Specification documentation: ISA-TAB 1.0

Status

ISA-TAB v1.0 specification document, version 13th January 2009.

List of changes: fixed typos, grammatical fixes, removed Derived Spectral Data File (in section 4.3.4 and 4.3.5) because of Derived Data File (as in section 4.3.1) usage, added recommendation for no spaces in file names.

Authors

Initially drafted by Philippe Rocca-Serra[1], Susanna-Assunta Sansone[1] and Marco Brandizi[1] (1) this document and the work described herein incorporates input from David Hancock[2], Stephen Harris[3], Allyson Lister[4], Michael Miller[5], Kieran O'Neill[6], Chris Taylor[1], Weida Tong[3] and other collaborators, listed at the ISA-TAB project page (2) under 'Contributors'.

Documents and contacts

The current document and **several example files** are available from the ISA-TAB project page (2). This documentation is primarily aimed at software engineers, to facilitate the development of automated export from databases, or import into analytical or other tools. A **mailing list** (isaforum@googlegroups.com) has been set up to facilitate discussion. Comments and suggestions on this document should be addressed to Philippe Rocca-Serra (proccaserra@gmail.com) and the mailing list.

Abstract

This document describes ISA-TAB, a general purpose framework with which to capture and communicate the complex metadata required to interpret experiments employing combinations of technologies, and the associated data files. Sections 1 to 3 introduce the ISA-TAB proposal, describe the rationale behind its development, provide an overview of its structure and relate it to other formats. Section 4 describes the specification in detail; section 5 provides examples of design patterns.

ISA-TAB builds on the existing paradigm that is MAGE-TAB - a tab-delimited format to exchange microarray data. ISA-TAB necessarily maintains backward compatibility with existing MAGE-TAB files to facilitate adoption; conserving the simplicity of MAGE-TAB for simple experimental designs, while incorporating new features to capture the full complexity of experiments employing a combination of technologies. Like MAGE-TAB before it, **ISA-TAB is simply a format**; the decision on how to regulate its use (*i.e.* enforcing completion of mandatory fields or use of a controlled terminology) is a matter for those communities, which will implement the format in their systems and for which submission and exchange of minimal information is critical. In this case, an additional layer or of constraints should be agreed and required on top of the ISA-TAB specification.

Prerequisites

Familiarity with the syntax and grammar defined in the MAGE-TAB specifications (3) is required, as the authors assume knowledge of the MAGE-TAB format when describing the structure of ISA-TAB in this document.

---

[1] EMBL-EBI The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

[2] NERC Bioinformatics Center (NEBC), Centre for Ecology and Hydrology, Oxford, UK, and the University of Manchester, School of Computer Science, Manchester, UK.

[3] FDA's National Center for Toxicological Research (NCTR), Center for Toxicoinformatics, Jefferson, Arkansas.

[4] CISBAN & School of Computing Science, Newcastle University, Newcastle upon Tyne, UK.

[5] Rosetta Biosoftware, Seattle, Washington.

[6] BC Cancer Research Centre (BCCRC), Vancouver, Canada.

1. ISA-TAB introduction

1.1 Rationale

The Investigation / Study / Assay (ISA) tab-delimited (TAB) format is a general purpose framework with which to collect and communicate complex metadata (*i.e.* sample characteristics, technologies used, type of measurements made) from experiments employing a combination of technologies. In particular, ISA-TAB has been developed for — but not limited to — experiments using genomics, transcriptomics, proteomics or metabol/nomics techniques (hereafter referred as 'omics-based' experiments). For example, consider an investigation into the effect of a compound that induces liver damage, which looks at changes both in (*i*) the metabolite profile of urine and (*ii*) gene expression in the liver (by mass spectrometry and microarray technologies, respectively). The general motivation for this work is the fulfillment of the needs of two groups:

- The BioInvestigation Index project (4) to create a common structured representation of the metadata — required to interpret an experiment — with which to facilitate combined submission to ArrayExpress (5), Pride (6), and in the near future, a metabolomics repository;
- Collaborative systems (2, 7, 8) committed to pipelining omics-based experiments into EBI public repositories, or willing to exchange data among themselves, or planning to enable their users to import data from public repositories into their local systems.

ISA-TAB has been developed in synergy with other existing efforts, complementing and extending where necessary, as follows:

- it builds on the existing MicroArray Gene Expression (MAGE)-TAB paradigm (9);
- it includes metadata as required by the Pride repository, supporting the Proteomics Standards Initiative (PSI, 10);
- it places its concepts in line with the Functional Genomics Experiment (FuGE) objects (11, 12);
- it complements the Study Data Tabulation Model (SDTM) (14, 15);
- it introduces a generic structure to accommodate experiment employing a combination of technologies.

As stated, ISA-TAB could be viewed as an extended version of MAGE-TAB, a format that supports the management, exchange and submission of microarray-based experiment data and metadata. This format was designed for use by laboratories with little or no bioinformatics support, which cannot therefore deal with the complexity of the MAGE Markup Language (ML) format (13). ISA-TAB builds on the MAGE-TAB paradigm, and shares its motivation for the use of tab-delimited text files; *i.e.*, that they can easily be created, viewed and edited by researchers using spreadsheet software such as Microsoft Excel. ISA-TAB also employs MAGE-TAB syntax as far as possible to ensure backward compatibility with existing MAGE-TAB files (as described further in section 3.1). ISA-TAB contains all the fields in MAGE-TAB; however, it has a number of additional features that make it a more general framework, able to accommodate experiments employing a combination of technologies. Concepts in ISA-TAB have also been aligned with some of the objects in the FuGE model, for several reasons: First, because FuGE is a formal published model intended for functional genomics. Second, because the FuGE model is integral to the development of MAGE-ML *v2*, which will be related to the MAGE-TAB format. Finally, because ISA-TAB is an effective way to render to FuGE -based documents in user-friendly readable format (as described further in section 3.2). The relation to SDTM biomedical tabular format in described in section 3.3.

The ISA-TAB format is described in detail in section 4.

1.2 Definitions

Investigation, Study and Assay are the three key entities (8) around which the ISA-TAB framework is built. They assist in structuring and classifying information relevant to the subject under study and the different technologies employed. Note that 'subject' as used above could refer *inter alia* to an organism, or tissue, or an

environmental sample. Study is the central unit, containing information on the subject under study, its characteristics and any treatments applied.

A Study has associated Assays; these are measurements performed either on the whole initial subject or on sample taken from the subject, which produce qualitative or quantitative data. Assays can be characterized as the smallest complete unit of experimentation producing data associated to a subject; *i.e.* one hybridization is treated as one assay; each technical replicate represents an additional assay; one LC-MS run equals one assay; a multiplexed microarray with n a layouts of the same design corresponds to *n* hybridizations; and a MALDI MS chip with *n* spots could perform up to *n* assays (*i.e.* all spots analyzed). Investigation is a higher-order object, whose primary role is to group related Studies.

It should be noted that the word 'experiment' has been deliberately avoided. A comparison of ArrayExpress and Pride revealed that 'experiment' is used to refer to objects at different levels of granularity in each; *i.e.* to refer to a *set* of related hybridizations in ArrayExpress, but only a *single* gel-based separation run in Pride. Following the abstractions proposed here, an experiment in ArrayExpress would be equivalent to a Study.

The choice of Study as the central unit of the ISA-TAB proposal is supported by its use in existing biomedical formats, such as the SDTM, which encompasses both the Standard for Exchange of Nonclinical Data (SEND, 14) and the Clinical Data Interchange Standards Consortium (CDISC, 15). SDTM has been endorsed by the US Food and Drug Administration (FDA) as the preferred way to organize, structure and format both clinical and non-clinical (toxicological) data submissions (16, 17). See also sections 3.3 and 5.3.

1.3 ISA-TAB development process

Created to fulfill the need of the BioInvestigation Index project (4), the first straw-man proposal, ISA-TAB v0.1, was presented and discussed in a workshop held at the European Bioinformatics Institute (EBI), Cambridge, UK, on 6-8 December, 2007, funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC BB/E025080/1) (1). The goal of this first workshop was the creation of an 'exchange network test bed' for a heterogeneous set of resources (*e.g.* public and proprietary repositories, public and commercial software tools), variously run by academic, industrial and governmental groups. The participants represented a range of groups interested in leveraging common standards to report omics-based experiments, bringing with them experience of building standards gained through their involvement in, or leadership of relevant efforts. The workshop report (8) is available from the ISA-TAB website (2). The workshop produced a general consensus on the important role for ISA-TAB in addressing the immediate need for a communications framework for multi-omics experiments. A series of technical challenges were also identified, the majority of which have been resolved in the ISA-TAB v0.3 release.

A second workshop was held at EBI on 16-18 June, 2008. On this occasion, the ISA-TAB v0.3 proposal was revised in the light of a series of real-life case examples - produced by the participating communities. These modifications were agreed on, resulting in the current ISA-TAB v1.0 version. Participants also discussed examples of tools that need to be developed to assist the users in the creation of ISA-TAB files and finalized the work plan to release the XSL templates described in section 3.2. A short summary of this second workshop is available as a PowerPoint presentation from the ISA-TAB website (2).

It is foreseen that, following the publication this 'Release candidate 1, ISA-TAB v1.0 specification' document on the ISA-TAB website, other workshops will be held to bring together those communities, which will implement the format in their systems. Based on their feedback the initial design patterns documentation (see section 5) will be expanded and eventually form an implementation guideline document.

2. ISA-TAB v1.0 structure overview

ISA-TAB uses three types of file to capture the experimental metadata:

- Investigation file
- Study file
- Assay file (with associated data files)

The Investigation file contains all the information needed to understand the overall goals and means used in an experiment; experimental steps (or sequences of events) are described in the Study and in the Assay file(s). For each Investigation file there may be one or more Study files; 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 and more fully in section 4.

2.1 Investigation file

In the investigation file, information is reported on a per-column basis and the fields are organized and divided in sections. The Investigation file is intended to meet three needs: (*i*) to define key entities, such as factors, protocols, parameters, which may be referenced in the other files; (*ii*) to relate Assay files to Study files; and optionally, (*iii*) to relate each Study file to an Investigation (this only becomes necessary when two or more Study files need to be grouped). The declarative sections cover general information such as contacts, protocols and equipment, and also – where applicable – the description of terminologies (controlled vocabularies or ontologies) and other annotation resources that were used.

The Investigation file is intended to meet four needs: (i) to define key entities, such as factors, protocols, which may be referenced in the other files; (ii) to track provenance of the terminologies (controlled vocabularies or ontologies) there are used, where applicable; (iii) to relate Assay files to Study files; and optionally, (iv) to relate each Study file to an Investigation (this only becomes necessary when two or more Study files need to be grouped).

The optional Investigation *section* of an Investigation file is a flexible solution to group two or more Study files, as required by several use cases. In the toxicogenomics domain, for example, acute toxicity studies are followed by long term toxicity studies and *in vitro* toxicity studies. For clarity, the users would link these Study files by filling the Investigation section. Another example comes from the environmental genomics domain, where several studies carried out in the same area can be usefully related under the same Investigation. See also section 4.1.

2.2 Study file

In this file, information is structured on a per-row basis with the first row being used for column headers. The Study file contains contextualizing information for one or more assays, for example; the subjects studied; their source(s); the sampling methodology; their characteristics; and any treatments or manipulations performed to prepare the specimens. See also section 4.2.

2.3 Assay file

In this file, as for the study files, fields are again organized on a per-row basis with the first row being used for column headers. The Assay file represents a portion of the experimental graph (*i.e.*, one part of the overall structure of the workflow); 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). See also section 4.3.

2.3.1 Relating Study and Assay files

In a study looking at the effect of a compound inducing liver damage in rats by characterizing the metabolic profile of urine (by NMR spectroscopy) and measuring protein and gene expression in the liver (by 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 and the steps undertaken to take urine and liver (samples) from the treated rats.
- The Assay file for the urine metabolic profile (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 resultant raw and derived data files.
- The Assay file for the gene expression profile (measurement) by DNA microarray (technology) will contain the (stepwise) description of how the RNA extract was prepared from the liver (or a section), how the extract was labeled, how the hybridization was performed and so on, and will also contain the links to the resultant raw and derived data files.
- The Assay file for the protein expression profile (measurement) by mass spectrometry (technology), will contain the (stepwise) description of how the protein extract was prepared from the liver (or a section), how the extract was labeled, how the hybridization was performed and so on, and will also contain the links to the resultant raw and derived data files.

2.4 Data files and the ISArchive

ISA-TAB focuses on structuring experimental metadata; raw and derived data files are considered as external files. The Assay file can refer to one or more of these external data files, see section 4.3. For guidelines on how to format these data files, users should refer to section 5.3 and the relevant standards group or reference repository (*i.e.* 5, 6, 10, 11, 12, 18, 19, 20). In addition to raw and derived data files, the Assay file for gene expression (measurement) by microarray (technology) will also refer to a Derived Array Data Matrix and an Array Description File (ADF), both described in the MAGE-TAB specifications (see section 4.3.2).

For submission or transfer, ISA-TAB files and associated data files can be packaged into an **ISArchive** as shown in Figure 1.

3. Relating ISA-TAB to other formats and requirements

3.1 Commonalities and difference with MAGE-TAB v1.1

ISA-TAB v1.0 contains all the fields in MAGE-TAB, but it introduces a generic structure to accommodate experiments employing a combination of technologies. Nonetheless it maintains backward compatibility with existing MAGE-TAB files, to facilitate conversion between the formats prior to the adoption of one common format (ISA-TAB). While no changes have been made to the ADF and the Derived Array Data Matrix files, the changes affect the MAGE-TAB Investigation Description Format (IDF) and Sample and Data Relationship Format (SDRF) as follows:

- While MAGE-TAB has been designed for microarrays, the ISA-TAB Investigation file allows the definition of any type of assay by specifying the measurement type and the technology. In turns, this determines the specific fields to be used in the Assay file(s).
- The content of the MAGE-TAB SDRF has been divided between the ISA-TAB Study and Assay files with their relationship being declared in the ISA-TAB Investigation file. The content of the Study file corresponds to the SDRF fields describing the subjects studied, source(s), characteristics and treatments, while the content of the Assay file contains the information related to the input material, technology employed and links to the resultant data files.
- The MAGE-TAB IDF corresponds largely to the ISA-TAB Investigation file. The latter reuses many fields from the IDF but in a slightly different order and divided into sections that can be repeated accordingly; the intent being to enhance readability and presentation, and to aid parsing.
- In the MAGE-TAB IDF, Investigation is used as a synonym for 'experiment' (*sensu* MIAME, 21) to link related hybridizations. In the ISA-TAB Investigation file, the homonymous section is used *sensu* FuGE, to relate (list and order) Study files, which also serves to facilitate alignment with FuGE's Investigation object.

3.2 ISA-TAB rendering of FuGE and other XML documents via XSL transformations

The ISA-TAB format should be seen as a complement to XML-based formats, whether existing or under development, such as FuGE-ML and its extensions. ISA-TAB addresses the immediate need for a general purpose framework with which to collect and communicate experiments employing a combination of technologies, while most existing FuGE-based modules are still under development, with the exception of Gel-ML (22). Even as extensions become available, ISA-TAB can continue serving users lacking bioinformatics support. It can also serve as a user-friendly presentation layer for XML-based formats *via* an eXtensible Stylesheet Language (XSL) transformation: XML formats are meant to be read by software; a tabular representation can considerably improve human readability. Initial work has been carried out to evaluate the feasibility of rendering FuGE-ML files (and FuGE-based extensions) in ISA-TAB. Examples are available from the ISA-TAB website under the 'Document' section (2), along with a report detailing the issues faced during these transformations. When finalized, the XSL templates will also be released, along with XPath expressions and a table 'mapping' FuGE objects to ISA-TAB labels in a separate document.

Similarly, other XML-based formats in the bio-domain can be rendered as ISA-TAB; examples for cell assays, flow cytometry and high-throughput sequencing are available from the ISA-TAB web page (2).

3.3 Relation to SDTM biomedical tabular format

Where experiments include clinical or non-clinical studies, ISA-TAB can complement existing biomedical formats such as the SDTM (14, 15), by formally capturing information about the interrelationship of the various parts they describe. A reference system has been created to allow SDTM source and sample(s) to be referenced from a Study file, as described in details in section 5.3.

3.4 Minimal content and terminology

ISA-TAB is no more than a format with which to communicate information. Both 'minimum information' requirements and the use of controlled vocabularies/ontologies are beyond the scope of this proposal, and are

the focus of related efforts. ISA-TAB has no mandatory fields and values can be either in free text or from controlled vocabularies or ontologies.

'Minimal information' checklists are under development both by individual communities for their particular domains of interest and collaboratively through the Minimal Information for Biological and Biomedical Investigations (MIBBI, 23, 24) project. Efforts proceeding under the Open Biomedical Ontology (OBO, 25, 26) Foundry umbrella will provide general terms to describe studies and assays, and specific terms relevant to particular domains (*i.e.* chemical, biological, clinical), *via* the Ontology for Biomedical Investigations (OBI, 27). However, the decision on how to regulate the use of the ISA-TAB (marking certain fields mandatory, or enforcing the use of controlled terminology) is a matter for those communities that implement the format in their system; for example, where the submission and exchange of minimal information is critical. In such cases, an additional layer of constraints should be agreed and combined with the ISA-TAB specification, most usually in an appropriately constructed data entry/validation tool, or in the form of written guidance.

4. ISA-TABv1.0 detailed structure

This section describes the structure of the Investigation, Study and Assay files in detail. Where relevant, the design decisions which influence the formatting rules are described below; these are in alignment with MAGE-TAB (3) to ensure that files are formatted correctly.

- **Field separator**
  The column delimiter is the Unicode Horizontal Tab character (Unicode code point 0009) is the official separator accepted in ISA-TAB.
- **File encoding**
  The UTF-8 encoding of the Unicode character set is the preferred encoding for ISA-TAB files. However, parsers should be able to recognize which Unicode encoding is being used and to behave accordingly.
- **File naming conventions**
  In order to facilitate identification of ISA-TAB components in an ISArchive, specific file naming patterns have been created as it 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-given name. For maximal portability, filenames should only contain ASCII characters without spaces, as many software utilities do not accept spaces in file paths and names.
- **Case sensitivity**
  All labels are case-sensitive:
  - In the Investigation file, section headers are 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), see section 4.1.
  - 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), see section 4.2 and section 4.3.

  This will facilitate visualization of headers and fields when viewing the file in spreadsheet software.
- **Dates**
  Dates should be supplied in the ISO 8601 format "YYYY-MM-DD".
- **Object identifiers**
  All values found in fields, whose headers containing the string 'Name' (*e.g. Source Name*) or 'File' (*e.g. Raw Data **File***) are considered object identifiers. In the Study and Assay file such object identifiers represent 'nodes' in the experimental graph (throughout this document 'node' denotes either [biological] material, such as a sample or an RNA extract, or a data object, while 'edges' show the relationships between nodes). All such object identifiers need to be locally unique within an ISA-TAB formatted file, and could also be fully qualified external accession numbers. Finally, object identifiers can follow structuring guidelines, for example, as documented in the MAGE-TAB specification (3, section 3.1.3) an LSID-like structure maybe used: `<authority>:[<namespace>]:<object>[:<revision>]`
- **Referencing**
  Objects defined in the Investigation file can be referenced in either the Study or Assay files. Two mechanisms are available to indicate a reference:
  - A 'REF' label component; for example, *Protocol REF* or *Term Source REF*.
  - Square brackets + object name is the alternate formalism for referencing *Parameters* and *Factors* declared in the Investigation File; *as in Parameter* [oven temperature] or *Factor Value* [compound].
- **Free text descriptions**

Since text is stored in a single tab-delimited field, any embedded tab or new line characters must be protected by enclosing the whole text within double quotes aka quotation mark (") Unicode U+0022. Also, any double quote contained by the text should be protected using the same mechanism.

- **Multiple value per field**
  In several sections of ISA-TAB, multiple values may be supplied in one given fields by relying a semi-colon (;) Unicode (U0003+B) as value separators.
- **Notes**
  In any of the Investigation, Study or Assay files, rows in which the first character in the first column is Unicode U+0023 (the # character) will be interpreted as comments. Such entries maybe used to provide notes; for example, to communicate a comment to curators but, ISA-TAB parsers should ignore those lines entirely. However, if found in any other position, the # character (U+0023) need not be escaped. So Sample #2 found in cell should not be ignored by parsers.

4.1 Investigation file

In this file the fields are organized on a per-column basis and divided in sections, described in detail below. Table 1 describes the types of values allowed in each field together with some comments implementers may find useful; Table 2 provides information on the number of items allowed in each field. However, as stated in section 3.4 Investigation files with all fields left empty *are* syntactically valid, as are those where all fields are filled with free text values rather than controlled vocabulary or ontology terms.

4.1.1 Ontology source section

This annotation section is identical to that in the MAGE-TAB format.

ONTOLOGY SOURCE REFERENCE

Term Source Name
The name of the source of a term; *i.e.* the source controlled vocabulary or ontology. These names will be used in all corresponding *Term Source REF* fields. For examples, the abbreviation OBI can be used to identify the Ontology for Biomedical Investigations (27) source.

Term Source File
A file name or a URI of an official resource.

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

Term Source Description
Use for disambiguating resources when homologous prefixes have been used.

4.1.2 Investigation section

This section is organized in several subsections, described in detail below. The Investigation section provides a flexible mechanism for grouping two or more Study files where required. When only one Study is created, this section may be left empty.

4.1.2.1 INVESTIGATION

Investigation Identifier
A locally unique identifier or an accession number provided by a repository.

Investigation Title
A concise name given to the investigation

Investigation Description
A textual description of the investigation

Investigation Submission Date
The date on which the investigation was reported to the repository.

Investigation Public Release Date
The date on which the investigation should be released publicly.

4.1.2.2 INVESTIGATION PUBLICATIONS

Each publication associated with an Investigation has its own column in the Investigation Publication section. Such publications are specifically dealing with the investigation as a whole. Publications relating to the specific Studies may 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 described publication(s) associated with this investigation.
Investigation Publication DOI
A Digital Object Identifier (DOI) for that publication (where available).
Investigation Publication Author List
The list of authors associated with that publication.
Investigation Publication Title
The title of publication associated with the investigation.
Investigation Publication Status
A term describing the status of that publication (*i.e.* submitted, in preparation, published).
Investigation Publication Status Term Accession Number
The accession number from the Term Source associated with the selected term.
Investigation Publication Status Term Source REF
Identifies the controlled vocabulary or ontology that this term comes from. The Source REF has to match one the *Term Source Name* declared in the in the ontology section 4.1.1.

4.1.2.3 INVESTIGATION CONTACTS

Investigation Person Last Name
The last name of a person associated with the investigation.
Investigation Person First Name
The first name of a person associated with the investigation.
Investigation Person Mid Initials
The middle initials of a person associated with the investigation.
Investigation Person Email
The email address of a person associated with the investigation.
Investigation Person Phone
The telephone number of a person associated with the investigation.
Investigation Person Fax
The fax number of a person associated with the investigation.
Investigation Person Address
The address of a person associated with the investigation.
Investigation Person Affiliation
The organization affiliation for a person associated with the investigation.
Investigation Person Roles
Term to classify the role(s) performed by this person in the context of the investigation, which means that the roles reported here need not correspond to roles held within their affiliated organization. Multiple annotations or values attached to one person can be provided by using a semicolon (";") Unicode (U0003+B) as a separator (e.g.: submitter;funder;sponsor) .The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.
Investigation Person Roles Term Accession Number
The accession number from the *Term Source* associated with the selected term.
Investigation Person Roles Term Source REF
Identifies the controlled vocabulary or ontology that this term comes from. The Source REF has to match one of the *Term Source Names* declared in the ontology section 4.1.1.

4.1.3 Study section

This section is organized in several subsections, described in detail below. This section also represents a **repeatable block**, which is replicated according to the number of Studies to report (*i.e.* two Studies, two Study blocks are represented in the Investigation file). The subsections in the block are arranged vertically; the intent being to enhance readability and presentation, and possibly to help with parsing. These subsections must remain within this repeatable block, although their order may vary; the fields must remain within their subsection.

4.1.3.1 STUDY

Study Identifier
A unique identifier: either a temporary identifier supplied by users or one generated by a repository or other database. For example, it could be an identifier complying with the LSID specification.
Study Title
A concise phrase used to encapsulate the purpose and goal of the study.
Study Description
A textual description of the study, with components such as objective or goals.
Study Submission Date
The date on which the study is submitted to an archive.
Study Public Release Date
The date on which the study should be released publicly.
Study File Name
A field to specify the name of the Study file corresponding the definition of that Study. There can be only one file per cell. In case, implementers wish to split the Study Files on their nodes (i.e Source Name and Sample Name), a process which results in multiple files being necessary to report the same information, they should create a bundle archive with files and report the name of the archive, thereby complying with the one file only rule.

4.1.3.2 STUDY DESIGN DESCRIPTORS

Study Design Type
A term allowing the classification of the study based on the overall experimental design, e.g cross-over design or parallel group design. The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.
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 that this term comes from. The Study Design Term Source REF has to match one the *Term Source Name* declared in the ontology section 4.1.1.

4.1.3.3 STUDY PUBLICATIONS

Study PubMed ID
The PubMed IDs of the publication(s) associated with this study (where available).
Study Publication DOI
A Digital Object Identifier (DOI) for this publication (where available).
Study Publication Author List
The list of authors associated with this publication.
Study Publication Title
The title of this publication.
Study Publication Status
A term describing the status of this publication (i.e. submitted, in preparation, published). The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.

Study Publication Status Term Accession Number
The accession number from the *Term Source* associated with the selected term.
Study Publication Status Term Source REF
Identifies the controlled vocabulary or ontology that this term comes from. The Source REF has to match one the *Term Source Name* declared in the in the ontology section 4.1.1.

4.1.3.4 STUDY FACTORS

Study Factor Name
The name of one 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 biological systems in a way that can be measured by an assay. The value of a factor is given in the Study or Assay file, accordingly. If both Study and Assay have a *Factor Value* (see section 4.2.5 and 4.3.1.5, respectively), these must be different.
Study Factor Type
A term allowing the classification of this factor into categories. The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.
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 that this term comes from. The Source REF has to match one of the *Term Source Name* declared in the ontology section 4.1.1.

4.1.3.5 STUDY ASSAYS

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

Study Assay Measurement Type
A term to qualify the endpoint, or what is being measured (*e.g.* gene expression profiling or protein identification). The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.
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 has to match one of the *Term Source Name* declared in the ontology section 4.1.1.
Study Assay Technology Type
Term to identify the technology used to perform the measurement, *e.g.* DNA microarray, mass spectrometry. The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the Term Accession Number and Term Source REF fields below are required.
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 that this term comes from. The Source REF has to match one of the *Term Source Names* declared in the ontology section 4.1.1.
Study Assay Technology Platform
Manufacturer and platform name, e.g. Bruker AVANCE
Study Assay File Name
A field to specify the name of the Assay file corresponding the definition of that assay. There can be only one file per cell. In case, implementers wish to split the Assay Files on their nodes, a process which results in multiple files being necessary to report the same information, they should create a bundle archive with files and report the name of the archive, thereby complying with the one file only rule.

4.1.3.6 STUDY PROTOCOLS

Study Protocol Name
The name of the protocols used within the ISA-TAB document. The names are used as identifiers within the ISA-TAB document and will be referenced in the Study and Assay files in the *Protocol REF* columns. Names can be either local identifiers, unique within the ISArchive which contains them, or fully qualified external accession numbers.
Study Protocol Type
Term to classify the protocol. The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.
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 that this term comes from. The Source REF has to match one of the *Term Source Name* declared in the ontology section 4.1.1.
Study Protocol Description
A free-text description of the protocol.
Study Protocol URI
Pointer to protocol resources external to the ISA-TAB that can be accessed by their Uniform Resource Identifier (URI).
Study Protocol Version
An identifier for the version to ensure protocol tracking.
Study Protocol Parameters Name
A semicolon-delimited (";") list of parameter names, used as an identifier within the ISA-TAB document. These names are used in the Study and Assay files (in the "Parameter Value [<parameter name>]" column heading) to list the values used for each protocol parameter. Refer to section Multiple values fields in the Investigation File on how to encode multiple values in one field and match term sources
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
Identifies the controlled vocabulary or ontology that this term comes from. The Source REF has to match one of the *Term Source Name* declared in the ontology section 4.1.1.
Study Protocol Components Name
A semicolon-delimited (";") list of a protocol's components; e.g. instrument names, software names, and reagents names. Refer to section Multiple values fields in the Investigation File on how to encode multiple components in one field and match term sources.
Study Protocol Components Type
Term to classify the protocol components listed for example, instrument, software, detector or reagent. The term can be free text or from, for example, a controlled vocabulary or an 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 to the selected terms.
Study Protocol Components Type Term Source REF
Identifies the controlled vocabulary or ontology that this term comes from. The Source REF has to match a *Term Source Name* previously declared in the ontology section 4.1.1.

4.1.3.7 STUDY CONTACTS

Study Person Last Name
The last name of a person associated with the study.
Study Person First Name
The first name of a person associated with the study.
Study Person Mid Initials

The middle initials of a person associated with the study.

Study Person Email

The email address of a person associated with the study

Study Person Phone

The telephone number of a person associated with the study.

Study Person Fax

The fax number of a person associated with the study.

Study Person Address

The address of a person associated with the study.

Study Person Affiliation

The organization affiliation for a person associated with the study.

Study Person Roles

Term to classify the role(s) performed by this person in the context of the study, which means that the roles reported here need not correspond to roles held within their affiliated organization. Multiple annotations or values attached to one person may be provided by using a semicolon (";") as a separator, for example: "submitter;funder;sponsor" .The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.

Study Person Roles Term Accession Number

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

Study Person Roles Term Source REF

Identifies the controlled vocabulary or ontology that this term comes from. The Source REF has to match one of the *Term Source Name* declared in the ontology section 4.1.1.

Several ISA-TAB example files are available from the project page (2). The **example** below shows an Investigation File with one Study and two Assays. The column A holds headers for sections (*e.g.*, *INVESTIGATION PUBLICATIONS*) and fields (*e.g.*, *Investigation PubMed ID*); while subsequent columns hold the value(s) for the fields named. In this example, where only one Study has been created, the Investigation section is left empty. It is important to note that each section is independent of any other, therefore the values in a column are related only within each section (*i.e.*, between headings), never between sections. For example, the values in column B in the *STUDY FACTORS* section are not necessarily connected to the values in the same column in the *STUDY ASSAYS* section.

A	B	C
ONTOLOGY SOURCE REFERENCE		
Term Source Name	CTO	MO
Term Source File	http://obo.sourceforge.net/cgi-bin/detail.cgi?cell	http://mged.sourceforge.net/ontologies/MGEDontology.php
Term Source Version		1.3.0.1
Term Source Description	The Cell Type Ontology	The MGED Ontology
INVESTIGATION		
Investigation Identifier		
Investigation Title		
Investigation Description		
Investigation Submission Date		
Investigation Public Release Date		
INVESTIGATION PUBLICATIONS		
Investigation PubMed ID		
Investigation Publication DOI		
Investigation Publication Author list		
Investigation Publication Title		

Investigation Publication Status		
Investigation Publication Status Term Accession Number		
Investigation Publication Status Term Source REF		
INVESTIGATION CONTACTS		
Investigation Person Last Name		
Investigation Person First Name		
Investigation Person Mid Initials		
Investigation Person Email		
Investigation Person Phone		
Investigation Person Fax		
Investigation Person Address		
Investigation Person Affiliation		
Investigation Person Roles		
Investigation Person Roles Term Accession Number		
Investigation Person Roles Term Source REF		
STUDY		
Study Identifier	STD1	
Study Title	The Influence of Pharmacogenetics on Fatty Liver Disease in the Wistar and Kyoto Rats: A Combined Transcriptomic and Metabonomic Study	
Study Submission Date	YYYY-MM-DD	
Study Public Release Date	YYYY-MM-DD	
Study Description	Analysis of liver tissue from rats exposed to orotic acid for 1, 3, and 14 days was performed by DNA microarrays and high resolution 1H NMR spectroscopy based metabonomics of both tissue extracts and intact tissue (n) 3).	
Study File Name	s_griffin123.txt	
STUDY DESIGN DESCRIPTORS		
Study Design Type	time course design	
Study Design Type Term Accession Number	OBI_11215	
Study Design Type Term Source REF	OBI	
STUDY PUBLICATIONS		
Study PubMed ID	17203948	
Study Publication DOI	10.1021/pr0601640	
Study Publication Author list	Griffin JL, Scott J, Nicholson JK.	
Study Publication Title	The influence of pharmacogenetics on fatty liver disease in the wistar and kyoto rats: a combined transcriptomic and metabonomic study.	
Study Publication Status	indexed for MEDLINE	
Study Publication Status Term Accession Number		

Study Publication Status Term Source REF		
STUDY FACTORS		
Study Factor Name	Time	Treatment with compound
Study Factor Type	Temporal	Stressor
Study Factor Type Term Accession Number	OBI_XXXx	OBI_XXXx
Study Factor Type Term Source REF	OBI	OBI
STUDY ASSAYS		
Study Assay Measurement Type	Gene Expression	Metabolite Characterization
Study Assay Measurement Type Term Accession Number	OBI_XXXx	OBI_XXXx
Study Assay Measurement Type Term Source REF	OBI	OBI
Study Assay Technology Type	DNA microarray	1D 1H NMR spectroscopy
Study Assay Technology Type Term Accession Number	OBI_XXXx	OBI_XXXx
Study Assay Technology Type Term Source REF	OBI	OBI
Study Assay Technology Platform	Affymetrix	Bruker AVANCE
Study Assay File Name	a_griffin123-Tx.txt	a-griffin123-Mx.txt
STUDY PROTOCOLS		
Study Protocol Name	standard procedure 1	griffin procedure 2
Study Protocol Type	animal procedure	nucleic acid extraction
Study Protocol Type Term Accession Number	OBI_XXXx	OBI_XXXx
Study Protocol Type Term Source REF	OBI	OBI
Study Protocol Description	All animal procedures conformed to Home Office, UK, guidelines for animal welfare. Male Wistar rats (n) 3 for each time point; control animals fed control diet for the same time period; Charles River UK Ltd.) were fed either standard laboratory chow , or chow supplemented with 1% orotic acid (Sigma Aldrich, UK) ad libitum.5-6 Rats were killed by cervical dislocation at days 0, 1, 3 and 14, and the left lateral lobe of the liver excised. Tissues were snap frozen and stored at -80 °C.	Total RNA was extracted by RNA Isolation Kit (Stratagene) from the livers of Wistar rats at day 0 (n) 3), day 1 and 3 (n) 2), and day 14 (n) 3).
Study Protocol URI		
Study Protocol Version		
Study Protocol Parameters Name	diet;population density	extracted product; amplification
Study Protocol Parameters Term Accession Number		
Study Protocol Parameters Term Source REF		
Study Protocol Components Name		
Study Protocol Components Type		
Study Protocol Components Type Term Accession Number		
Study Protocol Components Type Term Source REF		
STUDY CONTACTS		
Study Person Last Name	Griffin	Nicholson
Study Person First Name	JL	JK
Study Person Mid Initial		
Study Person Email	email@email.server.com	email@email.server.com
Study Person Phone	44(0)xxxxxx	44(0)xxxxxx
Study Person Fax		

Study Person Address	University of Cambridge, Cambridge, UK	Imperial College London, London, UK
Study Person Affiliation	Department of Molecular Biology	Genetics and Genomics Research Institute
Study Person Roles	submitter;investigator	investigator
Study Person Roles Term Accession Number	OBI_XXXx;OBI_YYYY	OBI_XXXx
Study Person Roles Term Source REF	OBI	OBI

The example below shows how to report multiple values in *Person Roles*, *Protocol Parameters* and *Components* (see also Table 2) and ensure accurate matching with *Term Accession Number* and *Term Source REF*:

A	B	C
Study Protocol Parameters Name	column temperature; column length; flow rate	extracted product; amplification
Study Protocol Parameters Term Accession Number	PSI:234;PSI:444;OBI_12132	;OBI_1212314
Study Protocol Parameters Term Source REF	PSI-CV;PSI-CV;OBI	;OBI

Column B shows the case where all parameters are annotated with a controlled term, using semi colon ";" (Unicode U+003B) character as value's separator within the cell. Column C illustrates the case where the first parameter is free text and the second is from a controlled vocabulary or ontology. Note the use of semi colon to ensure the *Term Accession Number* is matched and assignment to the relevant parameter.

4.2 Study file

In the Study file the fields are organized on a per-row basis, the first row containing column headers. The Study file contains contextualizing information for one or more assays. The sections below describe in details the Study file column's 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.

Before reading these sections, the reader is advised to familiarizes with several ISA-TAB example files available from the project page (2) and from section 5, which provides examples of design patterns for representing information in a Study file.

In addition, Table 3 shows the nodes together with the file, in which the node should be used, and their possible attributes, the number of allowed values, data type and dependency on a parent node. Table 4 lists the attributes that can be used to qualify nodes, the number of allowed values, data type and dependency on a parent node.

4.2.1 Study nodes

Source Name
Sources are considered as the starting biological material used in a study. Source items can be qualified using the following headers: *Characteristics [ ], Material Type, Term Source REF, Term Accession Number, Unit, Provider, Description, and Comment [ ]*.

Sample Name
Samples represent major outputs resulting from a protocol application other than the special case outputs of *Extract* or a *Labeled Extract*. *Sample* items can be qualified using the following headers: *Characteristics [ ], Material Type, Term Accession Number, Term Source REF, Unit and Comment [ ].*

4.2.2 Attributes of Study nodes

Material Type
Used as an attribute column following *Source Name*, or *Sample Name*. The term can be free text or from, for example, a controlled vocabulary or an ontology (e.g. whole organism, organism part). If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.

Characteristics [<category term>]

Used as an attribute column following *Source Name, Sample Name*. This column contains terms describing each material according to the characteristics category indicated in the column header. For example, a column header "Characteristics [organism part]" would contain terms describing an organism part. The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields are required. If the characteristic being reported is a measurement, a *Unit* column (with qualifying *Term Source REF* and *Term Accession Number*) may be also be used.

Provider ([STDYID])

Used as an attribute column following *Source Name*. It can be appended with the [STDYID] tag only in the context of referencing Sources used in a SDTM data submission. This mechanism would allow listing subject identifiers coming from several different SDTM submissions. See section 5.3 for implementation guidelines.

4.2.3 Attributes of processing events for Study nodes

Protocol REF

One or more Protocol REF columns should be used to specify the method used to transform a material or a data node. This column contains a reference to a *Protocol Name* (previously defined in the Investigation File) or to accession numbers of protocols already present in public repositories. Protocol REF can be further refined with the following elements:

Parameter Value [<parameter term>]

The field allows reporting on the values taken by the parameter when applying a protocol. Note that the term between [ ] must map to one (and only one) of parameters defined in the investigation file. Values can be qualitative or quantitative. Refer to section 5 on design pattern for additional information.

Performer

Name of the operator who carried out the protocol. This allows account to be taken of operator effects and can be part of a quality control data tracking.

Date

The date on which a protocol is performed. This allows account to be taken of day effects and can be part of a quality control data tracking. Dates should be reported in ISO format (YYYY-MM-DD).

4.2.4 Qualifiers for the Study nodes' attributes

Unit

Used if the terms provided in the *Characteristics [ ], Parameter Value [ ]* or *Factor Value [ ]* column classify data that are dimensional.

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 an ontology. Qualifies the following headers *Characteristics [ ], Material Type, Parameter Value [ ]* or *Factor Value [ ]* and *Unit*.

Term Source REF

Identifies the controlled vocabulary or ontology that the selected term comes from. The Source REF has to match a *Term Source Name* previously declared in the ontology section 4.1.1.

4.2.5 Other Study file fields

Factor Value [<factor name>]

A factor is an independent variable manipulated by an experimentalist with the intention to affect biological systems in a way that can be measured by an assay. This field holds the actual data for the *Factor Value* named between the square brackets (as declared in the Investigation file, section 4.1.3.4); for example, *Factor Value* [*compound*]. Qualifiers for Factor Values, are: *Unit, Unit Term Accession Number* and *Unit Term Source REF* in case of quantitative values, and *Term Accession Number* and *Term Source REF* in case of qualitative values. See section 5.2 for examples.

Comment

Comment columns can be added to provide additional information, but only when no other appropriate field exists.

4.3 Assay file

In the Assay file the fields are organized on a per-row basis, the first row containing column headers. The Assay file represents a set of assays, defined by the endpoint measured (*i.e.* gene expression) and the technology employed (*i.e.* DNA microarray), as described in the Investigation file. An assay file can refer to one or more external data files, see section 2.4.

Several ISA-TAB example files are available from the project page (2). In addition, Table 3 shows the nodes together with the file, in which the node should be used, and their possible attributes, the number of allowed values, data type and dependency on a parent node. Table 4 lists the attributes that can be used to qualify nodes, the number of allowed values, data type and dependency on a parent node. Section 5 provides examples of design patterns for representing information in an Assay file. As stated in section 3.4, Assay files with all columns left empty *are* syntactically valid, as are those where all columns are filled with free text values rather than controlled vocabulary or ontology terms.

Below, section 4.3.1 provides a list of (generic) column headers to describe several types of assays, the following sections, 4.3.2, 4.3.3, 4.3.4, 4.3.5 and 4.3.6 describe additional column headers (or provide more specific names) for several technology types, in line with the MAGE-TAB specification (3) and other domain-specific formats (5, 6, 10, 11, 12, 18, 19, 20, 22). It should be noted that for some of the technology types this is a work in progress, being carried forward in collaboration with the relevant standards groups.

4.3.1 Generic Assay file structure

This section holds a list of (generic) column headers to describe several types of assays, organized as: nodes (containing the string Name or File), attributes for these nodes, attributes for node processing events, qualifiers for the node's attributes and other valid fields.

4.3.1.1 Assay nodes

Sample Name (or Source Name)
Sample Name is used as an identifier to refer to from within the Study file. Thereby associating it with the Source Name in that Study file. Source Name in the Assay file can only be qualified with *Comment [ ]*. However, in experiments where the source (starting biological material used in a study) is also the sample (measured in the assay) the Source Name acts as the identifier to link the Assay to its Study file. For example, a whole body scan by NMR where the body is both source and sample, (*i.e.*, used *in toto*, rather than in part).

Extract Name (where applicable)
Used as an identifier within an Assay file. This column contains user-defined names for each portion of extracted material. Valid optional qualifying headers for *Extract* are *Characteristics [ ], Material Type, Description,* and *Comment [ ].*

Labeled Extract Name (where applicable)
Used as an identifier within an Assay file. Labeled Extract Name can be qualified using the following headers: *Label*, *Characteristics [ ], Material Type, Description, Comment [ ].*

Assay Name
Used as an identifier within the Assay file. This column contains user-defined names for each assay. Qualifying headers for Assay Name are *Performer*, *Date*, and *Comment [ ]*.

Image File (where applicable)
Column to provide names (or URIs) of the image files generated by an assay. The optional qualifying header for Image File is *Comment [ ]*. For submission or transfer, image files can be packaged with ISA-TAB files into an ISArchive, see section 2.4.

Raw Data File
Column to provide name (or URI) of raw data files. The optional qualifying header for Raw Data File is *Comment [ ].* For submission or transfer, data files can be packaged with ISA-TAB files into an ISArchive, see section 2.4.

Data Transformation Name
Used as an identifier within the Assay file. This column contains a user-defined name for each data transformation applied.
Normalization Name
Used as an identifier within the Assay file. This column contains a user-defined name for each normalization applied.
Derived Data File
Column to provide name (or URI) of files resulting from data transformation or processing. The optional qualifying header for Derived Data File is *Comment [ ].* For submission or transfer, data files can be packaged with ISA-TAB files into an ISArchive, see section 2.4.

4.3.1.2 Attributes of Assay nodes

Material Type
In the Assay file, this is used as an attribute column for *Sample Name* (unless the sample has already been described in the Study file), *Extract Name*, or *Labeled Extract Name*. The term can be free text or from, for example, a controlled vocabulary or an ontology (e.g. whole organism, organism part). If the latter source is used the *Term Accession Number* and *Term Source REF* fields below are required.
Characteristics [<category term>]
Used as a qualifying field following *Sample Name* (unless the sample has already been described in the Study file), *Extract Name*, or *Labeled Extract Name*. This column contains terms describing each material according to the characteristics category indicated in the column header. For example, a column header "Characteristics [purity]" would contain terms describing the purity of that portion of material. The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields are required. If the characteristic being reported is a measurement, a *Unit* column (with qualifying *Term Accession Number* and *Term Source REF*) may be also be used.
Label (where applicable)
Used as an attribute column following *Labeled Extract Name* to indicate a chemical or biological marker, such as a radioactive isotope or a fluorescent dye which is bound to a material in order to make it detectable by some assay technology (e.g. P33, biotin, GFP). The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields are required.

4.3.1.3 Attributes of processing events for Assay nodes

Protocol REF
One or more Protocol REF columns should be used to specify the method used to transform a material or a data node. Each column should contain a reference to a *Protocol Name* (defined in the Investigation File) or to accession numbers of protocols already present in public repositories. Protocol REF can be further refined with the following elements:

Parameter Value [<parameter term>]
The field allows reporting of the values taken by a parameter when applying a protocol, where Parameters have been declared. Values can be qualitative or quantitative. Refer to section 5 on design pattern for examples.
Performer
Name of the operator who carried out the protocol. This allows accounting for the operator effect and can be part of a quality control data tracking. If several operators need to be reported, a semi-colon (;) U+0033 separated list of values is accepted.
Date
To report of on the date on which a protocol is performed. This allows accounting for a day effect and can be part of a quality control data tracking. Date should be reported in ISO format (YYYY-MM-DD).

4.3.1.4 Qualifiers for the Assay nodes' attributes

Unit
Used if the terms provided in the *Characteristics [ ], Parameter Value [ ]* or *Factor Value [ ]* column classify data that are dimensional.
Term Accession Number
The accession number from the *Term Source* associated with the selected term, if this comes from, for example, a controlled vocabulary or an ontology. Qualifies the following headers; *Characteristics [ ], Material Type, Parameter Value [ ]* or *Factor Value [ ]* and *Unit*.
Term Source REF
Identifies the controlled vocabulary or ontology that the selected term comes from. The Source REF has to match a *Term Source Name* declared in the ontology section 4.1.1.

4.3.1.5 Other Assay file fields

Factor Value [<factor name>]
When Factor Value [] in an Assay File, they should reference technical variations (such as software, intrument or protocol variations) . In that sense, they are intended to allow reporting of studies where technical fine tuning plays a key role in the quality of the measured signal, for instance in the case of technique optimization.
This field holds the actual data for the *Factor Value* named between the square brackets (as declared in the Investigation file, section 4.1.3.4); for example, *Factor Value* [*compound*]. Qualifiers for Factor Values, are: *Unit, Unit Term Accession Number* and *Unit Term Source REF* in case of quantitative values, and *Term Accession Number* and *Term Source REF* in case of qualitative values. See section 5.2 for examples.
Comment
Comment columns can be added to provide additional information, only when an appropriate field does not exists.

4.3.2 Assay file with Technology Type: DNA microarray hybridization

Extending the list given in section 4.3.1, this type of Assay file contains additional nodes and nodes relabeled with more specific name, in line with MAGE-TAB (3):

Hybridization Assay Name (in place of Assay Name)
Used as an identifier within the Assay file. This column contains an user-defined name for each hybridization. Qualifying headers for Hybridization Assay Name item include *Array Design REF* or *Array Design File*.
Scan Name
Used as an identifier within the Assay file. This column contains a user-defined name for each Scan event.
Array Data File (in place of Raw Data File)
Column to provide name (or URI) of raw array data files.
***Derived Array Data File* (in place of Derived Data File)**
Column to provide name (or URI) of data files resulting from data transformation or processing.
Array Data Matrix File
Column to provide name (or URI) of raw data matrix files.
Derived Array Data Matrix File
Column to provide name (or URI) of processed data matrix files, resulting from data transformation or processing. Where data from multiple hybridizations is stored in a single file, the data should be mapped to the appropriate hybridization (or scan, or normalization) via the Data Matrix format itself [9].

In addition to the node attributes described in section 4.3.1, this Assay file may contains additions and/or replacements if certain headers. These changes in labeling allow more straightforward compatibility with MAGE-TAB (3):

Array Design File

Column to provide name of file containing the array design, used for a particular hybridization. For submission or transfer, ADF files can be packaged with ISA-TAB files into an ISArchive, see section 2.4.

Array Design REF
This column is used to reference the identifier (or accession number) of an existing array design.

4.3.3 Assay file with Technology Type: Gel electrophoresis

Extending the list given in section 4.3.1, this type of Assay file contains additional nodes and nodes relabeled with more specific name, in line with other domain-specific formats and associated reference repositories (6, 10, 12, 22).

Gel Electrophoresis Assay Name (in place of Assay Name)
Used as an identifier within the Assay file. This column contains user-defined names for each electrophoresis gel assay. For 2-dimensional gels, the following qualifying headers can be used instead:

First Dimension
The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields are required.

Second Dimension
The term can be free text or from, for example, a controlled vocabulary or an ontology. If the latter source is used the *Term Accession Number* and *Term Source REF* fields are required.

Scan Name
Used as an identifier within the Assay file. This column contains user-defined names for each Scan event.

Spot Picking File
Column to provide name (or URI) of files file holding protein spot coordinates and metadata for use by spot picking instruments.

The node attributes and remaining fields are as described in section 4.3.1.

4.3.4 Assay file with Technology Type: Mass Spectrometry (MS)

Extending the list given in section 4.3.1, this type of Assay file contains additional nodes and nodes relabeled with more specific name, in line with other domain-specific formats and associated reference repositories (6, 12,18).

MS Assay Name (in place of Assay Name)
Used as an identifier within the Assay file. This column contains user-defined names for each MS Assay.

Raw Spectral Data File (in place of Raw Data File)
Column to provide name (or URI) of 'raw' spectral data files.

When Mass Spectrometry is used in proteomics the following data files are required, according to PSI specifications and Pride submission requirements (6, 10):

Peptide Assignment File
Column to provide name (or URI) of file(s) containing peptide assignments.

Protein Assignment File
Column to provide name (or URI) of file(s) containing protein assignments.

Post Translational Modification Assignment File
Column to provide name (or URI) of file(s) containing posited post-translational modifications.

*Capturing data resulting from the use of mass spectrometry in metabol/nomics requires a settled definition for a **Metabolite Assignment File** (inter alia); such a file is currently under development in collaboration with the Metabolomics Standards Initiative (MSI, 18)*

The node attributes and remaining fields are as described in section 4.3.1.

4.3.5 Assay file with Technology Type: Nuclear Magnetic Resonance spectroscopy (NMR)

Extending the list given in section 4.3.1, this type of Assay file contains additional nodes and nodes relabeled with more specific name, in line with other domain-specific formats (12,18).

NMR Assay Name (in place of Assay Name)
Used as an identifier within the Assay file. This column contains user-defined names for each NMR Assay
Free Induction Decay Data File (in place of Raw Data File)
Column to provide name (or an URI) of files corresponding to the free induction decay data files.
Acquisition Parameter Data File
Column to provide name (or an URI) of files corresponding to the acquisition parameters data files.

*Capturing data resulting from the use of NMR spectrometry in metabol/nomics requires a settled definition for a **Metabolite Assignment File** (inter alia); such a file is currently under development in collaboration with the Metabolomics Standards Initiative (MSI, 18)*

The node attributes and remaining fields are as described in section 4.3.1.

4.3.6 Assay file with Technology Type: High throughput sequencing

Under development in collaboration with the Genomics Standards Consortium (GSC, 19) and reference repository (20).

5. Design patterns for Study and Assay files

Like MAGE-TAB before it, ISA-TAB is a framework with which to capture and communicate the complexity of the experimental metadata required to interpret an experiment. The fields in the Study and Assay files whose headers containing the string 'Name' (*e.g. Source Name*) and 'File' (e.g. *Raw Data File*) represent nodes in the experimental graph, corresponding to material (e.g., samples, RNA extracts, synthetic material etc.) or data objects. Edges show the relationships between nodes in the experimental graph.

Below, section 5.1 describes how to represent node (*i.e.*, material or data) processing events, and the following section 5.2 their associated qualifiers (*i.e.*, values); section 5.3 illustrate how to reference observations in an external SDTM file, and section 5.4 provides guidelines on how to represent several study designs. Further examples of how to represent experimental metadata in ISA-TAB are available from the project page (2).

5.1 Representing node processing

Conceptually, a protocol takes one or more inputs (biological material or data) and generates one or more outputs (biological material or data). Therefore protocols correspond to edges in the experimental graph, while materials and data correspond to the nodes. The following **example** shows how to represent the transformation of a node, the operator and the date by referencing its protocol (declared in the *STUDY PROTOCOLS* section of the Investigation file) and also shows the optional performer and date columns:

<Node>Name	Protocol REF	Performer	Date	<Node> Name
Node N	Protocol P	John Smith	2008-09-22	Node N'

If *Study Protocol Parameters* have been declared (in the *STUDY PROTOCOLS* section of the Investigation file) those must be referenced, for example:

<Node>Name	Protocol REF	Parameter Value [param1]	Performer	Date	<Node> Name
Node N	Protocol P	Value for param1	John Smith	2008-09-22	Node N'

In the case when two different protocols and different parameters have been declared, the transformation of a node is represented as follows:

<Node> Name	Protocol REF	Parameter Value [param1]	Parameter Value [param2]	Performer	Date	<Node> Name
Node N	Protocol P1	Value for param1	<empty cell>	John Smith	2008-09-22	Node N'
Node N	Protocol P2	<empty cell>	Value for param2	<empty cell>	<empty cell>	<empty cell>

5.1.1 Pooling or joining nodes

In a pooling event, two or more parent nodes contribute to the creation of the same child node. The following **example** shows how to represent several sources of biological material pooled together into one sample:

Source Name	Protocol REF	Sample Name
Source1	PoolP1	Sample1
Source2	PoolP1	Sample1
Source3	PoolP1	Sample1
Source4	PoolP1	Sample1

The following **example** shows how to represent several labeled extract materials that are put on the same gel.

Labeled Extract Name	Label	Protocol REF	Gel Electrophoresis Assay Name
LabeledExtract1	Cy3	GelP	Gel1
LabeledExtract2	Cy3	GelP	Gel1
LabeledExtract3	Cy5	GelP	Gel1

5.1.2 Splitting nodes

In a splitting event, one node is processed and generates two or more children nodes. The following **example** shows two splitting events, where two samples (children nodes) are derived from the same source (parent node), also using the optional descriptive columns *Term Source REF* and *Term Accession Number*:

Source Name	Material Type	Term Source REF	Term Accession Number	Protocol REF	Sample Name
Animal1	whole organism	MO	MO:1234	organ_removal_Pl	Animal1.liver
Animal1	whole organism	MO	MO:1234	organ_removal_Pk	Animal1.kidney
Animal2	whole organism	MO	MO:1234	organ_removal_Pl	Animal2.liver
Animal2	whole organism	MO	MO:1234	organ_removal_Pk	Animal2.kidney

Similarly, the following **example** show how to represent two images derived from the same scanning event:

Hybridization Assay Name	Scan Name	Image File
Hybridization1	Scan1	Image1.tiff
Hybridization1	Scan1	Image2.tiff

5.2 Representing node qualifiers

The following examples show how to represent qualitative values (Characteristics [ ], Factor Value [ ] and Parameter Value [ ]) for different nodes, when terms for the values are from a controlled vocabulary or ontology:

Characteristics [organism part]	Term Accession Number	Term Source REF
Liver	CARO:123424	CARO

Factor Value [compound]	Term Accession Number	Term Source REF
Aspirin	CHEBI:15365	CHEBI

Parameter Value [detector]	Term Accession Number	Term Source REF
Channeltron	PSI:1000107	PSI

If free text terms are used for the qualitative values, the *Term Accession Number* and *Term Source REF* columns are not necessary.

The following examples show how to represent quantitative values qualified by their units for the same three nodes, when terms for the values are from a controlled vocabulary or ontology:

Characteristics [body weight]	Unit	Term Accession Number	Term Source REF
56	kg	UO:0000009	UO

Factor Value [dose]	Unit	Term Accession Number	Term Source REF
100	mg/ml	UO:0000176	UO

Parameter Value [column temperature]	Unit	Term Accession Number	Term Source REF
70	degree celsius	UO:0000027	UO

When quantitative measurements have no units, such as pH, the *Unit* column is not required:

Parameter Value [pH]
4.3

Additional information that does not appropriately fit into any of the other node qualifiers may also be represented as free text by using a Comment [ ] column:

Comment [DBtype]
MySQL

5.3 Referencing SDTM source and sample from Study file

When required, observations in a SDTM file (14, 15) can be referenced from the ISA-TAB Study file by matching SDTM variables for source and sample(s) [in square brackets] to the corresponding *Source Name* and/or *Sample Name* column headers. The following **example** shows how to reference the SDTM subject id [USUBJID] and the laboratory reference for the sample [IDVAR=XXREFID]:

Source Name [USUBJID]	Provider [STUDYID]	Sample Name [IDVAR=LBREFID]
Source1	Study1	A
Source2	Study1	B
Source3	Study2	C
Source4	Study3	47

In this example, the sample IDVAR is set to LBREFID which points to a SDTM Laboratory Domain as identified by the 2 LB letters.

5.4 Representing design in the Study file

This section illustrates how to represent the design in the Study file for two *Study Design Type*: 'parallel group' and 'cross-over' designs. These designs are fairly common in intervention studies, where a perturbation or treatment is applied to the source or sample.

5.4.1 Study Design Type: parallel group design

In a 'parallel group' design there are two or more groups, each treated in a different and independent way. The treatments are defined as a combination of *Factor Values*. The **example** below shows a treatment consisting of three *Factor Values*: [compound], [dose] and [time of sampling post dose]:

Source Name	Protocol REF	Factor Value [compound]	Factor Value [dose]	Factor Value [time of sampling post dose]
Animal1	TreatmentP	aspirin	low	24h
Animal2	TreatmentP	aspirin	low	24h
Animal3	TreatmentP	aspirin	high	24h
Animal4	TreatmentP	aspirin	high	24h
Animal5	TreatmentP	acetaminophen	low	48h
Animal6	TreatmentP	acetaminophen	low	48h
Animal7	TreatmentP	acetaminophen	high	48h
Animal8	TreatmentP	acetaminophen	high	48h

The **example** below shows the same design type, but where several samples are derived from the same source:

Source Name	Protocol REF	Protocol REF	Sample Name	Characteristics [organ]	Factor Value [compound]	Factor Value [dose]	Factor Value [time of sampling post dose]
Animal1	TreatmentP	OrganCollectionP	Sample1	liver	aspirin	low	24h
Animal1	TreatmentP	OrganCollectionP	Sample2	kidney	aspirin	low	24h
Animal2	TreatmentP	OrganCollectionP	Sample3	liver	aspirin	low	24h
Animal2	TreatmentP	OrganCollectionP	Sample4	kidney	aspirin	low	24h
Animal3	TreatmentP	OrganCollectionP	Sample5	liver	aspirin	high	24h
Animal3	TreatmentP	OrganCollectionP	Sample6	kidney	aspirin	high	24h
Animal4	TreatmentP	OrganCollectionP	Sample7	liver	aspirin	high	24h
Animal4	TreatmentP	OrganCollectionP	Sample8	kidney	aspirin	high	24h
Animal5	TreatmentP	OrganCollectionP	Sample9	liver	acetaminophen	high	24h
Animal5	TreatmentP	OrganCollectionP	Sample10	kidney	acetaminophen	high	24h
Animal6	TreatmentP	OrganCollectionP	Sample11	liver	acetaminophen	high	24h
Animal6	TreatmentP	OrganCollectionP	Sample12	kidney	acetaminophen	high	24h

5.4.2 Study Design Type: cross-over design

In a 'cross-over' design a sequence of treatments is applied to the source or sample. To capture and order the sequence of treatments an additional label is required: [order= *integer*]. The **example** below shows an ordered sequence of treatments consisting of several *Factor Values*: [compound], [dose] and [duration], [washout period], and again [compound], [duration]:

Sample Name	Protocol REF	Factor Value [compound] [treatment order=1]	Factor Value [dose] [treatment order=1]	Factor Value [duration] [treatment order=1]	Factor Value [washout period] [treatment order=2]	Factor Value [compound] [treatment order=3]	Factor Value [dose] [treatment order=3]	Factor Value [duration] [treatment order=3]
Tissue1	TreatmentP	<empty cell>	<empty cell>	<empty cell>	<empty cell>	<empty cell>	<empty cell>	<empty cell>
Tissue1	TreatmentP	resveratrol	100mg/day	2 weeks	<empty cell>	<empty cell>	<empty cell>	<empty cell>
Tissue1	TreatmentP	resveratrol	100mg/day	2 weeks	4 weeks	<empty cell>	<empty cell>	<empty cell>
Tissue1	TreatmentP	resveratrol	100mg/day	2 weeks	4 weeks	quercetin	500mg/day>	2 weeks
Tissue2	TreatmentP	<empty cell>	<empty cell>	<empty cell>	<empty cell>	<empty cell>	<empty cell>	<empty cell>
Tissue2	TreatmentP	quercetin	500mg/day	2 weeks	<empty cell>	<empty cell>	<empty cell>	<empty cell>
Tissue2	TreatmentP	quercetin	500mg/day	2 weeks	4 weeks	<empty cell>	<empty cell>	<empty cell>
Tissue2	TreatmentP	quercetin	500mg/day	2 weeks	4 weeks	resveratrol	100mg/day	2 weeks

6. References

1. NET project members: http://www.ebi.ac.uk/net-project
2. ISA-TAB project page and documents: http://isatab.sourceforge.net
3. MAGE-TAB specification: http://www.mged.org/mage-tab
4. BioInvestigation Index project: http://www.ebi.ac.uk/bioinvindex
5. ArrayExpress: www.ebi.ac.uk/arrayexpress
6. Pride: www.ebi.ac.uk/pride
7. RSBI working group: http://www.mged.org/Workgroups/rsbi/index.html
8. Sansone SA, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S; Members of the Rsbi Working Group. The First RSBI (ISA-TAB) Workshop: "Can a Simple Format Work for Complex Studies?". OMICS. 2008 Jun;12(2):143-9
9. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, Irizarry RA, Liu J, Maier DS, Miller M, Petersen K, Quackenbush J, Sherlock G, Stoeckert CJ Jr, White J, Whetzel PL, Wymore F, Parkinson H, Sarkans U, Ball CA, Brazma A. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*. 2006 Nov 6;7:489.
10. PSI: http://www.psidev.info
11. Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, Degreef J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian RK Jr, Laursen K, Oliver SG, Paton NW, Sansone SA, Sarkans U, Stoeckert CJ Jr, Taylor CF, Whetzel PL, White JA, Spellman P, Pizarro A. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat Biotechnol*. 2007 Oct;25(10):1127-1133.
12. FuGE working group: http://fuge.sourceforge.net
13. Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ, Brazma A. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* 2002;**3**:RESEARCH0046. doi: 10.1186/gb-2002-3-9-research0046.
14. SEND: http://www.cdisc.org/models/send/v2.3
15. CDISC: http://www.cdisc.org/standards
16. FDA Data Standard Council: http://www.fda.gov/oc/datacouncil
17. Pharmacogenomic Data Submissions - Companion Guidance http://www.fda.gov/cder/guidance/7735dft.pdf
18. Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW, Kaddurah-Daouk R, Kristal BS, Lindon J, Mendes P, Morrison N, Nikolau B, Robertson D, Sumner LW, Taylor C, van der Werf M, van Ommen B, Fiehn O. The metabolomics standards initiative. Nat Biotechnol. 2007 Aug;25(8):846-8.
19. GSC: http://gensc.org
20. Short Read Metadata xml: http://www.ebi.ac.uk/embl/Documentation/ENA-Reads.html
21. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*. 2001 Dec;29(4):365-71.
22. GEL-ML: http://www.psidev.info/index.php?q=node/254
23. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK Jr, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Novère NL, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ Jr, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol*. 2008 Aug;26(8):889-96.
24. MIBBI: http://www.mibbi.org
25. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007 Nov;25(11):1251-5.
26. OBO Foundry: http://www.obofoundry.org
27. OBI: http://obi.sourceforge.net

Figure 1 - ISAarchive

For submission or transfer, files can be packaged into an ISArchive, as shown in this figure.

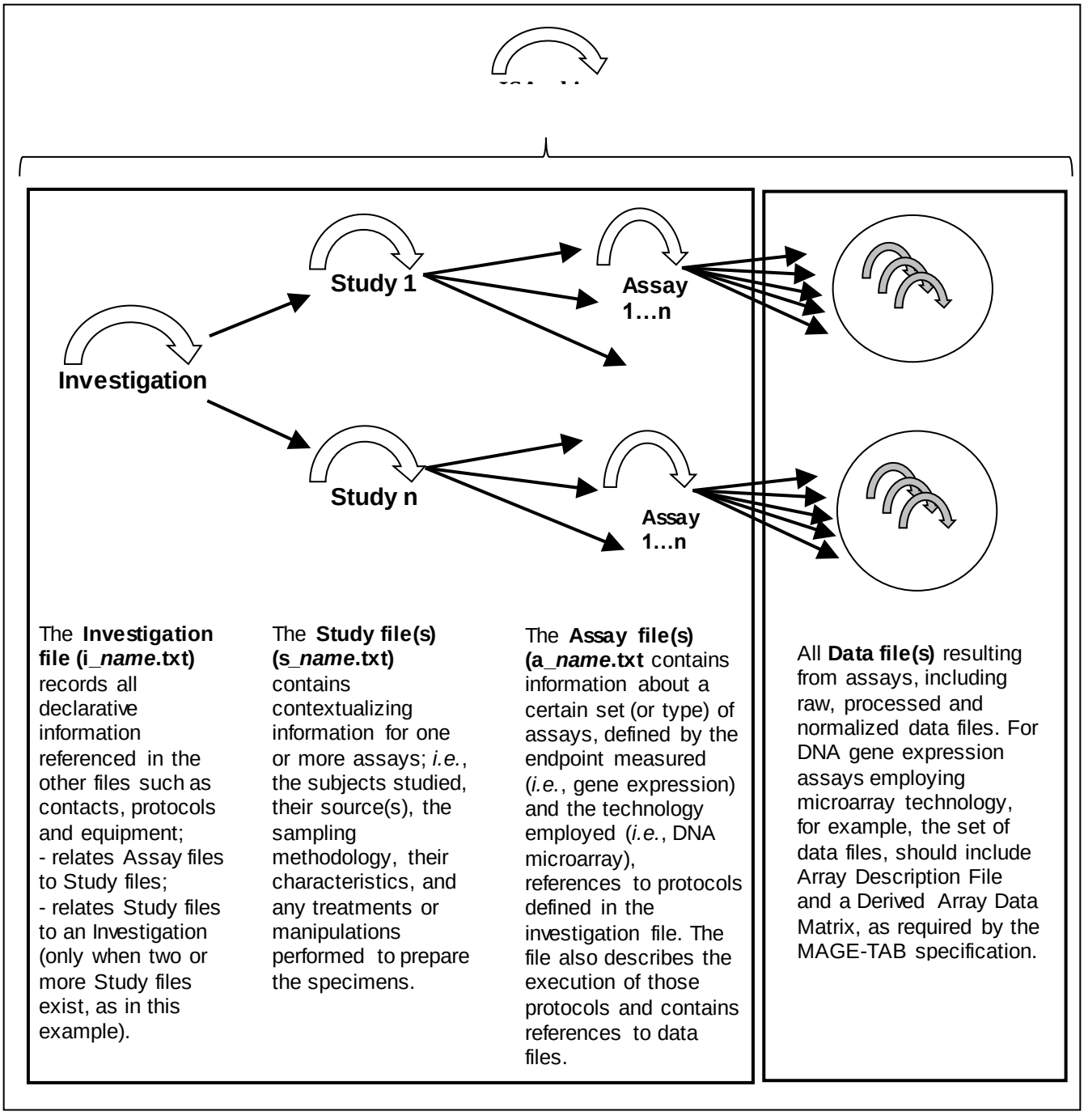

Table 1 - Types of values in the Investigation file

This table describes the types of values allowed in each field in the Investigation File's sections, together with some comments implementers may find useful. As stated in section 4.4 ISA-TAB files with all fields left empty are syntactically valid, as are those where all fields are filled with free text values rather than controlled vocabulary or ontology terms.

Investigation file fields	Type of value	Comments for implementations
ONTOLOGY SOURCE REFERENCE		
Term Source Name	text	
Term Source File	URI	
Term Source Version	text	
Term Source Description	text	
INVESTIGATION		
Investigation Identifier	text	
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	controlled terminology could be useful
Investigation Publication Status Term Accession Number	text	
Investigation Publication Status Term Source REF	Term Source Name	
INVESTIGATION CONTACTS		
Investigation Person Last Name	text	
Investigation Person Mid Initials	text	
Investigation Person First Name	text	
Investigation Person Address	text	
Investigation Person Phone	text	
Investigation Person Fax	text	
Investigation Person Email	text	
Investigation Person Affiliation	text	
Investigation Person Roles Type	text	controlled terminology could be useful
Investigation Person Roles Type Term Accession Number	text	
Investigation Person Roles Type Term Source REF	Term Source Name	
STUDY		
Study Identifier	text	Globally unique ID, *i.e.* LSID could be useful
Study Title	text	
Study Submission Date	date (YYYY-MM-DD)	

Study Public Release Date	date (YYYY-MM-DD)	
Study Description	text	
Study File Name	URI/text	
STUDY DESIGN DESCRIPTORS		
Study Design Type	text	controlled terminology could be useful
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	controlled terminology could be useful
Study Publication Status Term Accession Number	text	
Study Publication Status Term Source REF	Term Source Name	
STUDY FACTORS		
Study Factor Name	text	
Study Factor Type	text	controlled terminology could be useful
Study Factor Type Term Accession Number	text	
Study Factor Type Term Source REF	Term Source Name	
STUDY ASSAYS		
Study Assay Measurement Type	text	controlled terminology could be useful
Study Assay Measurement Type Term Accession Number	text	
Study Assay Measurement Type Term Source REF	Term Source Name	
Study Assay Technology Type	text	controlled terminology could be useful
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	
STUDY PROTOCOLS		
Study Protocol Name	text	
Study Protocol Type	text	controlled terminology could be useful
Study Protocol Type Term Accession Number	text	
Study Protocol Type Term Source REF	Term Source Name	
Study Protocol Description	text	
Study Protocol URI	URI	
Study Protocol Version	version number	
Study Protocol Parameters Name	text	
Study Protocol Parameters Term Accession Number	text	
Study Protocol Parameters Term Source REF	Term Source Name	
Study Protocol Components Name	text	
Study Protocol Components Type	text	controlled terminology could be useful
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	
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 Type	text	controlled terminology could be useful
Study Person Roles Type Term Accession Number	text	
Study Person Roles Type Term Source REF	Term Source Name	

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. However, as stated in section 4.4, ISA-TAB files with all fields left empty *are* syntactically valid.

	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 Publication Status Term Accession Number	1	As many columns as necessary
Investigation Publication Status Term Source REF	1	As many columns as necessary
INVESTIGATION CONTACTS		
Investigation Person Last Name	1	As many columns as necessary
Investigation Person Mid Initials	1	As many columns as necessary
Investigation Person First Name	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 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 Roles	1..n semi-colon (;) separated	As many columns as necessary
Investigation Person Roles Term Accession Number	1..n semi-colon (;) separated	As many columns as necessary
Investigation Person Roles Term Source REF	1..n semi-colon (;) separated	As many columns as necessary
STUDY		
Study Identifier	1	Only 1 within a study block
Study Title	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
Study Description	1	Only 1 within a study block
Study File Name	1	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 Publication Status Term Accession Number	1	As many columns as necessary
Study Publication Status Term Source REF	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 Number	1	As many columns as 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 Term Accession Number	1..n semi-colon (;) separated	As many columns as necessary
Study Protocol Parameters 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 Initial	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
Study Person Roles Term Accession Number	1..n semi-colon (;) separated	As many columns as necessary
Study Person Roles Term Source REF	1..n semi-colon (;) separated	As many columns as necessary

Table 3 - Nodes in the Study and Assay files

This table shows nodes together with the file in which the node should be used, related nodes and attributes, number of allowed values, data type and the nature of any parent node dependencies.

Node name	Related nodes and attributes	ISA-TAB file	Accept multiple values per field	Data type	Parent node dependency
Source Name	Characteristics, Provider, Material Type, Description, Comment	Study	NO	string	
	NO	Assay			
Sample Name	Characteristics, Material Type, Protocol REF, Description, Comment	Study	NO	string	Source Name
	Characteristics, Material Type, Protocol REF, Description, Comment	Assay			
Extract Name	Characteristics, Material Type, Protocol REF, Description, Comment	Assay	NO	string	Sample Name
Labeled Extract Name	Characteristics, Material Type, Protocol REF, Description, Label, Comment	Assay	NO	string	Extract Name
Assay Name	Protocol REF, Raw Data File, Derived Data File, Image File	Assay	NO	string	
Hybridization Assay Name	Protocol REF, Array Data File, Derived Array Data File, Array Data Matrix File, Derived Array Data Matrix File, Array Design or Array Design REF, Comment	Assay	NO	string	Labeled Extract Name
Gel Electrophoresis Assay Name	Protocol REF, Image File, Raw Data File, Derived Data File, Spot Picking File	Assay	NO	string	Sample Name or Extract Name or Labeled Extract Name,
MS Assay Name	Protocol REF, Raw Spectral Data File, Derived Spectral Data File, Protein Assignment File, Peptide Assignment File, Post Translation Modification Assignment File	Assay	NO	string	Sample Name or Extract Name or Labeled Extract Name
NMR Assay Name	Protocol REF, Free Induction Decay Data File, Acquisition Parameter Data File, Derived Spectral Data File	Assay	NO	string	Sample Name or Extract Name or Labeled Extract Name
Scan Name	Protocol REF, Image File, Raw Data File, Derived Array Data File, Comment	Assay	NO	string	Hybridization Assay Name or Gel Electrophoresis Assay Name
Normalization Name	Protocol REF, Derived Array Data File, Derived Data File, Derived Spectral Data File, Comment	Assay	NO	string	Assay Name, Gel Electrophoresis Assay Name, MS Assay Name, NMR Assay Name, Hybdridization Assay Name
Data Transformation Name	Protocol REF, Derived Array Data File, Derived Data File, Derived Spectral Data File, Comment	Assay	NO	string	Assay Name, Gel Electrophoresis Assay Name, MS Assay Name, NMR Assay Name, Hybdridization Assay Name
(Data) File	Comment	Assay			

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, Extract Name, Labeled Extract Name
Characteristics [ ]	Unit, Term Accession Number, Term Source REF	Study, Assay	NO	string or number	Source Name, Sample Name, Extract, Labeled Extract 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, Extract Name, Labeled Extract Name, Assay Name, Gel Electrophoresis Assay Name, MS Assay Name, NMR Assay Name, Hybdridization Assay Name
Label	Term Accession Number, Term Source REF	Assay	NO	string	Labeled Extract Name
Array Design File, Array Design File REF	Comment	Assay	NO	string	Hybridization Assay Name
Image File	Comment	Assay	NO	string	Assay Name, Hybridization Assay Name, Gel Electrophoresis Assay Name
Raw Data File	Comment	Assay	NO	string	Assay Name, Gel Electrophoresis Assay Name
Derived Data File	Comment	Assay	NO	string	Assay Name, Gel Electrophoresis Assay Name, MS Assay Name, NMR Assay Name
Array Data File	Comment	Assay	NO	string	Hybridization Assay Name
Derived Array Data File	Comment	Assay	NO	string	Hybridization Assay Name
Array Data Matrix File	Comment	Assay	NO	string	Hybridization Assay Name
Derived Array Data Matrix File	Comment	Assay	NO	string	Hybridization Assay Name
Spot Picking File	Comment	Assay	NO	string	Gel Electrophoresis Assay Name
Raw Spectral Data File	Comment	Assay	NO	string	MS Assay Name
Peptide Assignment File	Comment	Assay	NO	string	MS Assay Name
Protein Assignment File	Comment	Assay	NO	string	MS Assay Name
Post Translational Modification Assignment File	Comment	Assay	NO	string	MS Assay Name
Free Induction Decay Data File	Comment	Assay	NO	string	NMR Assay Name
Acquisition Parameter Data File	Comment	Assay	NO	string	NMR Assay Name

Performer	Comment	Study, Assay	NO	string	Protocol REF
Provider	Comment	Study, Assay	NO	string	Source Name
Date		Study, Assay	NO	date	Protocol REF
Description		Study, Assay	NO	string	Source Name, Sample Name, Extract Name, Labeled Extract Name
Comment [ ]		Study, Assay	NO	string	ALL nodes