**Interaction of the MDR with B2FIND EUDAT service – report**

Author(s): Sergey Goryanin, Christian Ohmann, Steve Canham, Maria Panagiotopoulou

Date: July 2020

**Report**

1. **Introduction and background  
     
   What is an MDR?**

An MDR, or Metadata Repository, brings together the metadata about clinical trials and the data objects generated by them, from a large variety of data sources. It standardizes that metadata into a consistent schema, and then enables users’ access to the metadata (with links back to the source material where appropriate) via an API interrogated by a web-portal.

**Background**

In recent years there has been a growing acceptance that to accurately assess the results of trials and other clinical research, and in particular to combine the results from different trials in meta-analyses, it is necessary to have access to the original source data, the “individual participant data” (IPD), as well as the results found in published papers. In addition, to make sure that the IPD can be fully understood and properly analysed, a variety of other study documents (protocols, analysis plans, etc.) are required. As a result, under pressure from funders and journal editors, more and more researchers are making such material (generically, “clinical trial data objects”) available for sharing with others. The datasets are rarely freely available - instead a variety of access mechanisms (e.g. individual request and review, membership of pre-authorized groups, or web based self-attestation), are used in combination with different access types (e.g. download versus in-situ perusal). Furthermore, the various data objects are stored in a wide variety of different locations: a rapidly growing number of general and specialized data repositories, trial registries, publications, the original researchers’ institutions, etc.

The researcher or reviewer wishing to locate relevant data objects for a study is therefore faced with a bewildering mosaic of possible source locations and access mechanisms, and this problem of ‘discoverability’ will almost certainly become much worse in the future as more and more materials are made available for sharing.

**Aims**

The principal aim of the MDR is to combat this discoverability problem, by making the data objects generated from clinical research easier to locate, and by describing how each of those data objects can be accessed, providing direct links to them where that is possible. The central idea is to develop systems that can collect the *metadata* about the data objects, including object provenance, location and access details, and aggregate it into a single **Metadata Repository** (or MDR). The MDR is therefore designed to assemble, standardize and display the metadata about clinical studies and the data objects generated by them, and provide access to that metadata through a single system, accessed via a web portal.

**Implementation**

The MDR system has been designed and developed by [**ECRIN**](https://www.ecrin.org/) (the European Clinical Research Infrastructure Network), in collaboration with ([**ONEDATA**](https://www.onedata.org/#/home)) and [**INFN**](http://www.bo.infn.it/) (Istituto Nazionale di Fisica Nucleare) at Bologna. Up to now development of the project has been within the H2020 **[eXtreme - DataCloud](http://www.extreme-datacloud.eu/the-project/)** (XDC) project, funded by the EU under grant agreement 777367.  
Metadata from a variety of data sources have been collected by ECRIN using different modalities (e.g. DB download, import of XML files through an API, scraping of web pages) and stored in a relational DB. Data is then exported as json file metadata to the OneData file management system and indexed via Elastic Search to make it available to the web portal.

**References**

The MDR web-portal: [crmdr.org](http://crmdr.org/)

The MDR Wiki: <https://ecrin-mdr.online/index.php/Project_Overview>

**What is an EUDAT B2FIND service?**

Metadata are important whenever research data are intended for serious use. In addition to being essential for understanding an individual dataset, systematically structured metadata are also a key element of data infrastructures, such as searchable catalogs of data, or automated systems for mapping and analysis.

The EUDAT metadata service B2FIND addresses these requirements by providing a comprehensive joint metadata catalogue and a powerful discovery portal. Metadata stored through EUDAT services such as B2SHARE are harvested from various research communities overarching a wide scope of research disciplines.

The B2FIND portal and API provide the user with advanced search functionalities and allow access to the data resources behind the metadata found in the catalogue.

**How does it work?**

Metadata that are made searchable in B2FIND are harvested from metadata providers using the standard OAI-PMH interface. The community itself decides which metadata are made available to EUDAT and how their metadata elements are mapped to EUDAT-specific facets. A sophisticated framework ensures that metadata from various providers are harvested regularly to display complete and up-to-date information. This framework also provides the translation from the community metadata schema to standard facets in the B2FIND metadata catalogue.

**References**

User documentation: <https://eudat.eu/services/userdoc/b2find>

Training: <https://github.com/EUDAT-Training/B2FIND-Training>

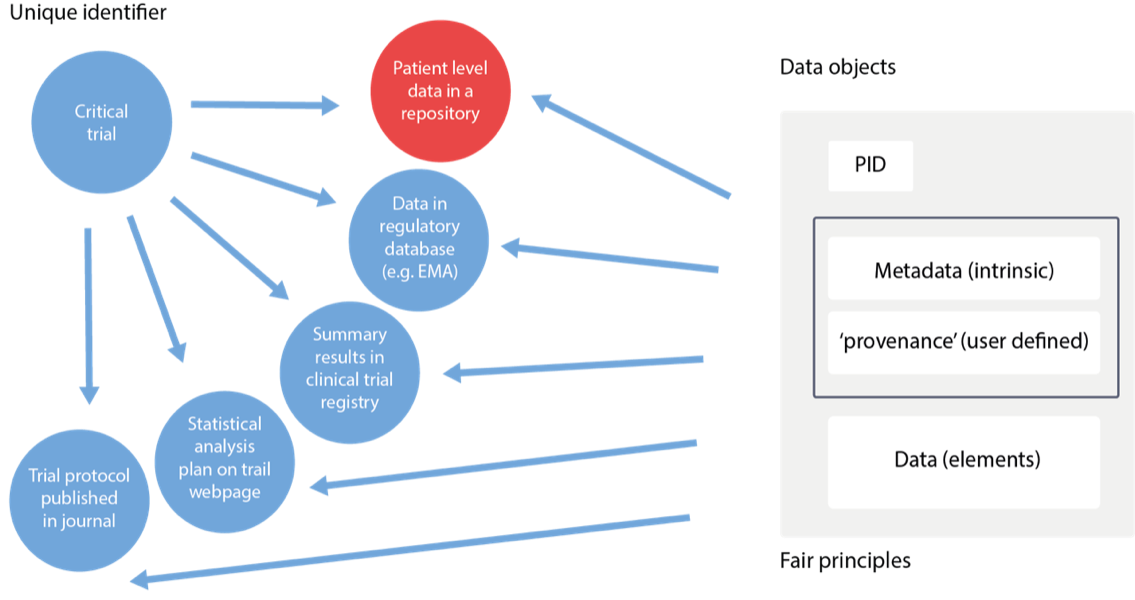
**2. Metadata schemas**

**2.1 MDR metadata schemes**

The MDR metadata system includes two main schemas – one for studies and the other for data objects created within those studies.

A ***study*** is any clinical research study with humans as study participants, and which is therefore subject to ethical approval, whether or not the study is interventional (a 'clinical trial'), or observational (including disease registry data), or a case study.

A ***data object*** is any information available in electronic form (a 'data stream') and may be a document (e.g. a pdf), a dataset in one of a variety of formats (spreadsheet, csv files, database file, XML etc.), a media (audio, video) file, an image (e.g. a conference poster) or simply a web page with useful text. As figure 1 illustrates:



***Figure 1. Data Object generation in a clinical trial***

Users are more familiar with studies, and in most cases would search for data objects by first searching for and selecting the source studies. In most cases studies have PIDs (e.g. trial registry ids) and titles, whereas the data objects normally do not have any identifier, or even a specific name. Their 'title' is therefore usually the study name plus the object type, e.g. <study name> final IPD datasets, or <study name> protocol.

At the moment the data is made available as json files, so there are two file types, one for studies and one for data objects. To provide a more accurate description of the files used JSON schema descriptions were developed, corresponding to each of the two schemas, and these are available within the MDR wiki.

**References**

Wiki: <https://ecrin-mdr.online/index.php/JSON_Schemas>

ECRIN Metadata Schema (v.3): <https://zenodo.org/record/3562911>

The current study JSON scheme: <https://ecrin-mdr.online/index.php/Study_JSON_v4>

The current data object JSON scheme: <https://ecrin-mdr.online/index.php/Data_Object_JSON_v3>

**2.2 B2FIND metadata requirements**

B2FIND is open to integrating any metadata scheme. Metadata formats that are already supported by B2FIND are listed in table1.

***Table 1: Metadata formats supported by B2FIND***

|  |  |  |  |
| --- | --- | --- | --- |
| Format | Specification Docs | Description | Users |
| **Dublin core** | <http://dublincore.org/specifications/> and in the following standard documents:   * IETF RFC 5013 * ISO Standard 15836-2009 * NISO Standard Z39.85 | The **Dublin Core Schema** is a small set of vocabulary terms that can be used to describe web resources as well as physical resources such as books or CDs, and objects like artworks. The full set of Dublin Core metadata terms can be found on the Dublin Core Metadata Initiative (DCMI) website. | DataCite, NARCIS, PanData, The European Library, SDL |
| **ISO 19115** | <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798> | **ISO 19115-1:2014** defines the metadata schema required for describing geographic information and services by means of metadata, e.g. the extent, the quality, the spatial and temporal aspects, and other properties of digital geographic data and services. | ENES |
| **Marc XML** | <http://www.loc.gov/standards/marcxml/> | **MARC** (**MA**chine-**R**eadable **C**ataloging) standards are a set of digital formats for the description of items catalogued by libraries, such as books. | B2SHARE and ALEPH |
| **CMDI** | <http://www.clarin.eu/content/component-metadata> | **CMDI** (**C**omponent **M**eta**D**ata **I**nfrastructure) was initiated by CLARIN to  provide a framework to describe and reuse metadata blueprints. | CLARIN |
| **DDI** | [http://www.ddialliance.org](http://www.ddialliance.org/) | **DDI**(**D**ata**D**ocumentation**I**nitiative**)** is an effort to create an international standard for describing data from the social, behavioural, and economic sciences. | CESSDA |

The harvested "raw" metadata records are community-specific with regards to the metadata format and to the content, i.e. the property definitions and values. As a result, a consultation needs to be held between B2FIND and any source community to determine how a mapping can be configured and adapted to the community-specific needs.

The core B2FIND schema is listed below. The only mandatory item is 'Title'. All other fields are optional but recommended.

***General Information***

*Title*: A name or title by which a resource is known.

*Description*: Additional information describing the content of the resource. Could be an abstract, a summary or a Table of Contents.

*Tags*: A subject, keyword, classification code, or key phrase describing the content.

***Data access***

*Source*: The Source is an identifier, therefore a unique string that identifies the resource.  It may link to the data resource itself or to a landing page that curates the data.

*PID*: The PID is an alternate identifier.

*DOI*: The DOI is an alternate identifier.

***Provenance***

*Community*: Research communities that provide research data to EUDAT. Could be an aggregator as well.

*Discipline*: A scientific discipline the resource originates from. A closed vocabulary based on a Wikipedia-classification is used.

*Creator*: The main researchers involved in producing the data, or the authors of the publication in priority order.

*Publisher*: The name of the entity that holds, archives, publishes, prints, distributes, releases, issues or produces the resource.

*Publication Year*: The year in which the resource was or will be made publicly available

*Formal Language*: The primary language of the resource. Codes are mapped to long names according to ISO 639.

*Temporal Coverage*: Period of time the research data resource is related to. Could be a date format or plain text or both.

*Spatial Coverage*: A geolocation the research data resource is related to. Could be geographic coordinates of the Earth's surface (e.g. longitude/latitude) or denomination of places.

*Format*: Technical format of the resource.

***Additional information***

*Contact*: Any contact information for this resource.

*MetadataAccess*: The [OAI-PMH GetRecord](http://www.openarchives.org/OAI/openarchivesprotocol.html#GetRecord) request.

1. **Interaction between MDR and B2FIND**

Unfortunately, it appears that there are some fundamental differences between the MDR as normally consumed within B2FIND and the metadata used within the MDR.

1. The MDR metadata is 2 level – studies and data objects. There is a many to many relationship between these – a study will generate many data objects, and some data objects will refer to more than one study. This seems very different from the single level of entities, i.e. data objects, used within B2FIND.
2. The many to many relationship makes it impossible to easily 'flatten' the MDR metadata into single entities – i.e. combined study and data object records. Such a process would also lead to a huge amount of unnecessary duplication, and there would be no obvious benefits for users if we did this.
3. In the MDR users do not in general search for data objects. They cannot because the data objects do not have unique identifiers or names. They search for studies, and then see what data objects are available for those studies. The exception would be users searching using a journal article, though in those cases the system first finds the studies associated with that article.
4. The metadata schemas used in the MDR are much richer than the basic B2FIND schema, partly because users need the additional information, partly because the source data is itself richer. For example, studies often have several names, some of which may be in different languages. Data objects may be available in multiple formats in multiple places. Study sponsors and other key organisations are included as an important possible search criterion, over and above creators. Studies have a 'type' (e.g. interventional versus observational), and a 'status', (e.g. recruiting, pending, completed, terminated). These are again key search criteria.
5. One of the key attributes of datasets linked to clinical trials (which of course contain sensitive personal data) is their legal status in terms of the GDPR and other data protection legislation. The MDR object schema includes items relating to the level of de-identification of data, the type of consent for secondary use (if any) associated with the data, and whether the data is seen as pseudonymised or anonymized. This information is critical for potential users of the data but seems to be missing from the B2Find schema.
6. A lot of the datasets associated with clinical research are not directly available – instead they are subject to a variety of controlled access measures, which often involve formal application for access. Details about these procedures also need to be stored as part of the provenance metadata. The B2FIND schema does not appear to include this. Without such details, however, the other metadata is of very limited use.

The MDR is also designed – eventually – to be updated frequently because the source data changes quickly, and new studies are being registered all the time. The current system has 0.5 million study records and 0.8 million data object records, but is likely to grow steadily over time. Any interaction with B2FIND would necessitate an additional data migration burden on top of what is already a heavy workload, for both organisations.

For these reasons it seems any direct integration between MDR data and B2FIND would be very difficult to implement. In fact, any interaction involving data is likely to be difficult and it is difficult to see how it would provide much additional value to users. At an organisational level, however, it will be important to maintain ongoing liaison between ECRIN and its MDR partners and EUDAT and the various services it offers, and explore possible ways in which the organisations could benefit each other. These could include providing access to relevant B2FIND linked data from the MDR, or registering the MDR as a whole into B2FIND as a resource, and future discussion around metadata schemas and APIs.