Creating a data platform for the humanities

Recommendations for platform creator

Developed in LINDAT/CLARIAH-CZ project, Brno, March 2022

Are you preparing a digital collection, archive, database or other research platform and thinking about the steps, approaches and challenges? This material is for you as a basic introduction to the issues. We'd be happy to meet with you and go deeper into your needs – please contact us at lindat-clariah@phil.muni.cz or book a personal consultation.

All original research begins with the preparation of a data backup. Sometimes existing information resources, factual databases, data sets, etc. can be used. If there is no such sources for the intended research, or if it does not have eligible coverage, suitable structure or appropriate validity, it comes down to creating your own data platform. The creator of such a system goes through a complex process in which he makes a series of decisions of a conceptual, professional, managerial, technical nature. We have prepared this material as a basic methodological support for this process.

The following recommendations are general, with no link to a specific technical solution. The first set of recommendations – SEVEN PRIMARY CONSIDERATIONS – is devoted to basic conceptual decisions, the second one – FIVE STEPS TO THE GOAL – presents recommendations for individual procedures, methods and practical solutions.

SEVEN INITIAL CONSIDERATIONS

This section provides suggestions for the data platform developer's initial reflection before embarking on the actual development. When planning the entire process, it is necessary to think not only about the content, scope and structure of the intended database, but also about the documentation of the steps and results, the legal solutions for public access to the database content or the user interaction with the information resource, and the integration of the newly created platform into the data universe.

1) **Specification of the purpose and the intended function**

The basic questions can be summarized as follows: Why should the platform exist? What should its content be? Who will be working with the platform and how?

The reason for the existence of the platform is probably due to the research project and the lack of an appropriate resource. The platform is usually used by its administrator (research team) when adding and editing its content. This research team also works with the platform after the content is provided. If the platform is publicly accessible, the group of potential users and their needs extends. The needs of users and the required functions can be described in more ways, from simple text descriptions to various types of models. There are also a number of methods for identifying these needs (see section 6). The decisions concerning the description of the content are also important for the functionality (see section 2).

Other requirements may pertain to the data about the use of the platform. For example, it is appropriate to determine whether we want to record the numbers of file downloads or individual record views and whether this data should be displayed publicly, directly in the system.

2) **Content description**

In order to be able to work with the collection objects (scanned materials, data, photos, texts, etc.), they need to be described appropriately. The complexity of the description depends on the needs of the users, primarily the research

team building the collection. The amount of time available for production is also a determining factor. In formulating the rules of description, it is necessary to decide what data to include, where to collect it from, and how to record it. All decisions about how to describe should be documented and accessible to all those who will work with the content. These decisions will be influenced by, inter alia, by the practices established in the respective discipline and should be made with a view to the future usability and sustainability of the platform and its content. The way in which descriptive information is stored will also vary depending on the technical solution chosen.

3) Documentation of the process and the result

Documentation is often a time-consuming task, yet it is essential to fully exploit the potential of a laboriously built platform. The technical documentation is essential for the eventual machine processing of the contained data or its transfer to another system (for example, if the current solution becomes obsolete and no longer works properly). For inspiration, you can use the offered Documentation Template (PDF) with the basic headings that should be addressed.

Documentation of the decisions that led to the creation of the platform and shaped its design is also important to make it more usable. Basic information on the content and objectives of the project should also be available on the platform's website if it has a public address.

4) **Ensuring the copyright of the content provided**

If the platform includes digitised content, it is important to know whether and how it can be handled. It is also necessary to indicate the copyright owners. Special attention should then be paid to personal and sensitive data if it is included in the collection.

According to Czech law, if more than 70 years have passed since the death of the author of the digitised content, then the work can automatically be treated as a free work. Otherwise, stipulations of Section 27b(3) of the Copyright Act should be observed. In the event that the copyright holder cannot be found in those resources, a licence for certain uses of orphan works should be requested pursuant to Section 37a.

5) Identification of licence for data access and citation

The question of further use of contained data does not pertain only to freely accessible collections. If you are the creators of the content or its parts, it is appropriate to specify, which licence this content is provided under. You can use the Creative Commons licence, for instance. The CC BY 4.0 variant is recommended as the least restrictive option.

The database as a whole can also be licensed. It may be protected by copyright, if is the so-called original database, or by the database right. For more information and a schematic, visit Otevřená data (Open Data; Czech version only).

For further work with the content, it is important to clearly identify who the creator is and when the content was created and published. It is also customary to indicate how the platform should be properly cited, either as a whole or the individual data contained.

6) User care

The users of research platforms are usually the creators themselves. If the platform has potential for further use, it is advisable to involve more people in its creation and to find out what their needs are. Accessibility for users with various types of disabilities is also an important part of the creation of web content and services.

Prototypes (interim, unfinished versions) can be used to test the platform and find out whether it is easy to use. For user testing, it is recommended to assign tasks the testers will be asked to perform. General information about user testing can be found, for example, on 100 metod (100 Methods; Czech version only), along with other methods for identifying user needs.

Once launched, it can be monitored how users use the platform. You can work with data generated by the system itself (e.g., using Google Analytics or HotJar tools) or methods such as user observation or shadowing.

7) **Sharing information and (meta)data**

In order to make potential users aware of the platform, it is worth finding out if there is a register of platforms with a similar focus where it could be listed. We collect platforms created at the Faculty of Arts of MU in a catalogue on the Digitalia MUNI ARTS infrastructure website. There are also common search services that collect metadata from multiple platforms. In this case, the metadata usually needs to meet specified requirements.

FIVE STEPS TO THE GOAL

This section is intended to be a guide on the journey from the initial research idea to a quality data platform, its access, presentation and use. It focuses on expert issues, but each step requires managerial and technical solutions – for which we also provide selected suggestions. The division into individual steps is to some extent arbitrary; in practice, the implementation of some steps may (partly) run in parallel.

We have attempted to approach specific examples through the fictional researcher Albert, whose approaches and doubts may prompt further questions.

Step 1: WHAT WE WANT

Or: The basic concept, design and purpose of the upcoming data platform.

Starting point: Clarifying the goals of the upcoming research – what, how, where, when, how we will research. Who will be involved in the research and how. What the research results will look like. What types of data the research will generate, how it will be described (metadata), structured, formatted, linked to other data sources. Who will use the data and how. How the presentation and promotion of the resulting data platform will be ensured. Who will maintain and update the data platform, how frequently and by what methods.

Potential obstacles and challenges: The scope or structure of the data cannot always be definitively, accurately and completely determined in advance, as it only emerges during the research. The research may reveal a parameter, property, characteristic of the data that was not initially known or considered as an essential aspect of the analysis.

Outcome of this step: A text document summarizing all the essential decisions listed above, including any ambiguities, doubts, etc.

Management questions: Do I have sufficient financial budget for my research? Is there a provision for the creation of a data platform? Has an appropriate timeframe been established for this procedure? Does my team include all the necessary professions? Have I prepared an effective communication platform for it?

Technical solution: The programmers are involved in the team from the beginning and are familiar with the content of the project. They propose alternative solutions to complex issues (choice of software, data backup, user interface, etc.).

The LINDAT/CLARIAH-CZ team will help by: consulting, focusing in particular on whether nothing essential has been omitted in this step, whether the extent of the professional and time investment in the preparation of the data platform is clear.

Researcher Albert's comments: So far it is clear. I have been working on the issue of the language of Czech children's comic magazines for a long time, and my overview of the primary and secondary literature is perfect. We will examine the language of the heroes' statements in Čtyřlístek [Quatrefoil] magazine from its foundation (1969) to 2020, tracing intertextual and extratextual links. My team includes a linguist, a literary scholar, an educationalist, a comics genre specialist, and a private researcher focused exclusively on the Čtyřlístek magazine. We'll also need a programmer. The output will be a structured corpus of texts with links to other information objects. Primarily it will serve linguists and literary scholars, but it will also

be useful for other researchers, and the lay audience for the Čtyřlístek is also large. The exact structure of the data and metadata has not yet been determined, but it is one of the important next steps for our team. We do not yet have a precise idea about the presentation and updating of the resulting database. I would definitely use the consultation with the LINDAT/CLARIAH-CZ team to know that nothing essential has been missed.

Step 2: WHERE WE WILL OBTAIN THE DATA

Or: Methods of data collecting, creating and structuring.

Starting point: Identifying all available data sources of various types (traditional and electronic) that can become constituent data sources. Preparing methods for data extraction, and establish a structure for data storing and ensuring validity checks. Determining the same for field research data. Carrying out a data process analysis – i.e. what happens to the data during processing (from their identification to their final storage in the database) in scope of their formalisation, unification, standardisation, etc., and determining the responsibility for the individual procedures. Ensuring the required data retrospectivity, timeliness, granularity, completeness, consistency.

Potential obstacles and challenges: Data, part of it or a certain parameter of it, cannot be obtained to the extent or quality required by source or field research, or is associated with disproportionate costs.

Output of this step: A) A structured text document describing the above parameters – sources and their mining potential, field survey methods, data characteristics and parameters, and associated procedures. B) A structured dataset based on the established criteria.

Management questions: Has the issue of copyright for external data been adequately addressed? Do the requirements set for the scope and quality of the data exceed the capabilities of the working team?

Technical solution: Preparation of a database whose structure reflects the requirements for the format and structure of data and the possibilities of working with them (editing, import/export, browsing, filtering, searching, statistical reports, visualization, analytical processing).

The LINDAT/CLARIAH-CZ team will help: in particular, in defining the properties and parameters of the data, the methods of their initial acquisition and subsequent processing.

Researcher Albert's comments: Some concepts I am encountering for the first time (e.g. data granularity). Otherwise, our source is clearly the individual issues of the Čtyřlístek magazine. We have yet to decide whether we will use the already digitized material (requiring collaboration with the magazine's editorial staff) or proceed with our own digitization. We are considering the possibility of OCR of the texts, which is an untried method due to the comic format, so it may be easier to use a "manual" transcription. So far we have structured the data as follows: statement (always one "balloon"), speaker, statement identification (year, number, page, sequence). The statements will be written exactly as they appear in the document; we need to work out what to do about typos, typographical errors and other obvious formal deficiencies, how to deal with meta- and polystatements, etc. We are considering a three-layer model of processors: "fillers" (primary data processors), editors (check the quality and completeness of the data), and an administrator (deals with system issues). With the LINDAT/CLARIAH-CZ team, I would particularly like to discuss the systemic handling of input data irregularities and inconsistencies.

Step 3: PREPARING METADATA

Or: Definition and parameterization of metadata.

Starting point: Defining the scope of descriptive, structural and administrative metadata. Determine the content, structure and format of each type of metadata, preferably taking into account existing standards. Considering the use of controlled vocabularies for the defined range of metadata fields. Planning the use of tools and procedures to ensure metadata completeness and consistency. Use modelling (e.g. oriented graphs) to represent the relationships between data and metadata and implementing it in the search interface.

Potential obstacles and challenges: Formal, poor or exhaustive descriptive metadata that does not allow the data to be handled in the expected way or with sufficient precision. Over-emphasis on administrative metadata, which may detract from the actual content (descriptive and structural) metadata.

Outcome of this step: A) A structured text document that describes the above parameters – the different metadata types (descriptive, structural, administrative), their format and structure, and the methods for processing them. B) A structured set of metadata joined to the previously processed data based on the specified criteria.

Management questions: Does the metadata meet all the requirements of technical and legal standardisation?

Technical solution: Prepare the storage and display of metadata in the database, linking it to the data.

The LINDAT/CLARIAH-CZ team will help: to distinguish between different types of metadata, to determine their specific structure and format, to recommend appropriate standards and efficient methods of metadata creation.

Researcher Albert's notes: Our descriptive metadata refers to the actual texts (statements), the subject of those statements, and the context of the statements. Each layer of metadata includes additional criteria and parameters, e.g., for the text, components of linguistic analysis, for the subject, gender, emotional/physical state, for the context, time of day, location of the statements (interior/exterior), etc. For the text layer we would like to deploy standard corpus tools for text segmentation, for the other layers controlled vocabularies, preferably standardized and parameterizable. I would like to discuss all aspects of the chosen metadata in detail with the LINDAT/CLARIAH-CZ team to make sure that we have not missed any aspects and that we have not been affected by "professional blindness". We would also be grateful for recommendations on the minimum level of administrative metadata.

Step 4: ACCESSING RESULTS

Or: User interface and search options.

Starting point: Determining the data and metadata that will be used most often and thus affect the basic structure of the user interface. Choosing the structure of the main UI page and the parameters of the sub-pages associated with displaying, viewing, filtering, and searching the data. Thoughtfully differentiate the capabilities and scope of full-text, unstructured and structured searching. Involve end-users in testing the system.

Potential obstacles and challenges: Structured search does not sufficiently exploit all metadata parameters and thus limits the user's analytical access to the database content. The search results are not clearly or comprehensibly displayed from the end-user's perspective.

Output of this step: A) A structured text document that describes the above parameters – individual interfaces, searchable fields and their contents, data display format. It becomes the basis of user documentation – help. B) A concrete user interface with functional and tested search options. C) Contextual help based on A and implemented in B.

Management questions: Will access to the data platform vary by user type? What licenses will be set for specific content and users? Do these licenses respect the publishing policy of our institution (open access)? How can the results of the project be applied to the evaluation of research and scholarly activities?

Technical solution: Based on the defined structure, prepare a prototype user interface taking into account accessibility rules and other standards. After repeated testing, make the final version of the data platform available.

The LINDAT/CLARIAH-CZ team will help: with setting up search filters, defining the search interface, choosing the appropriate display of search results and testing the user interface.

Researcher Albert's notes: We consider the objects of the statements, the subjects of the statements and the temporal aspect as the main filters of our data. These should be part of the main page as well as a full-text searching of the statements. At the same time, we would like to allow researchers to make analytical, combined, parameterizable queries, even if irrelevant or nonsensical from our point of view (e.g. how often a subject like man in an emotional state of annoyance uses extratextual references to objects like building). What I expect from the LINDAT/CLARIAH-CZ team in this step is mainly consultation on the possibilities of analytical search and intensive testing of the user interface with emphasis on advanced search.

Step 5: STAYING HERE FOR A LONG TIME

Or: Data platform management and maintenance, long term preservation.

Starting point: Defining a concept for further development of the data platform. Identifying the data and metadata types that will be updated and added at specified intervals. Identification of data and functional gaps that have

emerged from the initial design phase. Determine the long-term need for expansion of the database, metadata description and enhanced or new functionality.

Potential obstacles and challenges: Lack of motivation or resources for long-term management of the data platform. Inconsistent updating of data resulting in outdated or incomplete data.

Outcome of this step: A) A structured text document that sets out the rules and procedures for managing and updating the data platform. B) The continuously updated content of the data platform.

Management questions: Who will be responsible for the long-term management of the data platform? Will it be necessary to raise extra-budgetary financial resources for the maintenance of the database?

Technical solution: Implementation of tools to effectively monitor content and formal changes to the data platform.

The LINDAT/CLARIAH-CZ team will help: to define the parameters for the short, medium and long-term development of the data platform, to model the potential users of the database, to modify the structure, content and user interface.

Researcher Albert's comments: This is the phase I am most concerned about. We are creating the data platform as part of the project and after the project is over there is a risk that everything will run "idle" or the database will be preserved in its state at the end of the project, which we would not like to do, as we have many ideas for its further development. The advice of the LINDAT/CLARIAH-CZ team could therefore be useful to us with regard to setting development priorities and ensuring their implementation, including funding.

FEEDBACK

Please let us know how useful this document has been to you:

A – It substantially clarified for me the issues involved in creating a data platform.

B – It explained some ambiguities, provided partial suggestions, but overall did not change my insight.

C – He was quite general for me, I have a overview of the issue and I need rather to address specific steps.

D – It is not clear to me what is going on, I probably need the issue explained in a different way.

Comments: _____

Thank you for your opinion. We will take it into account in our future work.