DARIAH-EU

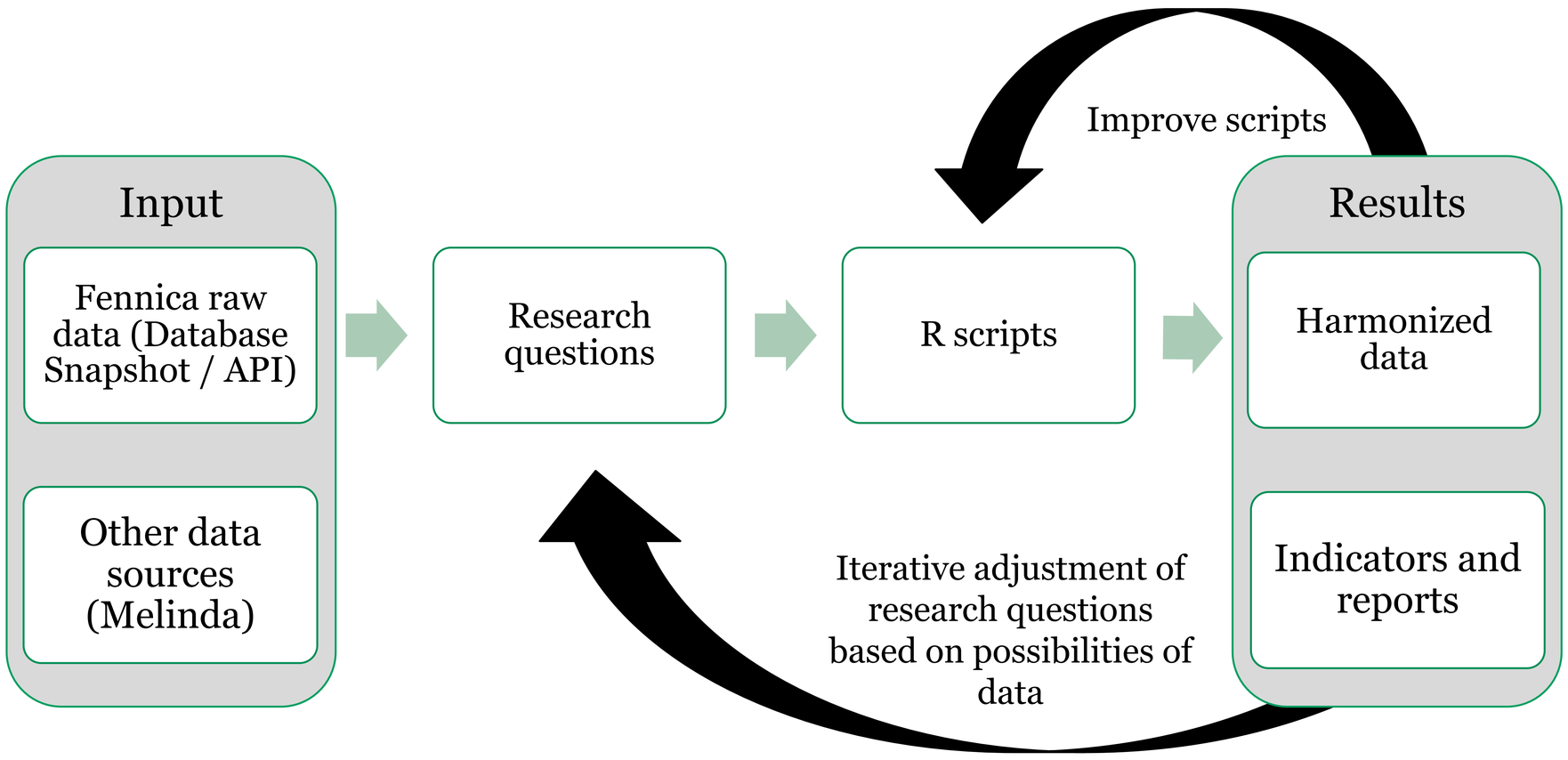
CLARIN

Common Language Resources and Technology Infrastructure

FIN-CLARIAH

WP 4.1 METADATA HARMONIZATION AND ANALYSIS

Leo Lahti
Department of Computing, Faculty of Technology
University of Turku
Pyry Kantanen
Department of Computing, Faculty of Technology
University of Turku



Visualization of the workflow.

INTRODUCTION

As part of the Metadata Harmonization And Analysis work package, we use the Finnish national bibliography (FNB) dataset to create a harmonized dataset that can be used for research purposes as well as provide the groundwork for infrastructure that can be iterated further.

Systematic organisation of bibliographic records and algorithmic tools to access, harmonize, and analyse bibliographic metadata collections are essential for complementing qualitative research with insights gained from quantitative analysis (Tolonen et al, 2019).

Bibliographic data science is an emerging method in the history of literature that has already produced information on topics such as publishing trends in the English-speaking world (Lahti, Ilomäki and Tolonen, 2015) and secularization in Sweden and Britain (Myrdal and Söderberg, 2012). It has the potential to provide a more complete picture of Finnish literature history as well.

METHODS

The used metadata includes the Finnish national bibliography Fennica and the National Metadata Repository Melinda. Fennica is maintained in the machine-readable cataloguing (MARC) format as a subset of Melinda and is the most comprehensive source of published Finnish literature during the study period. The Fennica bibliographic records have already been published as open data. (Tolonen et al, 2019.)

The FNB contains metadata for **71,919 documents** and spans over 450 years, more specifically consisting of records printed between 1488 and 1955. There are **28 data fields** in raw data in machine-readable cataloguing (MARC) format, which provide information on the document, for example, language, author name, author lifetime, publication title, publication place, publisher, and page count. Through harmonization and enriching the dataset with data from other sources, the abovementioned 28 fields can result in **96 data fields** per document. (Tolonen, et al 2019.)

RESULTS

An important output of our work is a harmonized version of the FNB. The FNB metadata is an example of a dataset consisting of structured data. The FNB dataset is not created for research purposes and needs refining before it can be used. Tools are needed for accessing, cleaning, refining, enriching and analysing this type of data, and printing documentation of this process.

In addition to the harmonized output, the developed open-source tools, documentation describing the steps and the structure of the workflow and automatically generated reports providing indicators describing the success rate of our harmonization scripts are important end products in their own right.

Our tools consist mainly of scripts in R programming language that can be run locally or as scheduled tasks in the cloud, e.g. a high-performance computing (HPC) environment. The scripts use a snapshot of FNB as input and output a harmonized version as well as indicators that describe the success rate of harmonization. In the future, these scripts could be improved by including the option to access always up-to-date raw metadata through Fennica and Melinda OAI-PMH API.

DISCUSSION

Improved infrastructure for studying bibliographic datasets can contribute to research in liberal arts and literature studies but also in various other fields, such as social science, linguistics, and science and technology studies. Increased familiarity in using such tools allows for iterative development process in which tools are iteratively adjusted to fit each discipline's typical research questions.

REFERENCES

Lahti, L., N. Ilomäki, and M. Tolonen. 2015. A quantitative study of history in the English short-title catalogue (ESTC), 1470–1800, *LIBER Quarterly* 25 (2):87–116.

Myrdal, J. and J. Söderberg. 2012. Bokproduktion och sekularisering 1500–1800. Agrarlitteraturen på 1700-talet som detaljexempel. In *Människans kunskap och kunskapen om människan*, ed. M. Wallenberg Bondeson, O. Husz, J. Myrdal, and M. Tydén. Lund, Sweden: Sekel.

Tolonen, M., Lahti, L., Roivainen, H., and Marjanen, J. 2019. A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History,* 52:1,57-78

UNIVERSITY OF TURKU

DARIAH-FI

KIELIPANKKI The Language Bank of Finland

FIN-CLARIN