

IMI2 101034344 – EPND
European Platform for Neurodegenerative Diseases
WP1 – Establishment of the platform

D1.5 – Online Demonstration Of Record-Level Discovery Service, Including At Least 20 Participating Cohorts/Biobanks

Lead contributor (partner organisation)	02 – ULEIC
Other contributors	26 – GV
	11 – BMD
	04 – DZNE
	03 – CHUV
	13 – UMCG
	06 – UNILU

WP1	
Due date	31/10/2025
Delivery date	29/10/2025
Delivery type¹	DEM
Dissemination level²	PU

¹ Use one of the following codes:

R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

² Use one of the following codes:

PU = Public, fully open, e.g. web;

CO = Confidential, restricted under conditions set out in Model Grant Agreement;

CI = Classified, information as referred to in Commission Decision 2001/844/EC.

Description of Action	
V2.0	27/10/2022

Document History

Version	Date	Status
V0.1	01/10/2025	Complete Draft (ULEIC)
V0.2	02/10/2025	Added content (BMD)
V0.3	13/10/2025	Final edits, to review
V1.0	28/10/2025	Final version

Table of Contents

Table of Contents 2

1. Abstract..... 3

2. Introduction 3

3. EPND Federated Discovery Concept..... 4

4. Building the Federated Discovery Pilot..... 6

 4.1 Metadata & Aggregate data models (URLs for both).....6

 4.2 Discovery API Based on the Beacon-2 standard.....7

 4.3 EMIF-AD Beacon Service8

 4.4 Cafe Variome Technology.....8

 4.5 Central and Network Nodes9

5. Future 12

1. Abstract

This Deliverable describes the EPND Federated Discovery Pilot, which has successfully established a robust and effective technology for scalable, federated discovery of biomedical datasets across multiple catalogues. To maximize wider acceptance and adoption, the system was designed around proven standards (GA4GH Beacon-2 API, Health-DCAT metadata, JSON serialization) and mature software (Café Variome v3 (CV3)). Rather than predefining the ultimate scope of future federated discovery networks, the pilot focused on technical feasibility, interoperability, precision of queries, and data security to ensure the solution had broad utility and could serve a wide range of use cases. The information types that were targeted for use in this pilot were core metadata and pre-aggregated record-level data, as these provide a reasonable balance between ease of creation, informativeness for interrogation, and safety with respect to subject privacy and data security. We report details of all the utilized (meta)data models, APIs, and design considerations, and explain how we used these to establish a fully operational federated discovery network spanning 5 nodes and 78 cohorts/datasets. A central query portal (an empty copy of CV3) was installed on a virtual server alongside the Technical Hub, and the Hub's Keycloak solution was employed for the user-authentication (login) service. These two decisions ensure that the technology has a clear path to future integration with the EPND Catalogue. Most data nodes (at ULEIC, DZNE, CHUV, UNILU) employed the CV3 software to make their datasets discoverable as part of the federated network, whilst one node (at BMD) built a new Beacon-2 discovery service based on the pilot's API specification. CV3 was adapted to interoperate with such services, as well as with other CV3 instances. The technology we have thus developed and validated offers a practical foundation for future expansion and adoption across biomedical research infrastructures.

2. Introduction

Catalogue-based sharing/analysis platforms, such as the EPND Technical Hub, aim to promote the extensive and optimal use of research data/samples (assets). This requires the integrated provision of a large number of functions and services, the first and foremost of which entails enabling users to easily and accurately identify ("discover") specific assets that are suitable for their needs. This discovery function is directly analogous to the role online marketplaces play in enabling users to find specific products from amongst a wide universe of listed items, so they can make an optimal and informed choice about what to buy.

The discovery functions provided by the EPND Technical Hub relate to "Datasets" and "Biosample Collections" relevant to neurodegeneration, created and made available by some form of "Study" (typically a cohort initiative, a biobank, or a research consortium). The usefulness of this system directly reflects the number and diversity of assets it is able to make discoverable, as well as the richness of information available on those assets and the convenience and effectiveness of the discovery interface. As such, EPND has built a powerful catalogue based on a carefully designed metadata model (WP1), and continually invests a lot of effort into populating that catalogue (WP4).

To date, the EPND catalogue has been designed to operate as a central catalogue, whereby metadata about research assets are entered directly (via online forms) or syndicated from other listings (via an API). A user interface provides content filters on multiple data fields to enable users to distill from a starting list of (currently) >100 assets down to those that may best serve their needs. In time, AI-based query functions will also be added. This works well, and represents a notably large resource, but ultimately this approach is limited by the number of assets that the support team can shepherd into (and regularly update within) this single catalogue. Of course, the same fundamental problem confronts all related catalogues across the globe, such as DPUK, MJFF, GAAIN, IADRP, NIAGADS, ADataViewer, NCRAD, EMIF-AD, DARWIN, Vivli, and many more. Therefore, central catalogues like these cannot and do not endeavor to list all possible data/sample assets relevant to neurodegeneration research. But this means that a user may need to visit and explore many different catalogues before they discover the assets they need.

To address this limitation, EPND has been exploring the idea of enabling discovery queries launched in one catalogue, to be propagated through a network of many federated catalogues, to move towards a comprehensive discovery ecosystem. The referred queries would be executed at each central catalogue, and minimal metadata plus corresponding weblinks about any discovered assets would be passed back from each one to the starting catalogue, to be presented to the user as an aggregated set of results. Data and technical standards, along with mature software solutions, now exist to bring this powerful concept to life. This Deliverable describes work within EPND to design and pilot such a federated discovery network, initially by connecting catalogues/listings operated by EPND Partners. Once established and validated, the project can then offer this solution to the external community of large data and sample catalogues across the globe.

3. EPND Federated Discovery Concept

The objective of the EPND Federated Discovery pilot work is to develop and test optimal solutions for enabling assets of interest to be found by means of a single query that is distributed onward to multiple different catalogues. The precise focus, scale and footprint of the network(s) and service(s) that this might ultimately be used to instantiate cannot yet be stated, as this will depend on lessons learned from undertaking the pilot, and also on the willingness of the wider community to engage with whatever solution(s) we may create. Instead, a focus was placed on establishing the technical practicality, required standards, optimal precision of queries, and data security - giving sufficient overall utility to make the system attractive. That said, the general idea of federated discovery based on safe content has received support in our initial discussions with other catalogue stakeholders (e.g., DPUK, MJFF, current and planned IHI projects). Their comments emphasized the perceived value of such a system - so long as it is not too onerous to operate, compatible with establishing multiple networks, and mutually beneficial (i.e., exposing minimal metadata and referring users back to the respective source of each metadata entry).

Given this framing of the pilot's mission, the first task was to determine an appropriate granularity for the information that would be queried within, and the results being returned from, the nodes

in a federated discovery network. Datasets, rather than Biosample Collections, were chosen as the target for the pilot, as prior EPND work had shown that only the former could be defined (and hence meaningfully queried) in a standardized way. In contrast, the latter typically requires extensive back-and-forth discussions between users and biobank operators to identify suitable assets of interest. As illustrated in Figure 1 (adapted from *Nat Rev Genet*, 2015: <https://doi.org/10.1038/nrg3932>), data can be instantiated and described in a spectrum of ways, which can variously be used for discovery and/or sharing. The EPND Catalogue covers a number of these options, whilst this Federated Discovery Pilot overlaps and extends this range.

Core metadata attributes concerning datasets offer a natural first target for safe discovery queries (explored in Task 1.3). Whilst this level of information is simple enough to be routinely generated, it unfortunately provides limited discriminatory information. At the other extreme, the greatest discovery potential would be afforded by targeting record-level data (explored in Task 1.4), but this raises major challenges around schema complexity, content diversity, and data sensitivity/privacy. Various intermediates between pure metadata and pure record-level data (e.g., pre-aggregated record-level information, data dictionaries, value ranges/distributions) can balance these considerations, hence offering a pragmatic and useful way forward for the pilot to explore. Core metadata plus pre-aggregated record-level information was therefore chosen as the target for this Federated Discovery Pilot.

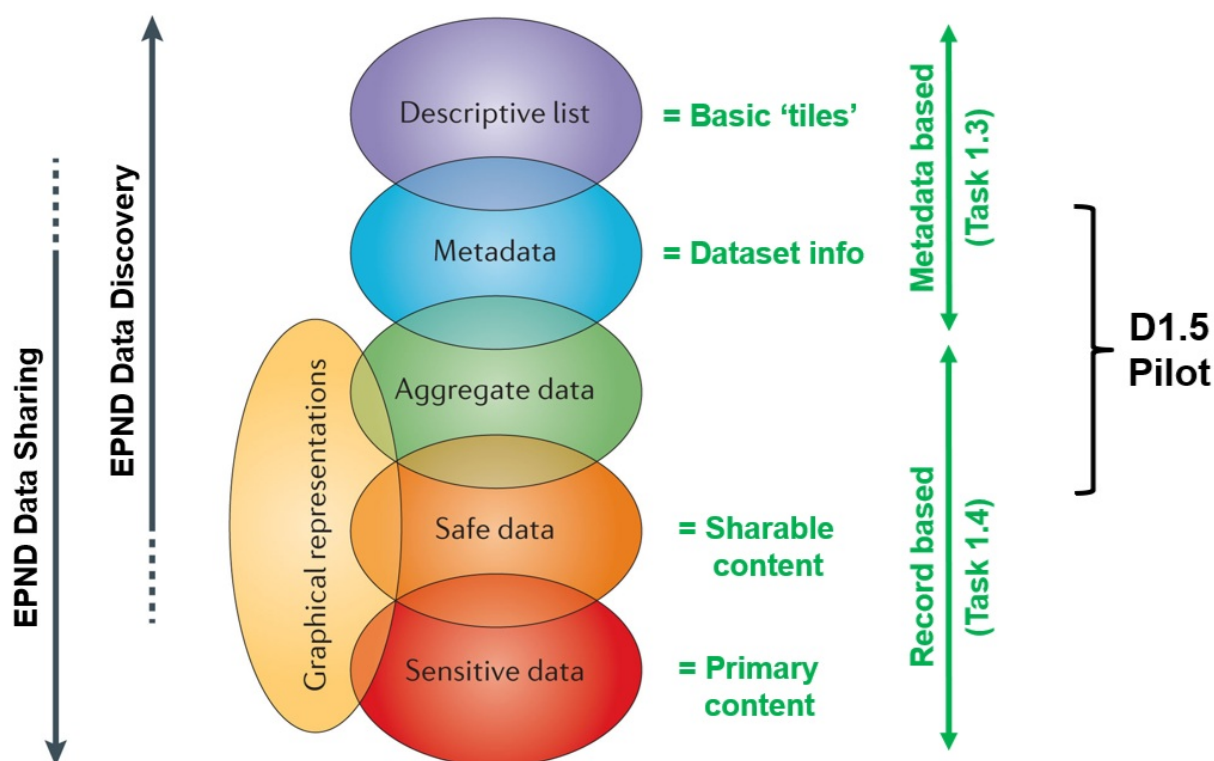


Figure 1

4. Building the Federated Discovery Pilot

4.1 Metadata & Aggregate data models (URLs for both)

```

sourceId
connectionId
sourceId
sourceName
sourceType [custom|cohort|catalog|biobank|registry|guideline|dataset|dataCollection]
resourceUrls
publisher
  publisherType [individual|organization|agency|other]
name
contactEmail
contactName
url
location
description
themes
releaseLicense
language
customFields
datasetVersions
  datasetDetails
    versionId
    versionName
    keywords
    publishedDate
    updateDate
  datasetContent
    numberOfSubjects
    minAge
    maxAge
    countries
    diseases [controlGroup|ad|pd|irbd|dlb|caa|ftd|als|psp|cbd|msa|hd|ataxia|other]
    sex [male|female|other|undifferential|unknown]
    clinical [comorbidities|medicationUse|familyHistory|ageOfSymptomOnset
              |clinicalDiagnosis|exposure|lifeStyle|infoVitalSigns]
    markers [amyloid|tau|neurofilamentLightChain|alphaSynuclein|dat]
    images [mri|petAmyloid|petTau|spect|ocular]
    electrophysiology [eeg|meg|erp]
    dataTypes [demographics|clinical|lifestyle|functionalRatings|motor
               |neuroPsychiatric|neuroPsychological|qualityOfLife|sleepScales
               |digitalData|imaging|electrophysiology|neuroPathology|other]

```

Figure 2

A federated discovery system necessarily requires that all participating nodes support filtering on the same range of queryable content (i.e., attributes and permitted values). To this end, we reviewed the Dataset metadata model from the EPND Catalogue, and consulted (via WP4) with dementia researchers and cohort creators, to identify metadata fields that would be present in most catalogue designs, and which also aligned with the standard DCAT metadata model. Additionally, to enhance the discovery potential, we considered record-level information typically available in cohort datasets, and identified common informative content and ways in which these could be pre-aggregated (e.g., average age, number of subjects by sex, list of biomarkers). Via further input from WP4, an initial list of suggestions was then culled to a final set of aggregate record-level values. The final range of dataset information to be targeted for federated discovery is available via [this link](#). Operationally, a matching JSON data structure was defined that this content would be packaged into, thereby ensuring completeness (required fields always present) and correctness (consistent data types, ranges and semantics) to enable federated interrogation. The current version of that JSON specification is always available at [this link](#) and [this link](#). Figure 2 shows a summary of the discoverable information.

4.2 Discovery API Based on the Beacon-2 standard

A global standard API Framework for federated discovery of biomedical data, called Beacon-2, was recently created by the GA4GH consortium [Rambla et al, 2022, doi:10.1002/humu.24369]. Its associated data model concerns genetic data and associated disease phenotypes, but a new initiative in that consortium (led by ULEIC and UMCG partners from EPND) was recently launched to develop an associated metadata discovery model. The EPND federated discovery pilot was based on this Beacon-2 standard, as this will facilitate widespread future uptake, whilst also being a key driver in advancing that standard itself towards metadata-based discovery.

The core GA4GH Beacon-2 standard we have employed is the [v2.0 schema](#). We tailored this specification to meet the needs of the EPND Discovery Pilot, and this API is described at [this link](#). The API is comprised of six informational endpoints that are used to retrieve details of each implemented discovery service, most notably /map and /filtering_terms that we modified to define metadata field filters (as this capability is not yet part of the Beacon-2 standard). Additionally, the main /dataset query endpoint provides the API's interface for receiving and responding to datasets query statements. This was adapted to accept filters relating to:

- Minimum Age
- Maximum Age
- Number of Control Group
- Average Follow-Up Duration
- Maximum Follow-Up Duration
- Published Date
- Start Date
- End Date
- Update Date

- Minimum Female Subject Count
- Minimum Male Subject Count
- Minimum Other Subject Count
- Minimum Unknown Subject Count
- Country
- Publisher Name
- Publisher Type
- Number of Subjects
- Dataset Types
- Free-text search, interrogating
 - o sourceName
 - o publisher.name
 - o publisher.contactName
 - o publisher.location
 - o description

4.3 EMIF-AD Beacon Service

Using the EPND Metadata Beacon specification, any Catalogue can now build their own federated discovery service - enabling any approved federated discovery network to interrogate their metadata content. Partner BMD took this approach to making many EMIF-AD Catalogue entries discoverable via the EPND pilot network.

Specifically, BMD developed extensions to their EPND-SN Catalogue, containing EMIF-AD cohort metadata, whereby the Beacon specification was implemented as part of the existing REST API provided by BMD's catalogues, which includes token-based authentication.

EMIF-AD entries consist of metadata, with little “real” data, which brought some initial challenges to WP1 regarding data presence/availability to the project. The implementation establishes a cross-mapping between the entries field values and structure to the Beacon data model, which permits forming the data models and allows querying via the agreed filtering terms.

4.4 Cafe Variome Technology

Many Catalogues may not be in a position to create their own Beacon-2 service, and would prefer to install a pre-made solution. Such software does exist, called Cafe Variome v3 (CV3), and this was employed to support this Pilot on Federated Discovery [Lancaster et al, 2015, doi:10.1002/humu.22841]. CV3 represents foreground technology from partner ULEIC, and it has been licensed to Gates Ventures (GV) to use for EPND Technical Hub sustainability (conditional on not charging academic users).

CV3 is a mature and powerful platform, which includes its own advanced discovery API, front-end and back-end components, authentication and authorisation system, security components, admin interfaces, and decentralised networking functions. It has been fully penetration tested and approved for use behind hospital firewalls. Nevertheless, some customisation of CV3 was needed to adapt it to the specific needs of the EPND Discovery Pilot. This was done, whereafter the software was provided to other EPND Partners, along with training on how to use the system.

These customizations included:

- Adding a Beacon-2 module with metadata-related endpoints, so that it could operate within networks that included other discovery software or DIY services that had implemented the project's Metadata Beacon-2 specification, rather than limiting this pilot (and future networks) to only supporting CV3 nodes.
- Code enhancements to support metadata handling, to accommodate aspects such as the specific metadata model, the query logic, metadata storage, query and result interfaces, and custom network messages.
- Setting up registration and security tokens with the Technical Hub's identity and access management solution (Keycloak), so that use of the federated discovery network is securely constrained to users having an EPND Catalogue account. This required establishing parallel connections to this Keycloak from CV3's front-end and back-end, and bypassing CV3's internal authorisation functions to grant all EPND account holders the same query permissions.

A more complete description of CV3, plus instructions for how to load metadata (in the required JSON format), are provided at [this link](#). This also details partly developed metadata models for "EPND Cohort", "Data collection", "Catalog", "Biobank", "Registry", "Guideline" and "Custom" objects, which were not used in this Pilot but will be valuable for subsequent extensions.

4.5 Central and Network Nodes

Using the above models, API standards, software and services, the Discovery Pilot network was created that spanned 78 discoverable dataset entries. Specifically:

- The EMIF-AD catalogue was made discoverable by Partner BMD using a purpose-built Beacon service. This was implemented, loaded with 65 entries, and progressively improved by iterative testing and development work assisted by ULEIC. Its Beacon service URL is https://epnd.bmd-software.com/api/beacon_v2
- An instance for CHUV was initially hosted by ULEIC (though administered by CHUV) containing 5 dataset entries. Local servers at CHUV were recently provided, so that they can host this instance locally. Its URL is <https://www027.lamp.le.ac.uk>
- An instance for DZNE was hosted by ULEIC (though administered by DZNE) containing 3 dataset entries. They have explored the possibility of hosting this instance locally, but

this has been held up with their IT services support group who are reluctant to provide suitable servers (nb. this is a general restriction there, not a particular concern about CV3). Their ULEIC hosted URL is <https://www028.lamp.le.ac.uk>

- UNILU produced metadata and aggregate data for 1 dataset (LuxPARK), and a CV3 instance was provided at ULEIC for this Partner. Its URL is <https://epnd.cafevariome.org>

Any and all nodes in this network could, in principle, be queried via the federated network from any of the other nodes, assuming permission is sought and granted, and authorised user accounts are established. However, for the purposes of this Pilot we established a single query portal centrally alongside the Technical Hub. Its URL is <https://cafevpilot.alzheimersdata.org>. Having this centrally located will enable us to make rapid progress in setting up whatever form of integration into the Technical Hub the project may wish to pursue, as the federated discovery service grows beyond the Pilot.

The central query portal is an instance of CV3. It can optionally carry additional discoverable entries, and for this Pilot we loaded it with 4 example entries, copied from (supplied by) the ADWB Catalogue. It coordinates user login with the Technical Hub Keycloak, and then federates query and result communications between the user and all CV3 nodes, using the internal CV3 API. It also converts the user's query into a Beacon-2 request which it sends to the EMIF-AD Beacon service. All returned results are compiled and presented to the user. Should any filter at any node not be available, that portion of the query is dropped from the actioned query, and the returned results are displayed along with a warning to alert the user about this limitation.

Figures 3, 4 and 5 present screenshots from the live federated discovery network. Figure 3 shows the overall Administrator interface, with the “Metadata Source” control and “View Metadata Sources” option being used to list 4 datasets populated into that node of the network. Figure 4 shows an example of a query result returned to the user, depicting the full set of discovered entries as a tiled set with core information shown for each (blue tiles), along with full details for one user-selected entry presented in a scrollable pop-up box. Figure 5 shows the interface via which a user constructs a query (using Boolean AND logic across filters), operating at the centrally installed CV3. At the top of the page the user can chose to find all sources or only datasets, with the latter selected in this example. Below this, the available functions support searching all text fields using any keyword, and faceted searching by dataset characteristics (Publisher Name/Type, dates, data types, 8 different aggregated content values, and country coverage). Many of these filters offer a pulldown list of options, based on real values that exist across all entries in the network, and the modular design of the interface allows all these features to be modified as needed.

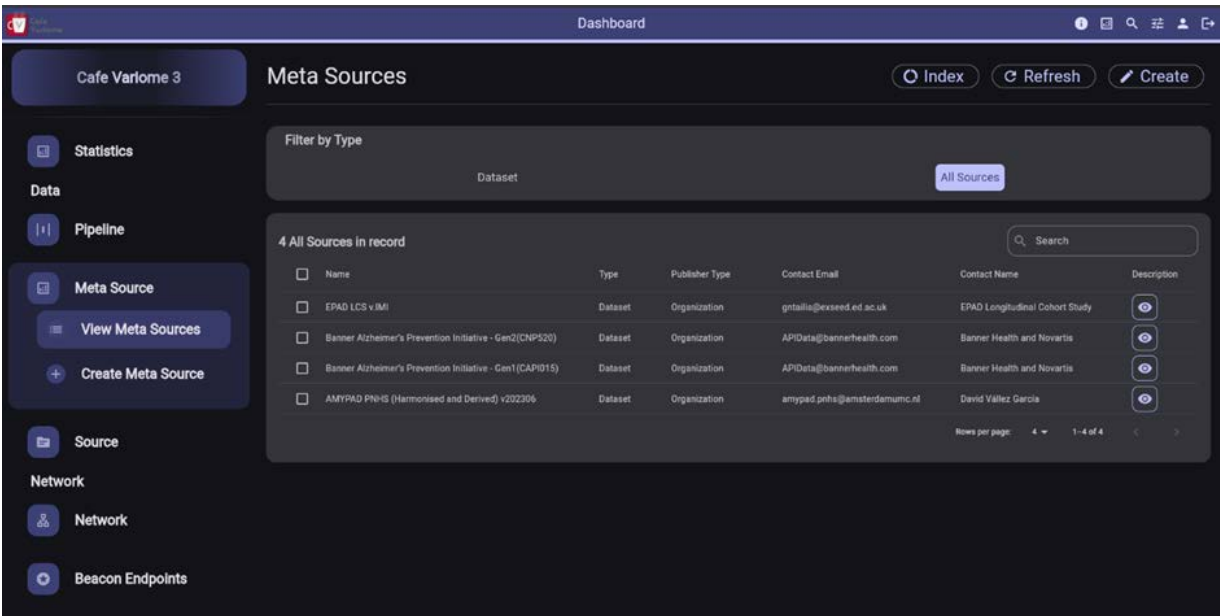


Figure 3

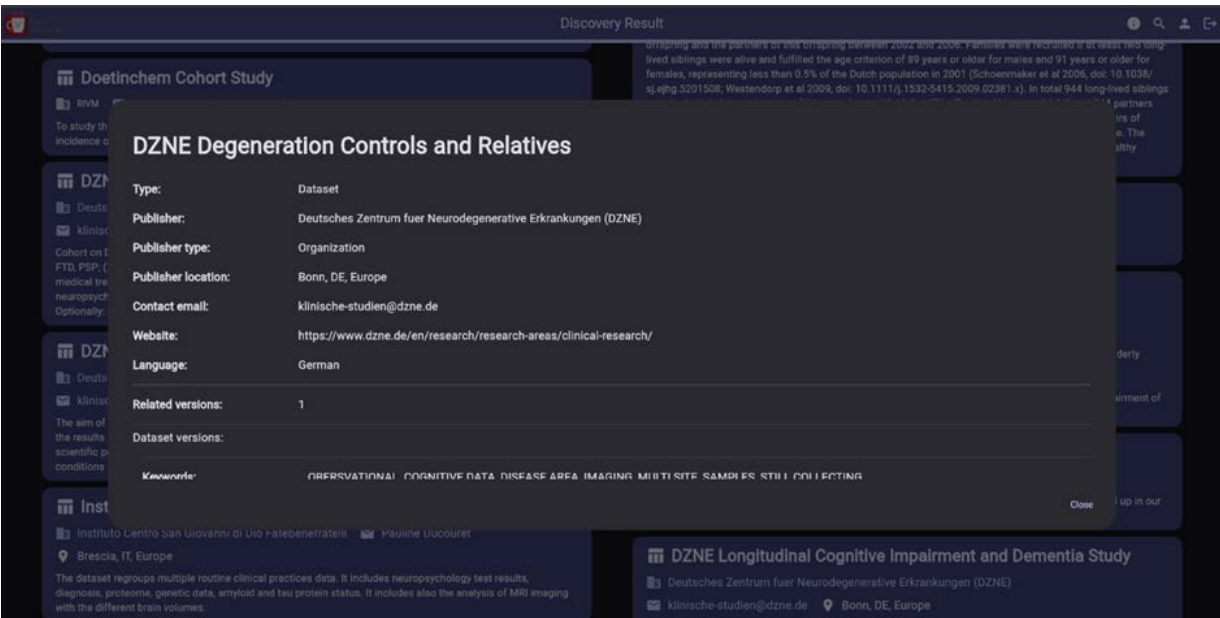
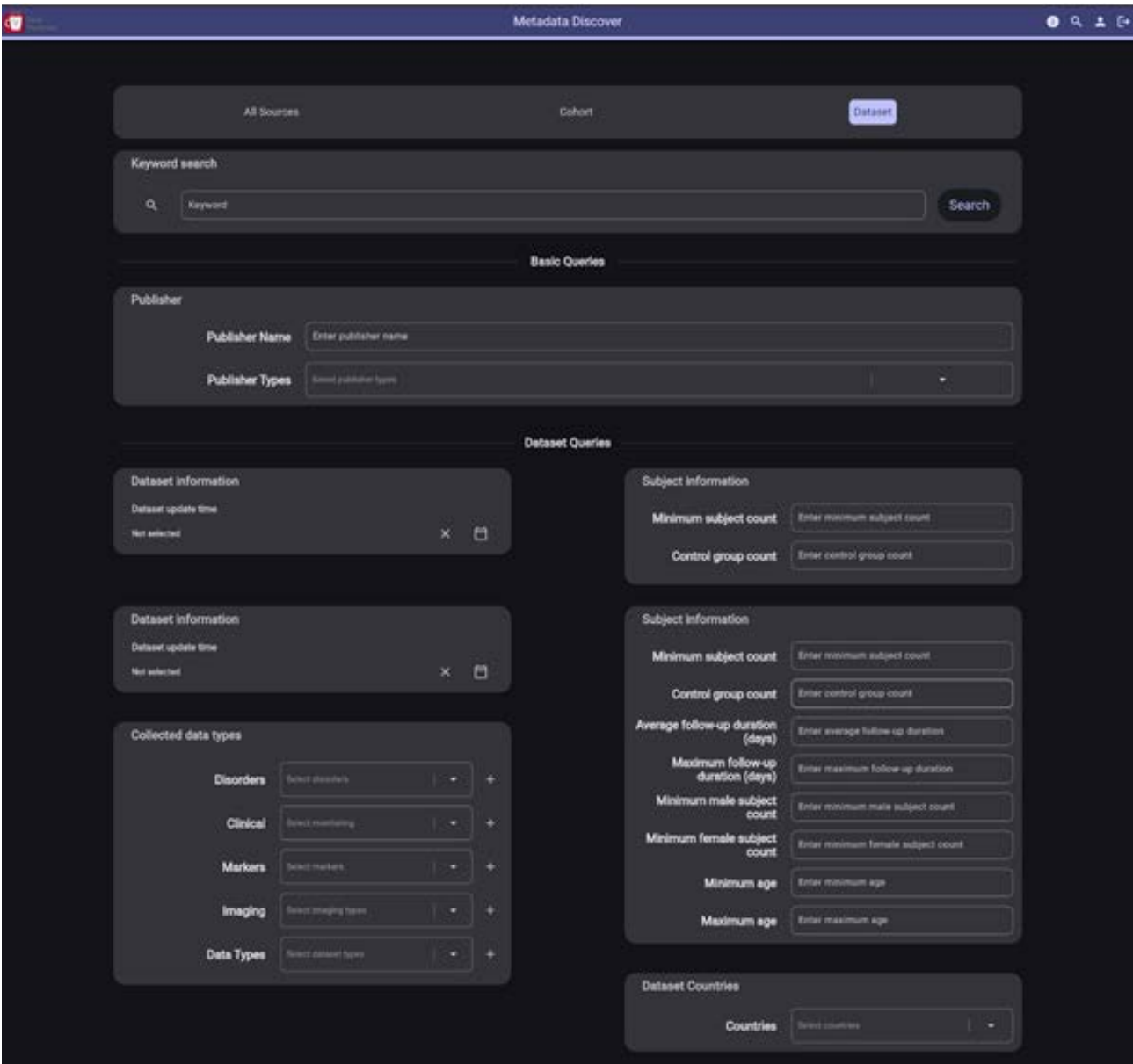


Figure 4



The screenshot shows the 'Metadata Discover' web application. At the top, there are tabs for 'All Sources', 'Cohort', and 'Dataset', with 'Dataset' being the active tab. Below the tabs is a 'Keyword search' section with a search bar and a 'Search' button. The main content area is divided into two columns. The left column contains a 'Basic Queries' section with 'Publisher' (Name and Types) and 'Dataset Queries' (two 'Dataset information' cards and a 'Collected data types' section with dropdowns for Disorders, Clinical, Markers, Imaging, and Data Types). The right column contains 'Subject Information' sections with input fields for 'Minimum subject count', 'Control group count', 'Average follow-up duration (days)', 'Maximum follow-up duration (days)', 'Minimum male subject count', 'Minimum female subject count', 'Minimum age', and 'Maximum age'. At the bottom right is a 'Dataset Countries' section with a 'Countries' dropdown.

Figure 5

5. Future

The progress described in this Deliverable endows EPND with a unique, powerful, proven technical ability via which multiple metadata catalogues/listings can be joined into a network for safe, federated discovery. Depending on the specifics of the catalogues, the datasets, the

research domain, and the intended users and use cases, various specifics of this technology may need to be modified (e.g., metadata/data models, utilized filters, result content, login/security design). Key extensions could also easily be built, such as querying on machine-readable use conditions, tunable searches for similar instead of perfect matches, querying on data dictionaries, or interrogating non-redundant value lists (e.g., list of diagnoses across all patients in a dataset). Critically, none of this should be confused with enabling data analysis, as that is completely counter to a safe discovery function. Use of data obfuscation and avoiding actual primary data querying, must therefore remain red lines. The main point is, EPND can now start to make decisions about how it wants to leverage federated discovery - with confidence that the core technology and concept work well. This allows us to give strong reassurances to external teams that might like to be part of forming a future federated discovery network.

To move forward from here, a decision will need to be made regarding how federated querying (and the user interfaces that support it) should be integrated into the discovery and display interfaces of the Technical Hub Catalogue. The work of this Pilot has ensured that the attributes and filters of the catalogue and the federated discovery API are compatible. However, since the latter must accommodate greater diversity (multiple sites) than the former (just the EPND site), the federated discovery models are necessarily less extensive. This means that the federated discovery process cannot be added on as a direct full-span extension to the catalogue discovery system. This leaves two options, which are not mutually exclusive. (A) The two systems could be joined and operate only through the catalogue discovery interface, with the filters that a user employs being used to whatever extent they can be by each remote system (i.e., the catalogue runs its own full query, while simultaneously creating a simpler sub-query to pass into the federated discovery network). The two sets of results could then be presented in parallel, or combined into one display with flags indicating those provided by the federated discovery network. Or (B) the two systems could be kept separate and utilize two distinct, optimised interfaces. In this case, the two result sets would most logically be presented separately. In scenarios A and B, the range of details provided for each dataset entry in the two sets of results would likely be different, with the federated discovery results presumably including links to the matching source node in the queried network.

A second key decision concerns whether the federated discovery network is to serve existing EPND partners (i.e., as an alternative to manual or API-syndicated central deposition of their metadata), and/or whether it connects the EPND catalogue with multiple other extant major catalogues so that all can query all. Discussions over the past years or so have suggested a need for both options. As an example for the former arrangement, BMD is currently making a large portion of the EMIF-AD catalogue available for federated discovery, whereas only a small fraction of their datasets have yet been entered (as metadata) into the EPND catalogue itself. Others may follow suit. The advantage here is that such sources only need to maintain one copy of the metadata (a single "source of truth"). As an example for the latter arrangement, other external catalogues (DPUK, MJFF (under construction)) have already expressed an interest in forming a federated discovery network with EPND. With such groups connected, one could easily imagine this getting a lot of attention, and many other catalogues then wanting to join in.

In summary, good progress has now been made towards a future of federated discovery. This will negate the need for users to visit and interrogate tens or even hundreds of different catalogues/listings to find optimal research assets of interest. The solution is based on data standards and uses proven technology, with the latter licensed to Gates Ventures so they can leverage it (always free for academics) as they grow the dementia data/sample sharing ecosystem over future years. As such, the capabilities we have developed and describe in this Deliverable can provide a valuable cornerstone to promote more effective research in dementia, and indeed, in any other domain.