nature publishing group npg

SCIENTIFIC DATA

ISA-Tab configuration specification for the experimental metadata component of a Data Descriptor

Status

Production version 1b[1,2].

Authors

This configuration has been defined by the *Scientific Data*[3] team (Ruth Wilson, Andrew Hufton, Victoria Newman and Susanna-Assunta Sansone) and Philippe Rocca-Serra, lead representative for the ISA community. The configuration is based on the original ISA-Tab specification document published by the ISA Commons[4] community, and also draws on discussions with several individuals, groups and members of *Scientific Data*'s Advisory Panel.

Abstract

Scientific Data publishes Data Descriptors (DDs)[5], articles that combine the *narrative* content typical of traditional scientific manuscripts with *structured* information (experimental metadata records) to provide detailed descriptions of biological, biomedical and environmental science datasets. DDs focus exclusively on how, and by whom, datasets were produced, and how they may be reused by independent investigators. *Scientific Data* provides authors with two templates[6], one for each manuscript component, to help authors to submit a DD. The experimental metadata component of the DD can also be submitted directly as an ISA-Tab file. Here we describe the configuration of this latter component, which follows the ISA syntax as defined in its original format specification[7] and also contains additional fields and restrictions to fulfil the requirements of *Scientific Data*.

Target users

This document is for ***advanced users*** or ***data service providers***, existing ISA-Tab users, or those interested in implementing ISA-Tab export from their data service. In addition, an ISA configuration file tailored to *Scientific Data* is also available for existing or prospective ISA tools users[8]; certain fields can be adapted and further extended according to the type of study and the study subjects, but only in the event that the rules set out in this technical specification are adhered to.

Other submitters should refer to the "Templates" page[9], where an Excel template is provided as an alternative for submitting the experimental metadata component of the DD. All submitters should also refer to the "Submission Guidelines"[10] for information on formatting and submitting the *narrative* component of the DD and/or associated data files.

Please note that Section 1 is a summary extracted from the original ISA-Tab specification[6] to introduce its elements; Section 2 describes the configuration in detail.

---

[1] http://www.nature.com/sdata/for-authors/submission-guidelines#metadata
[2] Released as CC BY 4.0; see http://creativecommons.org/licenses/by/4.0/deed.en_GB for terms of licence
[3] http://www.nature.com/sdata
[4] http://isacommons.org
[5] http://blogs.nature.com/scientificdata/2013/09/19/the-data-descriptor-making-your-data-reusable
[6] http://www.nature.com/sdata/for-authors/submission-guidelines#templates
[7] http://www.isa-tools.org
[8] http://www.isa-tools.org/
[9] http://www.nature.com/sdata/for-authors/submission-guidelines#templates
[10] http://www.nature.com/sdata/for-authors/submission-guidelines

Table of Contents

1. Introduction to ISA-Tab

1.1. Overview of the ISA-Tab structure

The three key entities around which the general-purpose ISA-Tab format for structuring and communicating experimental metadata is built are the **Investigation**, the **Study** and the **Assay**. The hierarchical structure of this format enables the representation of studies employing one or more technologies.

Three types of files are used to capture the experimental metadata, and together describe a complete experimental workflow (see Figure 1). These are:

- The Investigation file
- The Study file
- The Assay file.

The **Investigation** file contains all the information necessary to understand the design and overall goals of an experiment; experimental steps (or sequences of events) are described in the **Study** and **Assay** files. *Please note that for each Investigation file there may be one or more Study files, and for each Study file there may be one or more Assay files. Each file has a defined structure, with fields being organized on a per-column or per-row basis; each file is described briefly in the subsections below*.

The experimental metadata component of a *Scientific Data* DD can also be submitted directly as an ISA-Tab file following the instructions in this specific configuration of the ISA-Tab format.

1.1.1. Investigation file

The Investigation file is intended to meet four needs: (*i*) to define key entities, such as *Factors* or *Protocols*, that may be referenced in the Study or Assay files; (*ii*) to track the provenance of the terminology (from controlled vocabularies or ontologies) used in these files, where applicable; (*iii*) to relate Assay files to Study files; and, optionally (*iv*), to link each Study file to an Investigation—which only becomes necessary when two or more Study files must be grouped (for example, when related DD manuscripts are submitted).

In the Investigation file, information is reported on a **per-column basis** and the fields are organized and divided into various sections.

The declarative sections cover basic information such as the authors who have contributed to dataset generation and their institution(s); additional fields have been introduced, using the *Comment[ ]* element, to collect more information on the development of the dataset, such as grant number(s) and funder(s). This section is linked to the manuscript component of the DD, which also recapitulates the narrative content featured in the Investigation file (e.g. Abstract, Background and Summary, Methods, etc.).

The Investigation file also contains an optional Investigation section, but the two should not be confused (note that certain design decisions are legacy syntax from other formats with which ISA-Tab is devised to be compatible). This optional section is a flexible mechanism for grouping two or more Study files, as may be required by various use cases. In toxicogenomics, for example, acute toxicity studies are followed by long-term and *in vitro* toxicity studies. Another example comes from environmental genomics, where several studies carried out in the same location can be drawn together under one Investigation. Note that the experimental metadata component of a *Scientific Data* DD is equivalent to one Study. This means that the Investigation section should be left blank unless two DD manuscripts are submitted simultaneously and must be associated with one another.

1.1.2. Study file

The Study file contains contextualizing information for one or more assays: for example, the subjects studied; their source(s) and characteristics; the sampling methodology; and any treatments or manipulations performed to prepare the specimens. Note that "subject" as used above could refer, *inter alia*, to an organism, tissue or environmental sample.

In this file, information is structured on a **per-row basis**, with the first row used for column headers.

1.1.3. Assay file

The Assay file represents a portion of the experimental workflow (a related series of procedures used to generate a single dataset); each Assay file must contain assays of the same type, defined by the type of measurement (e.g. gene expression) and the technology employed (e.g. DNA microarray). Assay-related information includes protocols, additional information relating to the execution of those protocols, and references to data files (whether raw or derived).

In this file, as for the Study files, fields are organized on a **per-row basis**, with the first row being used for column headers.

Please note that this specific configuration (for *Scientific Data*; configuration version 1) only uses the Generic Assay file layout from the original ISA-Tab format specification (see section 2.4), and conversion to other supported formats is therefore limited.

1.2. Relating Study and Assay files

In a study that looks at the effect of a compound inducing liver damage in rats by characterizing the metabolic profile of urine (by NMR spectroscopy), as well as measuring protein and gene expression in the liver (with mass spectrometry and DNA microarrays, respectively), there will be one Study file and three Assay files in addition to the Investigation file.

- The Study file will contain information on the rats (the subjects studied), their source(s) and characteristics, the description of their treatment with the compound of interest, and the steps undertaken to collect samples of urine and liver from them.
- The Assay file for the metabolic profiling of urine (measurement) by NMR spectroscopy (technology) will contain the (stepwise) description of the methods by which the urine was processed for the assay, subsequent steps and protocols, and the link to the resulting raw and derived data files.
- The Assay file for the gene expression profiling (measurement) by DNA microarray (technology) will contain the (stepwise) description of how RNA extracts were prepared from the rats' livers (or sections thereof), how extracts were labelled, how hybridization was performed, and so on, and will also contain links to the resulting raw and derived data files.
- The Assay file for the protein expression profiling (measurement) by mass spectrometry (technology) will contain the (stepwise) description of how protein extracts were prepared from the livers (or sections thereof), the manner in which they were labelled, etc., and will also contain the links to the resulting raw and derived data files.

1.3. Data files and supplementary information

DDs link to and highlight both primary data (directly produced by an experimental or observational procedure) and supplementary files (for example code, models, workflows and summary tables). Like ISA-Tab records, the structured component of the DD focuses on modelling experimental metadata, and primary data files are considered external entities.

Scientific Data will provide a searchable publication platform to assist researchers in finding high-quality datasets archived within many different data repositories, but will not host primary research

data itself, as described in the "Data Deposition Policies" page on the *Scientific Data* website[11]. Any files presenting primary data should be submitted to an appropriate external repository and described in detail in the "Data Records" section of the DD manuscript. For guidelines on how to format and where to submit data, submitters should refer to the "Submission Guidelines" page[9].

1.4. Mandatory minimal content, terminology and special values

According to the original format specification[6], ISA-Tab has no mandatory fields, and values may be either in free text or from controlled vocabularies or ontologies; but all blocks and their sections, subsections and fields must be present (see section 2.1.10). The decision on how to regulate the use of the various fields is a matter for those communities implementing the format, and in such cases additional constraints should be agreed upon and incorporated into each community-specific configuration, as has been done for this *Scientific Data* configuration.

Initial submissions to Scientific Data *need not be accompanied by complete Investigation, Study or Assay files: files with values missing from any field may accompany DD manuscripts (please note that these files will not be ISA-compatible if the fields themselves are not present, see section* 2.1.10*). However, upon acceptance and before publication of DD manuscripts, the ISA-Tab files should be completed as much as possible.* In this document therefore selected fields are marked as **mandatory**, indicating that they must be filled in before publication when relevant; the use of terms from **controlled vocabularies / ontologies** may also be indicated for these fields. Additionally, some fields will require that the user **select from a list** of terms provided, while other fields will be populated with a **default value**. The *Scientific Data* curation team will provide assistance with terms from controlled vocabularies / ontologies and work to harmonize them internally. Open community ontologies must be used, e.g. those under the OBO Foundry[12] umbrella and accessible from the BioPortal[13] service (tagged as "Production Level"). For certain fields (such as DOIs and links to the manuscript component of the DD) the values will be added by *Scientific Data* curators, and these fields are marked accordingly. Furthermore, in the Study and Assay files, fields corresponding to certain required elements are marked as **core *Scientific Data* elements** rather than **additional *Scientific Data* elements**, as elsewhere.

1.4.1. Richer ISA-Tab files

Current ISA-Tab and ISA tools users planning to submit to *Scientific Data* should use this specific configuration of the ISA-Tab format or the provided ISA configuration file[7], which can, however, also be enriched—for example, using additional fields or more values from controlled vocabularies / ontologies—as long the enrichment follows the ISA-Tab syntax. All ISA-Tab records submitted to *Scientific Data* will be validated according to this particular ISA configuration[14].

Current ISA-Tab and ISA tools users should also be aware that, because this first version of the *Scientific Data* configuration uses only the Generic Assay file layout from the original ISA-Tab format specification (see section 2.4), *the conversion of the ISA-Tab file into other formats, such as MAGE-Tab, SRA-XML and PRIDE-ML, will be limited* (for further details, please refer to the original ISA-Tab format specification[6]).

Where experiments include clinical or non-clinical studies, ISA-Tab can be complemented by existing biomedical formats such as SDTM[15] by formally capturing information about the interrelationship of the various parts they describe. SDTM encompasses both the Standard for Exchange of Nonclinical Data (SEND) and the Clinical Data Interchange Standards Consortium (CDISC); SDTM has been endorsed by the US Food and Drug Administration (FDA) as the preferred way to organize, structure

[11] http://www.nature.com/sdata/data-policies
[12] http://www.obofoundry.org
[13] http://bioportal.bioontology.org
[14] http://www.isa-tools.org/
[15] http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm

and format both clinical and non-clinical (toxicological) data submissions. A reference system has been created in the original ISA-Tab format specification to allow SDTM source(s) and sample(s) to be referenced from a Study file. When required, observations in an SDTM file can be referenced from the Study file by matching SDTM variables for source(s) and sample(s) [in square brackets] to the corresponding Source Name and/or Sample Name column headers. For example, to reference the SDTM subject ID and the laboratory reference for the sample, the following specific column headers can be created: Source Name[USUBJID] and Sample Name[IDVAR=LBREFID] (the sample IDVAR is set to LBREFID which points to an SDTM Laboratory Domain as identified by the 2 LB letters).

2. DD configuration – detailed structure

2.1 Formatting rules

2.1.1. Field separator

The official column delimiter accepted is the Unicode Horizontal Tab character (Unicode codepoint 0009).

2.1.2. File encoding

The UTF-8 encoding of the Unicode character set is preferred. However, parsers should be able to recognize the Unicode encoding used and adapt accordingly.

2.1.3. File naming conventions

In order to facilitate identification of files, specific extensions have been created as follows:

- **i_**xxxxx**.txt** for identifying the Investigation file
- **s_**xxxxx**.txt** for identifying Study file(s)
- **a_**xxxxx**.txt** for identifying Assay file(s)

where 'xxxxx' is the user-assigned name or ID.

2.1.4. Case sensitivity and expected syntax

All labels are **case-sensitive**. In the Investigation file, section headers must be completely written in upper case (e.g. *STUDY*); field headers have the **first letter of each word in upper case** (e.g. *Study Identifier*), with the exception of the referencing label (*REF*). In the Study or Assay files, column headers also have the first letter of each word in upper case, with the exception of the referencing label (*REF*). This will facilitate visualization of headers and fields when the file is viewed in spreadsheet software.

Please also note that, where field or column headers in the Investigation, Study or Assay files reference, or are qualified by, additional elements given in square brackets '[]', <space> characters are not permitted immediately before the opening square bracket.

2.1.5. Dates

Dates should be supplied in the ISO 8601 format (YYYY-MM-DD).

2.1.6. Object identifiers

All values found in fields with headers that contain the string "Name" (e.g. *Source Name*) or "File" (e.g. *Raw Data File*) are considered object identifiers. In Study and Assay files such object identifiers represent "nodes" in the experimental workflow (throughout this document "node" denotes biological materials such as samples or data objects, while "edges" show the relationships between nodes). All such object identifiers must be locally unique within an ISA-Tab formatted file. They may also be fully qualified external accession numbers, Digital Object Identifiers (DOIs), etc.

2.1.7. Referencing

Objects defined in the Investigation file can be referenced in either the Study or the Assay files. Two mechanisms can be used to indicate a reference:

- A **REF** label component; for example, *Protocol REF* or *Term Source REF*.
- The notation **square brackets + object name** is used to reference *Parameters* and *Factors* declared in the Investigation file, as in *Parameter[oven temperature]* or *Factor Value[compound]*.

2.1.8. Free text descriptions

Since text is stored in a single tab-delimited field, any embedded tabs or new line characters must be protected by enclosing such text within double quotes (“ ”; Unicode U+0022). Any double quotes within the text should be protected using the same mechanism to ensure their preservation.

2.1.9. Multiple values per field

In several sections, multiple values may be supplied within a single field by using semi-colons (;) as value separators (Unicode U0003+B); see also Tables 2 and 4.

2.1.10. Order of blocks and fields in the Investigation file

In the Investigation file, the three main blocks must be ordered as shown in section 2.2: ‘Ontology Source Reference’, ‘Investigation’, ‘Study’. Investigation files missing one of these blocks or in a different order are incorrect. Within the same block, the subsections (e.g. in the ‘Study’ block in the ‘Study Contact’ subsection) may be moved around. Also within each subsection, the order of fields (e.g. *Study Contact Name*) is not fixed. However, all blocks and their sections, subsections and fields must be present.

2.1.11. Notes

In Investigation, Study and Assay files, rows in which the first character in the first column is Unicode U+0023 (the # character) will be interpreted as notes: for example, to communicate something to the *Scientific Data* curators. ISA-Tab parsers will ignore lines commencing with # symbols entirely. If the symbols are found in any other position, however—for example, a cell containing “Sample #2”—they will not be ignored.

2.2. Compiling an Investigation file

In the Investigation file the fields are organized on a **per-column basis** and divided into **sections** with specific **fields** that are described in detail below. Table 1 describes the types of values allowed in each field together with some comments that implementers may find useful; Table 2 provides information on the number of items allowed in each field. The DD configuration contains some mandatory fields, and may also stipulate that values in certain fields come from controlled vocabularies / ontologies. These requirements are marked in the relevant sections.

2.2.1. Ontology source section

This section is identical to that in the ISA-Tab format.

ONTOLOGY SOURCE REFERENCE

Term Source Name

The name of the source of a particular term: i.e., the source of the controlled vocabulary or ontology. These names will be used in all corresponding *Term Source REF* fields. Only abbreviations are accepted; these can be found in BioPortal[16]. For example, the abbreviation “OBI” can be used to identify the source “Ontology for Biomedical Investigations”.

Term Source File

A file name or Uniform Resource Identifier (URI) of an official resource.

Term Source Version

The version number of the *Term Source* to support term tracking.

Term Source Description

This term is used for disambiguating resources when homologous prefixes have been used.

[16] http://bioportal.bioontology.org

2.2.2. Investigation section

This section of the annotation follows the ISA-Tab format. However, please note that the Investigation section is required only to group two or more Study files: as explained in section 1.1.1., the experimental metadata component of the DD is equivalent to a single Study. Therefore, during preparation of an ISA-Tab submission to *Scientific Data*, this section should be left blank unless two DD manuscripts are submitted simultaneously and require linking via this section.

As the Investigation section will usually not be relevant to a single DD submission, *Scientific Data* will in general not attempt to verify the accuracy of any details input into the following fields (in grey font). Please contact us at scientificdata@nature.com for more information or with any questions, or advise us at the time of DD submission should you wish to use this section in linking multiple manuscripts.

INVESTIGATION

This section is organized into several subsections that are described in detail below.

Investigation Identifier
(optional)
An identifier (for example, a DOI) assigned to the Investigation.
Investigation Title
A concise name given to an Investigation that connects two or more Studies.
Investigation Description
A textual description of, or link to a description of, the Investigation, including how two or more Studies are related.
Investigation Submission Date
(left blank)
Investigation Public Release Date
(left blank)

INVESTIGATION PUBLICATIONS

Each publication associated with an Investigation has its own column in this section. Such publications must specifically deal with the Investigation as a whole: publications relating to individual Studies must instead be referenced in the Study sections. Information may be supplied using as many additional columns as needed.

Investigation PubMed ID
The PubMed IDs of the publication(s) listed as associated with this Investigation.
Investigation Publication DOI
A DOI for each publication (where available).
Investigation Publication Author List
The list of authors associated with each publication.
Investigation Publication Title
The title of each publication associated with the Investigation.
Investigation Publication Status
(select from list)
A term describing the status of the publication. Use one of the following options:

- in preparation
- submitted
- published.

INVESTIGATION CONTACTS

Investigation Person Last Name
The last name (surname) of the primary contact person associated with the Investigation.

Investigation Person First Name
The first name of the primary contact person associated with the Investigation.
Investigation Person Mid Initials
The middle initials of the primary contact person associated with the Investigation.
Comment[Investigation Person ORCID]
A unique Open Researcher and Contributor IDentifier (ORCID)[17]. Please note that a <space> character may not appear immediately before the opening square bracket in this field.
Investigation Person Email
The email address of the primary contact person associated with the Investigation.
Investigation Person Phone
The telephone number of the primary contact person associated with the Investigation.
Investigation Person Fax
The fax number of the primary contact person associated with the Investigation.
Investigation Person Address
The address of the primary contact person associated with the Investigation.
Investigation Person Affiliation
The organizational affiliation for the primary contact person associated with the Investigation.
Investigation Person Roles
Term(s) to classify the role(s) performed by the primary contact person in the context of the Investigation.

2.2.3. Study section

The Study section is organized into several subsections (described in detail below). This section represents a repeatable block that is replicated in accordance with the number of Studies reported (e.g., an Investigation with two Studies will necessitate an Investigation file with two Study blocks). The subsections are arranged vertically and their order may vary within this repeatable block, although each field must remain within its subsection.

Please note that many fields within this section will be completed by *Scientific Data* following information provided to our manuscript tracking system during DD submission, and are marked as such.

STUDY

Study Identifier
(to be added by *Scientific Data*)
A DOI, to be provided by *Scientific Data*, corresponding to that of the DD manuscript.
Study File Name
(mandatory)
A field to specify the name of the Study file corresponding to the definition of that Study.

 # Sections below enable linking to the relevant fields in the DD manuscript, in which further narrative description of the Study—including the Background and Summary, Technical Validation, Author Contributions and Usage Notes—is provided.

Study Title
(to be added by *Scientific Data*; can optionally be completed by authors)
A textual description encapsulating the purpose and goal of the Study. This should match the Title given in the DD manuscript and should not exceed 110 characters, including spaces.
Study Description
(to be added by *Scientific Data* during manuscript typesetting)
A URI pointing to the Abstract section of the DD manuscript.

[17] http://orcid.org

Study Submission Date
(to be added by *Scientific Data*)
The date on which the Study was submitted to *Scientific Data*.
Study Public Release Date
(to be added by *Scientific Data*)
The date on which the DD may be released publicly.

 # Please note that <space> characters may not appear immediately before the opening square bracket '[' in the following four blocks of 'Comment' sections.

Comment[Manuscript Licence]
(to be added by *Scientific Data*)
The Creative Commons licence selected for the DD manuscript in the *Scientific Data* manuscript tracking system. Use one of the following options:

- CC BY 4.0
- CC BY-NC 4.0
- CC BY-NC-SA 4.0.

Comment[Experimental Metadata Licence]
(mandatory; default value)
All ISA-Tab formatted experimental metadata within a DD will be made available under the CC0 protocol to promote maximum reuse; the only value accepted in this field is CC0.

 # Sections below describe data files associated with the DD.

Comment[Data Repository]
(mandatory)
The names of any external data repositories to which the primary data files have been submitted. Multiple repositories should be semicolon-separated.
Comment[Data Record Accession]
(mandatory)
Dataset identifier(s) assigned by any external data repositories. Multiple datasets should be semicolon-separated.
Comment[Data Record URI]
(mandatory)
URI(s) pointing to data record(s) in any external data repositories. Multiple URIs should be semicolon-separated.

 # Sections below describe supplementary files other than data files that might be associated with a DD. Please note that supplementary files should be uploaded to the Scientific Data *manuscript tracking system at the time of DD submission, and the fields below will be completed automatically during manuscript publication.*

Comment[Supplementary Information File Name]
(to be added by *Scientific Data*)
The name of a figure, table or other supporting file used in any of the above sections: for example, a table containing summary information such as sample numbers, demographics, etc. Multiple Supplementary Information file names should be semicolon-separated.
Comment[Supplementary Information File Type]
(to be added by *Scientific Data*)
Term(s) describing the type of supplementary file(s) appended to the DD manuscript in the *Scientific Data* manuscript tracking system. Multiple file-types should be semicolon-separated.
Comment[Supplementary Information File URI]
(to be added by *Scientific Data*)

A URI pointing to the location of the Supplementary Information file on the *Scientific Data* website. Multiple URIs should be semicolon-separated.

Comment[Subject Keywords]

(to be added by *Scientific Data*)

Terms(s) describing the scientific emphasis of the Data Descriptor. Multiple terms should be semicolon-separated.

STUDY DESIGN DESCRIPTORS

Study Design Type

(controlled vocabulary / ontology)

A term allowing the classification of the study based on the overall experimental design (e.g. crossover design; parallel group design). The term should be from a controlled vocabulary / ontology; the fields below are required to track its provenance.

Study Design Type Term Accession Number

The accession number from the *Term Source* associated with the selected term.

Study Design Type Term Source REF

Identifies the controlled vocabulary or ontology in which the term originates. The Study Design Term Source REF must match the *Term Source Name* declared in the ontology section.

STUDY PUBLICATIONS

Study PubMed ID

The PubMed ID(s) of the publication(s) associated with this DD (where available).

Study Publication DOI

The DOI(s) of the publication(s) associated with this DD (where available).

Study Publication Author List

The list of authors associated with the publication(s).

Study Publication Title

The title(s) of the publication(s).

Study Publication Status

(select from list)

A term describing the status of the publication(s). Use one of the following options:

- in preparation
- submitted
- published.

STUDY FACTORS

Study Factor Name

The name of a factor used in the Study and/or Assay files. A “factor” corresponds to an independent variable manipulated by the experimentalist with the intention to affect, for example, biological or environmental systems in a way that can be measured by an assay. The value of a factor is given in the Study or Assay file.

Study Factor Type

(controlled vocabulary / ontology)

A term allowing the classification of the factor into categories. The term should be from a controlled vocabulary / ontology, and the fields below are required to track its provenance.

Study Factor Type Term Accession Number

The accession number from the *Term Source* associated with the selected term.

Study Factor Type Term Source REF

Identifies the controlled vocabulary or ontology from which the term derives. The Source REF must match a *Term Source Name* declared in the ontology section.

STUDY ASSAYS

The Study Assay section declares and describes each Assay file associated with the current Study.

Study Assay Measurement Type

(mandatory; controlled vocabulary / ontology)

A term to qualify the endpoint, or what is being measured (e.g. gene expression profiling; protein identification). The term should be from a controlled vocabulary / ontology and the fields below are required to track provenance.

Study Assay Measurement Type Term Accession Number

The accession number from the *Term Source* associated with the selected term.

Study Assay Measurement Type Term Source REF

The Source REF must match a *Term Source Name* declared in the ontology section.

Study Assay Technology Type

(mandatory; controlled vocabulary / ontology)

A term to identify the technology used to perform the measurement (e.g. DNA microarray; mass spectrometry). The term should be from a controlled vocabulary / ontology and the fields below are required to track provenance.

Study Assay Technology Type Term Accession Number

The accession number from the *Term Source* associated with the selected term.

Study Assay Technology Type Term Source REF

Identifies the controlled vocabulary or ontology from which the term derives. The Source REF must match a *Term Source Name* declared in the ontology section.

Study Assay Technology Platform

The manufacturer and name of the technology platform used in the assay (e.g. Bruker AVANCE).

Study Assay File Name

(mandatory)

The name of the Assay file corresponding to the definition of that assay.

STUDY PROTOCOLS

Study Protocol Name

(mandatory)

The name(s) of the protocols used in the experimental sequence; these should match the names given in the Methods section of the manuscript component of the DD. The protocol names are used as identifiers within the document and will be referenced in the Study and Assay files in the *Protocol REF* columns. Names may be local identifiers, fully qualified external accession numbers or DOIs.

Study Protocol Type

(controlled vocabulary / ontology)

A term to classify the protocol. The term should be from a controlled vocabulary / ontology and the fields below are required to track provenance.

Study Protocol Type Term Accession Number

The accession number from the *Term Source* associated with the selected term.

Study Protocol Type Term Source REF

Identifies the controlled vocabulary or ontology from which the term derives. The Source REF must match a *Term Source Name* declared in the ontology section.

Study Protocol Description

(left blank)

This field will not be used by *Scientific Data* in this initial version of the configuration. Full descriptions of all protocols should be provided in the Methods section of the DD manuscript, and these are linked to using the 'Study Protocol URI' field.

Study Protocol URI

(to be added by *Scientific Data*)

A URI pointing to the Methods section of the DD manuscript.

Study Protocol Version
An identifier for the version of the protocol used to ensure adequate protocol tracking.
Study Protocol Parameters Name
(controlled vocabulary / ontology)
A list of the names used as identifiers for each protocol parameter within the ISA-Tab document; the list should be semicolon- (;-) delimited. Protocol parameter names are entered into the *Parameter Value[<parameter name>]* column in the Study and Assay files. Terms can be free text or from, for example, a controlled vocabulary / ontology. If the latter source is used, the *Term Accession Number* and *Term Source REF* fields below are required.

Study Protocol Parameters Name Term Accession Number
The accession number from the Term Source associated with the selected term.
Study Protocol Parameters Name Term Source REF
The controlled vocabulary or ontology from which the term derives. The Source REF must match a *Term Source Name* declared in the ontology section.

Study Protocol Components Name
A semicolon- (;-) delimited list of a protocol's components (e.g. instrument names, software names and reagent names).
Study Protocol Components Type
(controlled vocabulary / ontology)
A term to classify the protocol components listed (e.g. instrument, software, detector or reagent). Terms can be free text or from, for example, a controlled vocabulary or ontology. If the latter source is used, the *Term Accession Number* and *Term Source REF* fields below are required.

Study Protocol Components Type Term Accession Number
The accession number from the Source associated with the selected terms.
Study Protocol Components Type Term Source REF
Identifies the controlled vocabulary or ontology from which the term derives. The Source REF must match a *Term Source Name* previously declared in the ontology section.

STUDY CONTACTS

 # Authors may complete this section, but it is not required. At publication, it will be automatically populated and harmonized with author information from our manuscript tracking system and the manuscript component of the DD.

Study Person Last Name
The last name (surname) of the primary contact person associated with the Study.
Study Person First Name
The first name of the primary contact person associated with the Study.
Study Person Mid Initials
The middle initials of the primary contact person associated with the Study.
Comment[Study Person ORCID]
A unique Open Researcher and Contributor IDentifier (ORCID). Please note that a <space> character may not appear immediately before the opening square bracket in this field.
Study Person Email
The email address of the primary contact person associated with the Study.
Study Person Phone
The telephone number of the primary contact person associated with the Study.
Study Person Fax
The fax number of the primary contact person associated with the Study.
Study Person Address
The address of the primary contact person associated with the Study.

Study Person Affiliation
The organizational affiliation of the primary contact person associated with the Study.
Study Person Roles
Term to classify the role(s) performed by the contact person in the context of the Study (e.g. designed the experiment; analysed the data).

 # Please note that information contained within the following three sections should also be provided in the Acknowledgements section of the manuscript component of the DD. The fields below are optional but allow authors to provide this information in parallel in a structured, mineable fashion.

<Space> characters are not permitted immediately before the opening square bracket '[' in these three sections.

Comment[Funder]
The name of the funding agency(s) that have sponsored the author and the work detailed in the DD.
Comment[FundRef ID]
The unique FundRef[18] identifier for the funding agency(s).
Comment[Grant Identifier]
Unique identifiers provided by the funding agency(s).

2.3. Compiling a Study file

In the Study file the fields are organized on a **per-row basis** with the first row containing column headers. The Study file contains contextualizing information for one or more assays. The sections below describe in detail the Study file's column headers, organizing them as nodes (potentially containing the string *Name* or *File*; previously referred to as "sections") and attributes (previously referred to as "fields") for nodes and node-processing events, qualifiers for node attributes and other valid fields. Study files with all columns left empty are syntactically valid ISA-Tab files; however, those elements that must be present are marked **core *Scientific Data***. Within these elements, some fields are mandatory and/or values should come from controlled vocabulary / ontology terms, as indicated. Table 3 shows the nodes together with the (Study or Assay) file in which the node is used, as well as their possible attributes, the number of values allowed, data type and dependency on a parent node. Table 4 lists the attributes that can be used to qualify nodes, the number of values allowed, data type and dependency on a parent node.

2.3.1. Study file nodes

Source Name
(mandatory)
An identifier or name for the material, whether biological, environmental or other, that is considered the basis of a study: for example, a biological sample or a field site / habitat. *Source* items can be qualified using the following headers: *Characteristics[ ]*, *Material Type*, *Term Source REF*, *Term Accession Number*, *Unit*, *Description* and *Comment[ ]*. These attributes and their qualifiers are described in the sections below (2.3.2., 2.3.3., 2.3.4., 2.3.5.); core *Source* terms required by *Scientific Data* are clearly marked and distinguished from additional terms that can be used as needed to provide richer descriptions of the subjects studied.

Sample Name
Identifiers or names for samples that represent major outputs from a particular protocol. This column can be preceded by a *Protocol REF* block of elements (refer to section 2.3.3); see Figure 1. *Sample* items can be qualified using the following headers: *Characteristics[ ]*,

[18] http://www.crossref.org/fundref/fundref_registry.html

Material Type, *Term Accession Number*, *Term Source REF*, *Unit* and *Comment[ ].* These attributes and their qualifiers are described in the sections below (2.3.2., 2.3.3., 2.3.4., 2.3.5.); core *Sample* terms required by *Scientific Data* are clearly marked and distinguished from those that can be used as needed to provide richer descriptions of the subjects studied.

2.3.2. Attributes of Source Name and Sample Name nodes in the Study file

Core *Scientific Data* elements

Characteristics[<category term>]

(controlled vocabulary / ontology)

This column contains terms describing the material(s), sample(s) or subject(s) studied (e.g. organism, tissue or environmental samples) according to characteristics categories indicated in the column header. *Characteristics* that are specific to the samples, but not the source material, should be added as columns to the right of the *Sample Name* column.

If a term comes from a controlled vocabulary / ontology, the *Term Accession Number* and *Term Source REF* fields (see Qualifiers section) are required to track its provenance.

Because characteristics are domain- and organism-specific, this *Scientific Data* specification only provides lists of suggested characteristics categories. These lists will be expanded and refined progressively in collaboration with interested communities, according to their minimal information reporting requirements[19][20]. As the work progresses, the *Scientific Data* "Excel Templates"[8] (provided as an alternative for submitting the experimental metadata component of the DD) will also be updated.

For most biological and biomedical studies, the following characteristics categories are applicable; additional columns can be added as needed to provide richer descriptions of the *Source* and/or *Sample*.

Please note that <space> characters may not appear immediately before an opening square bracket '[' in *Characteristics* column headers.

Characteristics[organism]

(controlled vocabulary / ontology)

The taxonomic name of the organism used in a study or from which the starting biological material derives.

Characteristics[organism part]

(controlled vocabulary / ontology)

The anatomical part from which the source or sample derives.

Characteristics[cell line]

(controlled vocabulary / ontology)

The name of the cell line from which the source or sample derives.

For some environmental studies in which the environmental context of the source is pivotal, characteristics such as *Characteristics[habitat]* / *Characteristics[biome]*, *Characteristics[latitude]* / *Characteristics[longitude]* and *Characteristics[altitude] / Characteristics[elevation]* / *Characteristics[depth]* may be applicable. For other studies, *Characteristics[strain]*, *Characteristics[developmental stage]*, *Characteristics[disease status]*, etc. may be used. As for biological samples, additional columns may be added as needed to provide richer descriptions of the *Source* and/or *Sample* following the best-practice guidelines of the various communities.

[19] http://blogs.nature.com/scientificdata/2013/05/13/our-roadmap-to-engagement-your-call

[20] http://www.biosharing.org/standards/reporting_guideline

Additional *Scientific Data* elements

Material Type

This column may contain terms describing the *Material Type* (e.g. whole organism, organ, primary cells, immortalized cells) of the source or sample assayed. A material type that is specific to the sample(s), but not to the source material, should be added as a column to the right of the *Sample Name* column. If the term comes from a controlled vocabulary / ontology, the *Term Accession Number* and *Term Source REF* fields (see Qualifiers section) are required to track its provenance.

2.3.3. Attributes of processing events for Study file nodes

Core *Scientific Data* elements

Protocol REF

One or more *Protocol REF* columns should be used to specify the method used to transform a material. This column contains a reference to a *Protocol Name* (previously defined in the Investigation file) that should match the names in the Methods section of the DD manuscript.

Additional *Scientific Data* elements

A *Protocol REF* can be further refined with the following blocks of elements as needed, although these are very rarely used:

Parameter Value[<parameter term>]

This field allows reporting on the values assumed by the *Parameter* when a protocol is followed. Note that the term between the brackets ([ ]) must map to one (and only one) of the parameters defined in the Investigation file. Values can be qualitative or quantitative.

<Space> characters may not appear immediately before an opening square bracket '[' in *Parameter Value* column headers.

Performer

The name of the operator who carried out the protocol. This allows operator effects to be taken into account, and can serve as a form of quality control in data tracking.

Date

The date on which a protocol is performed. This allows possible variation in experimental results stemming from the day on which the data are collected to be taken into account, and can serve as quality control in data tracking. Dates should be reported in ISO format (YYYY-MM-DD).

2.3.4. Qualifiers for Study file nodes' attributes

Core *Scientific Data* elements

Term Accession Number

The accession number from the *Term Source* associated with the selected term, if this is from, for example, a controlled vocabulary or ontology. The accession number qualifies the headers *Characteristics[ ]*; *Material Type*; *Parameter Value[ ]* or *Factor Value[ ]*; and *Unit*.

Term Source REF

Identifies the controlled vocabulary or ontology from which the selected term derives. The Source REF must match a *Term Source Name* previously declared in the ontology section.

Additional *Scientific Data* elements

Unit

Unit is used if the terms provided in the *Characteristics[ ]*, *Parameter Value[ ]* or *Factor Value[ ]* columns classify data that are dimensional, i.e., quantitative values.

2.3.5 Other Study file fields

Core *Scientific Data* elements

Factor Value[<factor name>]

(controlled vocabulary / ontology)

A *"*factor*"* corresponds to an independent variable manipulated by the experimentalist with the intention to affect, for example, biological or environmental systems in a way that can be measured by an assay. This field holds the actual values for the *Factor Value* named between the square brackets (as declared in the Investigation file). For instance, for the *Factor Name* "genotype", the *Factor Value* might be the genotype information specific to each sample. *Factor Value* columns should be added, as needed, to explain all possible sources of variation among samples.

Qualifiers for *Factor Value* are also listed in section 2.3.4. Please note that <space> characters may not appear immediately before an opening square bracket '[' in *Factor Value* column headers.

Additional *Scientific Data* elements

Comment

As used in the Investigation file, *Comment* columns can be added to Study files when no other appropriate fields exist.

2.4. Compiling a Generic Assay file

In the Assay file the fields are organized on a **per-row basis**, with the first row containing column headers. The Assay file represents a set of assays defined by the endpoint measured (e.g. gene expression) and the technology employed (e.g. DNA microarray), as described in the Investigation file. An Assay file can refer to one or more external data files. Assay files with all columns left empty are syntactically valid ISA-Tab files; however, those elements that must be present are marked **core *Scientific Data***. Within these elements, some fields are mandatory and/or values should come from controlled vocabulary / ontology terms, as indicated. Table 3 shows the nodes together with the (Study or Assay) file in which the node should be used, as well as their possible attributes, the number of values allowed, data type and dependency on a parent node. Table 4 lists the attributes that can be used to qualify nodes, the number of values allowed, data type and dependency on a parent node.

As stated in section 1.1.4., this initial version of the *Scientific Data* configuration uses only the Generic Assay file layout—from the original ISA-Tab format specification—with a list of (generic) column headers to describe several types of assays. Additional specialized fields needed for conversion of ISA-Tab files into other formats, such as MAGE-Tab, SRA-XML and PRIDE-ML, are not included at this stage (for further details, refer to the original ISA-Tab format specification[6]).

The column headers are organized as nodes (containing the string *Name* or *File*), attributes for the nodes, attributes for node-processing events, qualifiers for nodes' attributes and other valid fields.

2.4.1. Assay file nodes

Core *Scientific Data* elements

Sample Name

(mandatory)

The user-defined unique identifier or name matching the value provided in the Study file, and thereby used to associate assay information with relevant study material. *Sample Names* in the Assay file can only be qualified with *Comment[ ]* fields.

Assay Name

(mandatory)

A user-defined unique identifier or name for each assay; this column can be preceded by a *Protocol REF* block of elements (refer to section 2.3.3). If the same sample was assayed multiple times (technical replicates), then each replicate should be assigned a separate row with a separate assay name.

Qualifying headers for *Assay Name* are *Performer*, *Date* and *Comment[ ]*.

Raw Data File

(mandatory)

A column to provide names of raw data files. The mandatory qualifying headers for a *Raw Data File* are *Comment[Data Repository]* and *Comment[Data Record Accession]*; others can be added as needed using additional *Comment[ ]* fields.

Additional *Scientific Data* elements

Data Transformation Name

This column contains a user-defined name for each transformation applied to the data.

Normalization Name

This column contains a user-defined name for each normalization applied to the data.

Derived Data File

This column provides names of files resulting from data transformation or processing, and can be preceded by a *Protocol REF* block of elements (refer to section 2.3.3); see also Figure 3.

The mandatory qualifying headers for Derived Data Files are *Comment[Data Repository]* and *Comment[Data Record Accession]*; others can be added as needed using additional *Comment[ ]* fields.

2.4.2. Attributes of Sample Name nodes in Assay files

Additional *Scientific Data* elements

Characteristics[<category term>]

This column contains terms describing each material used in an Assay according to the *Characteristics* category indicated in the column header. For example, a column header such as *Characteristics[purity]* would contain terms describing the purity of that portion of the material. If the term comes from a controlled vocabulary / ontology, the *Term Accession Number* and *Term Source REF* fields (see Qualifiers section) are required to track its provenance. If the characteristic being reported is a measurement, a *Unit* column (with qualifying *Term Accession Number* and *Term Source REF*) may also be used.

Please note that <space> characters may not appear immediately before an opening square bracket '[' in *Characteristics* column headers.

Material Type

This column may contain terms describing the *Material Type* (e.g. total RNA, protein extract). If the term comes from a controlled vocabulary / ontology, the *Term Accession Number* and *Term Source REF* fields (see Qualifiers section) are required to track its provenance.

2.4.3. Attributes of processing events for Assay file nodes

Core *Scientific Data* elements

Protocol REF

One or more *Protocol REF* columns should be used to specify the method used to transform a material. This column contains a reference to a *Protocol Name* (previously defined in the Investigation file) that should match the names in the Methods section of the DD manuscript.

Additional *Scientific Data* elements

Protocol REF can be further refined with the following elements, as needed, although these are very rarely used:

Parameter Value[<parameter name>]

This field allows reporting on the values assumed by the *Parameter* when a protocol is applied. Note that the term between the brackets ([ ]) must map to one (and only one) of the parameters defined in the Investigation file. Values can be qualitative or quantitative.

<Space> characters may not appear immediately before an opening square bracket '[' in *Parameter Value* column headers.

Performer

The name of the operator who carried out the protocol. This allows operator effects to be taken into account, and can serve as quality control in data tracking.

Date

The date on which a protocol was performed. This allows day-to-day variation in carrying out an experimental protocol to be taken into account, and can serve as quality control in data tracking. Dates should be reported in ISO format (YYYY-MM-DD).

2.4.4. Qualifiers for Assay file nodes' attributes

Additional *Scientific Data* elements

Unit

Unit is used if the terms provided in the *Characteristics[ ]*, *Parameter Value[ ]* or *Factor Value[ ]* columns classify data that are dimensional, i.e., quantitative values.

Term Accession Number

The accession number, if any, from the *Term Source* associated with the selected term (for example, from a controlled vocabulary or ontology). Qualifies the headers *Characteristics[ ]*, *Parameter Value[ ]* and *Unit*.

Term Source REF

Identifies the controlled vocabulary or ontology from which the selected term derives. The Source REF must match a *Term Source Name* previously declared in the ontology section.

2.4.5. Other Assay file fields

Core *Scientific Data* elements[21]

Comment[Data Repository]

(mandatory)

The name of the external data repository to which the primary data files have been submitted[22].

Comment[Data Record Accession]

(mandatory)

The identifier for the datasets provided by the external data repository.

Additional *Scientific Data* elements

Factor Value[<factor name>]

This field holds the actual values for the *Factor Value* named between the square brackets ([ ]; as declared in the Investigation file). Qualifiers for *Factor Value* are also those listed in section 2.4.4.

[21] <Space> characters may not appear immediately before an opening square bracket '[' in *Comment* or *Factor Value* column headers.

Comment
Further *Comment* columns can be added to provide additional information only when an appropriate alternative field does not exist.

Figure 1. ISA-Tab hierarchical structure

The hierarchical structure of the ISA-Tab format, which enables the representation of studies employing one or a combination of technologies.

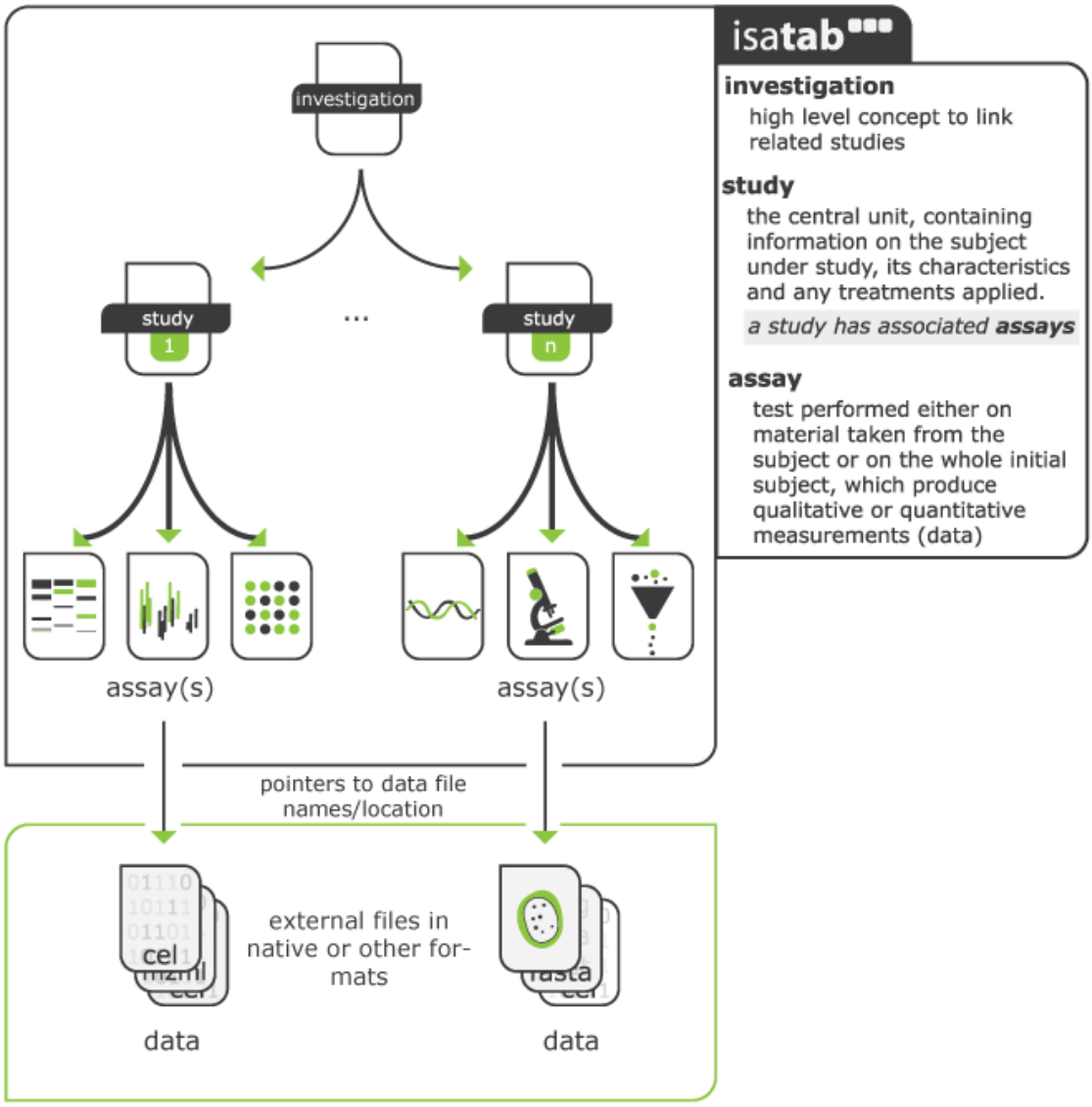

Figure 2. Transformation of material nodes in the Study file

Samples are the output of a protocol applied to a source material or to another sample. This figure illustrates the use of the *Protocol REF* block and its elements to describe the transformation of a material node entity into another material node entity. The side boxes show some of the fields allowed to qualify *Source Name* and *Sample Name* nodes (please refer to Table 4 for a full list of qualifiers). The central box shows how to use *Protocol REF* and its qualifiers. In this example, the source is transformed into a sample (e.g., tissue is collected from an organism according to a given method); a sample can also be transformed into another sample (e.g. primary cells are prepared from the collected tissue).

Protocols (e.g. tissue collection) must be declared in the *Protocol Name* field in the Study Protocol subsection of the Investigation file (see 2.3.3) and only referenced here in the *Protocol REF* field. The central box also shows how the *Parameter Value[<parameter name>]* qualifier allows the creation of customized column fields for protocol parameters (e.g. *Parameter Value[preservation buffer]*), and reporting of their values (e.g. saline buffer).

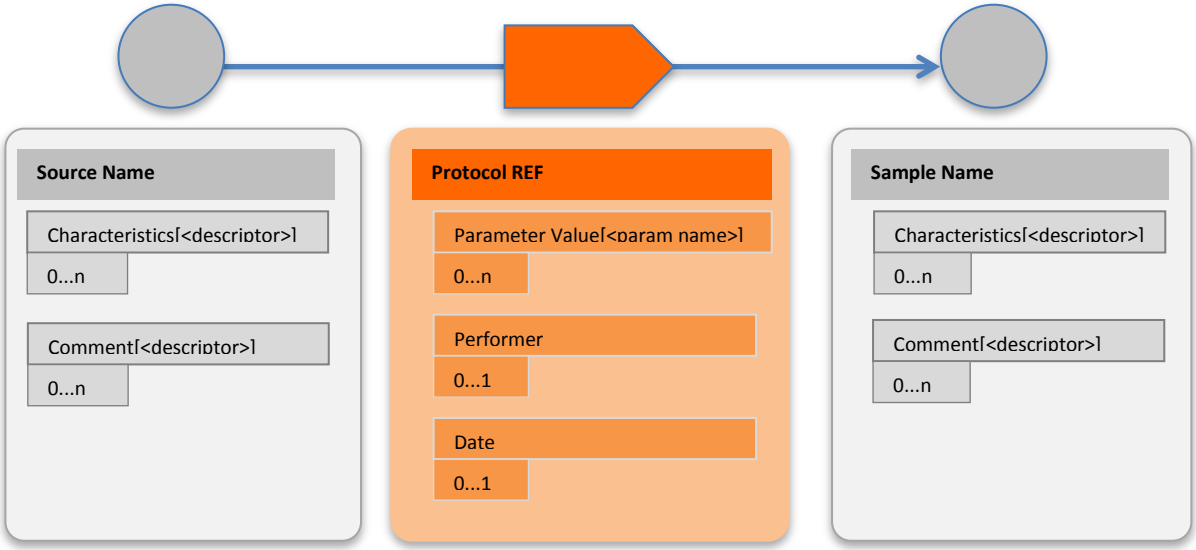

Figure 3. Transformation of data nodes in the Assay file

Derived data files are the output of a protocol applied to a raw data file. This figure illustrates the use of *Protocol REF* and its block elements to describe the transformation of a data node entity into another data node entity.

The side blocks show some of the fields allowed to qualify the *Raw Data File* and *Derived Data Files* nodes (refer to Table 4 for full list). The central block shows how to use *Protocol REF* and its qualifiers. *Raw Data File* and *Derived Data Files* must be preceded by the *Assay Name* and *Data Transformation Name* fields, respectively (see Table 3 and 4 for dependencies).

Protocols must be declared in the *Protocol Name* field (e.g. differential analysis) in the Study Protocol subsection of the Investigation file (see 2.3.3), and only referenced here in the *Protocol REF* field. The central box also shows how the *Parameter Value[<parameter name>]* qualifier allows the creation of customized column fields for protocol parameters (e.g. *Parameter Value[method]*), and reporting of their values (e.g. 2-way ANOVA).

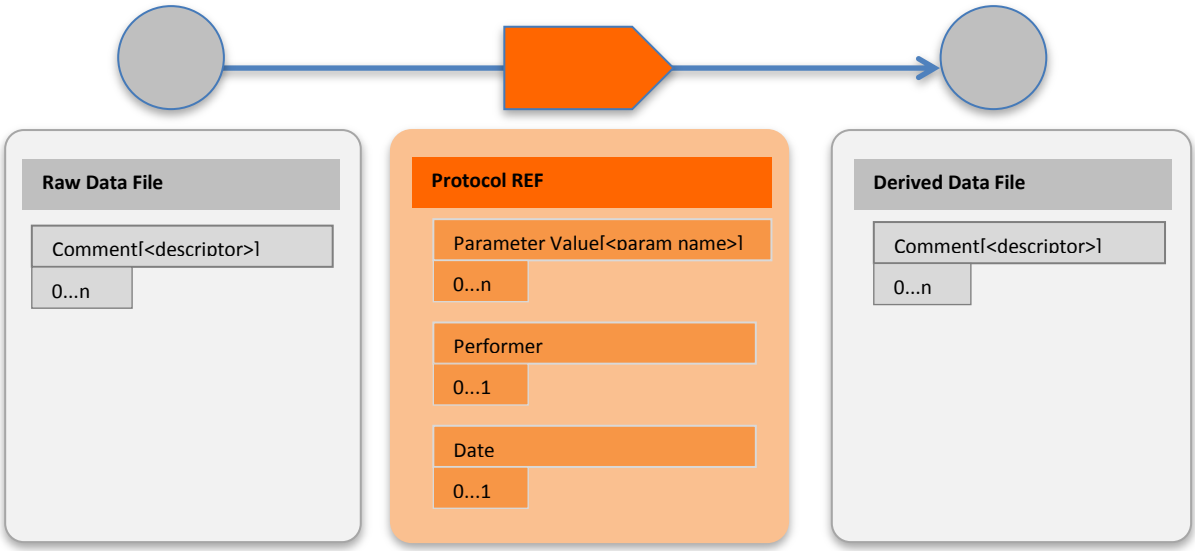

Table 1. Types of values in the Investigation file

This table describes the types of values allowed in each field in the Investigation file's sections, together with the specific requirements defined in the *Scientific Data* configuration, as outlined in section 1.4.

Investigation file fields	Type of value	Specific requirement and comments
ONTOLOGY SOURCE REFERENCE		
Term Source Name	text	
Term Source File	URI	
Term Source Version	text	
Term Source Description	text	
INVESTIGATION		
Investigation Identifier	n/a	
Investigation Title	text	
Investigation Description	text	
Investigation Submission Date	date (YYYY-MM-DD)	
Investigation Public Release Date	date (YYYY-MM-DD)	
INVESTIGATION PUBLICATIONS		
Investigation PubMed ID	accession number	
Investigation Publication DOI	accession number	
Investigation Publication Author List	text	
Investigation Publication Title	text	
Investigation Publication Status	text	select from the list of terms provided by *Scientific Data* (in preparation, submitted, published)
INVESTIGATION CONTACTS		
Investigation Person Last Name	text	
Investigation Person First Name	text	
Investigation Person Mid Initials	text	
Comment[Investigation Person ORCID]	accession number	refer to the ORCID website for how to register and obtain an ID
Investigation Person Email	text	
Investigation Person Phone	text	
Investigation Person Fax	text	
Investigation Person Address	text	
Investigation Person Affiliation	text	
Investigation Person Roles	text	
STUDY		
Study Identifier	n/a	leave blank, to be added by *Scientific Data*
Study File Name	URI/text	mandatory
Study Title	text	mandatory
Study Description	URI/text	leave blank; to be added by *Scientific Data*
Study Submission Date	date (YYYY-MM-DD)	leave blank, to be added by *Scientific Data*
Study Public Release Date	date (YYYY-MM-DD)	leave blank, to be added by *Scientific Data*
Comment[Manuscript Licence]	text	mandatory; select from the list of terms provided by *Scientific Data* (CC BY 3.0, CC BY-NC 3.0, CC BY-NC-SA 3.0)
Comment[Experimental Metadata Licence]	text	mandatory; default value: CC0
Comment[Data Repository]	text	mandatory
Comment[Data Record Accession]	accession number	mandatory
Comment[Data Record URI]	URI	mandatory
Comment[Supplementary Information File Name]	text	leave blank, to be added by *Scientific Data*
Comment[Supplementary Information File Type]	text	leave blank, to be added by *Scientific*

		Data
Comment[Supplementary Information File URI]	URI	leave blank, to be added by *Scientific Data*
Comment[Subject Keywords]	text	leave blank, to be added by *Scientific Data*
STUDY DESIGN DESCRIPTORS		
Study Design Type	text	controlled vocabulary / ontology
Study Design Type Term Accession Number	text	
Study Design Type Term Source REF	Term Source Name	
STUDY PUBLICATIONS		
Study PubMed ID	accession number	
Study Publication DOI	accession number	
Study Publication Author List	text	
Study Publication Title	text	
Study Publication Status	text	select from the list of terms provided by *Scientific Data* (in preparation, submitted, published)
STUDY FACTORS		
Study Factor Name	text	
Study Factor Type	text	controlled vocabulary / ontology
Study Factor Type Term Accession Number	text	
Study Factor Type Term Source REF	Term Source Name	
STUDY ASSAYS		
Study Assay Measurement Type	text	mandatory; controlled vocabulary / ontology
Study Assay Measurement Type Term Accession Number	text	
Study Assay Measurement Type Term Source REF	Term Source Name	
Study Assay Technology Type	text	mandatory; controlled vocabulary / ontology
Study Assay Technology Type Term Accession Number	text	
Study Assay Technology Type Term Source REF	Term Source Name	
Study Assay Technology Platform	text	
Study Assay File Name	URI/text	mandatory
STUDY PROTOCOLS		
Study Protocol Name	text	mandatory; should match a Methods subheader in the DD manuscript
Study Protocol Type	text	controlled vocabulary / ontology
Study Protocol Type Term Accession Number	text	
Study Protocol Type Term Source REF	Term Source Name	
Study Protocol Description	text	leave blank
Study Protocol URI	URI	leave blank, to be added by *Scientific Data*
Study Protocol Version	version number	
Study Protocol Parameters Name	text	controlled vocabulary / ontology
Study Protocol Parameters Name Term Accession Number	text	
Study Protocol Parameters Name Term Source REF	Term Source Name	
Study Protocol Components Name	text	
Study Protocol Components Type	text	controlled vocabulary / ontology
Study Protocol Components Type Term Accession Number	text	
Study Protocol Components Type Term Source REF	Term Source Name	
STUDY CONTACTS		
Study Person Last Name	text	
Study Person First Name	text	
Study Person Mid Initials	text	
Comment[Study Person ORCID]	accession number	refer to the ORCID website for how to register and obtain an ID
Study Person Email	text	
Study Person Phone	text or numeric	
Study Person Fax	text	

Study Person Address	text	
Study Person Affiliation	text	
Study Person Roles	text	
Comment[Funder]	text	
Comment[FundRef ID]	accession number	
Comment[Grant Identifier]	accession number	

Table 2. Multiplicity of values in the Investigation file

This table provides information on the number of items allowed in each field (also known as the multiplicity of each item) in the Investigation file's sections.

	Number of value(s) per column	To report multiple values
ONTOLOGY SOURCE REFERENCE		
Term Source Name	1	As many columns as necessary
Term Source File	1	As many columns as necessary
Term Source Version	1	As many columns as necessary
Term Source Description	1	As many columns as necessary
INVESTIGATION		
Investigation Identifier	1	n/a
Investigation Title	1	n/a
Investigation Description	1	n/a
Investigation Submission Date	1	n/a
Investigation Public Release Date	1	n/a
INVESTIGATION PUBLICATIONS		
Investigation PubMed ID	1	As many columns as necessary
Investigation Publication DOI	1	As many columns as necessary
Investigation Publication Author List	1	As many columns as necessary
Investigation Publication Title	1	As many columns as necessary
Investigation Publication Status	1	As many columns as necessary
INVESTIGATION CONTACTS		
Investigation Person Last Name	1	As many columns as necessary
Investigation Person First Name	1	As many columns as necessary
Investigation Person Mid Initials	1...n semi-colon (;) separated	As many columns as necessary
Comment[Investigation Person ORCID]	1	As many columns as necessary
Investigation Person Email	1	As many columns as necessary
Investigation Person Phone	1	As many columns as necessary
Investigation Person Fax	1	As many columns as necessary
Investigation Person Address	1	As many columns as necessary
Investigation Person Affiliation	1	As many columns as necessary
Investigation Person Roles	1...n semi-colon (;) separated	As many columns as necessary

STUDY		
Study Identifier	1	Only 1 within a Study block
Study File Name	1	Only 1 within a Study block
Study Title	1	Only 1 within a Study block
Study Description	1	Only 1 within a Study block
Study Submission Date	1	Only 1 within a Study block
Study Public Release Date	1	Only 1 within a Study block
Comment[Manuscript Licence]	1	Only 1 within a Study block
Comment[Experimental Metadata Licence]	1	Only 1 within a Study block
Comment[Data Repository]	1...n semi-colon (;) separated	Only 1 within a Study block
Comment[Data Record Accession]	1...n semi-colon (;) separated	Only 1 within a Study block
Comment[Data Record URI]	1...n semi-colon (;) separated	Only 1 within a Study block
Comment[Supplementary Information File Name]	1...n semi-colon (;) separated	Only 1 within a Study block
Comment[Supplementary Information File Type]	1...n semi-colon (;) separated	Only 1 within a Study block
Comment[Supplementary Information File URI]	1...n semi-colon (;) separated	Only 1 within a Study block
Comment[Subject Keywords]	1...n semi-colon (;) separated	Only 1 within a Study block
STUDY DESIGN DESCRIPTORS		
Study Design Type	1	As many columns as necessary
Study Design Type Term Accession Number	1	As many columns as necessary
Study Design Type Term Source REF	1	As many columns as necessary
STUDY PUBLICATIONS		
Study PubMed ID	1	As many columns as necessary
Study Publication DOI	1	As many columns as necessary
Study Publication Author List	1	As many columns as necessary
Study Publication Title	1	As many columns as necessary
Study Publication Status	1	As many columns as necessary
STUDY FACTORS		
Study Factor Name	1	As many columns as necessary
Study Factor Type	1	As many columns as necessary
Study Factor Type Term Accession Number	1	As many columns as necessary
Study Factor Type Term Source REF	1	As many columns as necessary
STUDY ASSAYS		
Study Assay Measurement Type	1	As many columns as necessary
Study Assay Measurement Type Term Accession Number	1	As many columns as necessary
Study Assay Measurement Type Term Source REF	1	As many columns as necessary
Study Assay Technology Type	1	As many columns as necessary
Study Assay Technology Type Term Accession	1	As many columns as

Number		necessary
Study Assay Technology Type Term Source REF	1	As many columns as necessary
Study Assay Technology Platform	1	As many columns as necessary
Study Assay File Name	1	As many columns as necessary
STUDY PROTOCOLS		
Study Protocol Name	1	As many columns as necessary
Study Protocol Type	1	As many columns as necessary
Study Protocol Type Term Accession Number	1	As many columns as necessary
Study Protocol Type Term Source REF	1	As many columns as necessary
Study Protocol Description	1	As many columns as necessary
Study Protocol URI	1	As many columns as necessary
Study Protocol Version	1	As many columns as necessary
Study Protocol Parameters Name	1...n semi-colon (;) separated	As many columns as necessary
Study Protocol Parameters Name Term Accession Number	1...n semi-colon (;) separated	As many columns as necessary
Study Protocol Parameters Name Term Source REF	1...n semi-colon (;) separated	As many columns as necessary
Study Protocol Components Name	1...n semi-colon (;) separated	As many columns as necessary
Study Protocol Components Type	1...n semi-colon (;) separated	As many columns as necessary
Study Protocol Components Type Term Accession Number	1...n semi-colon (;) separated	As many columns as necessary
Study Protocol Components Type Term Source REF	1...n semi-colon (;) separated	As many columns as necessary
STUDY CONTACTS		
Study Person Last Name	1	As many columns as necessary
Study Person First Name	1	As many columns as necessary
Study Person Mid Initials	1	As many columns as necessary
Comment[Study Person ORCID]	1	As many columns as necessary
Study Person Email	1	As many columns as necessary
Study Person Phone	1	As many columns as necessary
Study Person Fax	1	As many columns as necessary
Study Person Address	1	As many columns as necessary
Study Person Affiliation	1	As many columns as necessary
Study Person Roles	1...n semi-colon (;) separated	As many columns as necessary
Comment[Funder]	1...n semi-colon (;) separated	As many columns as necessary
Comment[FundRef ID]	1...n semi-colon (;) separated	As many columns as necessary
Comment[Grant Identifier]	1...n semi-colon (;) separated	As many columns as necessary

Table 3. Nodes in the Study and Assay files

This table shows nodes in the Study and Assay files and lists possible downstream nodes and allowed qualifiers, the number of allowed values and the nature of any parent node dependencies.

Node name	Downstream nodes and attributes	ISA-Tab file	Accept multiple values per field	Parent node dependency
Source Name	Characteristics, Material Type, Description, Comment	Study	NO	
	NO	Assay		
Sample Name	Characteristics, Material Type, Protocol REF, Description, Comment	Study	NO	Source Name
	Characteristics, Material Type, Protocol REF, Description, Comment	Assay		
Assay Name	Protocol REF, Raw Data File, Derived Data File	Assay	NO	
Normalization Name	Protocol REF, Derived Data File, Comment	Assay	NO	Assay Name
Data Transformation Name	Protocol REF, Derived Data File, Comment	Assay	NO	Assay Name
Raw Data File	Comment	Assay	NO	Assay Name
Derived Data File	Comment	Assay	NO	Assay Name, Data Transformation Name

Table 4. Node attributes in the Study and Assay files

List of attributes that can be used to qualify nodes and provide annotation to entities. The table also shows the file in which the node should be used, number of allowed values, data type and dependency on a parent node.

Attribute name	Attribute qualifiers	ISA-Tab file	Accept multiple values	Data type	Parent node dependency
Material Type	Term Accession Number, Term Source REF	Study, Assay	NO	string	Source Name, Sample Name
Characteristics[ ]	Unit, Term Accession Number, Term Source REF	Study, Assay	NO	string or number	Source Name, Sample Name
Factor Value[ ]	Unit, Term Accession Number, Term Source REF	Study, Assay	NO	string or number	
Parameter Value[ ]	Unit, Term Accession Number, Term Source REF	Study, Assay	NO	string or number	Protocol REF
Unit	Term Accession Number, Term Source REF	Study, Assay	NO	string	Characteristics[ ], Factor Value[ ], Parameter Value[ ]
Term Accession Number		Study, Assay	NO	string	Characteristics[ ], Material Type[ ], Factor Value[ ], Parameter Value[ ], Unit[ ]
Term Source REF		Study, Assay	NO	string	Characteristics[ ], Material Type[ ], Factor Value[ ], Parameter Value[ ], Unit[ ]
Protocol REF	Parameter Value[ ], Performer, Date, Comment	Study, Assay	NO	string	Source Name, Sample Name, Assay Name
Raw Data File	Comment	Assay	NO	string	Assay Name
Derived Data File	Comment	Assay	NO	string	Assay Name, Data Transformation Name
Performer	Comment	Study, Assay	NO	string	Protocol REF
Date		Study, Assay	NO	date	Protocol REF
Description		Study, Assay	NO	string	Source Name, Sample Name
Comment[ ]		Study, Assay	NO	string	ALL nodes