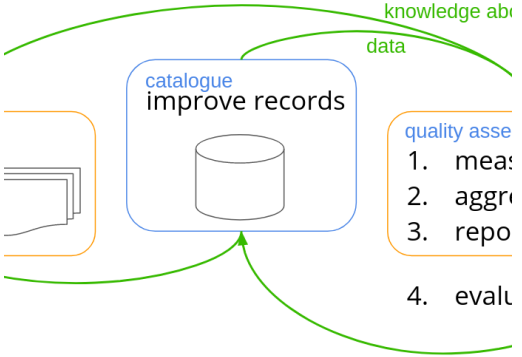
CA01 (Meta)data quality assessment

Background

Everybody recognizes bad data, but it is not easy to define what makes data good or bad. Quality assessment is a special data analysis process aiming to highlight some features of a dataset called quality dimensions. This analysis can be used in a later step of data analysis, such as data cleaning or exploratory data analysis. In this course we use cultural heritage metadata (library, archival and museum catalogues) as our research data. We will learn about theories such as data quality dimensions. We will use and contribute to the development of assessment tools to detect quality related problems. Finally we will discuss the results with metadata experts of the data provider institutions.



Goal

- Understanding the full life cycle of data quality assessment (study data quality dimensions, tools, and a metadata standard, assess quality with a relevant tool and communicate the result with data curators).

Tasks

- Review literature about (meta)data quality
- Understand the metadata schema of a selected cultural heritage data source
- Adapt a quality assessment tool (e.g. SHACL, JSON Schema, QA catalogue) to measure quality dimensions
- Visualize the results and communicate with metadata experts of the data provider

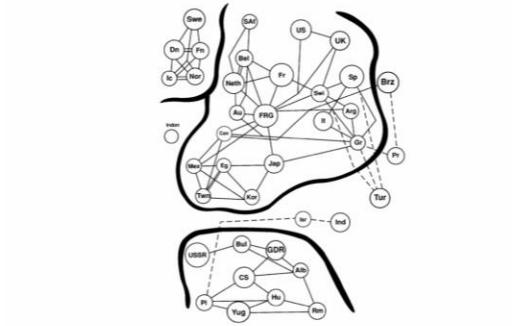
Péter Király

peter.kiraly@gwdg.de

CA02 Bibliographic data science

Background

Bibliographic data contains factual historical dimensions, such as personal names of authors and contributors (occasionally with additional properties), place and date of publication, name of publishers/printers/scriptors, genre, subject description of the content (keywords, classification terms), materiality, provenance (current and previous holding institutions, owners). After data cleaning and normalization all these information shed light to historical patterns, such as how the roles of different languages changed regionaly, how the literary canon evolved, who were the important authors and books in a particular periods, enduring and ephemeral best-sellers, how the media changed, and how all these correlated with each other?



Political and S-C divides in Europe in the 1970s & 1980s (Šajkevič 1992), Index Translationum data

Goal

- Run historical data analysis on library catalogues (understand, extract, normalize, analyze and visualize bibliographic data, compare result with qualitative sources).

Tasks

- Review literature about bibliographic data science
- Formulate research questions
- Understand the metadata schema of a selected cultural heritage data source
- Use R/Python/Java to clean, analyze and visualize data
- Check literature if your result is a novelty and compare with state-of-the-art research

Péter Király

peter.kiraly@gwdg.de