ECRIN Clinical Research Metadata Schema

Version 2 (April 2018)

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.

It also summarises a relational database structure that could be used to store the metadata within a central repository of such data.

For the original paper describing the development of this schema, (and which includes version 1), please see: Canham s and Ohmann C. A metadata schema for data objects in clinical research. Trials2016; **17**:557. <https://doi.org/10.1186/s13063-016-1686-5>

It is divided into the following sections:

[Summary table 2](#_Toc511940299)

[Description of Individual Data Fields 3](#_Toc511940300)

[Comparison with other Metadata schemes 10](#_Toc511940301)

[Differences between versions 2 and 1 12](#_Toc511940302)

[Data Structures 14](#_Toc511940303)

# Summary table

|  |  |  |
| --- | --- | --- |
| Mandatory | Recommended | Optional |
| The Source Study | | |
| A.1 Source Study Title | A.2 Study Identifiersº\*  A.3 Study Topicsº\* | A4. Other Study Titles\* |
| Data Object Identifiers | | |
| B.1 DOI (1)  B.2 Object Title | B.3 Version | B.4 Object Other Identifiersº\*  B.5 Object Additional Titlesº\* |
| Data Object Provenance | | |
| C.1 Creatorsº\* |  | C.2 Contributorsº\* |
| D.1 Creation Year |  | D.2 Datesº\* |
| Data Object Attributes and Descriptors | | |
| E.1 Resource Class  E.2 Resource Type  E.3 Record key typeº (3)  E.4 Identifier typeº (3)  E.5 Associated consentº (3) | E.6 Descriptionº\*  E.7 Language  E.8 Related Resource Identifiersº\* | E.9 Topics (of data object) º\* |
| Location and Access Details | | |
| F.1 Publisher / Provider  F.2 Access Type  F.3 Access Details (2)  F.4 Access Contact (2)  F.5 Resourcesº\* |  | F.6 Rightsº\* |

º Composite data point, **\*** May be multiple

(1) Mandatory for publicly accessible data objects, recommended for all others;

(2) Mandatory if access is non-public;

(3) Mandatory for datasets

# Description of Individual Data Fields

1. ***The Source Study***

The study or studies that the data object was generated by or describes (but not those it simply cites).

*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 title* as used on *version 1.0,* the first published version, of that 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, the scheme URI if there is one 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 ‘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 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. To ensure that the source system is clearly identified, any use of a controlled vocabulary term should be associated with a URI that identifies the scheme (and version) being used.

*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 to this field.

1. ***Data Object Identifiers***

Section B is based on the DataCite specification.

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

Data objects available publicly – e.g. journal articles, plus datasets and protocols available in repositories, should have a DOI (in line with the DataCite specification). Non-public data objects should also, wherever possible, also have a DOI. If a DOI is not possible, or has not yet been assigned, than the object should be identified either by an accession number from a metadata repository system, or by using the object’s name and version code, coupled with a unique identifier for the source study. The data object identifier (like study identifiers) therefore needs to be a composite, indicating its type and source as well as its value.

*B.2 Object Title (1)*

The default name of the data object. Within the context of the associated study or studies it should be unique.

*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.

*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. Again such IDs would be composite: and include the identifier type and assigning organisation, as well as its value, and optionally the identifier scheme and 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.

***C-D. Data Object Provenance***

Sections C and D are mainly based on the current DataCite metadata specification, though a few extensions have been added for datasets.

*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 citation purposes the given names are stored separately from the family (or organisation) 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 may need extending in the context of clinical research, to include (for example) drug supplier, drug distributor, device manufacturer, central laboratory, sponsor contact, recruitment contact, principal and chief (or co-ordinating) Investigator.

*D.1 Creation year (1)*

The year in which the object was created, expressed as 4 digits. Its precise definition will vary with the nature of the data object. For data sets it will be the year of their initial assembly, for published documents the year of their initial publication, and for internal documents the year of their approval for use. Note that ‘creation year’ is intended only to provide a broad indicator of the time something was created – e.g. in an on-screen listing or within a search. It is not an exact date, which are collected and stored separately – see D.2 below.

*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.

***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 Resource 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
* Interactive Resource

*E.2 Resource Type (0..1)*

A description of the resource. The format is open, but the preferred format is a single term, so that a pair can be formed with the ‘Resource Type General’ described above, e.g. Dataset/census data or Text/conference abstract. Existing DataCite / CASRAI resource types, e.g.

* book chapter
* journal Article
* conference Paper
* website

will need extending 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

In practice an expandable list would probably be needed – i.e. one where a user could supplement the supplied controlled vocabulary terms by free text, as and when necessary.

*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

If pseudonymised data, the keeper(s) of the additional, identifying material (not necessarily its location) should also be stored. If ‘None of the above’ further details should be given.

*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

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

For partial consent (and optionally for full), the exact wording of the consent given should be included as an additional 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). For published papers, the description would normally be the abstract. For clinical trial datasets, it will normally 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. Data objects that are documents may need

*E.7 Language (0..1)*

The primary 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, which must be globally unique identifiers. Related resources will normally be data objects themselves. The record is composite and must include the identifier itself, the related identifier type and the relation type. Relation types include*:* 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. 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.

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

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.

1. ***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 Publisher / Provider*

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 managing the repository. For journal papers it is 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*

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 description 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.

*F.4 Access Contact (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 could include an email address but more normally 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 resource URL

and, if downloadable, the

* resource file type (e.g. file extension or MIME type) and
* the resource size, usually in KB, MB or GB.

The resource host(s) would usually be obvious from the URL.

*F.6 Rights*

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.

# Comparison with other Metadata schemes

|  |  |  |  |
| --- | --- | --- | --- |
| ECRIN CR Metadata | DataCite | Dublin Core | DSpace |
| The Source Study | | | |
| A.1 Source Study Title | Award (title, URI, number) | -- not present -- | -- not present -- |
| A.2 Study Identifiers | -- not present -- | -- not present -- | -- not present -- |
| A.3 Study Topics | Subject | Subject | Subject keywords |
| A4. Other Study Titles | -- not present -- | -- not present -- | -- not present -- |
| *Comments*  Although other schemes can describe relationships *between data objects*, they do not normally explicitly describe any relationship between a data object and the *activity* (i.e. research project) that generated it. This linkage is, however, crucially important to a metadata repository for clinical research, where each study can generate a large number of objects, possibly located in different places, but very often discovered via the study (name, registry Id, etc.). In most cases, the study topics will be equivalent to the data object topics of other schemas. | | | |
| Data Object Identifiers | | | |
| B.1 DOI | Identifier (DOI) | -- not present -- | -- not present -- |
| B.2 Object Title | Title | Title | Title |
| B.3 Version | Version | -- not present -- | -- not present -- |
| B.4 Object Other Identifiers | Alternate Identifier | -- not present -- | -- not present -- |
| B.5 Object Additional Titles | Title, with type property | Title, alternative | Alternative title |
| Data Object Provenance | | | |
| C.1 Creators | Creator | Creator | Data Creator |
| C.2 Contributors | Contributor, Funding Reference | Contributor | Funder |
| D.1 Creation Year | Publication Year | -- not present -- | -- not present -- |
| D.2 Dates | Date | Period | Time period |

|  |  |  |  |
| --- | --- | --- | --- |
| ECRIN CR Metadata | DataCite | Dublin Core | DSpace |
| Data Object Attributes and Descriptors | | | |
| E.1 Resource Class | Resource General Type | Type | Type |
| E.2 Resource Type | Resource Type | -- not present -- | -- not present -- |
| E.3 Record key type | -- not present -- | -- not present -- | -- not present -- |
| E.4 Identifier type | -- not present -- | -- not present -- | -- not present -- |
| E.5 Associated consent | -- not present -- | -- not present -- | -- not present -- |
| E.6 Description (free text, suggested content) | Description, with type sub-property | Description; abstract,  coverage | Dataset Description;  Spatial coverage; |
| E.7 Language | Language | Language | Language |
| E.8 Related Resource Identifiers, with relation type indicator | Related Identifier, with relation type indicator | Relation, (is version of; is referenced by; replaces) | Relation (is version of, is referenced by, supersedes) |
| E.9 Topics (of data object) | Subject | Subject | Subject keywords |
| Location and Access Details | | | |
| F.1 Publisher / Provider | Publisher | Publisher | Data Publisher |
| F.2 Access Type | -- not present -- | -- not present -- | -- not present -- |
| F.3 Access Details | -- not present -- | -- not present -- | -- not present -- |
| F.4 Access Contact | -- not present -- | -- not present -- | -- not present -- |
| F.5 Resources | Size, Format | Description, table of contents | Dataset description (TOC) |
| F.6 Rights | Rights | Rights | Rights |

# Differences between versions 2 and 1

Version 1 of this schema was published in the journal Trials in 2016, as:

Canham s, Ohmann C, A metadata schema for data objects in clinical research. *Trials* **17**:557. (<https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-016-1686-5>)

The current version has four additional (mostly composite) data points, and has lost one data point, compared to version 1. Two elements have also – slightly – changed their names.

In several cases there has been a change only in the coding of the elements between the two versions, though these codes are for internal reference only, and are *not* designed to play any part in any system making use of the metadata schema. The differences and similarities between the systems are tabulated below., with green representing no change, yellow a change, blue an addition and red a removal.

|  |  |  |  |
| --- | --- | --- | --- |
| Version 2 | | Version 1 | |
| The Source Study | | | |
| A.1 | Source Study Title | A.1 | Source Study Title |
| A.2 | Study Identifiers | A.2 | Study Identifiers |
| A.3 | Study Topics | A.3 | Study Topics |
| A4. | Other Study Titles | **NEW in V2** | |
| Data Object Identifiers | | | |
| B.1 | DOI | B.1 | DOI |
| B.2 | Object Title | B.3 | Object Title |
| B.3 | Version | B.5 | Version |
| B.4 | Object Other Identifiers | B.2 | Object Other Identifiers |
| B.5 | Object Additional Titles | B.4 | Object Additional Titles |
| Data Object Provenance | | | |
| C.1 | Creators | C.1 | Creators |
| C.2 | Contributors | C.2 | Contributors |
| D.1 | Creation Year | D.1 | Creation Year |
| D.2 | Dates | D.2 | Dates |

|  |  |  |  |
| --- | --- | --- | --- |
| Version 2 | | Version 1 | |
| Data Object Attributes and Descriptors | | | |
| E.1 | Resource Class | E.1 | Resource Type general |
| E.2 | Resource Type | E.2 | Resource Type |
| E.3 | Record key type | **NEW in V2** | |
| E.4 | Identifier type | **NEW in V2** | |
| E.5 | Associated consent | **NEW in V2** | |
| E.6 | Description | E.3 | Description |
| E.7 | Language | E.5 | Language |
| E.8 | Related Resource Identifiers | E.6 | Related Resource Identifiers |
| E.9 | Topics (of data object) | E.4 | Topics (of data object) |
| Location and Access Details | | | |
| F.1 | Publisher / Provider | F.1 | Publisher |
| Removed in V2 | | F.2 | Other Hosting Institutions |
| F.2 | Access Type | F.3 | Access Type |
| F.3 | Access Details | F.4 | Access Details |
| F.4 | Access Contact | F.5 | Access Contact |
| F.5 | Resources | F.6 | Resources |
| F.6 | Rights | F.7 | Rights |

The ‘other Hosting Institutions’ was removed partly because of the difficulty of obtaining this data (at least when metadata was applied initially), and partly because the information should be available from the ‘Resources’ composite data point(s).

# Data Structures

The database required to store the metadata could be a traditional relational database, or a ‘noSql’ document based system, or use a system that included both types of data organisation (e.g. a relational system where some fields held XML or JSON data). For better support of searching a relational database structure is proposed here.

The two main entities in the system are a) Studies and b) Data Objects, with a join table linking the two, as shown below.

A screenshot of a cell phone

Description generated with very high confidence

The join table is necessary because not only can one study be associated with multiple objects, but one data object (e.g. a meta-analysis) may be associated with multiple studies.

Other data are either child entities or attributes of the studies, or child entities or attributes of the data objects, together with a few lookup tables to provide selection options for some of those attributes.

The study related data is shown below. There are none, one or more study identifiers, study topics and study ‘other titles’ associated with each study. In each case there are further links to look up tables. As the system develops the use of look up tables could be extended, e.g. in the study identifiers table by using OrganisationId, referencing an Organisations table, rather than the current textual Organisation field.

A screenshot of a computer

Description generated with very high confidence

The topics data is expected to focus on the disease or condition being investigated, the interventional agent(s) being used, and any methodological specialisms. The expectation is that the topics will be identified using controlled terminology from category systems (e.g. ICD10, ATC) rather than free text. The topic type Id references the list of topic vocabularies being used, which provides for each the name, including the version, and a URL that allows the specific coding system to be accessed. Similar studies could be identified by searching on specific topics.

The intention is to keep the data stored about the study relatively simple at this stage (further data could be obtained by following the links to registry entries, that should be included in the StudyIdentifiers table, at least for all trials).

The Data Object data has equivalent links to identifiers, topics and alternative titles as the study data, and they link back to the same look up tables, as shown below.

A screenshot of a computer

Description generated with very high confidence

The other identifiers are for any that a data object has which are not DOIs, as this is the default identifier for data objects (e.g. the sponsor’s code for a protocol document). The other titles include sub titles and translated titles, and the data object topics are for a collection of associated subjects or topics. As explained in the metadata description, however, such topics are more efficiently linked to the study, because they will tend to be shared by all data objects from a particular study.

The data object table includes those fields that are only present as singletons, if at all, for instance the DOI, title, and access type. Seven of these fields are integer foreign keys referencing look up tables, as shown below.

A screenshot of a computer

Description generated with very high confidence

This is the case for the IDs representing the data object language, the broad class of data object, the specific type, the access type, and (for datasets only) the type of record keys and identifiers within the data set and the associated consent. It might be possible in the future to replace the ‘Publishing Org’ text field with an integer referencing an organisations table.

The Data Object table is also linked to several ‘child’ tables. The two shown in the diagram above are the object descriptions (which may be multiple because it may be present in different languages, hence the link to the language lookup table), and the data object creators, which may be individuals or organisations, or a mix of the two. For individuals there is an opportunity to store ORCID Ids. The name is structured as family and given names to allow different citation patterns to be used.

The other child entity tables are shown in the diagram below. They include key dates, the relationship between the data object and other resources (many of which are also likely to be in the system, and therefore have an ‘internal’ system Id as well as a public one), any additional contributing persons or organisations (with the data organised similarly to data object creators but with an extra field to indicate the contribution type), and the URLs and physical descriptions of the data object(s), which may be available in multiple versions, considered as data object resource(s).

Again, many of these link themselves to simple look up tables that provide integer codes for key values, e.g. date type, contribution type.

A map of a computer

Description generated with high confidence

Note that this structure does *not* include the audit fields that will be required (date initially added, date most recently confirmed, source system etc.) nor does it include the tables that will be required to support the metadata harvesting process (source repository list, end-point details, staging tables, etc.). These will need adding as the system develops but are not necessary for the core metadata store, and have therefore been omitted for simplicity.

A script can be generated to create the database (in SQL Server) and that should be readily translatable into other dialects of SQL.