ECRIN Metadata Schema for Clinical Research Data Objects, Version 4 (September 2020)

Steve Canham, 14/09/2020

Contents

[Introduction 3](#_Toc50989093)

[Summary Tables 4](#_Toc50989094)

[Study Attributes 7](#_Toc50989095)

[A.1 Display Title (1) 7](#_Toc50989096)

[A.2 Study Identifiers (0...n) 7](#_Toc50989097)

[A.3 Study Titles (0..n) 7](#_Toc50989098)

[A.4 Brief Description (0…1) 7](#_Toc50989099)

[A.5 Data Sharing Statement (0..1) 7](#_Toc50989100)

[A.6 Study Features(0…n) 8](#_Toc50989101)

[A.7 Study Topics (0…n) 8](#_Toc50989102)

[A.8 Study Type (1) 8](#_Toc50989103)

[A.9 Study Status (1) 8](#_Toc50989104)

[A.10 Study Enrolment Number (0..1) 8](#_Toc50989105)

[A.11 Study Gender Eligibility (0..1) 9](#_Toc50989106)

[A.12 Study Minimum and Maximum ages (0..1) 9](#_Toc50989107)

[A.13 Related Studies (0..n) 9](#_Toc50989108)

[A.14 Linked Data Objects (1..n) 9](#_Toc50989109)

[A.15 Provenance Data (1) 9](#_Toc50989110)

[Data Object identifiers 10](#_Toc50989111)

[B.1 Data object identifier (0..1) 10](#_Toc50989112)

[B.2 Display Title (1) 10](#_Toc50989113)

[B.3 Version (0..1) 10](#_Toc50989114)

[B.4 Object Identifiers (0...n) 10](#_Toc50989115)

[B.5. Object Titles (0...n) 10](#_Toc50989116)

[B.6. Linked Studies (1...n) 11](#_Toc50989117)

[Creators and Contributors 11](#_Toc50989118)

[C.1 Creators (1...n) 11](#_Toc50989119)

[C.2 Contributors (0...n) 11](#_Toc50989120)

[Dates 12](#_Toc50989121)

[D.1 Publication year (1) 12](#_Toc50989122)

[D.2 Dates (0...n) 12](#_Toc50989123)

[Data Object Attributes 12](#_Toc50989124)

[E.1 Class (1) 12](#_Toc50989125)

[E.2 Type (1) 12](#_Toc50989126)

[E.3 Record key type (1, Datasets only) 12](#_Toc50989127)

[E.4 De-identification level (1, Datasets only) 13](#_Toc50989128)

[E.5 Associated consent (1, Datasets only) 13](#_Toc50989129)

[E.6 Description (0...n) 13](#_Toc50989130)

[E.7 EOSC Category (0..1) 13](#_Toc50989131)

[E.8 Language (1..n) 14](#_Toc50989132)

[E.9 Related Objects (0..n) 14](#_Toc50989133)

[E.10 Topic (0...n) 14](#_Toc50989134)

[Location and Access details 14](#_Toc50989135)

[F.1 Managing Organisation (1) 14](#_Toc50989136)

[F.2 Access Type (1) 14](#_Toc50989137)

[F.3 Access Details (Mandatory for any of the non-public access types) 15](#_Toc50989138)

[F.4 Access Details URL (Mandatory for any of the non-public access types) 15](#_Toc50989139)

[F.5 Resources (Mandatory unless case-by-case access) 15](#_Toc50989140)

[F.6 Rights (0..n) 15](#_Toc50989141)

[F.7 Provenance Data (1) 15](#_Toc50989142)

[Categorised Items 16](#_Toc50989143)

[JSON schemas 16](#_Toc50989144)

[References 16](#_Toc50989145)

# Introduction

A metadata schema for clinical research data objects was first developed by ECRIN in 2016 [1], as a mechanism for supporting increased discovery of the wide range of data objects, scattered across many different repositories, that are generated by clinical research activity, and in particular to support the development of a proposed metadata repository, or MDR, for clinical study data objects.

There are in fact 2 schemas, one for studies and one for data objects. The study schema is based on the main data points used by ClinicalTrials.gov, by far the largest trial registry in the world, with about 350,000 existing study entries. Those data points are themselves based around the core dataset required by the WHO and so – in broad terms – are also supported by the other 18 globally recognised trial registries. Trial registry data, and that from Clinicaltrials.gov in particular, represents the *de facto* standard data model for describing clinical research studies.

The data object schema is based on the DataCite standard (version 3.1), extended to cover the needs of clinical researchers, specifically to provide additional data covering:

* Location, ownership and access arrangements for data objects, many of which would not be immediately or publicly available, and instead require an application process, usually to the study investigator or sponsor, for access to be granted.
* Links to the generating studies. Apart from journal articles most of the data objects generated by clinical research are closely linked to the study or studies that generated them, and are usually discovered using the study’s name or identifiers.

Taken together the study and data object schemas provide a Discoverability – Access – Provenance (DAP) metadata schema. The ECRIN scheme does not attempt to provide detailed descriptive metadata, e.g. the detailed data dictionaries describing the structure of a dataset. (Such descriptive metadata files are of course themselves data objects, and ECRIN schema data could and should be used to describe them as well as the data to which they refer).

Note that the relationship between studies and data objects is many-to-many rather than one-to-one. Any system dealing with this information needs to take this into account by maintaining the data for studies and data objects separately, linking them as appropriate. Each element has to have a reference to the other element type – a study record has one or more references to linked data object records, whilst a data object includes one or more references to ‘parent’ studies.

The schemas have 42 main data points (though many of these are composite), split into six sections, A – F. Section A has 15 data points relating to study objects, while sections B - F have 27 data points relating to the data objects themselves.

Please note that this paper presents summaries of the metadata schemas and does not fully describe how the data would be stored, e.g. within databases or json files. In those contexts additional identifiers would be used to provide record keys and to link the data points. For example, in a database some form of join table would be used to link study and data object records, rather than the reference lists used in the schema.

In April 2018, this metadata schema was updated as version 2 [2], and a further version followed in February 2019 (version 2.2)[3]. Version 3.0 was developed in November 2019, [4] after extensive work with different data sources had revealed some deficiencies with the original schema, and this current version in September 2020

# Summary Tables

**The Study schema**

|  |  |  |
| --- | --- | --- |
| **Mandatory** | **Recommended** | **Optional** |
| **A. The Source Study** | | |
| **A.1 Display Title** *{display title, language code***⦁***}*  **A.8 Study Type ⦁**  **A.9 Study Status ⦁**  **A.14 Linked Data Objects** *{object identifiers}***\***  **A.15 Provenance data** | **A.2 Study Identifiers** *{identifier type***⦁***, Identifier value, source organisation, application date, identifier link}***\***  **A4. Brief Description** *{description text, ?contains html}*  **A.6 Study Features** *{feature type***⦁***, feature value***⦁***}***\***  **A.7 Study Topics** *{topic value, topic type***⦁***, topic vocabulary***⦁***, topic code}***\*** | **A3. Study Titles**  *{title text, title type***⦁***, language code***⦁***, comments}***\***  **A5. Data Sharing Statement** *{statement text, ?contains html}*  **A10. Study Enrolment Number**  **A11. Study Gender Eligibility**  **A12. Min and Max Ages** *{age, age units}*  **A13. Related Studies** *{relationship type***⦁***, target study}***\*** |

**\*** May be repeated **⦁** Categorised value

**The Data Object schema**

|  |  |  |
| --- | --- | --- |
| **Mandatory** | **Recommended** | **Optional** |
| **B. Data Object Identifiers** | | |
| **B.1 DOI**  **B.2 Display Title**  **B.6 Linked Studies\*** *{study identifiers}* | **B.3 Version** | **B.4 Object Identifiers\*** *{Identifier type***⦁***, Identifier value, source organisation, application date}*  **B.5 Object Titles\*** *{title text, title type***⦁***, language code***⦁***, comments}* |
| **C. Creators and Contributors** | | |
| **C.1 Creators\*** *{name type, person details OR organisation }* | For most data objects contributors should be the study contributors.  For journal articles creators will be authors, plus organisational study contributors  person details = given name, family name, full name, identifier, identifier scheme, affiliation, affiliation identifier, affiliation identifier scheme | **C.2 Contributors\***  *{contribution type***⦁***, name type, person details OR organisation }* |
| **D. Object Dates** | | |
| **D.1 Publication Year** |  | **D.2 Dates\***  *{date type***⦁***, Is range, date as string, start year, start month, start day, end year, end month, end day, comments}* |

**The Data Object schema (continued)**

|  |  |  |
| --- | --- | --- |
| **Mandatory** | **Recommended** | **Optional** |
| **E. Object Attributes and Descriptors** | | |
| **E.1 Class ⦁**  **E.2 Type ⦁** | **E.3 Record key type** (datasets) *{type***⦁***, text description }*  **E.4 De-identification levels** (datasets) *{type***⦁***, specific actions, details}*  **E.5 Associated consent** (datasets) *{type***⦁***, specific restrictions, details}*  **E.6 Description\*** *{description type***⦁***, label, description text, language code***⦁***, contains html?}*  **E.7 EOSC Category ⦁**  **E.8 Language\*⦁**   **E.9 Related Resources\*** *{ relationship type***⦁***, target object}* | **E.10 Topics\*** *{topic value, topic type***⦁***, topic vocabulary***⦁***, topic code}*  For most data objects topics should be the study topics. Journal articles will normally have their own listed topics |
| **F. Object Location and Access Details** | | |
| **F.1 Managing Organisation**  **F.2 Access Type ⦁**  **F.3 Access Details**  **F.4 Access Details URL** *{URL, Date last checked}*  **F.5 Resources\*** *{repository organisation, URL, URL accessible, date URL last checked, resource type***⦁***, resource size, size units**⦁, resource comments }*  **F.7 Provenance Data** | (F3 and F4 are mandatory if access is non-public) | **F.6 Rights\***  *{details, rights URI}* |

**\***May be repeated **⦁** Categorised value

# Study Attributes

Strictly speaking these data points are not metadata because they do not describe data – instead they summarise some key attributes of the study, especially those that promote its discoverability.

### A.1 Display Title (1)

This is by default, the shorter or 'public' study title. If there is no such title the full scientific or protocol title needs to be used. Whatever title is used it should also appear within the list of study titles (see A.3), where a fuller set of title attributes can be provided.

The language code indicates the language of the title using the two letter ISO language code, with default value 'en'.

### A.2 Study Identifiers (0...n)

None, one or more unique identifiers that have been assigned to the study. For studies entered into trial registries these should include, as a minimum, the registry ID(s), but any IDs that have been externally applied, and that might be useful in identifying the study, can be included, for instance funders' and / or sponsors' ids.

These IDs are composite. If provided, they must include

* the identifier value,
* the identifier type (categorised, as selected from a predetermined list),
* the assigning organisation
* (optionally), the date the identifier was assigned
* (optionally), any associated URL (for instance some public funder Ids in the US will link to a summary page about the grant and its use).

### A.3 Study Titles (0..n)

Studies usually have a short or ‘public’ title as well as a full scientific one (as used on the protocol document), and can also be described by an acronym. They may have titles in more than one language.

All titles should be included in this list. The type is composite, and should include:

* the title text,
* the title type (categorised, as selected from a predetermined list),
* the language of the title, as a 2 character ISO code,
* whether or not they contain embedded html, e.g. to indicate italics, super or subscripts (useful
* (optionally), any additional comments about their genesis (e.g. "authors' translation"),

### A.4 Brief Description (0…1)

Most study registry systems require a brief, non-specialist description of the study – which usually range from a few lines to a paragraph or two. This can be useful in assessing the relevance of studies to a particular search task and so is included in the study data points.

There should also be an indication of whether the description contains embedded html, so that display systems can interpret any tags correctly, rather than display them as 'raw' text.

### A.5 Data Sharing Statement (0..1)

In recent years several trial registries have requested study sponsors and / or leads to indicate if they will make individual participant data and related documents available for sharing, and if so how and when the data would be available. As such a statement is central to the purpose of the MDR it is captured within the study data, so that if present it can be displayed. This data point should also include an indication of whether the data sharing statement contains embedded html, so that the tags can be interpreted correctly.

### A.6 Study Features(0…n)

None, one or more design features of the study.  
The design features available will depend on whether the study is interventional or observational. Available types for interventional studies include Phase, Primary Purpose, Allocation method, Intervention Design and Masking. For observational studies the types include Observational Model, Time Perspective, and whether or not Specimens are retained.

In each case the possible values are categorised, and so restricted to a pre-defined set of values. This makes the feature types useful candidates for filtering of study records within a web portal and / or API.  
The composite study feature record is therefore

* feature type (categorised, as selected from a predetermined list), provided as a code-text pair.
* feature value, also categorised - each feature type has an associated list of options, each one available as a code-text pair.

### A.7 Study Topics (0…n)

None, one or more topic names or phrases, keywords, or classification codes describing the study or aspects of it. Topics is preferred to ‘Subjects’ because within clinical research ‘Study subjects’ is normally understood as referring to the study participants.

In the context of clinical research, most data objects – which will not have listed topics associated with them – would, for purposes of discoverability, take on the topics or keywords associated with their parent study. (The exception is journal articles, which almost always do have linked keywords).

The listed topics could be free text, but in many cases the text is structured, i.e. selected from a controlled vocabulary. There are a variety of such controlled vocabularies available (MESH, ICD 10, MedDRA, SnoMed CT etc.). In many such schemes the controlled term is associated with a code. Either topic name or code can be provided, but preferably both should be supplied.

Topics are also of a certain ‘type’ – determining the domain in which they apply, e.g. ‘condition’, ‘organism’, ‘chemical / biological agent’, ‘geographic’ etc.

The composite study topic record is therefore

* topic type (categorised, as selected from a predetermined list),
* topic value, the keyword or topic name
* (if applicable) topic vocabulary (categorised, as selected from a predetermined list),
* (if applicable) topic code in the vocabulary system.

### A.8 Study Type (1)

This is a single term representing – in very broad terms – the type of clinical research study, e.g. ‘interventional’ (= clinical trial), ‘observational’, ‘expanded access’. It is categorised and must be selected from a predefined list. It is included as an aid to filtering records.

### A.9 Study Status (1)

This is a single term representing the current status of the study in terms of its life-cycle, e.g. ‘not yet recruiting’, ‘recruiting’, ‘completed’, ‘terminated (early)’. It is categorised and must be selected from a predefined list. It is included as an aid to filtering records.

### A.10 Study Enrolment Number (0..1)

This is an integer representing the anticipated or actual number of study participants.

### A.11 Study Gender Eligibility (0..1)

This is a code / text pair that indicates whether the study is only open to male or female participants, or both.

### A.12 Study Minimum and Maximum ages (0..1)

These are integers representing the minimum and maximum age criteria for study participants, where they exist. In each case they are associated with a term indicating the time units associated with the integer. This is usually 'Years', but, for example for paediatric studies, may be months or weeks, or even days or hours.

### A.13 Related Studies (0..n)

Studies can have relationships between themselves, for instance one study can be a feasibility study for a later one, or a study can represent an ‘expanded access’ version of a clinical trial (when a new drug is available for compassionate reasons, even though recipients fail eligibility criteria for the study, and it use is reported on a case by case basis), or one study can represent a continuation of another, in an ongoing series. This data can be useful for tracking related studies and their data objects and so is included in the metadata scheme. It is composite, with

* the relationship type (categorised, as selected from a predetermined list)
* the identifier of the other or ‘target’ study (in a suitable system).

### A.14 Linked Data Objects (1..n)

The linked data objects (there should be at least one, representing the entry in a trial registry system) are listed as object identifiers, usually accession Ids within an appropriate database system.

### A.15 Provenance Data (1)

A string indicating the source or sources of the data (usually a trial registry) and the date-times on which the data was last downloaded from the source or sources.

# Data Object identifiers

### B.1 Data object identifier (0..1)

In line with the DataCite specification the principal identifier for data objects is seen as a Digital Object identifier or DOI, providing a persistent identifier that can be cited in other contexts. This applies to any objects that are available to others (whether publicly or under managed access).

Unfortunately a large number (the majority) of clinical research data objects, apart from journal articles, do not currently have a DOI. It remains to be seen if this situation improves. If it does not, consideration should be given to a mechanism for minting and applying one – if financially feasible and acceptable to the object creators – or alternative identifiers should be explored, especially if a resolvable URL exists which could be used to immediately linked to the resource.. The extent of this problem needs to be clarified.

### B.2 Display Title (1)

A title for the object. For a journal article it would be a citation of the article in a standard format (up to 3 authors, title, source journal information). For many other data objects the display title would need to be constructed from the study name followed by the object title or type, because in general such objects do not have unique names. In many situations the study name prefix could be dropped as it would be clear from the context (e.g. the study name would be a heading to the list of data objects). The study name and object type or name should therefore be separated by a clear indicator (‘ :: ’ is used within the MDR) so that if and when necessary the two parts of a composite title can be displayed separately.

### B.3 Version (0..1)

The version of the data object, in whatever notation was used by the original data object creators. Many versions of a particular dataset or document may have been created in the course of a clinical study, though only the version or versions that are made available for sharing are important in this context. The normal expectation would be that the final version of a data object (e.g. a protocol) would be the one that was shared with others.

In some cases multiple versions of the same document or dataset could be made available, or they might be specifically requested. Assuming the data objects have similar names, they will therefore need to be clearly differentiated using version codes (and relevant dates – see D.2 – and possibly descriptions – see E.6). E.8 describes how the relationship to previous or next versions can be made explicit. If multiple versions of the same dataset are available to access the version attribute should be completed and displayed with the name and other identifiers.

### B.4 Object Identifiers (0...n)

This refers to other unique identifiers that have been assigned to the data object in addition to its DOI primary identifier (for instance, for journal articles, a PubMed id). As with studies such IDs would be composite and include:

* the identifier value,
* the identifier type (categorised, as selected from a predetermined list),
* the assigning organisation
* (optionally), the date the identifier was assigned

### B.5. Object Titles (0...n)

The full description of the title(s) for the data object. In most cases there will only be one (the constructed display title), but journal papers may have titles in different languages, and in any case will be different from the display title (which is a full citation). The title description is composite , and should include

* the title text,
* the title type (categorised, as selected from a predetermined list),
* the language of the title, as a 2 character ISO code,
* (optionally), any additional comments about their genesis (e.g. "authors' translation"),

### B.6. Linked Studies (1...n)

The linked studies (there should be at least one, or the data object should not be included in the system) are listed as study identifiers, usually accession Ids within an appropriate database system.

# Creators and Contributors

### C.1 Creators (1...n)

The main personnel involved in producing the data, or the authors of a publication. It may be a set of institutional and / or personal names. Each creator description, which is composite, therefore needs to indicate whether or not it refers to an individual or an organisation (or collaboration). If it is a person then fields are available for names, identifiers, and affiliation details. If not the organisation name needs to be provided. The composite structure (which follows closely that in DataCite) is therefore

* whether an individual or not,

and if they are, the…

* + given name,
  + family name,
  + full name,
  + identifier, (ORCID id if available)
  + affiliation, (string description of department / organisation)
  + affiliation identifier, (if the organisation has a formal identifier)
  + affiliation identifier scheme (e.g. ‘ISNI’, ‘RINGGOLD’)

but if they are not, the….

* + organisation name

Most data objects, other than journal articles, are unlikely to have creators explicitly identified.

### C.2 Contributors (0...n)

From DataCite, contributors are “other institutions and / or persons responsible for collecting, managing, distributing, or otherwise contributing to the development of the data object.” A contributor record is composite and is essentially the same as that for creators, except that each needs to be prefixed with an indicator of

* contributor type (categorised, as selected from a predetermined list)

The types available include those defined within DataCite, but the list has been extended in the context of clinical research, to include (for example) trial sponsor, trial funder, device provider, central laboratory, public contact, study lead, (site) principal Investigator.

In general the contributor lists for data objects should be derived from the lists stored for their parent study, even though in the system contributors are only presented for data objects, not studies. These will usually include the study lead(s) and sponsors. Where data objects do record their contributors, they would normally take precedence over that of the study, but organisational contributors, in particular study sponsors and funders, should still be added to the data object list.

Any system retrieving creator / contributor data therefore needs to do so for the studies as well as for the data objects themselves.

# Dates

### D.1 Publication year (1)

The year in which the object is made available, i.e. in which it first becomes citable, expressed as 4 digits. Not the same as when an object becomes public – ‘available’ simply means that it can be accessed, but the conditions of that access remain in the control of the object’s owners or controllers, nor necessarily the year in which it was created (which may be present as one of the object’s dates).

### D.2 Dates (0...n)

None, one or more dates or date ranges that are relevant to the data object. It is composite and includes both string and integer representations of the date. Year, month and day data is held separately to make it easier to apply date filters when finding data objects. The elements of the composite record are:

* date type (categorised, as selected from a predetermined list),
* is range, whether or not it is a single date or a range,
* date as string, in a standard format yyyy MMM dd, e.g. “2018 Dec 12”, or “2012 Mar 7”
* start year, an integer
* start month, an integer – may not be present for partial dates
* start day, an integer – may not be present for partial dates
* end year, for date ranges only
* end month, for date ranges only, may not be present if date range partial
* end day, for date ranges only, may not be present if date range partial
* comments – any relevant / explanatory comments

# Data Object Attributes

This section is mainly based on the DataCite metadata specification, though a few extensions have been added for datasets (as opposed to document based data objects).

### E.1 Class (1)

A categorised value, one of the existing DataCite controlled list for ‘Resource Type General’. In most cases, for clinical research data objects, the class will usually be one of:

* Text
* Dataset

though other options include: Data Paper, Software, Service, Audiovisual, and Interactive Resource.

### E.2 Type (1)

A categorised description of the type of data object, at a more specific level than Class. The type and class should form a pair (as with DataCite), e.g. Dataset/census data or Text/conference abstract.

Unlike DataCite, both class and type are mandatory in the ECRIN schema. The types available include the CASRAI classifications of document objects, recommended by DataCite, together with additions to the list that represent object types of particular importance to clinical research (e.g. protocols, clinical study reports, statistical analysis plans, and datasets of various kinds).

### E.3 Record key type (1, Datasets only)

This is a composite item that indicates the type of record keys used within the dataset, and in particular if it is pseudonymised or anonymised. This is important for determining the legislation that might apply to the data set’s use. The contents are

* Record key type (categorised, as selected from a predetermined list)
* Details – text description to elaborate / clarify details

Note that the categorisation into 'pseudonymised'. 'anonymised', 'identifiable' etc. is based upon the description provided by the data controller / manager in the data (if one is supplied). The classification is therefore based on the data controller's understanding of the relevant terms. No attempt is made to apply a categorisation using standard criteria, as the meaning of the words used ('pseudonymised'. 'anonymised', etc.) may vary between different legal jurisdictions, over time, and in different usage contexts. The categorisation should therefore be read as only a very approximate guide to any legal requirements associated with the data.

### E.4 De-identification level (1, Datasets only)

An item that indicates the amount of de-identification that has been applied to the dataset. The item consists of :

* De-identification level (categorised, as selected from a predetermined list)
* Additional actions carried out - boolean data indicating if any of the following applies: a) direct identifiers have been removed, b) US HIPAA rules for de-identification have been applied, c) dates have been rebased or replaced with integers, d) narrative text fields have been removed , and e) k-anonymisation has been carried out.
* Details – text description to elaborate / clarify details.

### E.5 Associated consent (1, Datasets only)

The consent in question is for secondary use of the data - consent for primary use is assumed.  
The data item consists of:

* a coded field that indicates the range of application of consent (if any) available for re-use and sharing associated with the data, selected from a list.
* Possible additional restrictions, represented as a series of boolean data points: a) if use is limited to non-commercial research, b) if there any geographical restrictions on re-use, c) if only certain types of research are permitted, d) if only genetic research is allowed, and f) whether or not methodological or tool research (e.g. developing machine learning algorithms) is allowed.,
* Details – text description to elaborate / clarify details, in particular to expand upon any of the additional restrictions listed as being present.

### E.6 Description (0...n)

None, one or more pieces of additional general information about the data object, so far as that is publicly available (journal abstracts, although an obvious ‘descriptor’, remain the property of the publisher and cannot in general be reproduced within the system). The item is composite, consisting of:

* description type (categorised, as selected from a predetermined list)
* label, a heading that might be applied to the text
* description text, the description itself
* language code, the 2 character ISO code
* contains html?, useful to know for display purposes

### E.7 EOSC Category (0..1)

An integer (0, 1, 2 or 3) that conforms to an EOSC categorisation recommended for data objects. The classification is

0 = Non-personal data. Contains no information that refers to any identified or identifiable living individual.

1 = Anonymised data.

2 = Pseudonymised data.

3 = Sensitive pseudonymised data.

In general, almost all documents expected in the MDR will be categorised as 0, whilst all IPD datasets will be categorised as 3 - unless there is general agreement that they are fully anonymised, in which case they become 1.

### E.8 Language (1..n)

The language or languages of the data object itself (not of a description of the object), using the ISO language codes (e.g. en, de, fr). DataCite assumes a single language but some clinical research data objects (e.g. journal articles) are created in two or more languages. The record may therefore be multiple.

### E.9 Related Objects (0..n)

Data objects can be related to each other – for example one object can be a supplement to another, or a new version of an other, or be derived from, or the source of, one or more other data objects.

A particularly important relationship for clinical study data is the pairing of ‘Has Metadata’ and ‘Is Metadata for’. Metadata in clinical research can include, for example, a data dictionary that provides the metadata for a dataset. Note that the metadata in this context is itself a file, and a data object in its own right. Each record is composite and must include:

* the relationship type (categorised, as selected from a predetermined list)
* the identifier of the other or ‘target’ data object (in a suitable system).

To keep things simpler within the MDR, the requirement is that any related resource must also be indexed within the MDR. This allows the identifier to be an internal identifier within the MDR system, making navigation to it much simpler.

### E.10 Topic (0...n)

None, one or more topic names or phrases, keywords, or classification codes describing the object or aspects of it. In the context of clinical research, most data objects will not have listed topics associated with them (the exception is journal articles, which almost always do have linked keywords). Data Object topics for non journal articles should be those of the parent study or studies. This introduces a substantial amount of redundancy but it means that topics can be searched and used for filtering across all (rather than just some) data objects.

The structure of the each topic item is exactly the same as for study topics:

* topic type (categorised, as selected from a predetermined list),
* topic value, the keyword or topic name
* if applicable) topic vocabulary (categorised, as selected from a predetermined list),
* (if applicable) topic code in the vocabulary system.

# Location and Access details

An area where the existing DataCite schema needs to be extended is in providing a full description of the access arrangements for any data object. The following data points are proposed.

### F.1 Managing Organisation (1)

In this schema, this is the organisation that manages access to the document or data object, including making the overall decision about access type (see F.2). For data this would usually be the name of the organisation that was the data controller. For journal papers it would be the name of the company that publishes the journal, and which would normally run the primary web site on which it can be accessed.

### F.2 Access Type (1)

A categorised value that represents in broad terms the type of access under which the object is available, for example by publicly available download, or restricted download (restricted to members of a specific group) or on screen access after review on a case by case basis.

### F.3 Access Details (Mandatory for any of the non-public access types)

A textual summary of the access being offered, for example identifying the groups to which access is granted, the criteria on which a case-by-case decision would be based, any further restrictions on on-screen access, etc. It may reference the access details URL or other web based resources, on the object manager’s web site or elsewhere.

### F.4 Access Details URL (Mandatory for any of the non-public access types)

A link to a resource that explains how access may be gained, e.g. how a group can be joined, and / or how application can be made for access on an individual basis. This would normally be a link to a web page on the managing organisation’s site, that would explain access procedures or provide an application proforma.

The item is composite, and should include a date representing the last time the URL was checked to be valid (i.e. returned a 200 ‘success’ code rather than a 404).

### F.5 Resources (Mandatory unless case-by-case access)

The web based resources that represent this data object. Mandatory for public objects, when at least one resource should be listed. For data objects simply listed as existing, but under managed access, this information may not be available for harvesting. Each record is composite and includes

* the name of the organisation holding the resource
* the resource type (categorised, for downloadable resources normally based on the file extension)
* the resource URL
* whether or not the resource is directly accessible (i.e. is public and not behind a pay wall)
* the date the URL was last checked as valid

and, if downloadable, the

* the resource size,
* the resource size units, usually in KB, MB or GB.

In addition...

* resource comments, provides a free text field to hold further details of the resource, in particular to support machine processing. These could include the schema used for XML files, and / or the character coding used for text files (e.g. UTF-8 versus UTF-16) or the presence and types of any byte order marks.

### F.6 Rights (0..n)

Any intellectual property rights information for the data object, as a textual statement of the rights management associated with the resource. The item is composite, and should include the URI for the specific rights management scheme as well as a textual description.

### F.7 Provenance Data (1)

A string indicating the source or sources of the data and the date-times on which the data was last downloaded from the source or sources.

# Categorised Items

A variety of items in the schema are categorised, forcing a selection from a list of predefined values. The values available are listed in the appendix to this paper. The lists (or at least those exposed to external clients) cover:

|  |  |
| --- | --- |
| * Study types * Study statuses * Study gender eligibility types * Study features (types and categories) * Study relationship types * Object classes * Object types * Object filter types * Object access types * Object instance types * Object relationship types * Contribution types | * Dataset consent types * Dataset de-identification levels * Dataset record-key types * Date types * Description types * Identifier types * Language codes * Resource types * Time units * Title types * Topic types * Topic vocabularies |

# JSON schemas

Both schemas are also available defined as json schemas:

<https://ecrin-mdr.online/index.php/Study_JSON_v4>

<https://ecrin-mdr.online/index.php/Data_Object_JSON_v4>

# References

[1] Canham S & Ohmann C. A metadata schema for data objects in clinical research. Trials, volume 17, Article number: 557 (2016)

[2] ECRIN Clinical Research Metadata Schema Version 2 (April 2018). DOI: 10.5281/zenodo.1312539. Available at https://zenodo.org/record/1312539#.XefpFr62Kuc

[3] ECRIN Clinical Research Metadata Schema Version 2.2 (February 2019) DOI: 10.5281/zenodo.3534313. Available at <https://zenodo.org/record/3534313#.XefpZr62Kuc>

[4] ECRIN Clinical Research Metadata Schema Version 3 (November 2019) DOI: 10.5281/zenodo.3562911. Available at <https://zenodo.org/record/3562911>