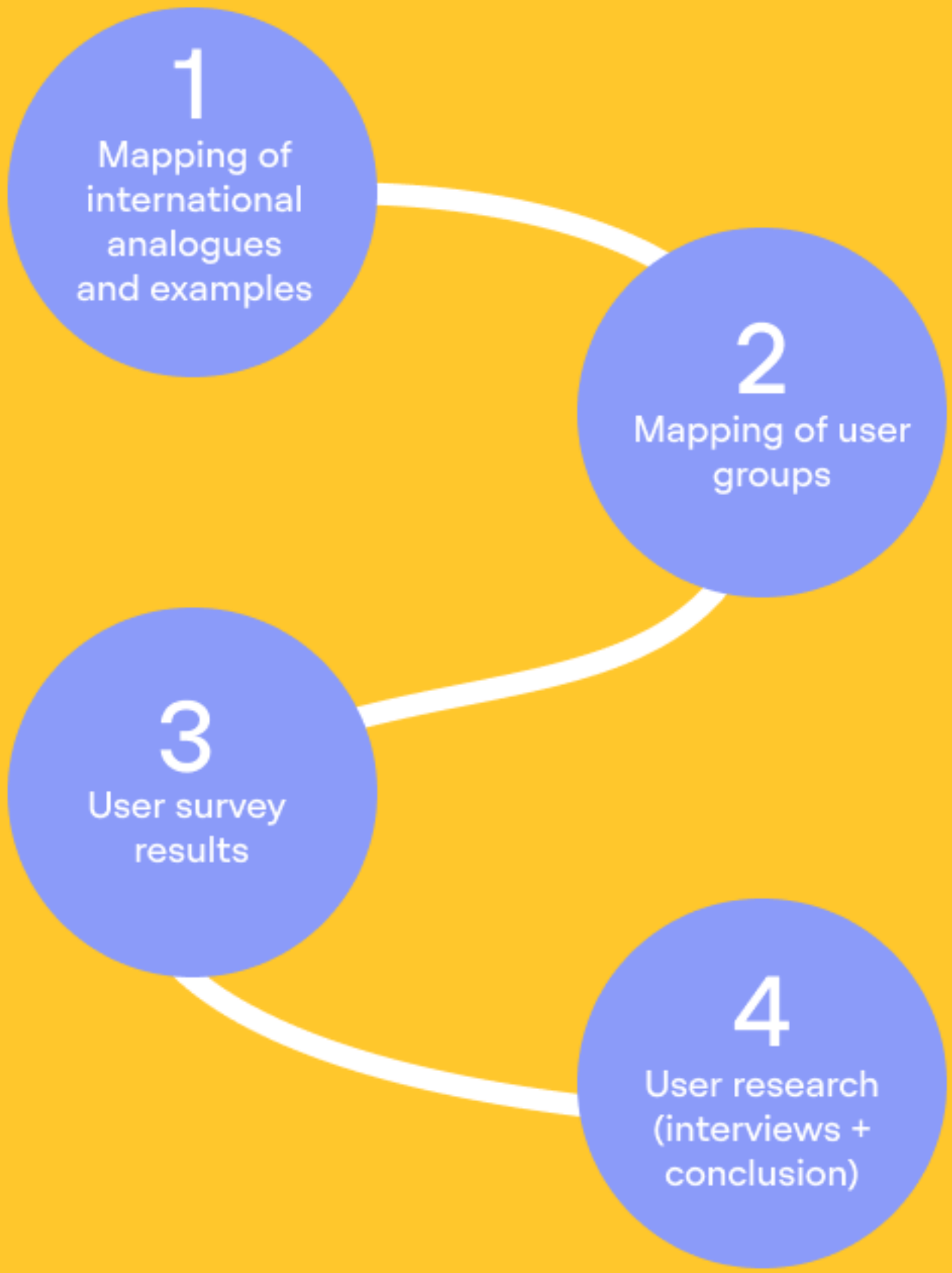
In most cases, a GLAM Lab comprises an environment with datasets derived from library collections as well as tools that assist users and facilitate access to the data, and in this way create new value. To engage with users, share expertise and test new methods, sectoral events (conferences, seminars, hackathons) are organised and/or cooperation networks (residencies, eg cooperation with universities) are developed.

NLE has decided to join the GLAM-Lab movement and further develop its open data portal. On the portal (data.digar.ee) you can download for free limited formats of metadata both from the Estonian National Bibliography (ERB) and NLE's digital archive DIGAR; full texts of open access publications are available in DIGAR as well.

The virtual lab, which we are currently developing, would give access to more data, provide processing tools and show how to use the data in a broader and more diverse way.

GLAM Lab or a virtual lab is a space where you can consume collections of digitised and/or digitally created data for experimental and creative means both onsite and online. (M. Mahey, 2020)

2 Design process



2.1. Mapping of international analogues and examples

Virtual laboratories or Glam Labs come together in an international network called the International GLAM Labs Community (glamlabs.io) where you can also find a guide to setting up a virtual lab.

Our main examples of creating a virtual lab service for NLE are:

- National Library of the Netherlands (KB);
- Austrian National Library (ÖNB);
- The British Library (BL);
- Digital Humanities Lab of the Royal Danish Library;
- National Library of Australia;
- Library of Congress;
- The project Digital Open Memory of the National Library of Finland;
- Atlas of Digitised Newspapers;
- NewsEye;
- Impresso;
- Europeana;
- ProQuest Text and Data Mining Studio;
- Gale Digital Scholar Lab.

2.2. Mapping of user groups

User groups were mapped based on NLE's previous overview of potential users who might be interested in using the lab and on international experience. These users were divided into distinct groups following the principles of service design. First, we tried to determine the largest possible group of potential users and created main user types for them. With each group, we tried to specify what they might be interested in and what their skill level would be. In this way, we created a list of the potential users of NLE's virtual lab, ie the list of interest groups.

- Digital humanities and digital sociology researchers

Researchers who are interested in digital methods and data that they would like to use in their research; their skill level varies but at least some of them can independently work with data.

- Students and future data scientists

Students and future data scientists could use NLE's materials for sample datasets and experimental projects that require unusual approaches. It is quite possible that we could implement some of their work results in our daily work at NLE (eg another clever AI solution such as the digital kratt).

- Instructors

Instructors could use our data to teach digital methods or illustrate their learning materials; with NLE's data they could, for example, open up the history of Estonian books from a new perspective. Instructors could also mediate communication with the students.

- Teachers

Teachers might be interested in using our data to illustrate their materials, or to introduce digital heritage that they could integrate into programmes of study to improve students' information and digital literacy. In achieving this, cooperation with NLE's educational network could also help. A great example in this field is the educational programme Cultural Remix (a poster presentation by Mahendra Mahey and Aija Sakova), which was shown to students at the Summer School for Memory Institutions in 2021.

- Students

Students could use NLE's data in schoolwork; in student research projects, for example. Here, cooperation with teachers is essential: they should introduce the materials in their lessons, assign tasks to look up for thematic articles etc.

- Journalists and other potential users

Some users are interested in data and want to compose texts, articles or general overviews based on this data, eg an overview of how much and in which manner people wrote about tennis in the Republic of Estonia in 1920–40.

- Interested societies, local historians and hobby scientists

Some users might need information from the NLE collection about a specific personal or family name, or wish to study NLE's collections of postcards and geographic maps to identify what has been depicted in the photos, where the letters were sent from etc.

- Artists and curators

Their interest was confirmed by the activities of KB and ÖNB (both partners of this project). There are also other examples: the tweeting machine created by Timo Toots at NLE in 2020, or the art project Crossroads on Curiosity at the British Library.

- International researchers

In the past, researchers have turned to the National Library as one of Estonia's main data managers to obtain information about the possibilities to analyse Estonian newspaper articles based on existing data.

- Data scientists/private sector

Data are necessary to develop Estonian language models, and for some publishers it may be important to have a constant overview of what is being published. Moreover, it is possible to create educational applications such as the augmented reality game written by the Innovation Centre EDUSPACE at Tallinn University in cooperation with MobiLab, a private enterprise. The purpose of the game is to present Estonia-related symbols and nature and encourage children to study these topics (see www.tlu.ee/armang).

- Open data enthusiasts

There is a Facebook group called OpenESTdata, which connects people who might want to work with data in a semi-professional way.

- Statistics Estonia and other state agencies

If the Estonian national bibliography were to be updated with as little delay as possible, Statistics Estonia and other state agencies could directly access the data they need, such as information about publishing houses, publications etc. For this purpose, NLE could offer user-friendly dashboards (see eg juhtimislauad.stat.ee/).

After considering time and other restrictions on resources, the workgroup appointed a priority degree to each potential interest group. This determined whether representatives from specific interest groups were interviewed at the current development stage of the virtual lab.

2.3. User survey results

The information collected from the interviews can be divided into two major thematic groups, which are divided into two subgroups.

- Firstly: aspects of functionality and user expectations for the development and good performance of a virtual lab environment. This includes themes such as the most popular datasets and access issues, data and file formats and the tools necessary for data processing.

- Secondly: a broader sectoral context or the virtual lab as a service. To develop this, we need different forms of cooperation that should increase awareness of the possibilities of a virtual lab and encourage different user groups to experiment with NLE's datasets.

Datasets of interest

Broadly speaking, people are interested in full texts, metadata and/or full texts with metadata.

Full texts, pictures and other datasets

- texts shared by metadata publications;
- texts from books – large text collections in Estonian are appreciated;
- digital publications in general;
- online archives;
- weekend politics and news shows with recordings and transcripts, even if already stored by media monitoring companies;

- pictorial material with information about the location and image captions;
- datasets about the reconstruction of the National Library.

What we saw from the interviews is that people mainly require access for two reasons:

- research in specific topics, eg policy research;
- research in language and word usage, here the context of a text remains unimportant. However, some researchers might be interested in language usage over time in specific fields, such as military, medical crises and law.

Full texts with metadata

The interviewees underlined repeatedly that when researching in a specific field, texts must be equipped with metadata.

Metadata: the author, titles of articles, names of periodicals, publication dates, keywords, subject areas, information about the location with images.

Derived data, ie data acquired during the analysis (geographical information about places of book publications, automatically or manually identified personal names in texts).

Among other things, NLE could provide users with pre-processed corpora ready for analysis, which would be thematically framed, digital, machine-readable and indexed.

What the interviewees are missing the most are materials with sufficient labelling. In other words, metadata is sometimes incomplete or missing some information, eg finished layers with names of organisations mentioned in texts.

Datasets should also include information about the quality of the data to provide researchers in advance with adequate details about the character of the data and the work in general.

If researchers already correct the data, these corrections should be sent to the data creator.

"On the other hand, even if metadata is at times incorrect, it should be open to access to provide at least some sort of basis."

"Cleansing various name forms takes considerable effort, but a field with multiple names could still be formatted in a beautiful set of metadata."

Legal framework

All respective interviewees pointed to a major problem concerning restricted access to full texts.

- Prominent media outlets often restrict access to their data.
- Full texts open to access are rarely complete: whether there are only fragments of texts available or all editions of a publication have not been fully digitised. Political scientists find this particularly problematic since their research requires continuity over time and entire text corpora.
- Access to full texts is available in separate computers at certain libraries only and the texts cannot be downloaded. Researchers are left without the necessary information.

There are three problem areas that the courts still have not been able to interpret:

- Where should the protective line between a text corpus and a model, such as a frequency list, be drawn? The material needed to create a model (a text corpus) should be protected, but the model itself should no longer be loaded with the original restrictions.
- To which extent should a text be legally protected? Would a sentence still be protected by copyright if the material in the corpus is scrambled sentence by sentence? What if those sentences include personal information, which is the case with most government documents?
- What sort of research and development activities provide rights to data and text mining? We received a proposal to regard data and text mining results on such data that are public and available to everybody as research and development activities, even if carried out by a private enterprise.

We must work out how to give researchers access to data without necessarily visiting the library, and still how to keep protected texts inside NLE.

Possible solutions that we have received:

- cloud-based access;

- ‘cleansing’ texts – excluding personal data;
- when working with language data, offer (scrambled) sentences as a minimal unit;
- contracts with users.

Technical access

Access to data can be provided either as a copy of the entire database (dump, eg zip files which can be downloaded), via APIs (eg OPAI-PMH is currently used at NLE for larger sets of data) or with SPARQL Endpoint (eg the solution created for the WikiData query page, see query.wikidata.org/).

Access to images could also provide cross-usage of images.

Data file formats

Different file formats should be available for download on virtual lab platforms. This would enable users to choose what kind of format meets their needs the best in technical or thematic terms.

From the answers we could see that even if NLE’s data are available in most common formats, these are unfamiliar to users accessing the data from outside of the library.

The interviewees listed the following formats as suitable:

- .json (Python users)
- .tsv
- .csv
- .xml

Tools for data processing

The choice of tools offered by a virtual lab largely depends on the different needs and skills of data enthusiasts.

Most of the interviewees emphasised the need for search and/or filtering functions that would enable them to consult specific data not only by individual years but also by topics, time periods, words etc.

For example: it would be great to see everything Andrus Ansip has said about the city of Kärdla. Or to sum up all texts written in the Russian Empire in the 1890s.

Important at this point would be the option to download individual records (piece by piece) since full datasets might be very large. This should be, for example, included in search options.

Users with beginner skills mostly want to read. In the extreme case, they might be interested in visualisation tools to experiment and play with data. Intermediate users know how to carry out qualitative as well as quantitative research, of which the latter is usually preferred. They would be the main users of the tools we would offer.

Our third interest group are advanced users who usually carry out quantitative research and have a better command of technological skills. Advanced users process data with their own coding and are therefore not in need of tools offered by the virtual lab. However, they must be given the chance to process data with their own coding.

In the interviews, advanced data researchers named the following application programming interfaces as necessary for data processing:

- Jupyter Notebook – previous knowledge and experience is necessary, but users can use it directly online without the need to install software on their computer. Both R and Python programming languages are possible.

- RStudio – a data analysis environment designed for the R language. Also suitable for Python.

The following were mentioned as suitable types of software for processing language datasets:

- Sketch Engine – licence given based on the number of users.

- NoSketch Engine – a free analogue with fewer options.

- KORP – open-source software, Estonia has the competence for further development. This remains insufficient for lexicographers.

- Other self-created software.

- The interviewees also mentioned tools that could be used to correct texts, analyse morphology and syntax, lemmatize or identify work stems. Python libraries, such as Stanford and EstNLTK, can be used for this purpose as they enable the user to automatically tag texts.

Most interviewees would prefer to download data. This is not by any means necessary; they would also appreciate the possibility to process data in the cloud and just download the results.

Cooperation and community

The interviewees would like to see more sectoral cooperation: institutions that collect and provide access to cultural data should share a common database registry and different datasets could be interlinked. Several interviewees have said that compiling separate datasets would enable users to compare or match them and thus create completely new solutions. This solution is something we should definitely consider working on.

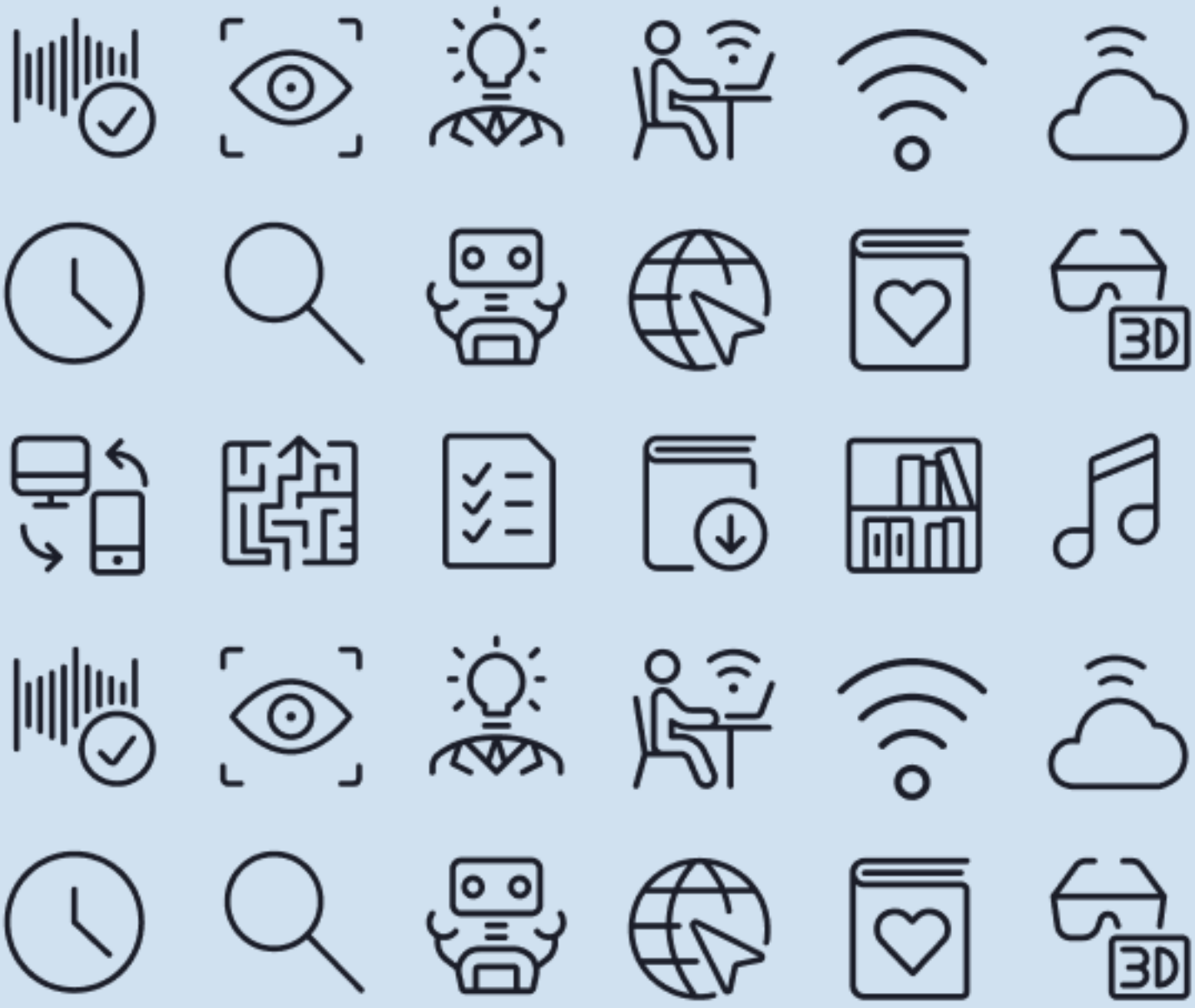
Engaging people to improve data quality

The interviewees saw the potential for cooperation in engaging users, university students and society in general to improve data quality. One possible way to organise and improve the quality of data is crowdsourcing, which has been the practice in several instances at BL. Schools, groups of students, educational programmes, volunteers and others were involved to control and improve data quality.

Many interviewees stressed the importance of raising awareness and having the necessary skills. First, we need to raise awareness about the data available at NLE that could be the object of text and data mining. Relevant disciplines could make use of NLE's open data as an input in teaching. Instructors could motivate students to access and use NLE's data. NLE's own staff could as well go to universities as visiting lecturers, organise workshops or seminars to introduce the data as well as the tools necessary for data processing, which can be accessed through the library.

Secondly, data processing skills must be improved, especially amongst students, and people in general should be encouraged to activate the skills they already have. User stories, guides and videos would surely help to achieve this.

Thirdly, the interviewees expressed hope that NLE could provide students and scientists who are processing their data with actual practical output. The virtual lab should work out and present specific projects that require accessing data. In any case, we should consider the possibility of receiving corrected data that could be published at the lab, with the author openly given credit.



Creating a service

The international experience of setting up virtual laboratories underlines five main aspects that should be focused upon:

- Understanding and support from upper management is important. Managers have to understand which goals are being pursued and how this is done.
- A detailed overview of a library's digital collections – what are its digital resources and reserves based on collections and data. Each collection should be numbered and provided with information on how the collection was created, in which format the data are presented etc.
- Within the library, clear workflows must be set up to determine how licences are granted for digital material under copyright and how potential usage is defined.
- Make sure how to provide access to: APIs, zip-files, data formats etc, whether to add DOIs to datasets etc.
- How to connect people with data – ie data created and managed by the library on the one hand, and on the other users and often data editors and correctors. We should call for cooperation between these two parties and send a clear signal that their work is valuable and well appreciated.

Instructions and examples

Most interviewees agreed that it is not enough to give access to data and provide these with different processing tools. While setting up a virtual lab we should still keep in mind that users' skill levels vary.

Datasets and tools have to be provided with content descriptions: which data can be accessed and what processing tools are available. The online environment should also include user stories: examples of creative data usage and other potential usage perspectives.

The library could offer exclusive help to researchers in their research and also provide research topics. Support and advice from our side would encourage researchers to consult NLE's collections more actively.

Counselling would open new unexpected perspectives to researchers, provide access to curious and fascinating materials that are normally known only to people well acquainted with the collections.

Media

The success of the virtual lab will depend on how well we will be able to brand it and how much public and media attention we would attract. The greater the possibility that a scientist's project attracts public attention, the bigger his or her interest would be in the datasets available at our lab.

3 Visions of a virtual lab at NLE

Based on the aforementioned international experience and the interviews carried out with potential future users, NLE's respective tasks concern three main areas.

- What does the user see (front-end) or activities directed at all users, the 'public image' of the lab: activities concerning the website, its content management and maintenance.
- Which tasks does the developer/administrator have (back-end) or the 'backstage work': there have to be enough high-quality resources available for access.
- The basis of a successful virtual lab lies in cooperation with the community. Engaging it requires regular activities as well as previous ready-made solutions.

Lab's web environment

The website has an important role in communicating and presenting the lab's tools, datasets and activities, and these have to be easy to find. The main elements of a lab are datasets and tools. The website should invite users to give their contribution as well as to enhance the data collections; every now and then, engagement campaigns addressing regular readers should be organised.

Datasets

All interested users should have unauthenticated access to datasets as well as tools via the digital lab's web environment. In this way, the largest possible user community would have the chance to experiment, try out different options and collect information.

However, giving access to some datasets might require authentication in order to define the rights and obligations the user must comply with when working with the materials. In this case, it should be possible to sign a contract online (on the web platform).

The main content of the lab are datasets that the library has to change into suitable formats, write introductions to, and update as well as manage. This can be done with NLE's own resources or with the help of some enthusiastic users.

Tools at the virtual lab

At least two types of data enthusiast should be regarded when developing tools for the virtual lab's web platform: those who know how to code and would like get access to data with coding, and those who prefer simpler tools.

We need to find a balance between what can also be learned by beginners and what amount of coding would scare potential users away.

In comparison with other labs, our tools include a broad spectrum of different options: there are graphs with data that can to some extent be played with but also datasets that can be examined with more sophisticated tools, such as correction algorithms.

Several tools may be limited to descriptions and links to external websites. Web environments suitable for data examination are often written in R using solutions available in Shiny (shiny.rstudio.com). It should be possible to install these on the website and, if necessary, link to larger data files stored in our server.

Special lab for word processing

NLE already has a special data processing tool that has been partly developed: a virtual environment where you can run code and carry out simpler data processing. NLE is negotiating with ETAIS on who could provide the web environment with Jupyter Notebook, data storage and secure login.

Currently, it is possible to access via this tool texts from the collection of digitised Estonian articles (DEA) available on an open network.

Datasets and blog posts created by the users

Examples of usage will continue to play an important role in introducing and illustrating the datasets and tools available at the lab. These examples are often added by the users themselves in the form of Markdown documents or blog posts with images.

The same applies to datasets that the users might have created. In this case, there should be a fairly standardised option of information input and this information should preferably be stored as plain text files (plaintext), which can be added to the dataset (eg references, tags etc, in a readme.md file).

If all the metadata of a dataset is included in one file and standardised (which the webpage can either read or which can be transferred to the webpage), it can be easily updated and will be available forever.

Building and engaging the community

Work with communities should go hand in hand with a virtual lab web platform. This work is constantly evolving and might be even more important than the platform itself if we want to increase the usage of digital collections and data.

We should try to avoid presenting a detailed and innovative web platform that no data enthusiast has been informed of; or one that requires technical skills and does not comply with what the users need, since their voice had not been heard in the first or the following development phase.

To build a community and keep in contact we need to do the following:

- Active correspondence with universities, especially with instructors and trainers, to have an overview of what is currently relevant in universities.
- Crowdsourcing and the engagement of interested parties.
- People creating and managing NLE's own collections are important. We should try to inform people who are responsible for creating NLE's datasets (eg the online archive, national bibliography, etc) about the benefits arising from their collections and giving open access to their data.
- NLE's virtual lab has to be a constant focus and invigorate interested parties to take action.
- The Lab team should be ready to consult and provide any further help to those who ask.
- We should also consider the possibility of combining the activities of the virtual lab with NLE's other undertakings (ie with digital kratts or crowdsourcing), eg whether it would be possible to apply automatic text tagging models also in student research.

The National Library will continue to educate on and promote important topics, such as using digital tools, writing code and general awareness on how many great possibilities open up when working with data.

Building and engaging the community

Work with communities should go hand in hand with a virtual lab web platform. This work is constantly evolving and might be even more important than the platform itself if we want to increase the usage of digital collections and data.

We should try to avoid presenting a detailed and innovative web platform that no data enthusiast has been informed of; or one that requires technical skills and does not comply with what the users need, since their voice had not been heard in the first or the following development phase.

To build a community and keep in contact we need to do the following:

- Active correspondence with universities, especially with instructors and trainers, to have an overview of what is currently relevant in universities.
- Crowdsourcing and the engagement of interested parties.
- People creating and managing NLE's own collections are important. We should try to inform people who are responsible for creating NLE's datasets (eg the online archive, national bibliography, etc) about the benefits arising from their collections and giving open access to their data.
- NLE's virtual lab has to be a constant focus and invigorate interested parties to take action.
- The Lab team should be ready to consult and provide any further help to those who ask.
- We should also consider the possibility of combining the activities of the virtual lab with NLE's other undertakings (ie with digital kratts or crowdsourcing), eg whether it would be possible to apply automatic text tagging models also in student research.

The National Library will continue to educate on and promote important topics, such as using digital tools, writing code and general awareness on how many great possibilities open up when working with data.

Building and engaging the community

Work with communities should go hand in hand with a virtual lab web platform. This work is constantly evolving and might be even more important than the platform itself if we want to increase the usage of digital collections and data.

We should try to avoid presenting a detailed and innovative web platform that no data enthusiast has been informed of; or one that requires technical skills and does not comply with what the users need, since their voice had not been heard in the first or the following development phase.

To build a community and keep in contact we need to do the following:

- Active correspondence with universities, especially with instructors and trainers, to have an overview of what is currently relevant in universities.
- Crowdsourcing and the engagement of interested parties.
- People creating and managing NLE's own collections are important. We should try to inform people who are responsible for creating NLE's datasets (eg the online archive, national bibliography, etc) about the benefits arising from their collections and giving open access to their data.
- NLE's virtual lab has to be a constant focus and invigorate interested parties to take action.
- The Lab team should be ready to consult and provide any further help to those who ask.
- We should also consider the possibility of combining the activities of the virtual lab with NLE's other undertakings (ie with digital kratts or crowdsourcing), eg whether it would be possible to apply automatic text tagging models also in student research.

The National Library will continue to educate on and promote important topics, such as using digital tools, writing code and general awareness on how many great possibilities open up when working with data.