

**IMI2 101034344 – EPND**

**European Platform for Neurodegenerative Diseases**

**WP1 – Establishment of the platform**

## **D1.3 – Online demonstration of the first ETL & FAIRification tools importing data into the platform**

|                                                |                   |
|------------------------------------------------|-------------------|
| <b>Lead contributor (partner organisation)</b> | <b>1 – UM</b>     |
| <b>Other contributors</b>                      | <b>2 – ULEIC</b>  |
|                                                | <b>4 – DZNE</b>   |
|                                                | <b>6 – UNILU</b>  |
|                                                | <b>22 – ROCHE</b> |
|                                                | <b>26 – GV</b>    |
|                                                |                   |

|                                        |                   |
|----------------------------------------|-------------------|
| <b>WP1</b>                             |                   |
| <b>Due date</b>                        | <b>31/10/2023</b> |
| <b>Delivery date</b>                   | <b>18/12/2024</b> |
| <b>Delivery type<sup>1</sup></b>       | <b>DEM</b>        |
| <b>Dissemination level<sup>2</sup></b> | <b>PU</b>         |

|                              |                   |
|------------------------------|-------------------|
| <b>Description of Action</b> |                   |
| <b>V3.0</b>                  | <b>14/07/2024</b> |

### **Document History**

|                |                   |                 |
|----------------|-------------------|-----------------|
| <b>Version</b> | <b>Date</b>       | <b>Status</b>   |
| <b>V1.0</b>    | <b>07/11/2024</b> | <b>Draft</b>    |
| <b>V2.0</b>    | <b>29/11/2024</b> | <b>Comments</b> |
| <b>V3.0</b>    | <b>17/12/2024</b> | <b>Final</b>    |

|                                          |   |
|------------------------------------------|---|
| Abstract .....                           | 2 |
| 1. Case Studies data harmonisation ..... | 2 |
| 1.1. Data collection/gathering .....     | 3 |
| 1.2. Data transformation .....           | 4 |
| 1.3. Data mapping .....                  | 5 |
| 2. Mapping tools .....                   | 5 |
| 3. Data harmonisation to OMOP .....      | 6 |
| 4. Metadata harmonisation .....          | 6 |
| 5. Discussion and conclusions .....      | 7 |
| References .....                         | 9 |

## Abstract

This deliverable report outlines the harmonisation process for importing data and metadata to the EPND platform. We explain the procedure and challenges of extracting and transforming data to fit a common data model, which can be standardised or project-specific. We use EPND Case Studies data as an experiment for identifying which parts of data harmonisation can be (semi-)automated or facilitated by tools that could be offered in the platform. Finally, we discuss the process of metadata harmonisation, which is relevant for increasing the number of entries in the EPND catalogue.

## 1. Case Studies data harmonisation

The process of harmonising data is a key step in any data analytics project. This procedure is rather time-consuming and it is estimated to take from 2-6 person-months for small scale efforts (10-15 variables across 4 cohorts), and 15-18 person-months for large scale efforts (150+ variables across 10+ cohorts) [1]. This time burden is often an obstacle for data sharing and consumes a lot of projects' resources. The development of tools and methods that can facilitate this task is therefore much needed. The EPND platform aims to facilitate data sharing, discovery and analytics, thus the provision of harmonisation services was thought to be a possible component of EPND's Technical Hub. We used the harmonisation of the EPND Case Studies data as a learning experiment to investigate what the Technical Hub could do to aid the community in standardising their data to allow their discovery in the platform and in harmonising multi-study data for data analytics.

A harmonisation task can be summarised into three main steps: (i) data collection/gathering, (ii) data transformation, and (iii) data mapping. The process can be performed centrally, executed separately by each data provider, or a hybrid approach can be taken where the initial steps are conducted by the data providers and the final stages by a central party. We decided to take a mostly centralised approach to perform the Case Studies data harmonisation for EPND. Besides, it was decided that the data harmonisation would be "manually" (no generic automation) done with WP4, since they have the best knowledge of the use cases needs, and used later to analyse which steps of the harmonisation process

could be generally (semi-)automated. The details of the manual harmonisation are not relevant for this report. In this document, we present general technical aspects of what to consider when performing harmonisation.

### 1.1. Data collection/gathering

Data were gathered from four internal partners and two external partners for the execution of Case Studies 1, 2 and 4. Around 120 data items were requested from each cohort. Most partners contributed data from one cohort, except for DZNE who supplied data from four different cohorts. Because of the diversity of sources, data harmonisation was necessary to ensure consistency and compatibility to allow for analysis.

A document explaining all the data variables and their definitions was shared with each partner, who had the task of assembling the requested items. Multiple iterations were necessary between the researcher executing the data harmonisation and the data providers to ensure that the supplied data matched what was requested.

The data providers were not expected to perform complex data transformations to deliver the desired data variables. It was acceptable to provide all the necessary information for variables to be computed later by the researcher harmonising the data.

Table 1 summarises the data files, codebooks (data dictionaries) and documentation received from each cohort. This table illustrates the diversity of sources within and across the different cohorts and the complexity involved in generating a final curated and harmonised dataset. DZNE and Stavanger University Hospital were the only partners that supplied documentation files explaining the data, but with information spread across many different documents. Most partners provided a codebook explaining the variables in the datafiles, except for UNILU. The only partners who were able to provide the requested data as a single file were UNIGE, Stavanger University Hospital and Motol University Hospital. The other partners supplied multiple data files, with DZNE-Delcode providing the data in 16 different data files, which demonstrates the complexity and variety of the data sources.

| Cohort                        | Data files | Codebooks | Documentation |
|-------------------------------|------------|-----------|---------------|
| DZNE-Dancer                   | 10         | 1         | 10            |
| DZNE-Delcode                  | 16         | 1         | 10            |
| DZNE-Describe                 | 9          | 1         | 10            |
| DZNE-Migap                    | 5          | 1         | 10            |
| UNIGE                         | 1          | 1         | 0             |
| UNILU                         | 2          | 0         | 0             |
| VUmc                          | 2          | 1         | 0             |
| Stavanger University Hospital | 1          | 1         | 7             |
| Motol University Hospital     | 1          | 1         | 0             |

*Table 1. Summary of the received files (documentation and data) from the different cohorts contributing data for Case Studies 1, 2 and 4.*

The generation of codebooks is a task that can be partially automated by simply computing data summaries, which can easily be done with programming languages like Python and R or other data software tools. Further explanation, such as units for numerical variables and explanation of categories, can be manually included later by someone with a good understanding of the data. Nevertheless, we observed that most cohorts did provide their data accompanied by a codebook, indicating that compiling a data dictionary is not usually an issue.

Documentation of the data is a rather valuable piece of information, as it usually describes how data were collected and for which purpose, being then essential for judging how the data can be (re-)used for analysis. We expected that such a document would be available for most cohorts, but only 2 out of 6 partners (5 out of 9 cohorts) provided documentation on their data. The generation of data documentation is something expected to be done during a study design (e.g., as a Study Protocol, typically submitted to the Data Access Body together with the request for data access for the purpose of that study) and/or execution and nothing where EPND could have an influence.

## 1.2. Data transformation

Data transformation is the process of preparing data for analysis. Especially in the case of multi-centric data analysis, a common data model (CDM) is normally either adopted or established. The CDM is a formal description of the data structure that is necessary to answer the research questions of a particular project or that can generally be used for representing data in a standardised way in a particular field [2].

There is currently no community-wide adopted CDM for research in neurodegenerative (ND) diseases. A few initiatives are trying or have tried to put forward a data model for ND [3, 4], but so far none have been successful in getting traction. An internal EPND CDM representing a few of the data items collected for the Case Studies analysis was created with WP4 and presented in the deliverable report D1.2. Logically, EPND should join some of the existing initiatives, as the ND community would highly benefit from a standard CDM. However, the creation of a standard CDM is a huge task that should involve a diverse pool of researchers and was beyond the scope of establishing the EPND Technical Hub.

For the scope of Case Studies 1, 2 and 4, data was partially transformed into the project-specific CDM developed jointly by WP3 and WP4. Harmonised dataset were the usual steps normally taken in any data analytics project and executed using the R programming language. It was necessary to write different scripts per cohort due to the different data structures of the received files. Below we give a short summary, with a few examples, of some of the data transformation steps.

- Data cleaning
  - Changing coding, format, units
- Data transformation
  - Computing time intervals from dates, z-scores, etc
- Data restructuring
  - Based on baseline choice

- Data merging
  - Different files (same/different cohorts)
  - Laboratory results

Developing generic tools that can automatically transform any data structure into a particular CDM is not an easy task. Healthcare and also ND data can be highly heterogeneous even when recording similar/same properties. Generic tools that can transform data between different CDMs is though a more manageable task and there are projects that offer solutions for this [5, 6, 7]. EPND does not plan to enforce any particular CDM to users willing to make their data findable through our catalogue. We can though provide a catalogue of tools that can be used to transform data between different CDMs or facilitate parts of the transformation of generic data into a particular CDM.

### 1.3. Data mapping

Data mapping is the process of applying a semantic standard (e.g. SNOMED-CT, LOINC, ICD) to the data, which can include for instance variables' names and categorical variables. It is essential to standardise the semantics of data items and can also facilitate record level discovery in a data catalogue and enable federated learning. In contrast to data transformation, it is a bit easier to develop tools that can partially automate the process of annotating data following a variety of structures. We have developed a mapping tool within EPND and experimented with a few others. In the next section we present more details about our outcomes and about using mapping tools.

## 2. Mapping tools

Annotating data is a process that should be conducted by people with a good understanding of what the data mean. It is often the case that the person preparing an ETL for data curation for research does not have the clinical knowledge to properly annotate the data. Thus, splitting the process of data mapping from general data transformation can facilitate the whole data curation process and possibly also increase the final data quality. Mapping, which standardises semantics, is an ideal task for clinicians or any other expert in the disease area. Whereas data transformation is an ideal task for technical people with a good understanding of how the data is stored. It is sometimes the case that a researcher has the expertise for executing both tasks, as it happened for the harmonisation of the Case Studies data, but we cannot assume that this is always the case.

Besides the expertise gap, annotating data can be rather time consuming. Thus, the development of tools that can aid this process is an asset for data harmonisation. We developed a mapping tool within WP1 (FAIRnotator [8]) and identified two other existing tools (NTK-Curation App [9] and the Harmonization tool that was previously on the AD Workbench) originating from other WP1 participants. We compared them and identified that each tool had different strengths and could be applicable in slightly different contexts.

The adoption of a tool for data mapping is a study-specific choice, thus we decided that rather than enforcing a specific tool within the EPND Technical Hub, we can provide a list of tools accompanied by description of their functionalities and let the researchers decide what fits best their needs. It became

clear that facilitating the deployment of a variety of mapping tools within the Technical Hub, rather than enforcing only one of them, is a better strategy for the establishment of the platform.

We tested the deployment and usage of these tools within testing workspaces in the ADWB. The tools were successfully deployed and worked as expected. There is available documentation for installation of custom software within the ADWB [10]. The process was seamless and did not require extensive interactions with the platform providers. It is though important to mention that users with little technical knowledge would have probably had a different experience.

Finally, it is worth mentioning that, during the course of the EPND project, the application of large language models (LLMs) has become increasingly common, and it is showing to be useful also for harmonisation tasks [11, 12]. From the above-mentioned list of tools, none offers recommendations for mappings based on similarity measures between local concepts and terms in provided vocabularies. It is expected that LLMs should be capable of giving recommendations after being exposed to only a handful of examples. Thus, further development of available or new tools should concentrate efforts in implementing LLMs for providing recommendations or automatic mapping of data elements.

### 3. Data harmonisation to OMOP

We evaluated the benefit of using the Observational Medical Outcomes Partnership (OMOP) Common Data Model within EPND. This model has seen increased adoption by providing an adaptable database schema to store clinical data. In addition, it includes standardised vocabularies (e.g., SNOMED) to represent the data. For EPND, the OMOP CDM can be a potential model to represent cohort data and enhance cohort interoperability.

We started by selecting eight variables from the Case Studies dataset. Then, we followed the last steps of a data harmonisation process for cohort data previously developed for the Netherlands Consortium of Dementia Cohorts [13]. For this, we mapped the variables to standard concepts from the SNOMED vocabularies and created a table with the dataset metadata (i.e., variable names and values used in the dataset). We then employed a Python tool [14] to create the OMOP database, transform the data, and store it.

With this exercise, we showed that the OMOP CDM can fit the Case Studies data. Although we effortlessly harmonised the data using the OMOP CDM, there are important caveats to consider: 1) the dataset had been harmonized manually, and 2) we included a small set of variables. These two factors can have a significant impact on the time and effort involved. In addition, we observed a lack of standard vocabularies for the ND field. According to the OMOP conventions, the solution would be to create project-specific concepts. However, following that route can hamper one of the major benefits, interoperability between cohorts.

### 4. Metadata harmonisation

In the beginning of the project, most efforts were concentrated in the harmonisation of datasets, which is essential for record level discovery. However, the expectation that users contributing with data for the EPND catalogue would be willing to undergo the effort of harmonising their data towards an EPND

oriented CDM is unreasonable. It would in fact hamper any new additions to the catalogue. We decided to shift the focus to metadata harmonisation, which is a more effective way of employing resources to extending the EPND catalogue.

The EPND catalogue currently has entries following the Bronze and Silver metadata models. The Silver model provides richer information, it has a “Study” and its description as the base, with “Datasets” and “Biosamples” being associated with it. An effort to onboard new cohorts is ongoing and being driven by WP4. Another way to expand the EPND catalogue is by means of federated discovery and syndication. The former is a scenario where queries are sent to external catalogues, whereas for the latter copies of external catalogues are periodically pulled. Similar to datasets, metadata coming from different parties, and sometimes even within a single party, often follow different metadata models. This means that harmonisation is a necessary step in order to enable the display and query of syndicated entries.

In order to evaluate the feasibility of inclusion of syndicated entries in the EPND catalogue, we first experimented with implementing syndication between ADWB datasets catalogue and EPND Silver catalogue entries. A crucial step in this exercise was the harmonisation of the different metadata models. ADWB has a metadata model which is based on the DCAT standard [15], whereas the Silver model was developed within EPND to capture what experts presume to be relevant for searching the catalogue for relevant information.

Metadata harmonisation was performed between the “Datasets” section of the Silver model and ADWB’s DCAT-based metadata model. In summary, it was necessary to identify overlapping attributes and analyse their coverage. The Silver-Datasets model has 11 required fields, 4 of which are also available in ADWB’s entries. However only three of them share the same semantics, while the fourth one (contact point) does not fully match with the Silver model, which requires PI information. This low overlap and coverage highlight the challenges of syndication.

The exercise of syndicating between ADWB and Silver showed that an initial effort is necessary for making the catalogues interoperable. Some requirements of the Silver model would need to be changed in order to accommodate ADWB entries. First, the concept of “Studies” is not part of the ADWB model and it was not possible to create Studies from the information captured in the original catalogue. Second, many required fields in the Silver’s dataset section were not available or poorly captured in ADWB’s entries. The chosen pragmatic solution was to display syndicated entries slightly differently from EPND’s original entries. Metadata attributes with compatible meanings and good coverage are queryable, but other information that is either not part of the Silver’s model or not captured in the same way are still displayed as “Other” information and can be viewed by users.

Syndication with external catalogues was proven to be possible. An initial manual step for harmonising the metadata models is necessary and we would like to emphasise that this step would have been easier if both catalogues implemented an established standard.

## 5. Discussion and conclusions

The preparation of data for general data sharing or answering project-specific research questions requires a process that resembles ETL (Extract, Transform, Load) for data integration. Data needs to be gathered, transformed and combined. This procedure is called data harmonisation and it is known to be

a rather time-consuming task. It is a very important component of any data project, as it will ensure data consistency, compatibility and quality. The aim of this deliverable was to investigate how the EPND Technical Hub could or should support data harmonisation for data sharing and data analytics within the platform. We used the harmonisation of the Case Studies data as an exercise for which services could or should be offered.

We started by analysing which components of data harmonisation could be (semi-)automated or aided by software tools. We identified three main stages in data harmonisation: 1) data gathering, 2) data transformation, and 3) data mapping. Due to a large variety in data structures coming from different entities, or even from the same entity, also when talking about similar or same data properties, we concluded that data mapping is the only component where targeted automation might aid the harmonisation process. We experimented with three different tools, each offering slightly different functionalities, and concluded that the choice of which tool to use should be based on the project needs. We could successfully deploy all the three tools within a testing environment in the ADWB and they all worked as expected, but we experienced that the procedure, despite being simple for technically-minded people, could be quite challenging for others. The main learning of this exercise was that instead of imposing a particular tool, the EPND Technical Hub should be flexible for accommodating a large variety of tools.

Next, we explored the topic of a common data model for ND. The OMOP CDM is increasingly being adopted for research with healthcare data. We examined its applicability for ND data, using the Case Studies data as a reference. The transformation of pre-harmonised ND data into the OMOP CDM was a reasonably effortless task. However, the main drawback is the lack of standard vocabularies for ND-related concepts. This issue can be addressed by the usage of local concepts but defeats the purpose of adopting OMOP for ensuring interoperability.

Due to the lack of a widely community accepted CDM for ND and the particularities of data harmonisation for project-specific aims, we concluded that rather than offering harmonisation services, the Technical Hub should facilitate the execution of harmonisation within the hub. This can be accomplished by allowing deployment of external or custom tools that can aid the harmonisation process and directing users to known solutions for common problems.

The main aim of the EPND platform is to facilitate data sharing, discovery and analytics for the ND community. Thus, increasing the number of entries in the EPND catalogue is a core aspect for EPND's success. We thus started looking into data catalogue syndication. Metadata harmonisation is a key component for presenting syndicated entries in the EPND catalogue, as different catalogues follow different metadata models. We experimented with syndication between ADWB and EPND. It became clear that overlap and coverage of attributes is the main issue when attempting syndication. We could identify only four attributes with good coverage. The pragmatic solution for going ahead with syndication was to adjust the Catalogue's UI to present syndicated entries with a slightly different model than the one used for the original catalogue's entries, which also means that they have less attributes that can be queried.

Finally, we would like to bring attention to the usage of LLMs going forward. LLMs are expected to be great in aiding (meta)data harmonisation, as it might learn the context of (meta)data by only seeing a few examples, being then able to execute components of (meta)data harmonisation. Besides, LLMs could be great to enable catalogue search going beyond the available filters, which could make the case for catalogue syndication more interesting and less cumbersome.

## References

- [1] Fortier I, Wey TW, Bergeron J, et al. Life course of retrospective harmonization initiatives: key elements to consider. *Journal of Developmental Origins of Health and Disease*. 2023;14(2):190-198. doi:10.1017/S2040174422000460
- [2] Ahmadi N, Zoch M, Kelbert P, et al. Methods Used in the Development of Common Data Models for Health Data: Scoping Review. *JMIR Med Inform*. 2023;11:e45116. Published 2023 Aug 3. doi:10.2196/45116
- [3] Bauermeister S, Bauermeister JR, Bridgman R, et al. Research-ready data: the C-Surv data model. *Eur J Epidemiol*. 2023;38(2):179-187. doi:10.1007/s10654-022-00916-y
- [4] Wegner P, Balabin H, Ay MC, et al. Semantic Harmonization of Alzheimer's Disease Datasets Using AD-Mapper. *J Alzheimers Dis*. 2024;99(4):1409-1423. doi:10.3233/JAD-240116
- [5] Rinaldi E, Thun S. From OpenEHR to FHIR and OMOP Data Model for Microbiology Findings. *Stud Health Technol Inform*. 2021;281:402-406. doi:10.3233/SHTI210189
- [6] Xiao G, Pfaff E, Prud'hommeaux E, et al. FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP Common data Model. *J Biomed Inform*. 2022;134:104201. doi:10.1016/j.jbi.2022.104201
- [7] Chytas A, Bassileiades N, Natsiavas P. Mapping OMOP-CDM to RDF: Bringing Real-World-Data to the Semantic Web Realm. *Stud Health Technol Inform*. 2024;316:1406-1410. doi:10.3233/SHTI240674
- [8] FAIRnotator. Accessed on 17 December 2024. <https://github.com/MaastrichtU-CDS/EPND-FAIRification>
- [9] NTK-Curation App. Accessed on 17 December 2024. <https://www.alzheimersdata.org/ntk/ntkapp-platform>
- [10] Installing Docker on your Virtual Machine. Accessed on 29 November 2024. <https://knowledgebase.aridhia.io/workspaces/analysing-data/virtual-machines/installing-software-on-virtual-machines/installing-docker-on-your-virtual-machine>
- [11] Choi I, Long Q, Getzen E. Filling the gaps: leveraging large language models for temporal harmonization of clinical text across multiple medical visits for clinical prediction. Preprint. *medRxiv*. 2024;2024.05.06.24306959. Published 2024 May 7. doi:10.1101/2024.05.06.24306959

[12] Higashi K, Nakagawa Z, Yamada T, Mori H. Automated Harmonization and Large-Scale Integration of Heterogeneous Biomedical Sample Metadata Using Large Language Models. Preprint. BioRxiv. 2024.10.26.620145. Published 2024 October 29. doi:10.1101/2024.10.26.620145

[13] Mateus P, Moonen J, Beran M, et al. Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: A Netherlands consortium of dementia cohorts case study. *J Biomed Inform.* 2024;155:104661. doi:10.1016/j.jbi.2024.104661

[14] OMOP ETL for file-based DB. Accessed on 29 November 2024. <https://github.com/MaastrichtU-CDS/omop-converter>

[15] Data Catalog Vocabulary (DCAT) - Version 3. Accessed on 29 November 2024. <https://www.w3.org/TR/vocab-dcat-3/>