Dec 12, 2025 Version 1

# Early Modern COMHIS BnF-Derived Dataset Production Workflow V.1

DOI
https://dx.doi.org/10.17504/protocols.io.kqdg31n5zl25/v1

Arianna Moretti[1], iiro.tiihonen[2], jonasp.jf[2]

[1]University of Bologna (FICLIT), University of Helsinki (Computational History Group);
[2]University of Helsinki (Computational History Group)

Arianna Moretti
University of Bologna





**DOI:** https://dx.doi.org/10.17504/protocols.io.kqdg31n5zl25/v1

**Protocol Citation:** Arianna Moretti, iiro.tiihonen , jonasp.jf 2025. Early Modern COMHIS BnF-Derived Dataset Production Workflow. **protocols.io** https://dx.doi.org/10.17504/protocols.io.kqdg31n5zl25/v1



**Protocol status:** In development
**We are still developing and optimizing this protocol**

**Created:** October 29, 2025

**Last Modified:** December 12, 2025

**Protocol Integer ID:** 231044

**Keywords:** dataset production workflow the early modern comhis bnf, dataset production, derived dataset production workflow, bibliographic dataset, derived dataset production, oriented bibliographic dataset, bibliographic data, data harmonisation workflow, interoperability throughout the entire data lifecycle, entire data lifecycle, dataset, compliant workflow for the production, rdf, suitable for exploratory analysis, compliant workflow, csv, bnf, bnf sparql service, data, early modern comhis bnf, exploratory analysis, workflow, replicability by the scientific community, interactive exploration of entity, bibliothèque nationale, reproducibility, entity, semantic consultation, publication, scientific community, collection to cleaning, multiple format

**Funders Acknowledgements:**
**European Union – NextGenerationEU, NRP Mission 4 Component 2 Investment 1.3, CHANGES**
Grant ID: CUP B53C22003780006
**European Union – Horizon Europe Programme, MSCA Doctoral Networks 2022, MECANO**
Grant ID: 101120349

## Abstract

The ***Early Modern COMHIS BnF-Derived Dataset Production Workflow"***is a FAIR and Open Science principle-compliant workflow for the production of a dataset derived from Bibliothèque nationale de France (BnF) data relating to the Early Modern period. The goal is to ensure transparency, reproducibility, tracking, and interoperability throughout the entire data lifecycle, from collection to cleaning, deduplication, publication, and reuse. The workflow is documented and made available on protocols.io to facilitate validation and replicability by the scientific community.
The dataset includes controlled, cleaned, and deduplicated bibliographic data, produced in accordance with FAIR principles to ensure long-term accessibility, findability, and reusability. To support different research contexts and disciplinary uses, the data will be distributed in multiple formats, machine-readable (CSV, JSON-LD, RDF) and suitable for exploratory analysis (CSV). The release will be under a CC0 license on Zenodo, while a SAMPO-based portal will allow semantic consultation and interactive exploration of entities.
This workflow is described in the article ***Introducing a data harmonisation workflow exploiting the BNF Sparql service to produce and disseminate a research-oriented bibliographic dataset concerning the Early Modern period*** by Arianna Moretti, Iiro L. I. Tiihonen and Jonas P. Fischer, submitted to the conference IRCDL 2026.

## SPARQL Data Retrieval

1 Data retrieval from BnF SPARQL endpoint.
Production of the RAW Datasets for Bibliographic entities and responsible agents.

The dataset is produced by retrieving editions first and then the actor information for these editions. This approach was preferred to the direct querying for actors with birth or death dates within the timeframe in question, succeeding in eliminating actors from the dataset that lived within the investigated timeframe but did not have their works published within that timeframe.

1.1 **Bibliographic Entities Retrieval**
Folder that contains the latest edition data and the script for querying it:

https://github.com/ariannamorettj/New-BnF-Data-Analysis-2/tree/5aa8bb4dba5c293ed58bccaf6683b50bdf974716/data/bnf_edition_data

1.2 **Responsible Agents Data Retrieval**
So here is the structure of the final version of the extraction step. No integration is needed, because all actor information comes from here:
https://github.com/ariannamorettj/New-BnF-Data-Analysis-2/tree/main/data/bnf_agents_data_querying
The script in the folder is a bit misleadingly named query_missing_agents.R, but it is responsible for querying all of the information about agents. It is based on the agents mentioned in some role in the edition data. **actor_queries_results.zip** should have all the agent-related information that we need.

[To do in future: Integrate a quantification of the loss among the two approaches, to investigate potential loss due to the chosen approach]

## Testing: SPARQL Data Retrieval

2 Run tests and verify that the query retrieves the expected data. This step is useful in terms of verifying that the structure of the SPARQL endpoint is consistent with that of the previous workflow iteration.

**In case tests are not passed, go back to refine step [1] and run the tests iteratively until they are passed.**

## Data Analysis - Phase 1 (identification of values to be harmonised)

3 Analysis of the raw data to find values to be harmonised:

3.1 **Script Run (customised)**
Execute scripts to retrieve values.

3.2 **Human analysis + identification of values to be harmonised**
Manual check of the inconsistencies in values collection in the following fields
- places
- names
- dates (with uncertainty degree)
- possible duplicates mapping

## Data Harmonisation

4 Execute the data cleaning

## Data Harmonisation

4.1 **Run previously developed harmonising scripts**

Run cleaning scripts concerning the management of the values about
- places
- names
- dates (with uncertainty degree)
- possible duplicates mapping

4.2 **Code extensiom (optional, if needed)**
In case the harmonising scripts provided do not cover new cases that emerged in step 2, expand the codebase to address the

## Testing: Data Harmonisation Process

5 Test that the harmonisation process is performed as expected on the input data.

**In case tests are not passed, go back to refine step [4] and run the tests iteratively until they are passed.**

## Dataset Analysis - Part 2 (Responsible Agents on Refined Data)

6 A domain expert involved in a specific task (in the BnF Early Modern Dataset production project, mapping BnF authors with the English Short Title Catalogue actors) suggests an optimal sub-dataset structure to be easily exploited to address real projects' requirements.

## Data Subsetting Optimisation

7 This step envisions the production of a subset of the initial dataset to meet the users' requests for an optimal version of the data meant to be used for mapping purposes.

The set of scripts to run this task is collected here: https://github.com/ariannamorettj/New-BnF-Data-Analysis-2/tree/main/subset_optimisation

7.1 **Roles Enrichment**

```
python subset_optimisation/roles_enricher.py
```

As in the new dataset version we have the "profession" information as an unstructured string that might contain whatever concerning what the person use to define as a job, i corss-queried the two datasets and retrieved the "roles" that the actor had in relation with the publications of the time of interest (that is the reverse process iiro used to retrieved the actors dataset in the first place.

The obtained output is stored at subset_optimisation/id_roles/actor_roles_links.csv
The structure of the file is as follows:

| | actor | contributed_to | roles |
|---|---|---|---|
| | | 33831605 | author |
| | | 30025965;30025966;30025967;30459965;31729151;31729152;33455210;36407468;41619706 | author;editor |
| | | 33499433 | editor |
| | | 30126448 | author |
| | | 30265699;30265700;30265701;39306928 | author |
| | | 30059699;30059716 | author |

| | actor | contributed_to | roles |
|---|---|---|---|
| | | 30059851;30195936 | author |

where actor is the BNF ID, contributed_to are the BNF IDs of the bibliographic resources they appear to have contributed to (separated by ";"), and roles is the ";"-separated set of all the roles the actor had in relation to the publications in question.

## 7.2 Authors Link Optimisation

```
Note

python subset_optimisation/gen_author_links_optm.py
```

this script takes in input the id-roles-contributed_to tabe AND the BNF Actors dataset produced by Iiro (INPUT_ZIP_DEFAULT = "data/bnf_agents_data_querying/actor_queries_results.zip") and returns three main things:

1. An optimized CSV with one row per actor (all fields as multi-value)
2. A statistical report on the dataset with merge information
3. A JSON file tracking all merged values per actor
4. A reduced version of the actors dataset, with only the mapping-relevant information ("BnF_ID", "actor_link_exact", "actor_link_close")

In detail:

(1) **subset_optimisation/output/bnf_actors_optimised.csv** (very complete and quite big, also due to the contributed_to field). The columns are: BnF_ID,actor_name,actor_first_name,actor_last_name,actor_birth,actor_death,actor_start,actor_end,first_year,entity_type,actor_gender,actor_profession,actor_country,actor_language,actor_link_exact,actor_link_close,contributed_to,roles

(2) **subset_optimisation/report/summary_report.txt**: reports some stats concerning the fields' values, the filling rate, mergings, and the multiple values rate

```
================================================================================
BnF DATASET OPTIMISATION REPORT
================================================================================

BASIC STATISTICS
--------------------------------------------------------------------------------
Total source rows processed: 403,973
Exact duplicate rows: 0
Net rows (excluding duplicates): 403,973
Total unique entities (actors): 124,446
Entities with merged records: 36,990
Merge rate: 29.72%
Average records per entity: 3.25

================================================================================
FIELD STATISTICS (ALL FIELDS)
--------------------------------------------------------------------------------
(All fields can have multiple values after merge)

Field                     Filled          Fill Rate       Avg/Entity
--------------------------------------------------------------------------------
actor_name                110,355          88.7%          0.8872
actor_first_name          107,865          86.7%          0.8668
actor_last_name           121,956          98.0%          0.9800
actor_birth                40,620          32.6%          0.3264
actor_death                40,694          32.7%          0.3270
actor_start                28,989          23.3%          0.2329
actor_end                  30,516          24.5%          0.2452
first_year                 28,989          23.3%          0.2329
entity_type               124,446         100.0%          1.0000
actor_gender               37,741          30.3%          0.3033
actor_profession           37,684          30.3%          0.3708
actor_country              39,491          31.7%          0.3173
actor_language             37,802          30.4%          0.3089
actor_link_exact           42,153          33.9%          1.4243
actor_link_close           12,705          10.2%          0.2455
```

```
================================================================================
OVERALL AVERAGE
--------------------------------------------------------------------------------
Average items per entity (across all fields): 0.5379

AVERAGE ITEMS PER ENTITY (BY FIELD)
--------------------------------------------------------------------------------
actor_name                  0.8872
actor_first_name            0.8668
actor_last_name             0.9800
actor_birth                 0.3264
actor_death                 0.3270
actor_start                 0.2329
actor_end                   0.2452
first_year                  0.2329
entity_type                 1.0000
actor_gender                0.3033
actor_profession            0.3708
actor_country               0.3173
actor_language              0.3089
actor_link_exact            1.4243
actor_link_close            0.2455

================================================================================
MERGE DETAILS - FIELDS WITH VARIATIONS
--------------------------------------------------------------------------------
Field                     Entities with variations
--------------------------------------------------------------------------------
actor_profession             5,502
actor_language                 534
actor_link_exact            36,537
actor_link_close            10,145

================================================================================
```

(3) **subset_optimisation/report/merged_entities.json** - provides some mapping information for what concerns the merging process of the entities sharing the same bnf

id. The keys of the dictionary are the merged bnf id actor entities (on the base of the shared bnf id). Each actor's sub-dictionary contains a "merge_count", stating how many entities were merged, and "fields_with_variations", exposing the different values found for a specific field, such as - for example - actor_link_exact

```
{
  "<http://data.bnf.fr/ark:/12148/cb124434672#about>": {
    "merge_count": 7,
    "fields_with_variations": {
      "actor_link_exact": [
        "<http://www.idref.fr/033582262>",
        "<http://wikidata.org/entity/Q134588>",
        "<http://viaf.org/viaf/120732037/>",
        "<http://fr.dbpedia.org/resource/Nicolas_V>",
        "<https://francearchives.gouv.fr/agent/27356835>",
        "<http://isni.org/isni/000000012149188X>",
        "<http://fr.wikipedia.org/wiki/Nicolas_V>"
      ]
    }
  },
  "<http://data.bnf.fr/ark:/12148/cb11926496p#about>": {
    "merge_count": 10,
    "fields_with_variations": {
      "actor_link_exact": [
        "<https://data.biblissima.fr/entity/Q7101>",
        "<http://fr.dbpedia.org/resource/Thomas_d%27Aquin>",
        "<http://fr.wikipedia.org/wiki/Thomas_d%27Aquin>",
        "<http://isni.org/isni/0000000453792101>",
        "<https://musicbrainz.org/artist/ea7f0f74-18fc-409c-ba6c-
ed1cae12cddc>",
        "<http://viaf.org/viaf/100910166/>",
        "<http://wikidata.org/entity/Q9438>"
      ],
      "actor_link_close": [
        "<http://id.loc.gov/authorities/n78095790>",
        "<https://imslp.org/wiki/Category:Aquinas%2C_Thomas>",
        "<http://datos.bne.es/resource/XX1767563>"
      ]
    }
  },
  "<http://data.bnf.fr/ark:/12148/cb119465711#about>": {
    "merge_count": 6,
    "fields_with_variations": {
      "actor_link_exact": [
        "<http://fr.dbpedia.org/resource/Nonius_Marcellus>",
        "<https://data.biblissima.fr/entity/Q12963>",
        "<http://www.idref.fr/033806853>",
        "<http://wikidata.org/entity/Q183420>",
        "<http://fr.wikipedia.org/wiki/Nonius_Marcellus>",
```

```
          "<http://isni.org/isni/0000000109058530>"
        ],
        "actor_link_close": [
          "<http://datos.bne.es/resource/XX1767300>",
          "<http://id.loc.gov/authorities/n86113590>"
        ]
      }
    },
....
}
```

(4) **subset_optimisation/output/bnf_actors_optimised_minimal.csv** (minimal version of (1), reduced for being handy in mapping tasks):
"BnF_ID","actor_link_exact","actor_link_close"

All main information about the usage of this module is also documented here: **subset_optimisation/usage.MD**

## Quantitative Analysis of the Harmonised Dataset

8 Perform quantitative analysis on the data production concerning the harmonisation results:
- How many entities were modified?
- Which type of inconsistency was the most recurrent?
- How much did the harmonisation contribute
Reports are produced (JSON + TXT format)

## Harmonised Datasets Publication

9 Harmonised Dataset Publication, including
- CSV dataset about responsible agents
- CSV dataset about bibliographic resources
- Optimised subsets of data for research aims
- Quantitative analysis results and supplementary materials concerning the performed harmonisation
- Reports about code execution, computational costs, and run settings.

The complete harmonised dataset is released and published alongside complementary data on an adequate platform (such as Zenodo, granting a versioning system).

A persistent identifier is provided (DOI), and the dataset is released, ideally under a CC0 or CCBY license.

## RDF Dataset Production

10 RDF graph production : reverse converter (Morph-KGC-based) to get a RDF graph from the BnF harmonised CSV dataset about the Early Modern bibliographic entities and involved actors.

## Testing: RDF Dataset Production

11 SHACL validation of the graph
**In case the validation is not passed, go back to refine step [9] and repeat the validation iteratively until feedback is positive.**

## TTL serialised RDF Graph Publication

12 The RDF harmonised dataset is released and published on an adequate platform (such as Zenodo, granting a versioning system).
A persistent identifier is provided (DOI), and the dataset is released, ideally under a CC0 or CCBY license.

## SAMPO-UI portal: RDF Data Interface publication

13 Develop a SAMPO portal to be integrated in the SAMPO-UI framework, meant to interconnect RDF dataste user interfaces to allow simple exploitation of RDF information.

## Testing: User Testing on the SAMPO portal usability

14 Perform User Testing to verify the usability of the portal for users without background knowledge about
**In case tests are not passed, go back to refine step [12] and repeat tests iteratively until feedback is positive.**

## Acknowledgements

This work was carried out within the academic activities of the Computational History Group (COMHIS) at the University of Helsinki. It 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. For more information, visit the project website: **https://mecano-dn.eu/**. This project has received funding from the Finnish Cultural Foundation.