ECRIN Clinical Research Metadata Schema

Version 2.2 (February 2019)

This paper provides an overview of this schema, which is designed to support the discoverability of data objects generated by clinical research and is an extension of the DataCite schema.

Contents

[Summary table 1](#_Toc24301422)

[Description of Individual Data Points 4](#_Toc24301423)

[A. The Source Study 4](#_Toc24301424)

[B. Data Object Identifiers 5](#_Toc24301425)

[C. Creators and Contributors 6](#_Toc24301426)

[D. Object Dates 6](#_Toc24301427)

[E. Data Object Attributes and Descriptors 7](#_Toc24301428)

[F. Identifying Location, Ownership and Access 10](#_Toc24301429)

[The ECRIN schema and DataCite 12](#_Toc24301430)

[Constructing a DataCite record from ECRIN metadata 12](#_Toc24301431)

[Fate of ECRIN metadata when translating to DataCite. 16](#_Toc24301432)

# Summary table

The proposed scheme has 28 main data points (though some of these are composite), split into six sections, A – F:

|  |  |  |
| --- | --- | --- |
| Mandatory | Recommended | Optional |
| 1. The Source Study | | |
| A.1 Source Study Title  *{Scientific title, Language code}* | **A.2 Study Identifiers**  *{Identifier type, Identifier value, Source organisation, Application date}* \*  **A.3 Study Topics**  *{Topic name, Topic code, Topic type}* \* | **A4. Other Study Titles**  *{Title, Title type, Language code}* \* |
| 1. Data Object Identifiers | | |
| B.1 DOI  *(If a DOI does not exist the local – i.e. ECRIN MDR – system accession number needs to be used instead)*  B.2 Object Title | **B.3 Version** | **B.4 Object Other Identifiers**  *{Identifier type, Identifier value, Source organisation, Application date}* \*  **B.5 Object Additional Titles**  *{Title, Title type, Language code}* \* |
| 1. Creators and Contributors | | |
| C.1 Creators  *{Name type, [Given name, Family name, Person Id, Email, Affiliation] | [Organisation name] } \** |  | **C.2 Contributors**  *{Name type, Contribution type, [Given name, Family name, Person Id, Email, Affiliation] | [Organisation name] } \** |
| 1. Object Dates | | |
| D.1 Publication Year |  | **D.2 Dates**  *{Date, Date type, Is range, Start date, End date, Details} \** |
| Mandatory | Recommended | Optional |
| 1. Object Attributes and Descriptors | | |
| E.1 Class  E.2 Type  E.3 Record key type  *{type, details}*  E.4 Identifier type  *{type, details}*  E.5 Associated consent  *{type, details}*  *(E3, E4 and E5 apply only to datasets)* | **E.6 Description**  *{description, language code, description type}* \*  **E.7 Language**  **E.8 Related Resource Identifiers**  *{target object, relationship type}* \* | **E.9 Topics** (of data object) *DEPRECATED*  *{Topic name, Topic code, Topic type}* \* |
| 1. Object Location and Access Details | | |
| F.1 Managing Organisation  F.2 Access Type  F.3 Access Details  F.4 Access Details URL  *(F3 and F4 are mandatory if access is non-public)*  F.5 Resources  *{Repository Organisation, URL, URL accessible, File type, Size, Size units} \** |  | **F.6 Rights**  *{details, rights URI} \** |

**\*** May be multiple

# Description of Individual Data Points

## A. The Source Study

This section deals with the study or studies that the data object was generated by or describes (but not those it simply cites). It is an addition to the DataCite specification. Strictly speaking it is not metadata because the data points do not describe data – instead they summarise some key attributes of the study as an activity.

***A.1 Source Study Title (1)***

The ‘title’ in this instance means the full or ‘scientific’ title, i.e. the title of the study protocol. For consistency it should be the *exact scientific title* as used within the initial trial registry entry, or where there is no registry entry the title on the protocol.

***A.2 Source Study Identifiers (0...n)***

None, one or more unique identifiers that have been assigned. 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.

These IDs are composite. If provided, they must include not just the identifier value and type, but also the assigning organisation and optionally the date the identifier was assigned.

***A.3 Source 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 data objects, it makes sense to include any topic data *with the study* rather than with each of the individual data objects relating to that study.

The listed topics could be free text, but it would be more useful if the text was structured, i.e. selected from a controlled vocabulary. There are a variety of such controlled vocabularies available for studies (ICD 10, MedDRA, SnoMed CT etc.). To ensure that the source system is clearly identified, use of a controlled vocabulary term should be associated with a URI that identifies the scheme (and version) being used.

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.

***A.4 Other Study Titles (0..n)***

Studies often have short or ‘public’ titles, and / or are described by an acronym. These can be added in this composite data point, which includes the name, the name type and the associated language code.

## B. Data Object Identifiers

***B.1 Data object identifier (1)***

All data objects that are available to others (whether publicly or under managed access) need to be citable – and should therefore have a DOI, in line with the DataCite specification.

Where a data object is identified which does not have a DOI (e.g. a trial registry entry, or a file in a repository that does not routinely apply them) consideration should be given to ECRIN or some other organisation minting and applying one – if financially feasible and acceptable to the object creators. The extent of this problem needs to be clarified.

When there is no DOI available the data object identifier is the local accession number of the data object within the MDR system, preceded by a suitable URL to guarantee uniqueness (e.g. “ECRIN/MDR/dataobject/…” but the exact URL to be determined).

***B.2 Object Title (1)***

This should be the title of the object. For a journal article it would be the name of the article. For many internal documents it would be the name on the document’s title page, which would often include a form of the study title, e.g. ‘<study name> Protocol’, or ‘<study name> Statistical Analysis Plan’.

If a data object does not have a pre-existing title it should be named to make the nature of the object clear, e.g. ’Final analysis data set’, ‘Consent forms’. Within the context of the associated study or studies that name should be unique. In a citation, and in some cases on display within a system, that name will then need prefixing by the study name to clearly differentiate it from other data objects – this will depend on the details of the display.

***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 focus here is on the version or versions that are made *available for sharing*. The data generators will need to make that selection, though 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. For instance datasets used for the primary analysis should normally be available, as well as possible later datasets that have additional follow up data. A protocol published before the trial began may need to be differentiated from the protocol as it existed at study end. 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 *must* be completed and displayed with the name and other identifiers.

***B.4 Other Object Identifiers (0...n)***

This refers to other unique identifiers that have been assigned to the data object in addition to its primary identifier (e.g. a Pub Med ID). As with studies such IDs would be composite and include the identifier type and assigning organisation, as well as the identifier value, and optionally the date of assignment.

***B.5. Additional Titles (0...n)***

Additional names for the data object can also be provided. If given they are composite: the title, plus one of ‘title type’, e.g. Translated Title, Alternative Title, Subtitle, and the language code.

## C. 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 or personal names. Each name in the list is a composite element, and can contain optional identifiers, e.g. ORCID IDs, and / or organisational affiliations, as well as the name itself. For flexibility of citation the given names are stored separately from the family name.

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

Optionally, other institutions and / or persons responsible for collecting, managing, distributing, or otherwise contributing to the development of the data object. If given, any contributor record is composite, with the same structure as the Creator data above, plus an additional data point specifying contributor type.

The latter has been extending in the context of clinical research, to include (for example) trial sponsor, sponsor contact, drug supplier, logistic support organisation, device provider, central laboratory, public contact, principal and co-ordinating Investigator. There is no suggestion that such information is available or need be collected for every study, only that when it exists in the source data it should be possible to capture it.

The creator and contributor lists for data objects created by the same study are likely to be very similar, but the data point remains attached to the data object rather than the parent study, to allow for variations at the level of the individual object.

## D. Object 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.)

The year is needed in order to be able to create an accurate citation from the metadata. The DataCite documentation provides more detailed guidance. The creation date can be referenced in the Dates item (see below), though if the creation year(s) are required in the citation they should be added to the title.

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

None, one or more dates or date ranges that are relevant to the data object, in the standard ISO 8601 format. Each date should be accompanied by a *date type* value that indicates what the date represents: e.g. the date accepted, available, copyrighted, collected, created, issued, submitted, updated, valid. There is also the opportunity of adding additional explanatory or clarifying information, as ‘date details’. An additional date (or date range) relevant to clinical research has been added: ‘Controlled access in force’.

## E. Data Object Attributes and Descriptors

Section E is mainly based on the current 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)***

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 description of the type of data object, a more specific categorisation than Class. The preferred format is a single term, so that a pair can be formed with the Class described above (as with DataCite), e.g. Dataset/census data or Text/conference abstract. Existing DataCite / CASRAI resource types, e.g.

* book chapter
* journal Article
* conference Paper
* website

are extended by a list of standard resource types for clinical research, e.g.

* protocol
* patient information sheet
* clinical studies report
* final analysis dataset
* participant reported questionnaire
* quality of life sub-study analysis dataset

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

This field is coded to indicate the type of record keys used within the dataset. The possible values are:

* 0: Unknown or not yet allocated – the default when a new data object is added to a system.
* 1: None: No keys present at all, apart from a possible sequential record number, generally only applicable to single flat files.
* 2: Anonymised: Anonymised data with arbitrary key values applied to records, used only to link the different records belonging to the same person, i.e. across different flat files. The keys used have no link with any other system or material.
* 3: Pseudonymised: Data belonging to the same individual share the same key, and that identifier references a separate file or material that allows the individual to be identified, though that file is kept securely and is *not* available to the data recipient.
* 4: Pseudonymised with identifying material: As pseudonymised data but in this case the data holder also has access to the identifying additional material. Would usually only apply to data maintained by the data generator or trial sponsor.
* 9: None of the above

A text details field exists to allow the record key type to be specified further. If data is classed as pseudonymised, the keeper(s) of the additional, identifying material (not necessarily its location) should be stored in the details field. If ‘None of the above’ further details must be provided.

***E.4 Identifier type (1, Datasets only)***

A coded field that indicates the identifiers present in the data (whether or not pseudonymising keys are present) and which thus gives an indication of the level of de-identification. The possible values are:

* 0: Unknown or not yet allocated – the default when a new data object is added to a system.
* 1: None: A dataset with no direct or indirect identifiers. Would be rare as scientific utility is likely to be severely affected, but could be a subset of data used for a particular purpose.
* 2: De-identified: A dataset with no direct identifiers, and with indirect identifiers modified by established de-identification steps (e.g. amalgamation of categories, rebasing of dates, removal of text comments) so that it is no longer possible to identify any individuals within the data set.
* 3: Has Indirect Identifiers: Dataset contains no direct identifiers, but does contain data fields that when considered in combination might be used to identify some of the individuals. In some cases, access would also be required to other systems.
* 4: Has Direct Identifiers: The dataset contains at least one direct identifier, i.e. a name, code, system Id or other data that allow the individual to the identified unambiguously – in some cases requiring access to an additional system. This would be very rare in the context of shared data.
* 9: None of the above

A text details field exists to allow the identifier type to be specified further. For de-identified data sets a summary of the de-identification process can be provided (if not here then as part of the descriptive metadata for the dataset). Datasets with direct and indirect identifiers should describe what those identifiers are. If ‘None of the above’ further details should be given.

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

A coded field that indicates the type of consent for re-use and sharing associated with the data.

* 0: Unknown or not yet allocated – the default when a new data object is added to a system.
* 1: None: No specific consent was given for the sharing of data or its re-use beyond the study in which it was originally collected.
* 2: Partial re-use: Consent was given explicitly, for sharing and re-use of the data, for specified purposes or work in a specified domain, e.g. ‘in cancer research’, or ‘for research into malaria’, or research within a named institution, country etc.
* 3: Full re-use for research: Broad consent was given explicitly, for sharing and re-use of the data for research purposes
* 9: None of the above

A text details field exists to allow the associated consent to be specified further. For partial consent (and optionally for full), the exact wording of the consent given should be included in the details field. If ‘None of the above’ further details should be given.

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

None, one or more pieces of additional general information. The format is open, though a language indicator is also required (as a separate field) as well as a ‘description type’ indicator. For published papers, the description would normally be the abstract. For clinical trial datasets, it might be useful to include an outline of:

1. The volume of the dataset, the number of participants to which it refers and the typical number of data items for each
2. The types of data in the set and its source, e.g. from the clinical area, from laboratories, from the participants themselves, or a mix of all three.
3. The population of participants to which the data refers , e.g. all trial participants or a subset of them
4. Whether the dataset is designed to support a particular paper or papers (and if so which) or is a general dataset with all the data collected.
5. The time point of the data. This will normally be study end, but it may, for example, represent or include longer term follow up data (post primary end-point).

For papers and datasets at least one description should normally be provided. Additional descriptions may be supplied in different languages – they do not need to be exact translations but should cover the same material.

***E.7 Language (1)***

The language of the data object (not the description), using the ISO language codes (e.g. en, de, fr).

***E.8 Related Resource Identifiers (0..n)***

These are the Identifiers of related resources. ‘Related’ covers a wide range of relationships, including*:* IsCitedBy, Cites, IsSupplementTo, IsSupplementedBy, IsContinuedBy, Continues, IsNewVersionOf, IsPreviousVersionOf, IsPartOf, HasPart, IsIdenticalTo, IsDerivedFrom, and IsSourceOf.

A particularly important relationship for clinical study data is the pairing of HasMetadata / IsMetadata for. Metadata in clinical research can include, for example, an ODM file or 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 identifier of the related data object and the relation type. Although in DataCite the related resource can be identified by one of a variety of identifier types, to keep things simpler within the MDR, the requirement is that any related resource *must also be indexed within the MDR*. This imposes an additional burden, of ensuring that this is the case, but it allows the identifier to be an internal identifier within the MDR system, making navigation to it much simpler.

***[E.9 Topic (0...n) DEPRECATED***

None, one or more subject names or phrases, keywords, classification codes describing the resource. In general, however, the recommendation is to include any subject / topic descriptors, keywords etc., with the *study* data points rather than the individual data objects (as A.3). This field is retained for compatibility with DataCite but its use is deprecated in this context.]

## F. Identifying Location, Ownership and Access

A major 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)***

One of

* unknown, not yet allocated
* public download (completely open access)
* public on-screen access (completely open access)
* public download (self-attestation required)
* public on-screen access (self-attestation required)
* restricted download
* restricted on-screen access
* case by case download
* case by case on-screen access
* none of the above

*Self-attestation* refers to the requirement for the data accessor to identify themselves and indicate some details about themselves (e.g. organisational affiliation, job role) and, possibly, their reason for wanting to access the data. In most cases self-attestation would be followed by a confirmation mechanism (e.g. activation via an email account).

*Restricted* means access would be dependent on membership of a predefined group, usually as determined by an authentication mechanism (e.g. username + password), for example as is the case with subscription to a journal, or membership of a collaborating organisation.

*Case by case* means that there is no predefined access but that applications for access to the data object will be considered by the object owners.

*On-screen access* means that a researcher can view and process data within a specified environment but cannot download a file of the raw data, though export of the results of re-analysis would be allowed.

***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 be a link to a web page on the publisher’s site, that would explain access procedures or provide an application proforma.

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

The web based resources that represent this data object. Mandatory for public or restricted access objects, when at least one resource should be listed. Each record would be composite and include

* The name of the organisation holding the resource
* 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 file type (normally based on the file extension) and
* the resource size,
* the resource size units, usually in KB, MB or GB.

***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 URI for the specific rights management scheme should also be given.

# The ECRIN schema and DataCite

## Constructing a DataCite record from ECRIN metadata

This section identifies how a DataCite record (and from that a citation) could be constructed from the ECRIN metadata. It does that by taking each of the defined data points of the DataCite schema (v4.1, available at <https://schema.datacite.org/meta/kernel-4.1/doc/DataCite-MetadataKernel_v4.1.pdf>) and indicates how it is supported by the ECRIN metadata scheme.

**1 Identifier (1)**

In DataCite this is a Mandatory field and is always a DOI.

With the ECRIN schema there is no guarantee, at the moment, that all listed objects will have a DOI, and although it is hoped and expected that the proportion of objects with DOIs will rise it may never attain 100%. The object identifier is therefore a DOI whenever the data object has one, otherwise it is the local (ECRIN MDR database) accession number. This means that the identifier type item (DC 1.1) will always be one of ‘DOI’ or ‘Local accession number’.

**2 Creator (1-n)**

In DataCite the full name of each creator is mandatory in a ‘family, given’ order, while organisations are represented by the organisation name. The given and family names can also be provided separately, and / or an optional ‘name identifier’ can be provided, in a specified schema (such as ORCID). In addition optional organisational affiliation(s) can be provided

In the ECRIN schema the full name of creators can be provided (assuming those names are given in the original source material). The underlying data store is also intended to include separate given and family names, affiliations, and ORCID identifiers. Where such information has been harvested, therefore, it is available to provide the corresponding data points in DataCite. The frequency of data objects without creator names is yet to be determined.

The only restriction is that the only identifier schema that is envisaged being used is ORCID. When an ORCID identifier is given therefore (DC 2.4) the ‘name identifier scheme’ (DC 2.4.1) will be ORCID and the scheme URI will be <https://orcid.org/>.

**3 Title (1-n)**

In DataCite at least one title must be provided, with additional titles optional, each with an optional ‘title type’ field.

In the ECRIN schema there is always at least one title, even if it has to be manufactured from the source study name and the object type. Additional titles can also be provided with the type specified. For the data objects in the system the controlled list values are the same as those in DataCite (studies may also have ‘public’ and ‘abbreviation or acronym’ title types).

**4 Publisher (1)**

In DataCite this is the name of the entity that ‘holds, archives, publishes prints, distributes, releases, issues, or produces the resource.‘

In the ECRIN schema there is a ‘Managing Organisation’ that fulfils the same function. This organisation *manages access* to the document or data object, including making the overall decision about access type. For personal data this would usually be the name of the organisation that was the data controller.

**5 PublicationYear (1)**

The year when the data object was or will be made publicly available. The same data is stored and can be made available within the ECRIN schema.

**6 Subject (0-n)**

In DataCite this is one or more of ‘subject, keyword, classification code, or key phrase describing the resource’. There is optionally a subject schema (DC 6.1) and a scheme URI (DC 6.2) that can be used to indicate a term taken from a published controlled vocabulary.

In the ECRIN scheme the same information exists but it is applied to the *source study* of the data object, as ‘study topics’. Where applicable the subject schema and schema URI are also stored. This means the information does not to be repeated for every data object associated with a study. The information is available for each data object, however, if and when required.

**7 Contributor (0-n)**

In DataCite this is ‘the institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.’ The information to be provided for each contributing person or organisation is exactly the same as for the object creator(s), with the addition of a categorised ‘contributor type’ (DC 7.1).

The ECRIN schema includes the same information, as long as it can be harvested from the source material (as with creators). The options available for ‘contributor type;’ have been extended, however, to reflect personal and organisational roles that are common within clinical research.

**8 Date** **(0-n)**

The DataCite schema includes the of providing specific ‘dates relevant to the work’, including date ranges if applicable. These should be categorised as one of a collection ‘date types’ (DC 8.1) and include supplementary ‘date information’ (DC 8.2).

All these data points are also present in the ECRIN schema and can be mapped directly to their DataCite equivalents.

**9 Language (0-1)**

The primary language of the resource. The same data is stored and can be made available within the ECRIN schema.

**10 Resource Type (1) and 10.1 Resource Type General (1)**

In DataCite the resource type and ‘resource type general’ (DC 10.1) are mandatory and combined provide a categorisation of the resource.

In the ECRIN schema the resource type general is represented by Object Class and has the same categories as in DataCite. The resource type is represented by object type. The available categories have been extended to include common types of textual documents and datasets found within clinical research.

**11 Alternate Identifier (0-n)**

DataCite allows for ‘an identifier or identifiers other than the primary Identifier applied to the resource being registered’, together with ‘alternate identifier type’ (DC 11.1).

In the ECRIN schema the same information is stored and can be provided, although the list of identifier types has been modified to match the common types found in clinical research.

**12 Related Identifier (0-n)**

DataCite allows the provision of identifiers of related resources (which must be globally unique identifiers) together with an indication of the relationship type (DC 12.2). In DataCite the identifier type can be one of a categorised list.

In the ECRIN schema identifiers of related resources can also be stored and described, but here the related resource *must also be in the MDR*, because internally the related identifier is the internal accession number. The ‘related identifier type’ (DC 12.1) is therefore always ‘ECRIN MDR ID’.

For data objects related by the relationships HasMetadata / IsMetadataFor, the name, URI and type of the metadata scheme is likely to refer either to CDISC standards for describing clinical dataset structures (e.g. ODM, define.XML) or proprietary structures created by specific data collection systems.

**13 Size (0-n)**

In DataCite this is the size (e.g. bytes, pages, inches, etc.) or duration (extent), e.g. hours, minutes, days, etc., of a resource.

The ECRIN schema has the same data points.

**14 Format (0-n)**

In DataCite this is free text describing the format of data objects, using the file extension or MIME type, e.g., PDF, XML, MPG or application/pdf, text/xml, video/mpeg.

In the ECRIN schema the same data points are available

**15 Version (0-1)**

The version number of the resource. The ECRIN schema has the same data point

**16 Rights (0-n)**

In DataCite this details any rights information for this resource and can optionally include the URI of the licence being used (DC 16.1).

In the ECRIN schema the same data points are available

**17 Description (0-n)**

DataCite includes this as ‘all additional information that does not fit in any of the other categories. May be used for technical information’ and states further that ‘It is best practice to supply a description.’ There is in addition a ‘description type’ field (DC 17.1), that provides a categorisation of the description(s) given.

In the ECRIN schema the same data points are available.

**18 Geo Location (0-n)**

DataCite has a range of optional data points to describe geographic location, usually about where the data was gathered or the are on which the data is focused.

There is no current plan to include such data within the ECRIN metadata scheme so this data cannot be generated.

**19 Funding Reference**

DataCite allows the name of the funder(s), (DC 19.1), or its ID and identifier type (DC 19.2, 19.2.1) to be included in the metadata. It also includes the option to include grant or financial award information: the award number (DC 19.3), the funder URI (DC 19.4) and the award title as free text (DC 19.4).

In the ECRIN schema there is currently no plan to store or use detailed information of the grant. Funder names could be identified from the Contributors data points if they were included in that listing.

**Summary**

The table below summarises how the data required for a DataCite record can be obtained from the ECRIN metadata. As described above, it can be seen that in most cases the mapping is unproblematic. Where the mapping is not straightforward this is indicated by differently coloured text.

The only serious problem would arise if and when the object did not have a DOI. The obvious response would be to try and generate the missing DOI, but this is likely to have financial implications that will need to be explored. This needs to be discussed further, once the extent of the problem is better defined.

|  |  |  |
| --- | --- | --- |
| DataCite |  | ECRIN Metadata Schema |
| 1 Identifier (1) | **<=** | Mapped from a DOI where one exists. If not a local accession number used instead, or a DOI created. |
| 2 Creator (1-n) | **<=** | Mapped directly from matching data point(s) |
| 3 Title (1-n) | **<=** | Mapped directly from matching data point(s) |
| 4 Publisher (1) | **<=** | Mapped directly from matching data point(s) |
| 5 PublicationYear (1) | **<=** | Mapped directly from matching data point(s) |
| 6 Subject (0-n) | **<=** | Mapped indirectly from source study topics |
| 7 Contributor (0-n) | **<=** | Mapped directly from matching data point(s) |
| 8 Date (0-n) | **<=** | Mapped directly from matching data point(s) |

|  |  |  |
| --- | --- | --- |
| DataCite |  | ECRIN Metadata Schema |
| 9 Language (0-1) | **<=** | Mapped directly from matching data point(s) |
| 10 Resource Type (1) and 10.1 Resource Type General (1) | **<=** | Mapped directly from matching data point(s) (types extended to match clinical research domain) |
| 11 Alternate Identifier (0-n) | **<=** | Mapped directly from matching data point(s) (types extended to match clinical research domain) |
| 12 Related Identifier (0-n) | **<=** | Mapped from matching data points, but restricted to data objects within the same system. |
| 13 Size (0-n) | **<=** | Mapped directly from matching data point(s) |
| 14 Format (0-n) | **<=** | Mapped directly from matching data point(s) (types extended to match clinical research domain) |
| 15 Version (0-1) | **<=** | Mapped directly from matching data point(s) |
| 16 Rights (0-n) | **<=** | Mapped directly from matching data point(s) |
| 17 Description (0-n) | **<=** | Mapped directly from matching data point(s) |
| 18 Geo Location (0-n) | **<=** | Not mapped – No data in ECRIN schema |
| 19 Funding Reference | **<=** | Funder name may be mapped from contributor data. Other data not present and not mapped. |

## Fate of ECRIN metadata when translating to DataCite.

This section examines translating the data starting with the ECRIN metadata record, and discusses the details of how the ECRIN data would need to be translated. Because the ECRIN metadata schema is essentially an extension of DataCite, however, some of that metadata will be lost unless it can be translated into a different data point.

The data stored about the source study, in particular, which provides essential information for the user, is difficult to translate into the DataCite schema. DataCite can store relationships between data objects, but has no facility to identify scientific *activity*, e.g. individual research studies, and therefore no easy way of storing relationships between data objects and their source activity.

One approach might be to identify the source study as a data object ‘subject’ (DC 6), i.e. a thing that the data object ‘is about’ or ‘associated with’. A subject entry might then be a trial registry Id, with the schema identified as the trial registry. This would allow unambiguous linkage of the study with the data object, as long as there was at least one registry entry, which is the case for most clinical trials but not for non-interventional research.

A second approach would be to make the study title and / or acronym (where one exists) also one of the data object’s ‘subjects’. This is potentially ambiguous (acronyms are not unique), and the key words of the title should probably be already listed within the object’s subjects, so while it may be easier to use it may be less useful in identifying the study.

A third approach would be to use the Related identifier option in DataCite (DC 12) and to consider key data objects such as the trial registry data or protocol as a proxy for the source study. If either of these had a DOI then a record could be constructed like:

DC 12 Related Identifier: protocol DOI, or registry DOI

DC 12.1 Identifier type: DOI

DC 12.2 ‘references’ or ‘is derived from’ (?)

to indicate the link to the study. If the registry entry does not have a DOI it still might be possible to use it:

DC 12 Related Identifier: registry ID

DC 12.1 Identifier type: Registry specific Id (i.e. with the registry identified)

DC 12.2 Relation type: ‘references’ or ‘is derived from’ (?)

Ideally a convention would be established to support consistent application of one or more of the mechanisms above.

The study topics could be mapped in straightforward fashion to DC 6 Subject, but would have to be applied top each of the data objects linked to the study,

The ECRIN data points E3, E4 and E5, which describe the level of pseudonymisation, de-identification and associated consent, are key information for datasets with sensitive personal data but have no obvious equivalents in DataCite. It is suggested that they should be included within a description field instead.

The same is true of the ECRIN metadata describing the arrangements for managed access data (F2, F3 and F4). Again these could be preserved by putting them into a description field.

The following table summarises the relationship between the ECRIN data points and the DataCite schema. In most cases the mapping is straightforward, but any issues are noted in red below.

|  |  |
| --- | --- |
| ECRIN CR Metadata | DataCite |
| A.1 Source Study Title  *{Scientific title, Language code}* | Possibly mapped, as a ‘subject’ (DC 6).  See discussion in text |
| A.2 Study Identifiers  *{Identifier type, Identifier value, Source organisation, Application date}* \* | Cannot be easily mapped, though registry value could be used as a ‘subject’ (DC 6), or as part of a Related identifier’ (DC 12).  See discussion in text |
| A.3 Study Topics  *{Topic name, Topic code, Topic type}* \* | Mapped, for each associated data object, to  DC 6 Subject,  and if applicable 6.1 Subject scheme and 6.2 scheme URI |
| A4. Other Study Titles  *{Title, Title type, Language code}* \* | Acronym and short titles possibly mapped, as a ‘subject’ (DC 6).  See discussion in text |
| ECRIN CR Metadata | DataCite |
| B.1 Identifier (usually DOI) | If the identifier is a DOI can be mapped to  DC 1 Identifier,  Otherwise would need to be mapped as  DC 11 Alternate identifier, with 11.1 Alternate identifier type, as ‘Local Accession Number’. |
| B.2 Object Title | Maps to initial DC 3 Title |
| B.3 Version | Maps to DC 15 Version |
| B.4 Object Other Identifiers  *{Identifier type, Identifier value, Source organisation, Application date}* \* | Maps to DC 11 Alternate identifier and  11.1 identifier type  Source organisation and date not mapped |
| B.5 Object Additional Titles  *{Title, Title type, Language code}* \* | Maps to initial DC 3, Titles and 3.1 Title type, as additional titles |
| C.1 Creators  *{Name type, [Given name, Family name, Person Id, Email, Affiliation] | [Organisation name] } \** | Maps to DC 2 creator:  2.1 Creator name, 2.1.1 Name type, 2.2 Given Name, 2.3 Family name, 2.4 Name identifier, 2.4.1 Name identifier scheme, 2.4.2 Scheme URI, 2.5 Affiliation |
| C.2 Contributors  *{Name type, Contribution type, [Given name, Family name, Person Id, Email, Affiliation] | [Organisation name] } \** | Maps to DC 7 creator:  7.1 Contributor type,  7.2 Creator name, 7.2.1 Name type, 7.3 Given Name, 7.4 Family name, 7.5 Name identifier, 7.5.1 Name identifier scheme, 7.5.2 Scheme URI, 7.6 Affiliation  If a funding organisation… could map to  DC19.1 Funder Name |
| D.1 Publication Year | Maps to DC 5 Publication Year |
| D.2 Dates  *{Date, Date type, Is range, Start date, End date, Details} \** | Maps to DC 8 Date  Including DC 8.1 Date Type  DC 8.2 Date information |
| E.1 Class | Maps to DC 10.1 Resource General Type |
| E.2 Type | Maps to DC 10 Resource Type |
| E.3 Record key type  *{type, details}*  E.4 Identifier type  *{type, details}*  E.5 Associated consent  *{type, details}*  *(E3, E4 and E5 apply only to datasets)* | -- not present in DataCite –  Cannot be easily mapped unless added to description data.  (A DataCite description type of ‘Personal Data parameters’ or similar might therefore be useful). |
| ECRIN CR Metadata | DataCite |
| E.6 Description  *{description, language code, description type}* \* | Maps to DC 17 Description,  with DC 17.1 Description type |
| E.7 Language | Maps to DC 9 Language |
| E.8 Related Resource Identifiers, with relation type indicator  *{target object, relationship type}* \* | Maps to DC 12 Related Identifier  DC 12.1 Identifier type  DC 12.2 Relation Type  But identifier type always ‘ECRIN MDR’ as object relationships restricted to objects within the same system |
| E.9 Topics (of data object) *DEPRECATED*  *{Topic name, Topic code, Topic type}* \* | Would map to DC 6 Subject,  and if applicable 6.1 Subject scheme and 6.2 scheme URI,  but deprecated.  Mapping should be from A.3 instead |
| F.1 Managing Organisation | Maps to D4 Publisher |
| F.2 Access Type  F.3 Access Details  F.4 Access Details URL  *(F3 and F4 are mandatory if access is non-public)* | -- not present in DataCite –  Cannot be easily mapped unless added to description data.  (A DataCite description type of ‘Access details’ might therefore be useful). |
| F.5 Resources  *{Repository Organisation, URL, URL accessible, File type, Size, Size units} \** | For each resource,  Can be mapped  To DC 13 Size, DC 14 Format  Repository Organisation and URL cannot be easily mapped |
| F.6 Rights  *{details, rights URI} \** | Maps to DC 16 Rights and 16.1 Rights schema |