ECRIN Metadata Schemas for Clinical Research, Version 8 (September 2023)

Steve Canham, 21/09/2023

Contents

[Introduction 3](#_Toc146232041)

[Interpreting the tables and Lists 4](#_Toc146232042)

[Summary Tables 5](#_Toc146232043)

[Study Attributes 9](#_Toc146232044)

[A.1 Display Title (1) 9](#_Toc146232045)

[A.2 Study Identifiers (0...n) 9](#_Toc146232046)

[A.3 Study Titles (0..n) 9](#_Toc146232047)

[A.4 Brief Description (0…1) 9](#_Toc146232048)

[A.5 Data Sharing Statement (0..1) 9](#_Toc146232049)

[A.6 Study Features (0…n) 9](#_Toc146232050)

[A.7 Study Topics (0…n) 10](#_Toc146232051)

[A.8 Study Type (1) 10](#_Toc146232052)

[A.9 Study Status (1) 10](#_Toc146232053)

[A.10 Study Enrolment (0..1) 10](#_Toc146232054)

[A.11 Study Gender Eligibility (0..1) 11](#_Toc146232055)

[A.12 Study Minimum and Maximum ages (0..1) 11](#_Toc146232056)

[A.13 Study Inter-Relationships (0..n) 11](#_Toc146232057)

[A.14 Linked Data Objects (1..n) 11](#_Toc146232058)

[A.15 Provenance String (1) 11](#_Toc146232059)

[A.16 Study Countries (0..n) 11](#_Toc146232060)

[A.17 Study Sites (0..n) 11](#_Toc146232061)

[A.18 Study Start Time (0..1) 12](#_Toc146232062)

[A.20 Study Conditions (0..n) 12](#_Toc146232063)

[A21. Study ICD (0..n) 12](#_Toc146232064)

[A.22 Study People (0..n) 12](#_Toc146232065)

[A23. Study Organisations (0..n) 12](#_Toc146232066)

[A.24 Study IEC (0..n) 13](#_Toc146232067)

[A25. Study IEC level (1) 13](#_Toc146232068)

[Data Object identifiers 14](#_Toc146232069)

[B.1 Data object identifier (0..1) 14](#_Toc146232070)

[B.2 Display Title (1) 14](#_Toc146232071)

[B.3 Version (0..1) 14](#_Toc146232072)

[B.4 Object Identifiers (0...n) 14](#_Toc146232073)

[B.5. Object Titles (0...n) 14](#_Toc146232074)

[B.6. Linked Studies (1...n) 15](#_Toc146232075)

[B.7 Provenance String (1) 15](#_Toc146232076)

[Creators and Contributors 15](#_Toc146232077)

[C.3 Object People (0..n) 15](#_Toc146232078)

[C.4. Study Organisations (0..n) 15](#_Toc146232079)

[Dates 15](#_Toc146232080)

[D.1 Publication year (0..1) 15](#_Toc146232081)

[D.2 Dates (0...n) 15](#_Toc146232082)

[Data Object Attributes 16](#_Toc146232083)

[E.1 Class (1) 16](#_Toc146232084)

[E.2 Type (1) 16](#_Toc146232085)

[E.3 Record key type (0..1, applies to datasets only) 16](#_Toc146232086)

[E.4 De-identification level (0..1, applies to datasets only) 16](#_Toc146232087)

[E.5 Associated consent (0..1, applies to datasets only) 17](#_Toc146232088)

[E.6 Description (0...n) 17](#_Toc146232089)

[E.7 EOSC Category (0..1) 17](#_Toc146232090)

[E.8 Language (1) 17](#_Toc146232091)

[E.9 Object Inter-Relationships (0..n) 17](#_Toc146232092)

[E.10 Topic (0...n) 18](#_Toc146232093)

[Location and Access details 18](#_Toc146232094)

[F.1 Managing Organisation (1) 18](#_Toc146232095)

[F.2 Access Type (1) 18](#_Toc146232096)

[F.3 Access Details (should be present for any of the non-public access types) 18](#_Toc146232097)

[F.4 Resources (should be present if access is not managed) 19](#_Toc146232098)

[F.5 Rights (0..n) 19](#_Toc146232099)

[Categorised Items 19](#_Toc146232100)

[MDR Implementation: JSON schemas 20](#_Toc146232101)

[References 21](#_Toc146232102)

# 

# Introduction

This paper presents summaries of the ECRIN metadata schemas, firstly in a concise tabular form, and then in the form of lists of the main data points that go into a little more detail about each data point.

This schema for clinical research 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, the largest trial registry in the world, with about 466,000 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 metadata schema focusing on discoverability, access and provenance (DAP). The ECRIN scheme does not attempt to cover *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 structure their DAP metadata 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-many. 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.

There are, altogether, 24 main data points relating to studies in the current version of the schema, and 26 main points relating to data objects. Very many of these are composite, however, and are themselves made up of 2 to 6 different attributes.

In April 2018, the 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 slightly revised versions 4.0 [5] and 5.0 [6] were produced in September and October 2020. Version 6.0 [7] was created in August 2021. Versions 7.0 and this current version, v8, were created as part of an extensive redesign, extension and rewrite of the MDR system in late 2022 and early 2023.

# Interpreting the tables and Lists

While this paper presents summaries of the metadata schemas it 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 addition, the records representing the various multiple-instance entities (titles, identifiers, topics etc.) would need a foreign key linking them back to the parent study or object, as well as identifier fields (primary keys) of their own.

Appendix 1 provides more details on the data points required in any practical implementation of the schema, in the form of JSON definition files for both studies and data objects.

Some metadata schemas (e.g. DataCite) divide their data points up into ‘Mandatory’, ‘Recommended’. ‘Optional’, as a guide to those applying the schema. The ECRIN schema, however, is applied to metadata that has already been created by other agencies, for example trial registries. There is almost no generation of schema data *de novo*, so ‘Mandatory’ and similar terms have little relevance. The process of applying the schema is one of matching as many as possible of the listed data points to the ‘source’ metadata.

The study data points are presented in groups related to their main function, although in practice these groups overlap. Object data points are grouped according to their purpose. The order of data points in the tables does not imply any relative importance, nor does it indicate the ways data points might be grouped and arranged in a practical implementation. The numbering of the data points (A.1, E.3, etc) is at least partly a function of the order in which they were added to the schema, and is retained only to support comparison with earlier schema versions. It has no significance in itself.

**Symbols used for Data Point Type:**

**T** indicates a text string

**N** indicates an integer

**C** indicates a composite point – the structure of which is described in the description / comments column

**#**indicates that the data point is categorised, i.e. uses a controlled vocabulary***.*** Such pointsmay be represented in data as a code, or a textual decode, or both. In some cases, such as MeSH or Geonames, these controlled vocabularies are external to the MDR, in others (e.g. topic type, identifier type) they have been developed as look up tables within the MDR.

**##** indicates organisation data, that includes a name, usually also an MDR Id, and possibly a ROR id.

**Abbreviations used:**

CT = Controlled Terminology / Vocabulary

GN = Geonames

MeSH = Medical Subject Headings (as used by the US National Library of Medicine)

ICD = International Classification of Diseases (version 11)

ROR = Research Organisation Registry

# Summary Tables

**The Study schema**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Findability Metadata: D**ata points that are important for searching for and identifying studies that meet particular criteria. Machine based searches would be the normal approach. | | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** | |
| A.1 Display Title | **T** | 1 | Title text from the Study Title designated as default | |
| A.2 Study Identifiers | **C** | 1..n | *Composite*, as {identifier type **#** , Identifier value, organisation **##**, date, url link} | |
| A.3 Study Titles | **C** | 1..n | *Composite*, as { title type **#,** title text, language code **#**, comments} | |
| A.8 Study Type | **#** | 0..1 |  | |
| A.9 Study Status | **#** | 0..1 |  | |
| A.6 Study Features | **C** | 0..n | *Composite*, as *{feature type***#***, feature value***#***}* | |
| A.7 Study Topics | **C** | 0..n | *Composite*, as { topic type**#,**  original value, original CT **#**, original CT code **#**, MeSH value **#** } | |
| A.18 Study Start Time | **C** | 0..1 | *Composite*, as {year , month} |
| A.20 Study Conditions | **C** | 0..n | *Composite*, as { original value, original CT **#**, original CT code **#**} | |
| A.21 Study ICD | **#** | 0..n | ICD value (4 character stem code) | |
| **Descriptive Metadata:** Data points that are useful in further characterising and selecting studies. Some points would allow machine based searching - others would be more normally examined manually. | | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| A.4 Brief Description | **T** | 1 | May need to be constructed from stated aims, hypotheses, endpoints in some cases. |
| A.5 Data Sharing Statement | **T** | 0..1 |  |
| A.10 Study Enrolment | **T** | 0..1 | A string rather than a number, to allow more complete information |
| A.11 Study Gender Eligibility | **#** | 0..1 |  |
| A.12 Min and Max Ages | **C** | 0..2 | *Composite*, as {age, age units **#**} |
| A.16 Study Countries | **C** | 0..n | *Composite*, as {Geonames County Value **#**, Recruitment status **#**} |
| A.17 Study Sites | **C** | 0..n | *Composite*, as {Facility **##,** GN City Value **#,** GN County Value **#**, Recruitment status **#** } |

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| A.22 Study People | **C** | 0..n | *Composite*, as {contribution type **#**, full name, ORCID id, affiliation, affiliation organisation **##**} |
| A.23 Study Organisations | **C** | 0..n | *Composite*, as {{contribution type **#**, organisation **##**} |
| A.24 Study IEC | **C** | 0..n | *Composite*, as {source id **#,** study registry id, sequence number, statement type **#**, split method **#**, leader, indent level, sequence string, criterion text} |
| A.25 Study IEC Level | **#** | 1 |  |
| **Linkage Metadata:** These data pointsprovide information on a study’s relationships with other studies, with data objects, and with the MDR respectively. | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| A.13 Inter-study relationships | **C** | 0..n | *Composite*, as {relationship type**#** , target study id} |
| A.14 Linked Data Objects | **N** | 1..n | *Integer object identifiers* |
| A.15 Provenance string | **T** | 1 |  |

**The Data Object schema**

|  |  |  |  |
| --- | --- | --- | --- |
| **B. Data Object Identifiers** | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| B.1 DOI | **T** | 0..1 | Apart from journal articles, relatively uncommon for the data objects in the MDR, even though considered mandatory by DataCite |
| B.2 Display Title | **T** | 1 | Apart from journal articles, often has to be constructed as an amalgam of parent study name and object type |
| B.3 Version | **T** | 0..1 | Relatively uncommon |
| B.4 Object Identifiers | **C** | 0..n | *Composite*, as {identifier type **#** , Identifier value, organisation **##**, date} |
| B.5 Object Titles | **C** | 0..n | *Composite*, as {title type**#** , title text, language code**#** , default?, comments} |
| B.6 Linked Studies | **N** | 1..n | Integer study identifiers |
| B.7 Provenance string | **T** | 1 |  |
| **C. Creators and Contributors** | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| C.3 Object People | **C** | 0..n | *Composite*, as {contribution type **#**, full name, ORCID id, affiliation, affiliation organisation **##**} |
| C.4 Object Organisations | **C** | 0..n | *Composite*, as {{contribution type **#**, organisation **##**}  Optionally, people and organisations linked to the parent study can be linked to the data objects associated with that study.  People and Organisation data can be optionally combined, e.g. on data export, to form generic ‘contributor’ records. |
| **D. Object Dates** | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| D.1 Publication Year | **N** | 0..1 |  |
| D.2 Dates | **C** | 0..n | *Composite*, as {{date type **#**, Is range, date as string, start year, start month, start day, end year, end month, end day, comments} |

|  |  |  |  |
| --- | --- | --- | --- |
| **E. Object Attributes and Descriptors** | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| E.1 Class | **#** | 1 | Broad categorisation of object / medium types |
| E.2 Type | **#** | 1 | Sub-types defined within each class |
| E.3 Record key type | **C** | 0..1 | For datasets only. *Composite*, as {*type***#***, text* description} |
| E.4 De-identification levels | **C** | 0..1 | For datasets only. *Composite*, as {type**#** , specific techniques used? (x5 Booleans), textdescription} |
| E.5 Associated consent | **C** | 0..1 | For datasets only. Composite, as {type**#** , specific conditions apply? (x5 Booleans), text description} |
| E.6 Descriptions | **C** | 0..n | *Composite*, as {description type**#**, label, description text, language code*}* |
| E.7 EOSC category | **#** | 1 |  |
| E.8 Language Code | **#** | 1 | Almost always one language code, very rarely a comma separated list |
| E.9 Object Inter-Relationships | **C** | 0..n | *Composite*, as {relationship type**#** , target study id}  None of this data present in the source and therefore none in the MDR |
| E.10 Object Topics | **C** | 0..n | *Composite*, as {topic type**#,**  original value, original CT **#**, original CT code **#**, MeSH value **#** }  Optionally, topics linked to the parent study can be linked to the data objects associated with that study. |
| **F. Object Location and Access Details** | | | |
| **Data Points** | **Type** | **Mult** | **Description / Comments** |
| F.1 Managing Organisation | **C** | 1 | organisation **##** |
| F.2 Access Type | **#** | 1 |  |
| F.3 Access Details | **C** | 0..1 | *Composite*, as {description, url of details, date url last checked} (should be applied if access is non-public) |
| F.4 Physical Resources | **C** | 0..n | *Composite*, as {repository organisation **##**, resource URL, URL accessible?, date url last checked, resource type**#** , resource size, size units, comments } |
| F.5 Rights | **C** | 0..n | *Composite*, as {description, url of details, date url last checked}(F3 is mandatory if access is non-public)  None of this data present in the source and therefore none in the MDR |

# 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.

### 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 type (categorised, as selected from a predetermined list of code-text pairs),
* the identifier value,
* the assigning organisation (name and where available the organisation's Id(s) - e.g. ROR Id, ECRIN MDR Id).
* (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 type (categorised, as selected from a predetermined list of code-text pairs),
* the title text,
* the language of the title, as a 2 character ISO code,
* (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. If not present in the source data the MDR constructs such a description by concatenating descriptions of study hypotheses, endpoints and study design features.

### 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.

### 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 code-text 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 code-text, as selected from a predetermined list.
* feature value, also categorised code-text. Each feature type has an associated list of options.

### 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.

The topics can be free text, but in many cases the text is structured, i.e. selected from a controlled vocabulary. The vocabulary that is used the most – by a large margin – is the MESH code system developed by the US Library of Medicine. This is because MESH codes are applied to both PubMed records and ClinicalTrials.gov trial registry entries. MedDRA and ICD10 are also used by some sources but in relatively small amounts. To try and provide a more consistent coding scheme for topics non coded terms are also matched, wherever possible, to MESH terms. Further work is required to reduce the proportion of non-coded items.

The study topic record is composite and has the following structure:

* topic type (categorised, as selected from a predetermined list of code-text pairs. Topic types include, ‘organism’, ‘chemical / biological agent’, and ‘geographic’.
* the original value, i.e. the topic as originally expressed.
* If present the original code from a controlled terminology (CT) or vocabulary, as well as an identifier for the CT itself.
* the MESH code and term, if present

Note that conditions are now treated separately from other topic types.

In the context of clinical research, most data objects – which will not have listed topics associated with them – could, 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). In the MDR at the moment, however, searching by topic is done only on *study topics*, not object topics.

### 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 of code-text pairs. 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 (0..1)

This is a string representing the anticipated or actual number of study participants. Usually a simple number but may be a short sentence providing enrolment details (e.g. for different sub-protocols).

### 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 categorised 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 Study Inter-Relationships (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 its 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 code-text list)
* the identifier of the other or ‘target’ study (within a suitable system, normally the same system in which the ‘subject’ study is found).

### 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. Normally this would be the same system (e.g. the ECRIN MDR) in which the study record is found.

### A.15 Provenance String (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.

### A.16 Study Countries (0..n)

One or more countries where recruitment took place for the study. It is composite, and includes a Geonames Id / Name for each country, as well as a coded value representing recruitment status (ongoing, completed, etc.) although the latter is not always present in the study.

In the MDR database the country name fields are preserved in the data because of the very small percentage of country names that cannot be directly Geonames coded (usually because they do not refer to countries but to other geographic regions).

### A.17 Study Sites (0..n)

One or more clinical sites, usually hospitals, where recruitment took place for the study. The data is only available for some sources. It is composite, and includes the site or ‘Facility’ details, as a three part organisation record (name, MDR Id, ROR Id), a Geonames Id / Name for the city in which the site is located, a Geonames Id / Name for the country, and a coded value representing recruitment status (ongoing, completed, etc.) although the latter is not always present in the study.

In the MDR database the city and country name fields are preserved in the data because of the very small percentage of such names that cannot be directly Geonames coded.

### A.18 Study Start Time (0..1)

Two integers that indicate the year (as 4 digits) and the month when the study began, usually defined as ‘first-patient-first-visit’.

*N.B. There is now no A.19. This did refer to Study Contributors, but these have been replaced by Study People and Study Organisations.*

### A.20 Study Conditions (0..n)

These are the subset of study topics that deals specifically with the conditions, or diagnoses, with which the study is concerned, for example for which a new treatment is being tested. They are treated differently in the system because they are coded, where possible, against ICD rather than MeSH.

In the data as exported, however, the ICD codes are split off (see A21., below) and the study conditions data is essentially a list of the original terms used, plus any coded equivalent. More formally the data is composite, and consists of

* the original value, i.e. the condition as originally expressed.
* If present the original code from a controlled terminology (CT) or vocabulary, as well as an identifier for the CT itself.

### A21. Study ICD (0..n)

This is data derived from the conditions data described above. It lists the ICD 11 codes / terms of the conditions linked to the study. It is generated separately from A.20 because often several of the conditions as originally listed match a single ICD 11 code. (The codes used are the 4 character ‘stem’ codes which each have a relatively wide scope, and so which may subsume several more granular condition descriptions.)

The data is essentially a list of ICD11 values. It is prepared with the eventual aim of using it as the basis of searching using an ICD browser.

### A.22 Study People (0..n)

This lists the people involved with the study in various roles (e.g. as study lead, as a public contact). Each person record is composite and includes:

* the contribution type, (categorised, as selected from a predetermined code-text list),
* the person’s full name,
* the person’s ORCID identifier, if they have one,
* the ‘affiliation string’ as presented in the original data,
* the affiliation organisation details (name, MDR Id, ROR Id) if that can be derived form the affiliation string.

### A23. Study Organisations (0..n)

This lists the organisation involved with the study in various roles (e.g. as sponsor or funder). Each record is composite and includes:

* the contribution type, (categorised, as selected from a predetermined code-text list)
* the organisation details( name, MDR Id, ROR Id)

### A.24 Study IEC (0..n)

The inclusion / exclusion criteria (IEC) statements associated with a study. The data is usually originally presented as one or two blocks of text in a trial registry entry, and the MDR attempts to identify the individual statements within that text.

Unusually for the schema’s study data, where the identity of the parent study is implied, the source registry and trial id are part of the IEC data record. This is because IEC data, for a variety of reasons, is not aggregated across different versions of the same study’s data (as is the case with all other parts of the study data). IEC data for those 30-40,000 studies that are registered in more than one registry may therefore be present in the system twice or more times. Each registry’s version of the study’s IEC must therefore be identifiable.

Each record is composite and consists of

* the source id, for the record being used as the source of IEC data, usually a trial registry (categorised, as selected from a predetermined code-text list)
* The Id of the source study, within the registry being used as the source of IEC data, often known as the ‘trial id’.
* A sequence number for each statement, in relation to the full set of criteria statements for that study
* A coded value indicating the type of statement – e.g. inclusion or exclusion criterion, header, supplementary statement.
* A coded value representing the way in which the criterion was split from the block of text in which it was embedded.
* Any identified leading characters, such as numerals, letters, bullets
* A number representing the logical indentation level of the statement, a lists often embed sub-lists and even sub-sub-lists within themselves.
* A ‘sequence string’ which combines data on the indentation level and sequence within the current level, which acts as an overall guide to writing the criterion out in list form.
* the text of the criterion.

### A25. Study IEC level (1)

This is a coded integer which indicates the amount and granularity of IEC data available. It ranges from 0 (no IEC present) to 10 (multiple statements available for both inclusion and exclusion criteria).

# Data Object identifiers

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

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

Unfortunately the great majority of clinical research data objects, apart from journal articles, do *not* currently have a DOI. If this situation does not improve, consideration may need to be given to a mechanism for minting and applying DOIs – if financially feasible and acceptable to the object creators – or alternative identifiers should be explored, especially if a persistent, resolvable URL exists which could be used to immediately linked to the resource.

### 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 needs 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 some 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 are 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 but 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.

Although it is relatively rare for more than one version of a data object to be made available, if that is the case they should 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 a version item exists it should be displayed with the name and other identifiers.

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

This refers to any unique identifiers that have been assigned to the data object in addition to a DOI (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 of code-text pairs),
* the assigning organisation (name and where available Id(s) - e.g. ROR Id, ECRIN MDR Id).
* (optionally), the date the identifier was assigned.

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

A description of any real, rather than constructed, title(s) for the data object. For most web based objects there will not be a title, but some documents have names (even if often generic) and journal papers will have a name, sometimes multiple titles in different languages. Titles in all cases will be different from the display title. 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 as accession Ids within the MDR.

### B.7 Provenance String (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.

# Creators and Contributors

N.B. The previous data points of C.1 Creators and C.2 Contributors have been replaced by C.3 and C.4 below.

### C.3 Object People (0..n)

This lists the people involved with creating the object in various roles (e.g. as an author). Each person record is composite and includes:

* the contribution type, (categorised, as selected from a predetermined code-text list)
* the person’s full name,
* the person’s ORCID identifier, if they have one,
* the ‘affiliation string’ as presented in the original data,
* the affiliation organisation details (name, MDR Id, ROR Id) if that can be derived form the affiliation string.

### C.4. Study Organisations (0..n)

This lists the organisation involved with the creating the object in various roles (e.g. as funder). Each record is composite and includes:

* the contribution type, (categorised, as selected from a predetermined code-text list)
* the organisation details( name, MDR Id, ROR Id)

If required, people and organisation lists for data objects can be derived from the same lists for their parent study, though of course some data objects, such as journal papers, have linked people and organisations in their own right.

It is also possible for People and Organisation data to be optionally combined, e.g. on data export, to form generic ‘contributor’ records. Such an approach would be more compliant with the DataCite standard, but ignores the very different data associated with each type of contributor.

# Dates

### D.1 Publication year (0..1)

The year in which the object is made available, i.e. in which it first becomes citable, expressed as 4 digits. Not always 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 is it necessarily the year in which it was created.

### 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 of code-text pairs),
* a boolean indicating whether or not it is a single date or a range,
* date as string, in a standard format dd MMM yyyy, e.g. “12 Dec 2018”, or “7 Mar 2012”,
* 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 (E3 – E5) 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 Physical Object, 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 substantial 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 (0..1, applies to 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 (at least as claimed by the dataset’s generators). The contents are:

* Record key type (categorised, as selected from a predetermined code-text pair 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. 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 (0..1, applies to 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 (0..1, applies to 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. They are therefore not included in the MDR). The item is composite, consisting of:

* description type (categorised, as selected from a predetermined list of code-text pairs)
* label, a heading that might be applied to the text (e.g. as a sub-heading).
* description text, the description itself
* language code, the 2 character ISO code

### 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 identifiable living individual.

1 = Anonymised data.

2 = Pseudonymised data.

3 = Sensitive pseudonymised data.

In general, almost all clinical research *documents* made available for sharing would be categorised as 0, whilst all *IPD (individual participant data) datasets* would be categorised as 3 - unless there is general agreement that they are fully anonymised, in which case they become 1.

### E.8 Language (1)

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). English is assumed as a default.

DataCite assumes a single language but some clinical research data objects (e.g. journal articles) can be created in two or more languages. The record may therefore occasionally consist of a comma separated pair of ISO codes, very rarely three or more.

### E.9 Object Inter-Relationships (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 another, or be derived from, or the source of, one or more other data objects.

A potentially 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 code-text list)
* the identifier of the other or ‘target’ data object (in a suitable system).

Because few data objects have DOIs, it is usually a requirement that both subject and target objects are stored within the same system. This allows the identifier to be an internal identifier within that system, making navigation to it much simpler. Unfortunately at the moment none of the MDR sources includes this type of data, soi it is not present in the MDR.

### 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 can, optionally, be considered as those of the parent study or studies.

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

* topic type (categorised, as selected from a predetermined list of code-text pairs. Topic types include, ‘organism’, ‘chemical / biological agent’, and ‘geographic’.
* the original value, i.e. the topic as originally expressed.
* If present the original code from a controlled terminology (CT) or vocabulary, as well as an identifier for the CT itself.
* the MESH code and term, if present

# Location and Access details

This area extends the existing DataCite schema to provide 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. In both cases the organisation name could be associated with both an MDR Id and a ROR Id.

### F.2 Access Type (1)

A categorised value (code-text pair) 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 (should be present for any of the non-public access types)

This is a composite element with three elements:

* 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 web based resources, on the object manager’s web site or elsewhere (see below).
* 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.
* A date, if one is available, representing the last time the URL was checked to be in existence (i.e. returned a 200 ‘success’ code rather than a 404). Practical considerations, with over a million potential urls to check, mean that this field is not systematically collected.

### F.4 Resources (should be present if access is not managed)

The web based resources that represent this data object. For public objects 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 and Id of the system 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 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.5 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 name of the rights being applied,
* a uri that identifies an information source, usually a url to a web page,
* any additional comments or description of the rights regime.

Unfortunate this data is not currently included in the sources used for the MDR, and so is not yet to be found in the MDR.

# Categorised Items

A large number of items in the schema are categorised, forcing a selection from a list of predefined values. The codes and text values, and where appropriate descriptions, of the category systems developed within the MDR are all available within the ‘About’ wiki of the MDR, that is embedded in the new MDR web site.

To view them, please go to

<https://newmdr.ecrin.org/About>

and in the left-hand tree select and expand The Data/MDR lookup data item. This will display the various category systems. Selecting any of these will display the relevant organisation. There are 26 listed category systems, plus an additional 7 listed under The Data/Context lookup data.

The screen shot on the following page shows a portion of the data for ‘Contribution types’.

***Please note*** that while the lookup data pages are functional, as of October 2023 the whole of the new MDR site is not yet fully functional so other pages may not yet work correctly.

A screenshot of a computer

Description automatically generated

**Part of the Contribution Types categories shown on the MDR web portal**

MDR Implementation: JSON schemas

More detailed, ‘implementation level’ information on the schema can be obtained from the same MDR wiki web site at

<https://newmdr.ecrin.org/About>

Once there expand:

Metadata schemas/Study Schema JSON, or

Metadata schemas/Object Schema JSON

to view the JSON definitions, expressed using json-schema.

The wiki on these page provides a wide range of detailed information about the MDR, including, in the Metadata schemas section, all the material in this document, as well as JSON definitions for all previous versions of the schemas, and summaries of the changes made between each version.

# 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>

[5] ECRIN Clinical Research Metadata Schema Version 4 (November 2019) DOI: 10.5281/zenodo.3562911. Available at [https://zenodo.org/record/4028900](https://zenodo.org/record/4028900%20)

[6] ECRIN Clinical Research Metadata Schema Version 5 (November 2019) DOI: 10.5281/zenodo.3562911. Available at [https://zenodo.org/record/4133889](https://zenodo.org/record/4133889%20)

[7] ECRIN Clinical Research Metadata Schema Version 6 (November 2019) DOI: 10.5281/zenodo.3562911. Available at [https://zenodo.org/record/5554961](https://zenodo.org/record/5554961%20)