Czech Literary Bibliography
Research Infrastructure
Institute of Czech Literature of the CAS

Data Management Plan

Administrative information

Project number: LM2023043

Project title: Czech Literary Bibliography

Period in which project was carried out: 1. 1. 2023 – 31. 12. 2026

DMP Version: 1.0

DMP Publication date: October 2023

DMP Licence: Attribution 4.0 International (CC BY 4.0)

The project Czech Literary Bibliography (LM2023043) is supported by the Ministry of Education, Youth and Sports within the specific grant scheme aimed at financial support of the large research infrastructures.

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Contents

Introduction 4
1. Description of data and data processing 5
Data creation and collection 5
Data processing 7
Continuous and acquired data 10
Measures for ensuring data quality 10
2. FAIR data 12
Making data findable 12
Data dissemination 12
Data interoperability 13
Data reuse 14
3. Data storage and backup 15
4. Ethical and legal issues 16
5. Other CLB outputs 17
6. Responsibility for data management 18
Index of tables 19

Introduction

This data management plan (Data Management Plan – DMP) was created to address the needs of the Czech Literary Bibliography research infrastructure (CLB), in particular those of the project LM2023043, which is being carried out over the period 2023–2026. The main task of the CLB is the preparation and ongoing development of an analytical bibliography of Czech literature, which continuously maps the discussion of Czech and (selectively) world literature and culture in the periodical press and book production published within the Czech lands, in all historical languages of the region, as well as post-1948 exile production. The period covered, in chronological order, by the CLB database is the last third of the 18th century to the present day. **The CLB produces data on a long-term and continuous basis. CLB data is not static (immutable) but continuously updated and augmented by new information and other data** (links to full texts, persistent identifiers, etc.). Due to this fact, the CLB undertakes regular versioning of data sets, which are archived and shared among those interested in their further use.

The CLB considers the DMP to be an important part of proper research and documentation practice. The DMP presented here describes data management in the CLB and explains how the CLB deals with open access requirements. The goal of the DMP is not only to provide information about the lifecycle of data as it is generated, collected, processed, stored, and shared within the CLB, but also to ensure that data is made available in accordance with FAIR principles.[1] The DMP is meant to facilitate the visibility of data and make it easier to use by third parties. The DMP is a living document that is updated as needed.

[1] The FAIR principles explain how data should be processed to make it more findable, accessible, interoperable, and reusable.

1. Description of data and data processing

Data creation and collection

The CLB collects, processes, stores and makes available metadata, i.e. a structured description of printed and digital documents that map literary life and literary science in the Czech Lands and circulation of Czech literature abroad, from a period starting at the end of the 18th century and continuing to present day, regardless of language, format, or medium. These typically take the form of bibliographic or other knowledge bases. The CLB considers such metadata to be real data as managed within the institution. This data is created through manual excerpting, which takes place as the recording of excerpts from primary and secondary literary sources. Data is created by CLB employees primarily in Aleph[2] and Koha[3] software, and (historically) also in the ICL's own proprietary system (RETROBI), in which they continue to manage it. Data from this software is regularly exported in open format and stored in the ASEP[4] institutional repository, where it is archived long-term while remaining available for download.[5]

The CLB has collected more than 2.3 million records to date, and continues to add approx. 20,000 records annually.

[2] Aleph is an integrated library system by Ex Libri.

[3] Koha is an open-source integrated library system.

[4] ASEP is an institutional repository for the Czech Academy of Sciences, serving to record all its relevant outputs.

[5] Starting in October 2023, the data sets of all individual databases listed in Table 1 have been stored in the ASEP repository, with the exception of the Czech literary figure database, which, due to data cleaning, will be inserted into the repository later in the project.

Data type	Title of database	Method for obtaining data	Data processing: inputs	Data processing: outputs	
Bibliographic and biographical data	Retrospective Bibliography of Czech Literature 1770–1945	Digitized card catalogue	RETROBI	RETROBI	VUFIND
		(Semi)automatically edited database records	RETROBI	RETROBI	
		Manually entered database records	ALEPH	ALEPH	
	ALKARO – Almanacs, Calendars, and Yearbooks in the Czech Lands 1801–1945	Manual extraction of public resources	ALEPH	ALEPH	VUFIND
	Current Bibliography (since 1945)	Manual extraction of public resources	ALEPH	ALEPH	VUFIND
	Czech Literary Web Bibliography				
	Czech Literary Samizdat Bibliography				
	Czech Literature of Exile Bibliography				
	Czech Literature in Translation	Created by an expert on the given topic + manual excerpts of public resources	Excel file	VUFIND	
	Czech Literary Biographical Database	Manual excerpts of public sources and questionnaires	Koha	CLO	

Table 1: Data processing

Table 1 above shows the methods for obtaining data and systems in which bibliographic and biographical data is processed and presented in the case of each individual database. In terms of structure, the CLB is divided into eight databases; however, the Retrospective Bibliography of Czech Literature also includes ALKARO, and the Current Bibliography includes Czech Literary Web, Samizdat and Exile Bibliography. The following table (Table 2) gives an overview of all possible formats for managing this data in the CLB. The data size of individual databases ranges from tens to lower hundreds of megabytes, depending on the given format. Software data and their formats are described in section 5.

Target group	System	Data format
Publicly available formats	VUFIND	JSON
		MARCXML
		MARC21
		DOCX
		PDF
	ALEPH OPAC	SAV
		END
Formats for internal use	ALEPH client	MARC21
		MARCXML
		ALEPH sequential
		TXT
	Internal scripts (python)	CSV
		JSON
		XML
	KOHA	MARC
		XML
	EXCEL	XLSX

Table 2: Formats of bibliographic and biographical data in individual systems

Data processing

The CLB processes a set of eight separate databases that all share the MARC21[6] metadata standard for bibliographic and authority records. The data structure of the databases is harmonized, including the material description systems (national authority file and local add-ons, UDC[7], conspectus group), and supplemented with persistent identifiers (authority identification number, ISBN[8], ISSN[9], DOI[10], CNB number[11], and local CLB identifiers[12]). The

[6] MARC21 is an internationally recognized format for the interchange of bibliographic information.
[7] UDC (Universal Decimal Classification) is a universal classification language combining hierarchical and faceted types of classification.
[8] ISBN is an international identifier for books and non-periodical publications.
[9] ISSN is an international identifier for periodicals and other serial publications.
[10] DOI is an international identifier for objects in the digital environment.

databases are technologically internally connected and covered by a unified search system on the VuFind[13] software platform, which also serves as the primary access point for end users. The standardized international exchange format and unified cataloguing make it technically possible to interface CLB data to other scientific information exchange systems.

Library standards

Data is processed as much as possible in accordance with the following standards and norms:

- MARC21 cataloguing format (for bibliographic and authority records)
    - Cataloguing according to RDA rules in the MARC21 format[14]
- International Standardized Bibliographic Description Rules (ISBD)[15]
- Name and subject description standards – Czech National Authority

These standards are the basic starting point for the processing of any data in modern library database systems, and compliance with them significantly facilitates the mutual sharing of data between individual institutions. It should be emphasized, however, that the CLB's primary goal is to create a comprehensive bibliographic description that contains all available information about the given document, thus serving the needs of the scholarly community, including those interested in quantitative research of bibliographic data. To this end, CLB records usually contain information beyond the scope of normal library cataloguing practice (mainly with regard to data on accountability and the amount of detail given by subject descriptions; all records are consistently annotated), and they have a more formalized registration method, especially with regard to search options. Because the CLB processes types of records for which there are no satisfactory cataloguing manuals (articles) at the national level, or for which rules are not available at all (analytical breakdown of books, processing of online documents, samizdat and exile periodicals, audiovisual materials, retrospective excerpts of continued publications, etc.), it has created its own methodological materials in the form of internal cataloguing manuals and publicly available certified methodologies.

---

[11] The CNB number is the Czech identifier for the production of texts in the Czech context during the 19th, 20th, and 21st centuries.

[12] The CLB creates local files of persistent identifiers for those values that are not represented in general files.

[13] VuFind is an open library search engine.

[14] The CLB's practice is based on the national interpretation of RDA cataloguing rules. Given that there are no standard rules for the description of articles within the Czech national context, the CLB uses its own interpretation, which is described in the data processing manuals named below.

[15] 'ISBD: International Standard Bibliographic Description' (Consolidated Edition). Berlin: De Gruyter Saur, 2011.

Methodology handbooks

- MACEK, Emanuel: *Bibliografie české beletrie a literární vědy. Metodická příručka* ['Bibliography of Czech Fiction and Literary Science: A Methodology Handbook']. ICL CSAS, 1969.
- [MIKA, Jiří]: *B97. Databáze české literární bibliografie od roku 1997* ['B97: The Czech Literary Bibliography Database Since 1997', internal document] Prague: ICL CAS 1996.
- MALÍNEK, Vojtěch. *Oborová analytická bibliografie. Metodika zpracování* ['Specialized Analytical Bibliography: Processing Methodology']. ICL CAS, 2020. ISBN 978-80-7658-008-4. Available online at https://kramerius.lib.cas.cz/uuid/uuid:2679143f-a41f-443d-be8a-34f7be85fc91.
- *Metodika tvorby a kontroly jmenných autorit podle RDA*. ['Methodology for the Creation and Control of Name Authorities According to the RDA']. Available online at https://autority.nkp.cz/jmenne-autority/metodicke-materialy/metodika-jmena-cvicne-2.
- *MARC 21. Pracovní manuál pro potřeby zpracování České literární bibliografie* ['Working Manual for Processing the Czech Literary Bibliography', internal digital document]. Version 6.0. Czech Literary Bibliography, 2023.
- *MARC 21. Pracovní manuál pro potřeby katalogizace jmenných autorit v databázi v České literární osobnosti* ['Marc 21: Working Manual for Cataloguing Name Authorities in the Czech Literary Figures Database', internal digital document]. Version 3.0. Czech Literary Bibliography, 2023.
- ŘEHÁK, Daniel. *Metodika zjišťování autorů u šifer a pseudonymů v periodickém tisku* ['Methodology for Identifying Authors in Ciphers and Pseudonyms in Periodicals']. ICL CAS, 2015. Certified methodology MK-S 13094/2015 OVV. Available online at http://hdl.handle.net/11104/0253681.

The two older works at the top of the list were written with the purpose of standardizing the processing of the field-specific article bibliography at the Institute of Czech Literature of the CSAS/CAS[16]. Macek's manual primarily deals with the processing of article-based bibliographic records for the Retrospective Bibliography of Czech Literature. This bibliography was created in the form of a card file, which was standard at the time. Jiří Mika's manual for excerpts in the period 1997–2012 was the first methodology text on the processing of field-specific article bibliographies in the form of a database, to be implemented with the original ISIS database system. Malínek's methodology for processing field-specific bibliographies was created under the auspices of a NAKI project and describes the processing of field-specific analytical bibliographies in MARC21 format, with an emphasis on the needs of literary bibliographies.

[16] Czechoslovak Academy of Sciences / Czech Academy of Sciences.

The processing of biographical data for the needs of the Database of Czech Literary Figures is based on original standards for the processing of a proprietary biographical database, which were supplemented during the conversion to MARC21 format according to the principles of the Methodology for the Creation and Control of Name Authorities According to the RDA, drawn up by the Department of National Name Authorities of the National Library of the Czech Republic (the 'ONJA handbook'). These rules were further supplemented and adapted to meet the needs of biographical research in the field of literary science. A definitive version of current description rules has been created in the form of an internal manual.

Continuous and acquired data

This DMP has been created to meet the needs of the Czech Literary Bibliography project, which is being implemented over the period 2023–2026. The data that figures in this project was created at the CLB during previous years of operation. We refer to them as 'continuous data', because they have been generated continuously since the CLB was first founded, and the CLB has continued to work with them in all subsequent periods of operation.

In addition to data obtained through its own excerpts, the CLB also processes data it receives from collaborating institutions (open datasets provided by the National Library in Prague, for example). The CLB also currently uses the following datasets in MARC21 format:

- dataset of the Comprehensive Catalogue of the Czech Republic, released under CC0 license.

For experimental purposes within development of the Literarybibliography.eu portal (see chapter 5), the CLB also uses the following datasets:

- dataset of the Polish literary bibliography, released under CC BY license,
- Fennica dataset of the Finnish National Library, released under CC0 license,
- Arto dataset of the Finnish National Library, released under CC0 license,
- dataset of the National Library of Spain (BNE), released under CC0 license.

Measures for ensuring data quality

Data management of the CLB corresponds to industry standards. CLB data quality is achieved by introducing a systematic three-phase data verification process that includes automated data validation. Records are also continuously redacted and harmonized at the level of individual field registers in order to remove discrepancies introduced by the implementation of various description rules in the past and deviations in handwriting among various processors.

The CLB uses an advanced system carried out on several levels for verifying processed data. The data structure (punctuation, proper notation, etc.) is verified by (a) a basic automatic validation at the moment the record is saved, (b) a daily detailed batch script that monitors

all newly created or modified records, (c) the same validation script, applied once per week, and (d) discrete one-off tests carried out at least once per year. In parallel, content verification is carried out by (a) each worker, who verifies all records he or she has created during the previous month, (b) a randomly selected employee, who subsequently verifies the same records, and finally (c) a database editor, who also verifies the records.

Database editing is carried out mainly according to the registers of individual database fields, whose data is automatically or manually corrected, updated, and harmonized as necessary in order to achieve uniform entry across the entire database. In particular, the extent to which records are connected to authority databases and persistent identifiers is systematically verified.

2. FAIR data

Conformation of research data to FAIR principles – regarding its findability, accessibility, interoperability, and reusability – is one of the CLB's priorities.

Making data findable

Making data and research findings available is one of the main goals of the research infrastructure. Data can be accessed by means of a wide variety of channels in open and standard formats, primarily through the VuFind web search engine, as well as tables and graphs exportable from the Statistical and Analytical Module (SAM) and through national collaborative networks such as Knihovny.cz[17] and the Union Catalogue of the Czech Academy of Sciences. For those interested in computational data processing, CLB data is available at the ASEP institutional repository in standardized MARC21 format, the customary format for the field of bibliography and library science. The CLB actively enables connections with other data portals and releases its data to meet the needs of cooperative networks for the exchange of scholarly information.

International handle and DOI identifiers are assigned to individual datasets from each database, which are provided with metadata structured on the basis of the UNIMARC format. Indexing and searchability of metadata is provided by the operator of the repository in which the data is stored and is also available on the CLB website.

Data dissemination

The CLB follows the general principle of open access in its dissemination of data. In practice, this means that all research data and metadata is made available, within the limits of legal regulations, under a Creative Commons license. Personal and sensitive data regarding living persons is only published if it has already been published, or with consent of the person concerned, who may also request that part of the data be archived and made available only after contractual conditions are fulfilled.

Whole datasets are made available in the following ways:

1) through the ASEP institutional repository,
2) through CLB websites and services, including
    a) the CLB VuFind central access point,
    b) proprietary interfaces of individual databases (OPAC Aleph, RETROBI, CLO),
    c) the Literarybibliography.eu portal,
    d) the SAM export module,

[17] The Knihovny.cz portal provides uniform access to Czech and Moravian libraries services.

3) the cooperation networks and collective portals (Knihovny.cz, CAS Union Catalogue, the INDIHU.cz platform).

Repository

Data is stored at the ASEP institutional data repository: https://asep-portal.lib.cas.cz/, which is registered in the RE3DATA data repository register. Data stored in this way is published in the online catalogue and provided with handle and DOI persistent identifiers. The description of individual data records is based on the MARC21 standard. The metadata of all records is freely accessible.

Data

CLB datasets are regularly published in the ASEP repository twice a year (in March and October) in open access mode. This publicly available data does not contain any personal, confidential, or sensitive data. Data with limited access is released in accordance with the applicable legal regulations (see *Ethical and legal issues* below).

Metadata

Basic metadata for datasets is published on individual database webpages of the CLB website. An extended form of metadata for individual datasets is available in the repository of the Library of the Czech Academy of Sciences, which is undertaking its further development. Metadata structure in the repository is based on the UNIMARC format. The CLB cooperates with the Library of the Czech Academy of Sciences to address data repository needs, with an eye to the long-term availability of metadata and data.

Data interoperability

The CLB aims to collect and document data in a standardized manner. Interoperability is ensured by processing the data in the MARC21 metadata standard. The data can then be machine-processed and linked to other existing datasets. The CLB also actively enables connections with other data portals. When processing data, the CLB follows standardized thesauri that include both name (figures and corporate bodies, events and works) and subject (thematic, geographical, chronological, and formal) authorities, which are supported by the professional library and bibliographic community. Due to a high degree of standardization and connection to existing systems of persistent identifiers, as well as the implementation of tools for sharing and exchanging data, it is exceedingly easy to link CLB data to data outside the CLB.

Data reuse

Data is offered to users throughout CLB research infrastructure activities. All published data is available in open access mode and provided with Creative Commons – namely CC BY-NC-SA 4.0 – licenses, which allow reuse in accordance with open science requirements.

CC BY-NC-SA 4.0 licenses define the scope of authority with regard to Czech Literary Bibliography databases (hereinafter referred to as the Work) as follows:

The User is entitled to:

- share the Work: that is, to extract, exploit, reproduce, and share the entire content of the database or a substantial part of the content of the database.
- modify the Work: that is, to change or build on the content of the database.

It is stipulated that the User may undertake these actions on the condition that he or she:

- indicate the origin of the Work: the User is obliged to indicate the origin of the Work in a predetermined manner. Citations must include information regarding the title and creator of the Work, source of data, and link to the full text of the CC BY-NC-SA 4.0 International public license terms.
- not use the Work commercially: it is forbidden to use the Work for commercial purposes.
- maintain license: if the User modifies, alters, or builds upon this Work, he or she must exhibit derivative Works under the same license as the original Work.

The use of databases is conditional on the User's proper citation and inclusion of information regarding the use of CLB for research and development outputs following the recommended template:

The resources of the Czech Literary Bibliography research infrastructure were used in the creation of this work [book/study] – https://clb.ucl.cas.cz/ (code ORJ: 90243).

3. Data storage and backup

All Data is backed up directly at the Institute of Czech Literature (ICL) in at least two copies each one stored in a different physical location. The main part of CLB data (bibliographic database) is processed in the Aleph system, operated by the Library of the Czech Academy of Sciences. This data is backed up by the CAS Library in five copies each one at a different physical location in the Czech Republic. CLB data is also displayed in other shared catalogues (e.g. Knihovny.cz) with continuous updating and with its own backup system. For all intents and purposes, there is no possibility for major damage or loss to CLB data.

CLB data management together with IT operation and development takes place primarily within the digitization and IT section of the CLB and in cooperation with the Computing Centre at the Institute of Czech Literature. All network applications are run on virtual servers located in a secure data centre that is owned exclusively by the ICL and under the administration and direct management of ICL IT Centre employees. Storage capacity is designed to ensure a sufficient reserve (at least 30%) over the planned lifespan, and is regularly updated and increased in two-year cycles. Data and application servers are backed up in accordance with current security standards, primarily on several levels at our own storage facilities. In addition to this, a standby copy of servers is maintained in several previous versions within the building, and an archive version is stored in encrypted form on the CESNET repository (e-INFRA CZ) outside ICL premises.

Long-term archiving of data is also ensured by storage at the ASEP institutional repository, operated by the Library of the Czech Academy of Sciences (see section 2, *Data dissemination*).

The sustainability of CLB data is currently determined by its size and the implementation of international standardized formats, and is also greatly strengthened by moving to the MARC21 international standard, which can be easily converted into practically any other format. A number of specialized and free-access tools are made available for this purpose.

4. Ethical and legal issues

The kind of data processed and enriched within the framework of the CLB is mainly public data. Some data however belongs to the category of personal and sensitive data (especially the content of the CLO database), and therefore requires a special access framework in accordance with GDPR regulations. The CLB does not make data available where there is any risk that their publication would result in an unreasonable interference with the right to privacy and personal data protection. The protection of personal data is a central ethical requirement, and CLB employees have a duty to provide information about how this data is handled. Information on the protection and processing of personal data is published on the CLB website, which contains an overview of the processed data, including the purpose and form of their processing, as well as the method of storage, accessibility of personal data, security methods, and other important information.

Access for the submission of data is permitted only from defined IP addresses and secured by a CLB employee user account. From a legal point of view of, the CLB is a workplace authorized to process personal data, with a certificate to do so issued by the Personal Data Protection Office. Personal data is published only in accordance with legal standards (Personal Data Processing Act, GDPR). Personal data is not published, (a) if the person concerned does not give consent, (b) if the data in question has not been published elsewhere and the CLB does not have the express consent of the person in question to publish it. All collection, processing, and storage of information is strictly governed by applicable legislation, with all due consideration for the protection of personal data of living persons (Personal Data Processing Act, GDPR).

5. Other CLB outputs

The CLB also actively develops software and IT tools. In addition to software for the digitization and online presentation of RETROBI ticket files, this mainly concerns adapting the VuFind open source discovery system, which CLB users have been able to use as central data access point since 2020. All CLB bibliographic databases are now accessible there for joint searching, and other CLB databases are gradually being prepared for inclusion. As a separate VuFind module, the CLB is developing a Statistical and analytical module (SAM), which will enable visualization and statistical analysis of CLB bibliographic data.

The CLB works closely with its nearest foreign partner, the Department of Polish Literary Bibliography, on the development of software and IT tools. Both national infrastructures are involved in the creation of the Literarybibliography.eu portal, which makes available bibliographic data on individual national literatures. At present, individual databases can be selected for Czech and Polish literary bibliographies, as well as Finnish and Spanish National Libraries.

The CLB develops software under a free license and publishes documentation on GitHub.

Data type	Data description	Formats	Available at
Scripts	Scripts for managing, checking, cleaning, and reusing data	.py, ipynb	https://github.com/CzechLitBib/UCL
Software	SAM	php	https://github.com/uclavcr/sam
	RETROBI	java	https://github.com/uclavcr/retrobi

Table 3: Other CLB outputs

Open science principles are also taken into account in all outputs of the CLB's publishing activities. Specialized book bibliographies are published as part of the Bibliographica series, and are archived in the ASEP institutional repository, as well as the Digital Library of the Czech Academy of Sciences, where they are made available to users after a three-year embargo. Starting in 2022, these publications have been assigned a DOI identifier. All other types of documents are published by CLB staff in accordance with the 'Open-Access Strategy for Scientific Information, 2022–2027'[18]. The CLB recommends that its employees consider the possibility of publishing in the open-access mode, and provides support to authors who wish to archive background data for their works.

[18] Internal ICL document ('Strategie otevřeného přístupu k vědeckým informacím pro roky 2022–2027').

6. Responsibility for data management

The producer of data is the Czech Literary Bibliography research infrastructure, operated by the Institute of Czech Literature of the Czech Academy of Sciences, which is the owner of all data.

Each member of the research infrastructure is responsible for correct processing of data collected by the CLB, and to proceed in accordance with established standards and work procedures.

Starting in 2020, a coordinator has overseen open access and related imperatives for the CLB. During the monitored period, an open access strategy was drawn up at the level of the ICL, which also manages the relevant activities of the CLB.

Index of tables

Table 1: Data processing..... 6
Table 2: Formats of bibliographic and biographical data in individual systems..... 7
Table 3: Other CLB outputs 17