Péter Király (GWDG, Göttingen)

An outline of an imagined training course on bibliographic data science[1]

Bibliographic data science is a relatively new interdisciplinary field of research that lies at the intersection of library science (or, more broadly, cultural heritage science), history and social sciences, and certain components of computer science. The objective of bibliographic data science is to establish previously hidden or possibly only suspected historical or collection trends based on data sources containing a (typically but not exclusively) large number of bibliographic records, ideally all those related to a given topic (e.g., national bibliographies), and on data science methods. Some of the field's research questions:

– What was the spatial distribution and prosopography of 17th-century German legal dissertations?[2]
– What degree of interdisciplinarity can be observed based on the metadata of philosophical dissertations?[3]
– How did the format and language of books change over time in different regions?[4]
– What is the pattern of translations from a given language, how has it changed, and which languages were super-central, central, and peripheral in a given era?[5]
– What impact do publishers have on fiction?[6]
– What were the profiles of the various book collections?
– Is there a correlation between the genre and format of the book?[7]
– How have genre proportions changed?[8]
– How many early modern publications could have been destroyed without a trace?[9]
– How can the reception of works be examined using bibliographic data?[10]

---

[1] The outline is a further elaboration of the work process described in the article *Bibliographical Data Science: from Catalogues to Research Data* (by Tolonen, Mikko, Vimr, Ondřej, Király, Péter, Panušková, Charlotte, = Social Sciences & Humanities Open Marketplace, 2023. https://marketplace.sshopencloud.eu/workflow/tE2HiC). I would like to thank Szilvia Maróthy (L'Harmattan Publishing House), Károly Kokas (University of Szeged), János Káldos (National Széchényi Library), and Thomas Wallning (University of Vienna), as well as Julia Damerow (Arizona State University) for her advice and for her paper *Training and Education Resources for Research Software Engineering in DH* (by Damerow, Julia, Nelson, David Ragnar, Hernandez, Jose, Developing = Anthology of Computers and the Humanities, 2025. 2 (October): pp. 13–18. https://doi.org/10.63744/X9dgVaDTvJHY)
[2] Heßbrüggen-Walter, Stefan, *Early Modern Dissertations in French Libraries: The EMDFL Dataset* = Journal of Open Humanities Data 2025. 11 (June): 36. https://doi.org/10.5334/johd.307.
[3] Heßbrüggen-Walter, Stefan, *Interdisciplinarity in the 17th Century? A Co-Occurrence Analysis of Early Modern German Dissertation Titles* = Synthese 2024. 203 (2): 67. https://doi.org/10.1007/s11229-024-04494-2.
[4] Lahti, Leo, Marjanen, Jani, Roivainen, Hege, Tolonen, Mikko, *Bibliographic Data Science and the History of the Book (c. 1500–1800)* = Cataloging & Classification Quarterly 2019. 57 (1): 5–23. https://doi.org/10.1080/01639374.2018.1543747.
[5] Heilbron, Johan, *Towards a Sociology of Translation: Book Translations as a Cultural World-System* = European Journal of Social Theory 1999. 2 (4): 429–44. https://doi.org/10.1177/136843199002004002.
[6] Bourdieu, Pierre, *A Conservative Revolution in Publishing* = Translation Studies 2008. 1 (2): 123–53. https://doi.org/10.1080/14781700802113465.
[7] Lahti et al. op. cit.
[8] Péter Király, András Kiséry, *'Mór Jókai, alas': the most successful Hungarian writer. A quantitative analysis* = Patterns of Translation blog 2025. https://translationpatterns.substack.com/p/mor-jokai-alas-the-most-successful
[9] Farkas Gábor Farkas, Káldos János, Király Péter, *A régi magyarországi kiadványok „sötét anyaga"* = Magyar Könyvszemle 2025. 141:2. pp. 226-266. https://doi.org/10.17167/MKSZ.2025.2.226-266
[10] Szemes Botond, Dobás Kata, *A visegrádi országok digitális irodalmi emlékezete: Wikipedia, Wikidata – a regionális irodalomtörténet új alakzatai* = Irodalomtörténeti Közlemények 2025. 129:2. pp. 191-212. https://doi.org/10.56232/itk.2025.2.04

– What is the quality of cultural heritage data, and what improvement strategies can be developed?[11]
– How do cultural heritage data, data structures, and standards help (or hinder) answering the above questions? What development opportunities does the research suggest for cultural heritage data standards?[12]

Although digital humanities education has developed dynamically in recent years, computer-based analysis of bibliographic sources is unfortunately rarely featured, and similarly absent from library science and IT education. In my opinion, this gap could be remedied by a new informal vocational training program that would appeal to those who are interested in some of the above issues and who already have some knowledge in one of the relevant fields (e.g., library science, cultural history, literary sociology, information technology). The analysis of records based on library bibliographic standards would probably also be of interest in library training. The training may take the form of a summer university or a seminar/course jointly organized by several university departments. Participants in the training could be university students or practicing professionals.

Brief outline of the curriculum:

1) the objective of bibliographic data science
2) theoretical models
    a) scientific theories and models (Darnton and his followers, Bourdieu, quantitative history, computational and data models of historical events – Thibodeau and Thaller)
    b) cultural heritage data models (the work-expression-manifestation-item model and its branches, ontologies, archival data models)
3) the data analysis workflow
    a) data acquisition. Main types of bibliographic data and data sources (library catalogs, citation databases, research data repositories, historical sources). Methods of data acquisition (standards, application programming interfaces, and tools), information about data structure (metadata schemas and serialization formats), terms of use.
    b) data validation. Technical validity (XML, JSON validity check) and quality assessment.
    c) preprocessing. File formats, data structures, conversion, and data loss control.
    d) data harmonisation (normalization and data enrichment). The reproducible conversion into a data set suitable for quantitative humanities analysis.
    e) data analysis and data visualization with programming (Python, R) and specialized tools.
    f) dissemination of results. Publication of software and research data for reuse.
4) After the research: the broader context. Professional communities, conferences, journals, continuing education opportunities.

This study merely outlines the data analysis workflow.

The bulk of the curriculum consists of the section on data analysis workflow. We distinguish between six interrelated stages in the workflow: data collection, pre-processing, data harmonization, analysis,

[11] Péter Király, *Measuring metadata quality*. PhD dissertation, University of Göttingen, 2019. https://doi.org/10.13140/RG.2.2.33177.77920.
[12] Király, Péter, Umerle, Tomasz, Malínek, Vojtěch, Herden, Elzbieta, Koper, Beata, Colavizza, Giovanni, Jagersma, Rindert, Lahti, Leo, Lindemann, David, Maciej Łubocki, Jakub, Milanova, Alexandra, Péter, Róbert, Rißler-Pipka, Nanette, Siwecka, Dorota, Romanello, Matteo, Roszkowski, Marcin, Tolonen, Mikko, Vimr, Ondřej. *Effects of Open Science and the Digital Transformation on the Bibliographical Data Landscape* = Gooding, Paul, Terras, Melissa, Ames, Sarah (eds.), Library Catalogues as Data: Research, Practice and Usage. Facet Publishing, 2025. ISBN: 9781783306589. pp. 19-44.

validation, and finally dissemination. The above stages do not necessarily appear in every research project. The Estonian National Library has described its data publication workflow,[13] which aims to make the Estonian national bibliography available for research purposes. The data is obtained via the OAI-PMH interface of the catalog, then converted from MARCXML format into tabular format, harmonized and enriched with data from external sources, and finally published with documentation of the process. This process lacks analysis, which, they hope, will be carried out by researchers rather than the library. The available data and methods have an impact on and limit the research questions. For example, if we want to study translations but our database lacks data on translators or attributes indicating the existence of a translation, we need to turn to other data sources—provided that they are available for the given period and language pair. Similarly, in order to analyze typographical features (print area, number and size of columns, fonts), we need to have digital photographs of the pages; if we only have access to the full text file, we need to modify the research question. Certain issues require high performance or specialized computational background. These conditions are not always apparent at the beginning of the research. It is important for researchers to bring any problems they encounter with data and data structures to the attention of the data provider;[14] dialogue helps to understand the differences in expectations regarding data and, in the longer term, can lead to wider use of public collection data.

3.a Digital data can be retrieved in three main methods. The most convenient option is for these to be available as downloadable files (e.g., as reusable research data). I collect and share information about such resources while developing the QA Catalogue,[15] but this type of data sharing is relatively rare. It is more common for data sources to be accessed through some kind of application programming interface. Various applications are available for the most common interfaces (OAI-PMH, Z39.50, SRW/SRU, SPARQL), so programming is not necessarily required, but time must be set aside to study the institution-specific settings and parameters of the interfaces. Finally, it is often the case that no previous opportunity was available. At this point, we extract the data from the HTML source of the website, assuming that the typographical formatting consistently indicates certain semantic elements[16] – but in this case, it is worth consulting with the website operator to see if there are any other options not documented on the site. Whichever solution you choose, make sure that the data license allows reuse. After downloading the data, the first step is to import it. Programming libraries supporting various bibliographic formats are available; for MARC21, for example, there are ones for Java, Python, Go, JavaScript, R, PHP, and other programming languages. There are four main types of data sources for library history research: i) library catalogs (e.g., national libraries or general purpose union catalogs, as well as catalogs of specifically old books, such as the VD16, VD17 and VD18 series that register 16th, 17th, and 18th century German books, their Italian counterpart EDIT16, and the Heritage of the Printed Book database), ii) digital library catalogs (Europeana, Gallica, German Digital Library, Hungarian Electronic Library, HathiTrust), iii) citation databases and research data repositories (DataCite, Zenodo, OpenAlex, Open Citation), and finally iv) databases of digitized (book) historical sources (the database of the Société Typographique de Neuchâtel, the database of 18th-century Dutch auction book catalogs MEDIATE, or the no defunct Eruditio in Hungary).

---

[13] Kruusmaa, Krister, Tinits, Peeter, Nemvalts, Laura, *Curated Bibliographic Data: The Case of the Estonian National Bibliography* = Journal of Open Humanities Data 2025. 11 (February): 16. https://doi.org/10.5334/johd.280.
[14] In addition to ad hoc problem reporting, research focusing on data quality and data structures is also present in bibliographic data science.
[15] https://github.com/pkiraly/qa-catalogue#datasources
[16] Király Péter, *Fetching Index Translationum* = Bibliographic Data blog 2025. https://bibliodata.substack.com/p/fetching-index-translationum

3.b The validity and quality of data are usually checked in two ways. Although there are schemas describing the structure of data and software that can be used to check whether a document complies with these schemas, in most cases these schemas are limited to describing only a general structure (for example, a MARC record contains control and data fields, the latter may contain indicators and subfields), i.e., they only affect the outermost layer of the data. It is therefore worth performing further checks – either using software available for the given format or using the Exploratory Data Analysis methodology. The simplest method is the so-called completeness check, which examines what data elements are found in the database and in what proportions.[17] It is also worth examining the content of the most important data elements covered by the analyses to see how consistent they are in terms of form and content: how many different forms does the same person or geographical name appear in, or how were the dates recorded? By browsing through a frequency list of values occurring in a given data element, we can gain an understanding of the nature of the data and the harmonization tasks to be performed in the subsequent processing steps. Such a list can be coded; for example, Harald Klinke grouped the dates in the Museum of Modern Art database according to format patterns (four numbers, four numbers-four numbers, four numbers-two numbers, etc.), thus obtaining a more manageable sample list instead of many individual dates.[18]

3.c During preprocessing, we convert the imported files into a data structure that is more suitable for processing with standard data analysis methods (in Python, Pandas is the most widely used, while in R, it is the Tibble "data frame"). It may happen that we do not transform all data, but only certain records (for example, only 17th-century books from a national bibliography) or certain data elements (for example, we omit library identifiers and other administrative data elements).

3.d Data maintained by others rarely fit in every respect to the specific analytical purpose for which we are preparing them. The data that are important to us must be harmonized, i.e., normalized (standardize, resolve contradictions, convert certain data types—such as particular text variables to numeric ones) and enriched (calculate derived data, such as page numbers, import data from external data sources). Below, we examine four such harmonization steps: the harmonization of dates, place names, persons, and concepts. The dates show a high degree of variation not only in the MoMe collection, but in almost every library catalog we find dates that differ from the format that is easy for programs to handle. For example, dates given in Roman numerals ("MDCCLXXX. [1780]"), in text form ("druk janvier 2016."), according to the reign of a monarch ("Meiji 40") or according to another calendar. Another problem is the handling of uncertain dates (e.g., "18--" and "18uu" in library catalogs both mean that the publication is from the 19th century). Due to the variety, the conversion is not trivial, but neither is deciding what to convert the dates to in the end. There are different approaches in certain areas (see, for example, archival standards or the practice of Europeana). As the latest proposal, the *undate* Python library[19] created by the DHTech community stores the following data elements: the unchanged form of the date in the source, the calendar, the accuracy of the date, the earliest and latest normalized dates, and the duration – i.e., for the sake of consistency in retrieval, the date is always a time range. For place names, gazeeters are available for

[17] see Kruusmaa et al. op. cit.
[18] https://x.com/HxxxKxxx/status/1066805548866289664
[19] Koeser, Rebecca Sutton, Damerow, Julia, Casties, Robert, Crawford, Cole, *Undate: Humanistic Dates for Computation: Because Reality Is Frequently Inaccurate* = Computational Humanities Research 2025. 1: e5. https://doi.org/10.1017/chr.2025.10006.

identification and, if necessary, for extra data elements required for map representation or the display of language variants. Among the most important ones are CERL Thesaurus,[20] Getty Thesaurus of Geographic Names,[21] and Geonames,[22] which can be queried via APIs. Although these are rich databases built from many sources and thoroughly checked, practice shows that in almost every bibliographic source we will find name forms that are not recognized by these services, so these can be incorporated into our own database with some non-automated manual data refinement. The same procedure can be followed for individuals, but naturally using different services: VIAF (Virtual International Authority File),[23] the CERL Thesaurus[24] personal name database, ISNI (International Standard Name Identifier),[25] Wikidata.[26] It is important to note that any given database will naturally contain many more personal names than geographical names, so the hit rate is likely to be lower. The world of concepts is much more diverse than that of geographical and personal names. Although there are universal conceptual dictionaries (knowledge organization systems), there is virtually no library catalog whose records contain only the concepts of a single dictionary. Instead of specific dictionaries, we recommend using the BARTOC service (Basic Register of Thesauri, Ontologies and Classification)[27] to find the dictionary that best suits your research questions. When discussing harmonization, it is essential to mention the categories of inaccurate, incomplete, subjective, and uncertain data.[28] We have seen an example of inaccurate data and its handling in the case of dates. Incomplete data is when we do not know all the details, for example, not all authors of a work are listed, or there are gaps in the provenance history of an object. We can deduce some data, but it is very difficult to describe what does not exist. Subjective data refers to provenance, i.e., who made the statement in question. Such statements are often hypothetical and may even be contradictory. Finally, uncertain data is when the truthfulness of a statement is doubtful. An important part of the theories cited above is that the past is constructed, and the interpretation of sources also depends on the interpreter's prior knowledge. Consequently, historical information systems must necessarily allow for the coexistence of contradictory interpretations, and instead of binary (true-false) logic, uncertainties could be described using probability values.[29] For example, "Alexandre Dumas" (if no other information is available) could refer to either the father or the son (both writers)—the former being more likely, the value of which can be recorded in the database and used, for example, when sorting search results.

3.e The most spectacular phase of the work process is data analysis.[30] Here, the researcher translates his or her own research questions into the operations offered by the tools used. Two types of data are used in the analyses: on the one hand, the original or enriched data available after data harmonization (title, place of publication, language, number of pages), and on the other hand, calculated data based on their analysis (including font type and size, amount of paper used to produce the form, number of words, number of

[20] https://data.cerl.org/thesaurus/
[21] https://www.getty.edu/research/tools/vocabularies/tgn/index.html
[22] https://geonames.org
[23] https://viaf.org/
[24] https://data.cerl.org/thesaurus/
[25] https://isni.org/
[26] https://wikidata.org
[27] https://bartoc.org
[28] Mariani, Fabio, *Introducing VISU: Vagueness, Incompleteness, Subjectivity, and Uncertainty in Art Provenance Data* = Proceedings of the Workshop on Computational Methods in the Humanities 2022, Vol-3602. Workshop Proceedings, 2023. pp. 63–84. https://ceur-ws.org/Vol-3602/paper5.pdf.
[29] Thaller, Manfred, *On vagueness and uncertainty in historical data* = Ivory Tower blog, 2020. https://ivorytower.hypotheses.org/88.
[30] see, among other things, the literature cited in connection with the research questions.

printed words per capita). Data analysis methods are not unique to bibliographic data science; they are general methods for which general and (to a lesser extent) specialized data science teaching materials in the humanities[31] are available, as well as textbooks on quantitative history.[32] The curriculum cannot, of course, cover every conceivable data science technique, but it should use examples to introduce the most common procedures used in the history of the discipline (the basics of statistics, time series analysis, data visualization including map representation, text analysis, network analysis).

3.f The final step in the work process is the dissemination of results, which includes traditional publication methods (papers, books, conference presentations) as well as newer approaches, such as the publication of software used in the process, the generated data, and data and software studies focusing specifically on these, blogging and microblogging, sharing presentation slides and recordings, and participating in professional organizations. I would like to emphasize the importance of special data sharing. Imagine the following scenario: a researcher has worked hard to enrich a popular data source with high research potential that is maintained by a public collection. Later, another researcher would like to use the same database for their research. If she is not familiar with the previous researcher's work, she can start the data enrichment process from scratch. But even if the first researcher published his data enrichment, it is much more likely that the subsequent researchers will find and use the original database. To prevent this, researchers would need to return the modified data to the original data provider. Fortunately, MARC21, introduced in the 34th update in 2022[33] a data provenance subfield to distinguish between data recorded by the library and data recorded by the researcher (and it is available in most fields), which could be a theoretical remedy for the library's legitimate demand to take responsibility for its own data. In the life sciences, researchers can use nanopublications to share data enrichment steps with libraries, which can then incorporate them into their catalogs without compromising their own responsibility and credibility. The second researcher can then work on the data-enriched version. In order to realize this vision, communication between the parties must be standardized, and the research community can play a coordinating role in this process.

Technical methodology. The curriculum covers a fairly broad range of topics, for which there is no single expert, so it is necessary to bring together experts in the various sub-fields. I recommend using a platform that allows collaborative editing for your writing. Like in other similar initiatives (DHTech, Programming Historian, etc.), in the final phase, it is advisable to transfer the materials to a Git-based platform and a manageable format (e.g., markdown) and automate the process of converting them into publication formats (HTML, PDF, epub). The completed teaching materials must be archived in a repository where version control is ensured and they are assigned a permanent identifier (DOI). For programming exercises, we also create Jupyter notebooks that can be edited and run in a browser, and place them on a platform that allows them to be executed.

---

[31] some of these: Karsdorp, Folgert, Kestemont, Mike, Riddell, Allen, *Humanities Data Analysis: Case Studies with Python*. Princeton University Press, 2021. ISBN 9780691172361; Klinke, Harald. *Cultural Data Science: An Introduction to R.* Springer, 2025. ISBN 9783031881305.; Taylor, Arnold, Tilton, Lauren, *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text.* 2nd edition. Springer, 2024. ISBN 9783031625657; Bátorfy Attila, *Adatvizualizáció. Elmélet, rendszer, módszer. Bevezetés az adatok grafikus ábrázolásának elméletébe és gyakorlatába.* Budapest: ELTE Eötvös Kiadó, 2024. ISBN 9789633123973; and the *Programming Historian* site: https://programminghistorian.org.

[32] Lemercier, Claire, Zalc, Claire, *Quantitative Methods in the Humanities: An Introduction*. University of Virginia Press, 2019. ISBN 9780813942681; Feinstein, Charles, Thomas, Mark, *Making History Count: A Primer in Quantitative Methods for Historians.* Cambridge University Press, 2002. ISBN 9780521001373.; Hudson, Pat, Ishizu, Mina, *History by Numbers: An introduction to quantitative approaches.* 2nd edition. Bloomsbury, 2016. ISBN 9781849665377.

[33] https://www.loc.gov/marc/up34bibliographic/bdapndxg.html

Target audience. The curriculum is designed for students and practicing professionals who would like to learn about digital methods in library history and bibliographic data science, and is primarily applicable in the following areas: a) digital humanities b) public collection studies (library and information science, archival science, museology), c) computer science. Each training session should be tailored to the needs of the target audience. For example, training for metadata specialists familiar with metadata standards should focus more on digital methods, while training for IT specialists should cover the specifics of bibliographic data programming.

Quality assessment and sustainability. Digital methods change rapidly, and there is a high probability that the relevant part of the curriculum at the time of writing will be based on tools or methods that will be obsolete a few years later. The creators must decide what to do with the outdated parts. The Programming Historian, for example, has developed a principle for extracting teaching material.[34] It is also important to prevent rapid obsolescence, so best practices in the field should be followed, such as the Programming Historian guidelines.[35] The key to long-term maintenance of the curriculum is regular use, so its developers must be committed to applying it in teaching, thereby ensuring its regular review.

Collaboration. In the final stage of curriculum development, it is worth contacting potential educational partners: on the one hand, universities, and on the other hand, informal educational initiatives such as Programming Historian, the DHTech education working group,[36] the LIBER Data Science in Libraries working group,[37] Library Carpentry,[38] GLAM Labs,[39] and the Research Software Engineering communities.[40] Adapting and incorporating the material into some of these initiatives should be considered.

---

[34] https://programminghistorian.org/en/lesson-retirement-policy
[35] Sustainable Writing: https://programminghistorian.org/en/author-guidelines#sustainable-writing, Reviewer Guidelines for Assessing Lesson Sustainability: https://programminghistorian.org/en/reviewer-guidelines#sustainability, Editor Guidelines for Fostering Lesson Sustainability: https://programminghistorian.org/en/editor-guidelines#c-sustainability--internationalization-review.
[36] DHTech DHTech Training & Education Working Group, https://dh-tech.github.io/wg-education-training/
[37] LIBER Data Science in Libraries Working Group, https://libereurope.eu/working-group/liber-data-science-in-libraries-working-group/
[38] https://librarycarpentry.org/
[39] https://www.glamlabs.io/
[40] Research Software Engineering, https://society-rse.org/, https://us-rse.org/, https://de-rse.org/