From a FAIR Workflow Blueprint to an Executable Data Harmonisation Pipeline: Exploiting the BnF SPARQL Service to Produce and Disseminate an Early Modern Research-Oriented Bibliographic Dataset

Arianna Moretti[1,2], Iiro L. I. Tiihonen[2] and Jonas P. Fischer[2]

[1]Università di Bologna, Dipartimento di Filologia Classica e Italianistica (FICLIT), 40126 Bologna, Italy
[2]University of Helsinki, Computational History Group, 00014 Helsinki, Finland

Abstract

This article introduces the design of a workflow blueprint compliant with FAIR principles and grounded in Open Science practices, developed for the retrieval, harmonisation, and publication of curated bibliographic metadata. In the documented case study, the blueprint was tested and refined on metadata retrieved from the Bibliothèque nationale de France (BnF) SPARQL endpoint, focusing on bibliographic resources from the Early Modern period and the agents involved in their production and distribution. The aim is to ensure transparency, traceability, and interoperability throughout the entire data lifecycle, fostering an approach that integrates automated steps with human intervention. The work resulted in two main outcomes: (1) a reusable workflow blueprint for the production of harmonised bibliographic datasets derived from open repositories, and (2) an executable workflow designed to ensure the reproducibility of the Early Modern BnF dataset production pipeline, supporting its reuse and re-execution over time.

Keywords

Workflow, Reproducibility, Data Harmonisation, Linked Open Data

1. Introduction

This paper presents a workflow blueprint for producing a harmonised subset of historical bibliographic data from an open data source and disseminating it in interoperable formats suitable for research, including Linked Open Data (LOD). The blueprint is then implemented as an executable workflow on a case study that focuses on the Early Modern bibliographic metadata extracted from the Bibliothèque nationale de France (BnF)[1] SPARQL endpoint. In this paper, we use the term *workflow blueprint* to refer to a reusable conceptual model, general enough to provide guidance and best practices across a broad range of case studies, independent of specific technologies and datasets, aiming to describe logical steps, roles of the actors involved, input and output types. On the other hand, by the term *executable workflow* we mean a specific, operational implementation of the blueprint in a real-world case study. It includes instructions to re-run the process, expressed as ordered steps in a dedicated environment, and provides links to the code-based actionable operations required to execute the steps.

Aligned with the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) and the Open Science movement [1, 2], this work encourages collaboration, reuse, and third-party validation of results. Consistently, the project is contextualised in the Open Science framework, fostering the recognition of research products not only for the final output but also for methodologies, tools, and intermediate process products. In this sense, workflows are naturally well-suited tools for promoting reproducibility and replicability in research. Their adoption not only optimises process speed and reduces the possibility of error, but also increases confidence in results and strengthens the systematic

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

arianna.moretti4@unibo.it (A. Moretti); iiro.tiihonen@helsinki.fi (I. L. I. Tiihonen); jonas.fischer@helsinki.fi (J. P. Fischer)
0000-0001-5486-7070 (A. Moretti); 0000-0003-0703-4556 (I. L. I. Tiihonen); 0009-0008-6283-3637 (J. P. Fischer)

[1]https://data.bnf.fr/sparql/

integration of complementary skills, ranging from the humanities domain to information technologies and library science. Documenting each stage and making it accessible also improves transparency and builds a more sustainable and verifiable scholarly ecosystem. Addressing this matter, although the case study focuses on the executable workflow implementation for the production of the Early Modern bibliographic dataset from the BnF, the underlying workflow blueprint is based on the sequential execution of general steps that can be reused entirely or partially to replicate the experiment with new input material.

The study presented in this paper is part of a broader project conducted within the Helsinki Computational History Group (COMHIS), aimed at building a curated dataset of information extracted from the BnF on Early Modern publications, together with metadata on the agents responsible for their production and dissemination. Within this framework, the research pursues two complementary objectives: (1) to develop a general tool that guides the adoption of good practices, ensures methodological soundness, and provides operational support for historical and bibliographic research; and (2) to assess its effectiveness through a case study by implementing an executable and transparent workflow capable of generating a high-quality, consistent, and interoperable dataset that can be semantically queried and readily reused in entity-mapping projects.

This paper primarily focuses on the first objective and is structured as follows. After the introduction, Section 2 illustrates the context and motivations behind the work, while Section 3 presents the literature review and the state of the art, divided into subsections dedicated to: (3.1) Open Science and FAIR principles, (3.2) bibliographic data-related reproducible workflows, (3.3) human participation in Digital Humanities (DH) semi-automatic research pipelines, (3.4) and open tools for the execution and dissemination of FAIR methodologies. Section 4 describes the methodology adopted for the definition of the proposed workflow; Section 5 presents the blueprint and the executable workflow as results, including a graphical representation and the operational implementation documented on Protocols[2]; Section 6 is dedicated to the qualitative evaluation of the product; while Section 7 discusses limitations and future development, such as the ongoing activities for the case study output dissemination, including the publication of a portal in the Sampo-UI framework[3], as well as possible extensions to test the model on other datasets. Finally, Section 8 concludes the paper by summarising the main highlights of the project.

2. Context

Bibliographic metadata has become a valuable resource for the quantitative study of Early Modern history. Typically collected by libraries hosting old books, this information describes properties of Early Modern publications such as their formats, publication places, languages and other characteristics. Some collections have information about hundreds of thousands or even millions of books printed with the movable-type press. As such resources have increasingly become accessible in a digital and machine-readable format, they have enabled historians and social scientists to analyse questions related to the large-scale development of Early Modern societies, like the relationship between printing and economic development [3], formation of cultural canons [4], and the rise of vernacular languages in print culture [5]. This has also made bibliographic data science, the computational aspect of working on historical bibliographic metadata, a topic of research in itself. Harmonisation and validation of data are especially important aspects of this strand of research, as the underlying data is often incomplete, its resolution varies (*i.e.*, we might only know an approximate year of publication), and its original format was typically not intended for numerical analyses [6]. The attention that different datasets of bibliographic metadata have received has spread unevenly. While some collections like the English Short Title Catalogue (ESTC) and the Finnish Fennica have been harmonised extensively [6], others have not been processed similarly, making it very hard to utilise them in quantitative analyses. For example, no one has yet described a workflow to process the vast Early Modern collections of the French National Library. The lack of work done on important national collections, that are typically

[2] http://protocols.io

[3] https://seco.cs.aalto.fi/tools/sampo-ui/

less ambiguous in terms of their contents than transnational catalogues, naturally hinders the study of the related region (in this case, France). Nonetheless, it also limits our ability to study transnational processes of print culture like the cultural influence that France and Britain had on each other during the eighteenth century [7, 8], as the study of the question would require the comparative use of harmonised data from both the BnF and ESTC.

An important part of our motivation to extract and harmonise the BnF data is in the possibilities that it offers for comparative analyses. In our case, the combination of using it together with the ESTC will enable us to approach what is sometimes described as the French 'anglomania' and British 'francophilia' of the eighteenth century [7, 8] in a data-driven way. The ESTC is a union catalogue that aims to cover all books published in the Early Modern anglosphere. It contains almost 500,000 records, and it is the most comprehensive bibliographic resource available for studying books printed in Britain before the nineteenth century. Moreover, it has been extensively harmonised and enriched by COMHIS. In addition to the work done to make information regarding basic properties (number of pages, language, publication year, etc.) more suitable for quantitative analyses [6], the ESTC has been further enriched with information like a work-id that maps together different versions (*i.e.*, reprints) of the same text [9] and an actor field that identifies distinct individuals involved in the production of books [10]. By combining the results of our workflow for the BnF with the long-developed ESTC data, we can, by and large, link together authors that appear in both datasets and deduce their 'primary cultural context' (*i.e.*, British or French). This enables us to study who and what was translated and published beyond their immediate cultural domain and to better understand the patterns of cultural influence between the two countries during the Early Modern era. Hence, our workflow contributes towards the wider effort to build an ecosystem of datasets that enable the study of transnational cultural processes during the Early Modern period.

3. Literature Review

This paper's literature review is structured at multiple levels to provide a concise overview of the multidisciplinary state of the art and its key concepts, presented in dedicated subsections. Specifically, the topics covered include: (3.1) Open Science practices and FAIR principles; (3.2) bibliographic data workflows and FAIR principles; (3.3) human participation in semi-automatic research pipelines; and (3.4) open tools for implementing and disseminating FAIR methodologies.

3.1. Open Science Framework and FAIR Principles

In the definition provided by UNESCO, Open Science is an interdisciplinary and inclusive framework of movements and practices based on quality, integrity, fairness, sustainability, collective benefit, respect for diversity and inclusivity, aiming to create, evaluate, share, and communicate knowledge to all social actors, even beyond traditional academic boundaries [1]. Rather than focusing on results, this scientific philosophy focuses on research processes and the conditions in which they are carried out, encouraging transparency and good practices of critical re-adoption - either partial or total - designed to minimise the use of resources [11].

In the context of a more affordable and sustainable science, although not mandatory, it is logically consistent to adopt the FAIR principles. These emphasise the importance of making research data (as well as ancillary tools and outputs) not only available for human access, but also adequately manageable by machines, so that they can be understood, communicated and easily integrated across scientific research pipelines [2].

Nonetheless, integrating research work into this virtuous framework is not automatic: it requires awareness, planning, and active commitment to overcome obstacles posed by established scientific practices within the various academic fields. Despite the existence of barriers and costs perceived by researchers in sharing tools and data, in predominantly quantitative research fields the importance of adopting FAIR and open approaches is generally acknowledged [12]. Indeed, the reuse of freely accessible research products and methodologies in these disciplines has a tangible impact on accelerating

discoveries and scientific progress [13]. On the other hand, the same cannot be said for research in areas where qualitative aspects prevail over quantitative ones, and there is no single interpretation of good methodological practices or, first and foremost, of what constitutes the research data to which they should be applied [14].

A pivotal research area in the DH domain where the usefulness of conforming to the above-mentioned paradigms is generally recognised is software development for metadata management. Since these tools are useful in professional contexts where technical expertise may be relatively limited, reusing a properly documented codebase can enable results to be achieved in less time, allowing for more resources to be devoted to qualitative research.

Among the projects dedicated to raising awareness on this issue, we can mention the Open Source Initiative[4], a Californian non-profit corporation that works to promote the use of good distribution practices, development transparency, and peer review. With similar aims, rules for ethical and functional software development have been published [15], which encourage (1) avoiding reinventing the wheel; (2) following conventions for good systematic and organised development; (3) producing useful tools and empathising with potential users; (4) being transparent in development to encourage feedback; (5) opting, where possible, for simplicity; (6) avoiding waiting until the code is considered complete to release the first version and choosing to publish frequent - even suboptimal - releases; (7) fostering the growth of a community that uses the software; (8) investing in promoting the project, also paying attention to usability and aesthetics; (9) applying for funding for the project's growth and maintenance; (10) taking into account the processes that fuel academic scientific cycles, acknowledging that maintaining open source code requires resources and thus social collaboration is crucial. Accordingly, there is growing awareness that open-software development should be incentivised through scientific recognition, in line with traditional academic reward structures. This has led to initiatives such as the FORCE11 Software Citation Working Group to promote the practice of citing software as fully-fledged research products, in accordance with commonly recognised policies [16].

Compliance with these practices positively contributes to the possibility of making research reproducible and replicable. These two concepts are cornerstones of Open Science and are defined as (1) the possibility of confirming the results obtained in previous research by repeating the experiments under similar conditions, reusing the same input data, procedural steps, methods, and tools; and (2) the possibility of obtaining results compatible with those previously achieved using different data and/or methodologies, thereby confirming the reliability of the research previously carried out [17, 18]. Reproducibility should be considered a priority requirement over replicability because it does not involve the introduction of new variables with respect to the execution of previous experiments. However, this does not mean that it is easy or obvious to achieve, as reproducibility is closely related to issues of transparency and free access to research data, conditions that cannot always be met in contexts not devoted to Open Science. Furthermore, verifying research reproducibility and replicability requires considerable work, and an adequate system of reward and recognition for researchers carrying out such tasks is currently lacking [18].

In recent years, there has been a growing effort to produce workflows, which are recognised as crucial tools for pursuing the objectives of reproducibility and replicability in the Galleries, Libraries, Archives, and Museums (GLAM) fields. Accordingly, FAIR and Open Science-compliant pipelines have been released, spanning from data retrieval and harmonisation to data publishing. However, it has been demonstrated that, to ensure the effectiveness of the total or partial reuse of procedural steps, making workflows freely accessible is not enough: supplementing workflows with additional resources - including descriptive annotations, testing datasets, and traces of previous executions - is also necessary [19]. To facilitate the effective reuse of research blueprints, guidelines have been proposed for applying the FAIR principles to these tools. Indeed, although the principles were initially developed for data management, they proved effectively applicable to a broad set of research products and tools [13].

[4] https://opensource.org/osd

3.2. Reproducible Workflows in the Bibliographic Domain

The bibliographic domain is characterised by recurring formalisation tasks, which mainly concern the harmonisation of a defined set of fields (*i.e.*, names of entities), the validation of persistent identifiers, the normalisation of dates, the exact attribution of locations, the deduplication of entities, and translation mapping tasks. For this reason, the open distribution of workflows allows for effective comparisons between alternative methodologies depending on the requirements of the data source.

Among the initiatives dedicated to this topic, the Bibliographical Data Working Group of the Digital Research Infrastructure for the Arts and Humanities in Europe (DARIAH-EU) brought together representatives from various European research groups for a workshop meeting to develop workflows for processing bibliographic metadata. The workflows developed were published in the Social Sciences & Humanities (SSH) Open Marketplace, a platform that collects research resources. Following this meeting, a collective article was written, focusing on the challenge of multilingual management of bibliographic datasets [20]. On the same topic, OpenCitations, an open scholarly infrastructure dedicated to citation and bibliographic data, has documented the harmonisation process and metadata crosswalk towards the OpenCitations Data Model for ingesting citation data from the Japan Link Centre[5] into its database [21]. Other significant challenges in the development of workflows for bibliographic data management include the metadata crosswalk between different data models for structuring information into LOD. In the `datos.bne.es` project, this task was addressed with an iterative methodology for extracting Linked Library Data (LLD) from MARC 21 records, supported by the MARiMbA tool, designed to facilitate the involvement of domain experts. Testing the workflow highlighted the importance of collaboration between Semantic Web developers and library experts, as it ensures better mapping quality and contributes to the establishment of a sustainable and scalable approach [22].

As for the studies explicitly dedicated to historical bibliographic metadata, and therefore in line with the purpose of this article, we can mention the development of work pipelines to transform national library data into datasets for quantitative analysis [23]. This research explicitly mentions biases, a lack of conventions and rigour in data collection, and obvious inaccuracies, which significantly undermine the possibility of actual reuse of the data. Accordingly, best practices were defined to make historical bibliographic metadata interoperable, so as to improve quality, consistency, and completeness, and ultimately enhance the potential for quantitative analysis on top of it [6].

3.3. Human-in-the-Loop in Humanities Workflows

The evolution of digital tools for managing Cultural Heritage (CH) catalogue metadata has greatly contributed to the formalisation of knowledge, making it possible to apply semi-automatic statistical and computational approaches to data management. However, the prospect of making GLAM metadata digitisation processes fully automated is unrealistic. In fact, the entropy introduced by the manual collection of unstructured information in the past decades and centuries to the present day often causes the necessity of human oversight of the digitisation and conversion processes, constituting intermediate steps of iterative workflows, in particular for tasks such as data model alignment, metadata curation, and data validation. The involvement of human operators in the execution of research methodologies and pipelines, known by various names in different research contexts and referred to here as *human-in-the-loop* [24], mitigates the impossibility of performing specific tasks in a fully automated manner, or improves their performance through supervision, feedback or direct intervention, to produce verified-quality data.

Methods of human intervention are diverse and depend on their purpose, ranging from annotation, targeting, metadata curation and enrichment, data validation and quality control, semantic mapping, classification, and knowledge extraction. More specifically, these approaches can be distinguished according to whether the task requires expert intervention (expert-driven) or a contribution from a general user base (crowd-driven). An example of the expert-driven approach is provided by Bernasconi *et al.* [25], who describe a system in which experts collaborate to validate information previously extracted

[5] https://japanlinkcenter.org/top/english.html

automatically from a Digital Library (DL). On the other hand, projects such as CrowdHeritage [26] offer examples of non-sectoral human intervention, in which semantic enrichment of metadata occurs through crowdsourcing campaigns supported by automatic annotation tools and user validation. A complementary contribution is provided by Kaldeli *et al.* [27], who propose a methodology for semantic enrichment of CH metadata within the Europeana Data Space, integrating automatic text analysis techniques with human validation and differentiation according to expertise level.

Another key point is how human decision-making authority is allocated, and whether it is centralised around specific people or groups, like in cases where the intervention is traceable and structured within a controlled flow [28], or delegated to end users, possibly relying on data such as a user reliability score to appropriately weight the contribution [29]. In most cases, crowdsourcing involves professionals who perform quality control functions, while decision-making processes are delegated to end users. In most cases, it can be stated that a hybrid approach integrating human feedback, supervision, or direct involvement outperforms a fully automated process in terms of efficiency, especially when implemented in an iterative and multidisciplinary manner [28]. Furthermore, approaches such as the one developed by Santoro *et al.* [29] show how human interaction can also progressively improve the performance of automated systems, for example, in the recognition of handwritten text, where user validation of words is used to train and refine the model over time. Similarly, Shabani *et al.* [30] demonstrate that combining neural networks and annotations from users with weighted reliability ratings produces better results than either fully automated or fully human-dependent models.

These examples contribute to proving that the human-in-the-loop approach is key to improving quality, transparency, and traceability in the workflows for the creation, enrichment, and validation of digital CH data.

3.4. Open Tools for FAIR Workflow Execution and Dissemination

To enable the effective and efficient reuse of workflows as research instruments, it is necessary to share them through tools that maximise their comprehension, clearly separating the stages and explaining the interaction between the various procedural components, including the management of inputs and outputs for each stage. Although not guaranteed by all tools available for workflow management in the humanities or CH domains, a relevant feature is *actionability*, *i.e.*, the ability to execute the various steps within the tool itself, generally being supported by sample data.

Among the tools and infrastructures for managing and disseminating FAIR workflows, WorkflowHub[6] is an open-source registry for publishing, describing, and reusing FAIR workflows. It is based on the Workflow-RO-Crate [31] standard and can be integrated with execution environments such as Galaxy and CWL. As an alternative, the aforementioned Protocols is a collaborative platform for interactively creating, sharing, and running research protocols, although internal software code execution functionalities are not provided. It enables the recording of inputs, outputs, and deviations, promoting human-in-the-loop reproducibility. Similarly, the SSH Open Marketplace[7], a catalogue curated by DARIAH and CLARIN, gathers tools, services, and workflows for the humanities and social sciences, providing FAIR metadata and semantic links to resources. One of the strengths of the SSH Open Marketplace is the enhanced possibility of interconnection among the resources, so that, in the case of workflows, input data, output datasets, released software components, and any other relevant tool or research product can be adequately connected. In order to guarantee open access, metadata management, findability, and a good versioning system, Zenodo is a reliable and popular platform for any type of research product hosting and publication, including workflows externally developed with dedicated tools. Thus, its use is often taken into account for disseminating products developed elsewhere.

Moving to the workflow standards and execution engines, RO-Crate[8] is a JSON-LD standard designed to package data, software, and workflows together with their FAIR metadata, thereby enabling portability

[6] https://workflowhub.eu
[7] https://marketplace.sshopencloud.eu
[8] https://www.researchobject.org/ro-crate

and re-execution in compatible environments [31]. On this standard is based ROHub[9], a repository for Research Objects (RO) that enables versioning, semantic linkage, and interoperability with WorkflowHub. Another popular standard is CWL (Common Workflow Language[10]), which is an interoperable standard language for describing reusable and executable workflows in different environments (Galaxy, Nextflow, CWLtool) [32].

As for the execution engines, Nextflow is a portable product designed for scientific workflows, compatible with CWL and containers such as Docker, which ensures reproducibility and actionability across different infrastructures [33]. As an option, Galaxy is a popular open-source platform for building, executing, and sharing scientific workflows[11], which nonetheless can also be adopted in digital humanities and CH contexts [34]. Finally, one of the most popular actionable tools for defining, running and sharing executable workflows is Jupyter Notebook[12], a fully interactive environment for documentation, execution and sharing of code and data, useful for creating readable and reusable computational workflows [35].

4. Methodology

As the main research output presented in this paper is the workflow blueprint, the Methodology section addresses the conceptual steps to plan and design a procedure leveraging bibliographic data from trusted public sources, such as those provided by National DLs, supporting the work of interdisciplinary research groups in the construction of harmonised datasets, intended both for academic use (analysis, mapping, datasets comparisons) and for LOD exploration by technically inexperienced audiences. In terms of structure, the methodology was developed in several phases, corresponding to the main steps in workflow design: (1) defining the overall purpose of the workflow, (2) identifying the actors involved, (3) defining the minimum outputs according to the target, (4) choosing the ethical, legal and conceptual framework, (5) integrating reliability and control components, (6) contextualising the human supervision and feedback cycle.

4.1. Workflow Purpose

As the first methodological step, the general purpose of the workflow's development was identified: to enhance the production of harmonised thematic bibliographic datasets that can be easily mapped with other resources and analysed in research contexts, starting from open and institutional data as input material. More in detail, the workflow was designed to facilitate three dimensions of use: (a) mappability with external datasets, (b) statistical analysability of harmonised data, (c) semantic explorability of the result, even by non-specialist users.

4.2. Involved Actors Identification

The workflow was designed to be orchestrated by multidisciplinary research groups involved in GLAM digitisation projects. Collaborators are expected to have different backgrounds and, collectively, to address diverse tasks with a varied skill set. In detail, domain specialists can contribute background knowledge and evaluate the informational content of the data, while metadata experts handle tasks related to the correct formalisation of data entries. Digital humanists cover hybrid roles and generally bridge between professionals with expertise in CH and in digital technologies. Since the blueprint is purpose-oriented and general, it allows executable implementations to be adapted to the specific context and reflect the actual composition of the research team, ensuring the operational sustainability of the process.

[9] https://www.rohub.org
[10] https://www.commonwl.org
[11] https://galaxyproject.org
[12] https://jupyter.org

4.3. Minimum Output Definition Based on the Target

Under the banner of FAIRness, the workflow aims to facilitate the production of highly usable outputs for both expert users and the general public. Consequently, two main outputs were identified: (a) a global dataset (provided both in tabular format and as Linked Open Data in Resource Description Framework - RDF), accompanied by ancillary data related to the workflow execution; and (b) an interface for exploring and querying semantic data, designed for discovery and semantic exploration by non-expert users.

4.4. Choice of Ethical, Legal and Conceptual Framework

The workflow was developed in accordance with FAIR principles and Open Science values, ensuring transparency and reproducibility of results. In selecting technologies and software components, compliance with open-source licences and regulations governing data from public entities was also considered to support modular adaptation across different case studies. Indeed, different sources may impose different data-reuse requirements (*e.g.*, the BnF reuse conditions declaration[13]), and the blueprint should allow users to adapt the executable workflow implementation to the conditions set by their data source.

4.5. Integration of Reliability and Control Components

To increase traceability, validity, and reproducibility, the workflow blueprint was iteratively refined to add components aimed at: (a) computational cost monitoring; (b) automated code and intermediate data validation testing; (c) scheduling of manual sampling phases and verification of harmonised data to ensure variety and consistency of values; (d) generation of intermediate outputs to promote explainability and transparency of the workflow.

4.6. Human Supervision and Feedback Loop

The definition of the workflow blueprint was strongly shaped by an iterative methodology, involving testing on the BnF Early Modern workflow implementation and moving back and forth between the case study and the blueprint, formalising lessons learnt into good practices.

5. Results

The result of this research is twofold: first, a workflow blueprint was developed and published as a Protocols workflow [36], providing reusable guidelines for the production and publication of harmonised Open and FAIR datasets across projects with different input data. Second, an executable workflow tailored to the BnF Early Modern dataset was formalised (Figure 1), supporting future updates and enabling the reuse of individual components in other bibliographic data-related contexts. The blueprint constitutes a transferable operating model independent of specific technologies, formats, or venues, while the BnF workflow represents its first concrete implementation and experimental validation. Across the blueprint stages, monitoring, evaluation, and testing are treated as quality-control sub-steps, implemented in the executable workflow through dedicated scripts.

The workflow blueprint, formalised using Protocols [36], is articulated into general stages, without mandating specific technologies, formats, or publication venues. Each main data production, harmonisation, and optimisation stage is accompanied by quality-control activities (monitoring, evaluation, and testing), which are specified as recommended sub-steps in the blueprint and implemented as dedicated scripts within the executable workflow. Ancillary comparison and sampling stages produce diagnostic outputs to be inspected by a human operator and used to guide subsequent decisions. The blueprint stages are: (1) preliminary input strategy comparison and baseline checks, which may either support input selection when starting from scratch or verify consistency with previous releases when

[13] https://www.bnf.fr/fr/conditions-de-reutilisations-des-donnees-de-la-bnf

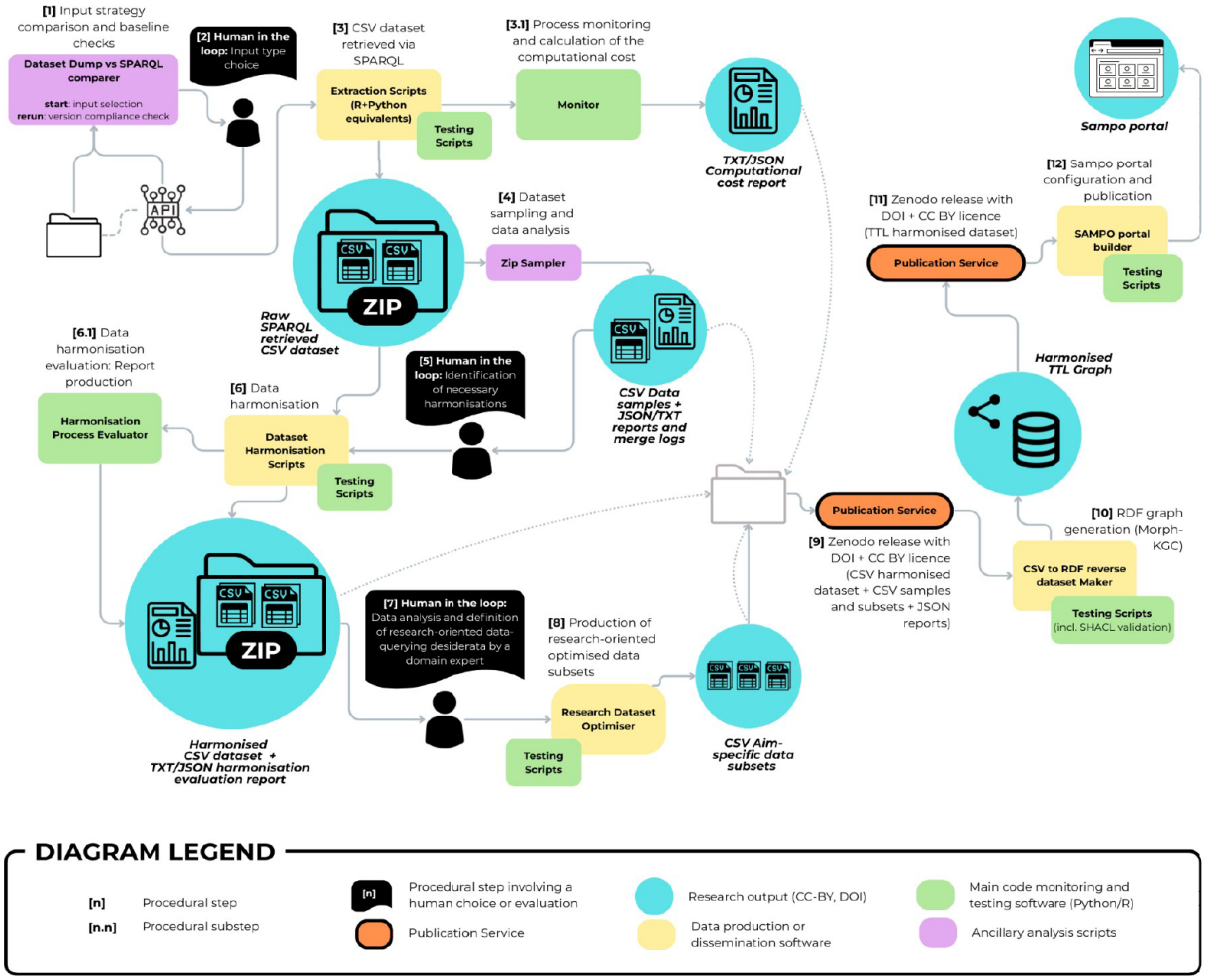

Figure 1: Visual representation of the executable workflow for producing the Early Modern BnF dataset, including a legend of diagram components.

re-running the workflow; (2) a human decision on the selected input/access strategy; (3) data extraction and production of an analysable intermediate dataset, accompanied by monitoring and automated tests; (4) sampling and preparatory analysis to produce diagnostic outputs for human inspection; (5) identification of harmonisation needs based on sample inspection; (6) harmonisation and cleaning, accompanied by evaluation and tests; (7) definition of research-oriented requirements and querying desiderata; (8) production of purpose-oriented optimised subsets together with traceability artefacts; (9) publication of the resulting tabular datasets and complementary materials with persistent identifiers under an open licence (*e.g.*, CC BY 4.0); (10) RDF graph generation from curated data, accompanied by validation and testing; (11) publication of the RDF serialisation (*e.g.*, TTL) to provide a FAIR, citable release; and (12) configuration and publication of a semantic access layer for data dissemination (*e.g.*, a portal).

The executable workflow[14] implements the blueprint on the BnF Early Modern case study (Figure 1[15]); the underlying software is available on GitHub[16]. Extraction is performed via SPARQL through R/Python scripts, while monitoring, evaluation, and testing are implemented as dedicated scripts and modules (monitor, harmonisation evaluator, testing scripts) associated with the main steps. Comparison and sampling are implemented in Python and generate outputs inspected by a human operator at human-

[14] https://www.protocols.io/view/early-modern-comhis-bnf-derived-dataset-production-kqdg31n5zl25
[15] https://github.com/ariannamorettj/Early_Modern_BnF_Harmonisation/blob/352d4471b2d4de946b756ec3da276d4205712c83/docs/bnf_exe_workflow.png
[16] https://github.com/ariannamorettj/Early_Modern_BnF_Harmonisation

in-the-loop decision points. All research outputs produced throughout the workflow (blue nodes in Figure 1) are released on Zenodo under the CC BY 4.0 licence with a DOI, together with complementary materials and documentation. In practical terms, first of all, (1) input strategy comparison and baseline checks are performed. Then, (2) a human operator selects the input/access strategy; in our case, SPARQL was preferred over data dumps to preserve control over extraction logic and updates. (3) Data extraction via SPARQL produces a raw CSV dataset (ZIP) through extraction scripts, accompanied by automated tests and by (3.1) process monitoring and computational-cost reporting (TXT/JSON). (4) A sampling step produces CSV data samples and JSON/TXT reports for exploratory inspection, which inform (5) the human identification of harmonisation needs (*e.g.*, uncertain dates, entity names consistency, potential duplicates). (6) Harmonisation is executed via dedicated scripts and accompanied by (6.1) a harmonisation evaluation report (TXT/JSON) and automated tests, producing a harmonised CSV dataset (ZIP). (7) A domain expert defines research-oriented requirements and querying desiderata, and (8) mapping-oriented, aim-specific CSV data subsets are produced through a research dataset optimiser. (9) A first Zenodo release is created, collecting the harmonised CSV dataset, CSV samples and research-oriented subsets, and the associated JSON reports, under CC BY 4.0 licence with a DOI. (10) An RDF graph is generated from the curated tabular data using Morph-KGC[17] via a CSV-to-RDF conversion module and tested through the workflow testing scripts (including SHACL validation). (11) The harmonised RDF is released on Zenodo as a dedicated record in TTL format under CC BY 4.0 with a DOI. (12) Finally, a Sampo-UI [37] portal is configured and published on top of the released RDF dataset, supported by automated tests.

6. Evaluation

At this stage of the research, it is not yet possible to conduct a comprehensive empirical evaluation of the final dataset obtained by executing the workflow implemented for the BnF data, as the pipeline is still being refined and has been tested only on limited data samples. However, the work has been structured so that workflow execution integrates checks to verify the local correctness of the results. The main components are covered by unit tests to ensure predictable behaviour for inputs with known characteristics, and automatic monitoring modules generate execution reports, useful for detecting technical anomalies. Similarly, the sampling and human-control steps integrate automated checks with expert critical analysis. Consequently, the evaluation here focuses more on the qualitative compliance of the workflow blueprint and its implementation with FAIR principles, which are the process requirements for producing FAIR datasets once production is complete.

1. **Findability.** The workflow is designed to release research results (harmonised datasets, optimised subsets, reports, merge logs and complementary materials) as versioned and documented artefacts. The adoption of recommended publication services, such as Zenodo, enables the assignment of versioned DOIs to datasets and the selection of an open licence. This ensures that each release is individually traceable and citable, facilitating comparisons across executions and subsequent versions. The blueprint itself, conceived as a research output, is available on Protocols, where it is released under CC BY 4.0 and identified by a versioned DOI.

2. **Accessibility.** The workflow blueprint involves three steps for releasing datasets and supporting materials (in tabular and semantic formats), including a semantic access level (portal) that facilitates consultation and data exploration. In this way, the workflow clearly separates the publication of data from its dissemination, reducing access barriers for non-technical users.

3. **Interoperability.** The entire research project aims to facilitate the conversion of bibliographic metadata into RDF, a format that enhances interoperability by enabling machine-readable interconnections among datasets. To make this feasible, a validation step, as suggested by SHACL, has been added as a quality-control measure for semantic data. Interoperability is therefore treated as an

[17] https://morph-kgc.readthedocs.io/en/stable/

operational requirement of the process, supported by a machine-readable representation and formal constraints that enhance consistency and compatibility with Semantic Web tools and practices.

4. **Reusability.** To maximise reusability, the blueprint workflow steps have been isolated in Protocols and complemented with examples, linking the blueprint steps to the GitHub BnF implementation. The adoption of an open licence fosters the possibility of reusing the tool in new research projects.

Whilst the contribution that can be assessed at this stage is the robustness of the workflow as a FAIR-oriented production tool, an empirical evaluation with quantitative metrics - to assess data accuracy, impact of performed harmonisation, and stability compared to previous versions - will be released following the official publication of the harmonised BnF Early Modern bibliographic dataset, thus completing the research project.

7. Limitations and Future Development

The experimental implementation of the blueprint was conducted using bibliographic data on the Early Modern period retrieved from the BnF SPARQL endpoint. The work described in this paper should be considered preliminary and complementary to the final production of the dataset, which will enable the integration of qualitative evaluation with quantitative data extracted from a complete case study. In the next steps of the research, the workflow will be re-run in its entirety to produce and publish a first complete version of the primary dataset (including information about bibliographic entities and responsible agents), systematically including materials supporting the transparency of the process, such as execution reports, quality controls, and traceability of merge operations in the deduplication processes. Within the Open Science framework in which this research is being developed, it is both predictable and desirable that the testing phases will highlight weaknesses and potential areas for development in the workflow blueprint, which will be followed by adjustments and updates to be incorporated into subsequent published versions.

A future development related to the implementation of the executable workflow in the case study concerns the possibility of conducting an empirical test of repeatability under varying conditions, such as different network connectivity, different computing resources, and updated input data. Since the code that executes the workflow automatically generates reports on performance and execution characteristics, such re-executions will enable comparisons between runs and more solid evidence of reproducibility and robustness.

The scalability of the blueprint layer will then be tested by implementing new executable workflows designed to manage other input datasets. Given that many of the operations managed by the software implemented for the case study are common to the management of several bibliographic datasets - such as harmonisation of personal names, professions and roles, places and dates; deduplication tasks; production of task-specific subsets - the next steps in the research also include the integration of already developed partial components into new executable workflows, to test the effectiveness of the modules in new scenarios.

8. Conclusions

This paper presented a workflow blueprint to manage open-source extraction, analysis, harmonisation, publication, and LOD dissemination of bibliographic data and associated responsible agents information, relating to a specific historical period. The data used to implement an executable workflow to test the blueprint came from the BnF and relates to the Early Modern period. The project was developed within the Open Science framework and in accordance with the FAIR principles.

The purpose behind the research was twofold: on the one hand, the development of a general tool providing guidance in the adoption of good research practices, ensuring methodological soundness to the projects that adopt it; on the other hand, testing its effectiveness in developing a transparent executable workflow in the case study. The tasks supported by the blueprint include: (1) tracking

activities performed on the original dataset to exercise greater control over the final product quality; (2) increasing the reliability of the research process; (3) enabling external feedback to be gathered on easily identifiable and isolatable steps to improve performance; and thus (4) accelerating the research process. More specifically, the heterogeneous nature of bibliographic data implies the need to manage a wide range of formalisations and conventions, especially when it comes from historical periods in which the quality and structure of catalogue information vary significantly even within the same collection. One of the project's contributions was therefore the development of a tool capable of proposing a robust, structured approach to address this variety without limiting the freedom of implementation management within specific executable workflows. The blueprint also encourages the structured organisation of human intervention throughout the scholarly pipelines. This distinctive feature makes the blueprint particularly suitable for humanities research contexts, where the adoption of new technologies cannot be separated from the intervention of domain experts, who remain crucial in steps that require qualitative choices among alternative strategies, supervision functions, feedback and dialogue, and manual analysis of unstructured information.

Acknowledgments

This work was partially funded by Project PE 0000020 CHANGES - CUP B53C22003780006, NRP Mission 4 Component 2 Investment 1.3, funded by the European Union - NextGenerationEU, and also received funding from the Horizon Europe Program for Research and Innovation under MSCA Doctoral Networks 2022, Grant Agreement No. 101120349. This project has received funding from the Finnish Cultural Foundation. This article is part of the ongoing efforts of the Helsinki Computational History Group (COMHIS) to harmonise and enrich historical bibliographic metadata for research purposes.

Declaration on Generative AI

According to the GenAI Usage Taxonomy, AI was ethically adopted to support Text Translation (DeepL) and Paraphrase and reword (ChatGPT - GPT-5, Grammarly).

References

[1] UNESCO, UNESCO Recommendation on Open Science, 2021. doi:10.54677/MNMH8546.

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[3] J. Dittmar, Information Technology and Economic Change: The Impact of the Printing Press, The Quarterly Journal of Economics 126 (2011) 1133–1172. doi:10.1093/qje/qjr002.

[4] M. Tolonen, M. J. Hill, A. Z. Ijaz, V. Vaara, L. Lahti, Data Visualization in Enlightenment Literature and Culture, in: I. Baird (Ed.), The Enlightenment in Literature and Culture, Springer International Publishing, 2021.

[5] J. Marjanen, T. Tahko, L. Lahti, M. Tolonen, Book Printing in Latin and Vernacular Languages in Northern Europe, 1500–1800, in: J.-M. Hanssen, S. Furuseth (Eds.), The Hermeneutics of Bibliographic Data and Cultural Metadata, Notabene, National Library of Norway, 2025.

[6] L. Lahti, J. Marjanen, H. Roivainen, M. Tolonen, Bibliographic Data Science and the History of the Book (c. 1500–1800), Cataloging & Classification Quarterly 57 (2019) 5–23. doi:10.1080/01639374.2018.1543747.

[7] R. Eagles, Francophilia in English Society, 1748–1815, Macmillan Press; St. Martin's Press, 2000.

[8] J. Grieder, Anglomania in France, 1740–1789: Fact, Fiction, and Political Discourse, Librairie Droz, 1985.

[9] A. Z. Ijaz, M. Tolonen, L. Lahti, I. Tiihonen, Analytical Determination of Editions from Bibliographic Metadata, in: H. Jantunen, S. Brunni, N. Kunnas, S. Palviainen, K. Västi (Eds.), Proceedings of

the Research Data and Humanities (RDHUM) 2019 Conference, Studia Humaniora Ouluensia, University of Oulu, 2019.
[10] M. J. Hill, V. Vaara, T. Säily, L. Lahti, M. Tolonen, Reconstructing Intellectual Networks: From the ESTC's Bibliographic Metadata to Historical Material, in: Digital Humanities in the Nordic Countries, CEUR Workshop Proceedings, CEUR-WS.org, 2019.
[11] S. Leonelli, Philosophy of Open Science, Cambridge University Press, 2023. doi:10.1017/9781009416368.
[12] D. G. E. Gomes, P. Pottier, R. Crystal-Ornelas, et al., Why Don't We Share Data and Code? Perceived Barriers and Benefits to Public Archiving Practices, Proceedings of the Royal Society B: Biological Sciences 289 (2022) 20221113. doi:10.1098/rspb.2022.1113.
[13] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, et al., FAIR Computational Workflows, Data Intelligence 2 (2024) 108–121. doi:10.1162/dint_a_00033.
[14] B. Gualandi, L. Pareschi, S. Peroni, What Do We Mean by "Data"? A Proposed Classification of Data Types in the Arts and Humanities, Journal of Documentation 79 (2022) 51–71. doi:10.1108/JD-07-2022-0146.
[15] A. Prlić, J. B. Procter, Ten Simple Rules for the Open Development of Scientific Software, PLOS Computational Biology 8 (2012) e1002802. doi:10.1371/journal.pcbi.1002802.
[16] A. M. Smith, D. S. Katz, K. E. Niemeyer, FORCE11 Software Citation Working Group, Software Citation Principles, PeerJ Computer Science 2 (2016) e86. doi:10.7717/peerj-cs.86.
[17] Committee on Reproducibility and Replicability in Science, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, et al., Reproducibility and Replicability in Science, National Academies Press, 2019. doi:10.17226/25303.
[18] X.-L. Meng, Reproducibility, Replicability, and Reliability, Harvard Data Science Review 2 (2020). doi:10.1162/99608f92.dbfce7f9.
[19] K. Belhajjame, J. Zhao, D. Garijo, et al., Using a Suite of Ontologies for Preserving Workflow-Centric Research Objects, Journal of Web Semantics 32 (2015) 16–42. doi:10.1016/j.websem.2015.01.003.
[20] V. Malínek, T. Umerle, E. Gray, et al., Open Bibliographical Data Workflows and the Multilinguality Challenge, Journal of Open Humanities Data 10 (2024) 27. doi:10.5334/johd.190.
[21] A. Moretti, M. Soricetti, I. Heibi, A. Massari, S. Peroni, E. Rizzetto, The Integration of the Japan Link Center's Bibliographic Data into OpenCitations, Journal of Open Humanities Data 10 (2024) 21. doi:10.5334/johd.178.
[22] D. Vila-Suero, A. Gómez-Pérez, Datos.Bne.Es and MARiMbA: An Insight into Library Linked Data, Library Hi Tech 31 (2013) 575–601. doi:10.1108/LHT-03-2013-0031.
[23] M. Koho, T. Burrows, E. Hyvönen, et al., Harmonizing and Publishing Heterogeneous Premodern Manuscript Metadata as Linked Open Data, Journal of the Association for Information Science and Technology 73 (2022) 240–257. doi:10.1002/asi.24499.
[24] M. Anderson, K. Fort, Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint, The International Review of Information Ethics 31 (2022). doi:10.29173/irie477.
[25] E. Bernasconi, M. Ceriani, M. Mecella, A. Morvillo, Automatic Knowledge Extraction from a Digital Library and Collaborative Validation, in: Linking Theory and Practice of Digital Libraries: 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022), Padua, Italy, 2022, pp. 480–484. doi:10.1007/978-3-031-16802-4_49.
[26] E. Kaldeli, O. Menis-Mastromichalakis, S. Bekiaris, et al., CrowdHeritage: Crowdsourcing for Improving the Quality of Cultural Heritage Metadata, Information 12 (2021). doi:10.3390/info12020064.
[27] E. Kaldeli, A. Chortaras, V. Lyberatos, J. Liartis, S. Kantarelis, G. Stamou, Combining Automatic Annotation with Human Validation for the Semantic Enrichment of Cultural Heritage Metadata, in: W. Haverals, M. Koolen, L. Thompson (Eds.), Proceedings of the Computational Humanities Research Conference 2024 (CHR 2024), volume 3834 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3834/.

[28] L. Lahti, V. Vaara, J. Marjanen, M. Tolonen, Studying Digital Humanities with Research Data: Perspectives from Finland, in: J. H. Jantunen, S. Brunni, N. Kunnas, S. Palviainen, K. Västi (Eds.), Proceedings of the Research Data and Humanities (RDHUM) 2019 Conference: Data, Methods and Tools, volume 17 of *Studia Humaniora Ouluensia*, University of Oulu, 2019. URL: http://urn.fi/urn:isbn:9789526223216.

[29] A. Santoro, A. Parziale, A. Marcelli, A Human in the Loop Approach to Historical Handwritten Documents Transcription, in: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016, pp. 222–227. doi:10.1109/ICFHR.2016.0051.

[30] S. Shabani, M. Sokhn, H. Schuldt, Hybrid Human-Machine Classification System for Cultural Heritage Data, in: Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC'20), ACM, New York, NY, USA, 2020, pp. 49–56. doi:10.1145/3423323.3423413.

[31] S. Soiland-Reyes, P. Sefton, M. Crosas, et al., Packaging Research Artefacts with RO-Crate, Data Science 5 (2022) 97–138. doi:10.3233/DS-210053.

[32] M. R. Crusoe, S. Abeln, A. Iosup, et al., Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language, Communications of the ACM 65 (2022) 54–63. doi:10.1145/3486897.

[33] P. Di Tommaso, M. Chatzou, E. W. Floden, P. Prieto Barja, E. Palumbo, C. Notredame, Nextflow Enables Reproducible Computational Workflows, Nature Biotechnology 35 (2017) 316–319. doi:10.1038/nbt.3820.

[34] Galaxy Community, The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2022 Update, Nucleic Acids Research 50 (2022) W345–W351. doi:10.1093/nar/gkac247.

[35] M. Beg, J. Taka, T. Kluyver, et al., Using Jupyter for Reproducible Scientific Workflows, Computing in Science & Engineering 23 (2021) 36–46. doi:10.1109/MCSE.2021.3052101.

[36] A. Moretti, I. L. I. Tiihonen, J. P. Fischer, FAIR Workflow Blueprint for Open Bibliographic Data Harmonisation and Dissemination, 2026. doi:10.17504/protocols.io.81wgbo97qlpk/v1, preprint.

[37] E. Hyvönen, P. Leskinen, A. Ahola, H. Rantala, J. Tuominen, SampoSampo: A Portal for Studying Enriched Data and Semantic Connections on a Cultural Heritage Linked Open Data Cloud, in: E. Curry, V. Presutti, J. McCrae, et al. (Eds.), The Semantic Web: ESWC 2025 Satellite Events, Springer Nature Switzerland, 2026. doi:10.1007/978-3-031-99554-5_13.