## Proteomics data annotation with ISA-Tab

# Vasileios Tsiamis and Veit Schwämmle

**Objective**

Investigation of the potential adequacy of ISA-Tab format for *a posteriori* annotation of public PRIDE proteomics datasets at the level of experimental design as described in the session’s minutes regarding to the ELIXIR Implementation Study entitled: “*Crowd-sourcing the annotation of public proteomics datasets to improve data reusability*”. Additionally, the available ISA tools have been evaluated according to their functionalities for manual creation, mapping, validation, configuration and conversion of ISA-Tab files, and their capability to be adapted with the least possible investment to a proteomics metadata annotation framework.

**Our conclusions**

- ISA-Tab allows annotation of proteomics data sets including multiplexed designs and extensive usage of controlled vocabularies.

- The available ISA Creator provides a functional tool for data annotation (ISA).

- A basic scheme for proteomics (MS) data exists but needs further specifications.

**Suggestion for future tasks**

- Define minimal sets for the annotation of different data types and incorporate them into Assay file definitions.

- Decide which ontologies and other CVs to use.

- Define gold standard for annotations including all terms and CV we want to consider.

**Brief introduction to ISA model and ISA-Tab format**

The ISA metadata framework consists of three main entities:

* The *Investigation entity*, which contains general information, goals and concepts necessary to understand the purpose and nature of the experiment.
* The *Study entity*, which includes all the information about the subject under study, along with material characteristics and protocols used for the sample extraction.
* The *Assay entity* that contains all the sample processing operations and protocols involved in an experiment that yields quantitative or qualitative measurements.

Each *Investigation entity* can contain one or more *Studies* and each *Study* one or more *Assays*. Additionally, the *Investigation entity* is a structural linker for its contained *Studies* and the *Study entity* is a structural linker for the corresponding *Assays*.

Each *Study* and *Assay* can be represented as acyclic directed graphs. The graphs may consist of specific type of nodes. The *material nodes*, that describe the characteristics of source or sample materials, the *process nodes*, that describe the application and the parameter of operations or protocols on specific materials or material products, and the *data nodes* that correspond to the output of the applied processes. All the types of nodes and their attributes (characteristics or parameters) can be annotated by ontology terms. An example of an Assay graph is illustrated in Figure 1.

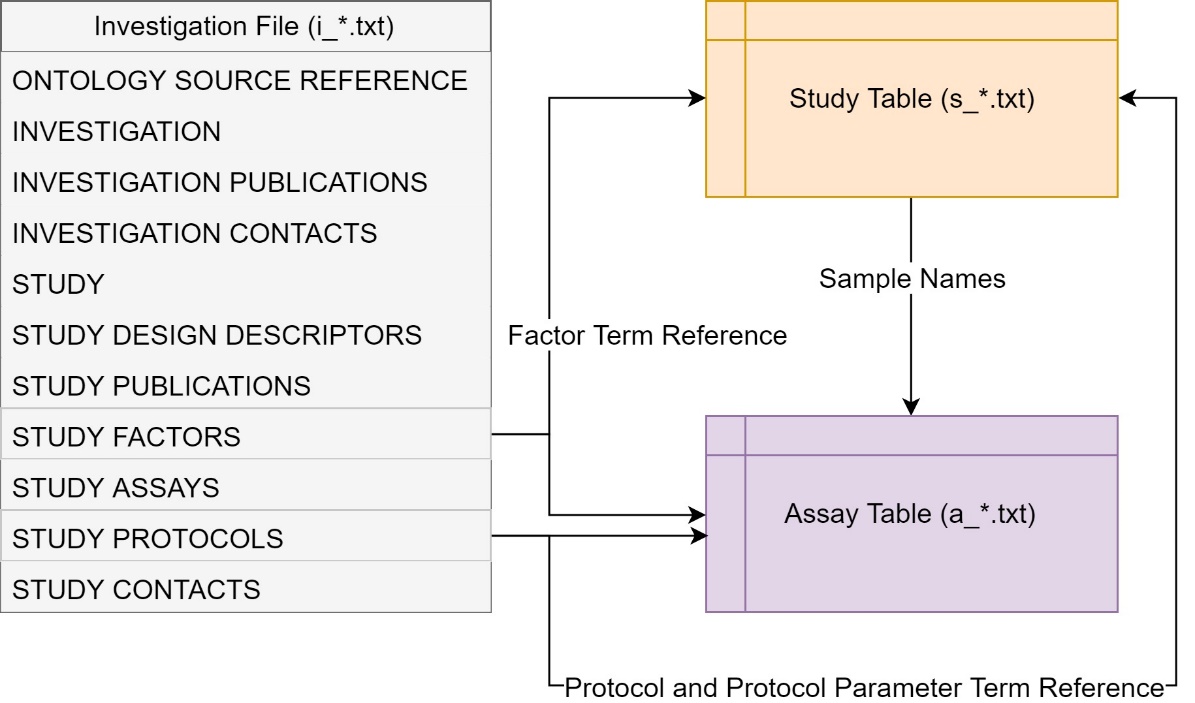


**Figure 1:** An example of an Assay graph. Material, process and data nodes are colored pink, blue and purple respectively.

The ISA-Tab format, in turn, consists of three types of TSV files, one for each of the main ISA model entities:

* The Investigation file contains several sections about the general information of the experiment, the declaration of key entities, such as factors, protocols, protocol parameters and protocol components, the determination of ontologies and CVs used in the ISA-Tab files and the connection between related Assay files to Studies and Study files to the Investigation. More specifically, the STUDY FACTORS and STUDY PROTOCOLS sections can provide factor, protocol and protocol parameter type ontology terms that can be used as reference identifiers in Study and Assay tables, as it is shown in Figure 2.
* The Study file contains information about the material and processes involved in the extraction of the biological samples. The Study file is a table structure. Each column identifier can be either a material or process node or a node attribute (characteristic, parameter value or factor).
* The Assay file describes all the experimental processes applied on the extracted samples and offers information about the provenance of the output data. The Assay file is a similar structure to Study file; however, it contains additional identifiers for data nodes and files.

In addition to ontology terms declared in the Investigation file and used as column identifiers in Study and Assay tables, more ontology terms can be declared in the two tables for the annotation of additional characteristics or parameter values.



**Figure 2:** Schematic representation of ISA-Tab files structure and interplay.

**ISA-Tab format for Proteomics metadata annotation**

ISA model is a very dynamic framework for metadata annotation and it can be easily applied to any type of experimental datasets. The main two aspects that were investigated about its adaptation capacities to MS based Proteomics experiment standards are:

1. The ability to model Multi-channel MS runs.

ISA model supports splitting and combining functionalities for all material nodes. Additionally, it offers custom column identifiers in the context of comments. Thus, it is applicable for all label-based MS assays by using ISA framework standard column identifiers, as it is shown in Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Name** | **Protocol REF** | **Extract Name** | **Comment[label]** | **Raw Spectral Data File** |
| Sample 1 | iTRAQ | Run 1 | reagent 114 | F1.raw |
| Sample 2 | iTRAQ | Run 1 | reagent 115 | F1.raw |
| Sample 3 | iTRAQ | Run 1 | reagent 116 | F1.raw |
| Sample 4 | iTRAQ | Run 1 | reagent 117 | F1.raw |
| Sample 1 | iTRAQ | Run 2 | reagent 114 | F2.raw |
| Sample 2 | iTRAQ | Run 2 | reagent 115 | F2.raw |
| Sample 3 | iTRAQ | Run 2 | reagent 116 | F2.raw |
| Sample 4 | iTRAQ | Run 2 | reagent 117 | F2.raw |
| Sample 1 | iTRAQ | Run 2 | reagent 114 | F3.raw |
| Sample 2 | iTRAQ | Run 2 | reagent 115 | F3.raw |
| Sample 3 | iTRAQ | Run 2 | reagent 116 | F3.raw |
| Sample 4 | iTRAQ | Run 2 | reagent 117 | F3.raw |

**Table 1:** Combining and splitting functionality for an iTRAQ protocol on four samples. A material node is followed by a process node and results in two combined MS runs. Run 2 is split in two raw spectral files.

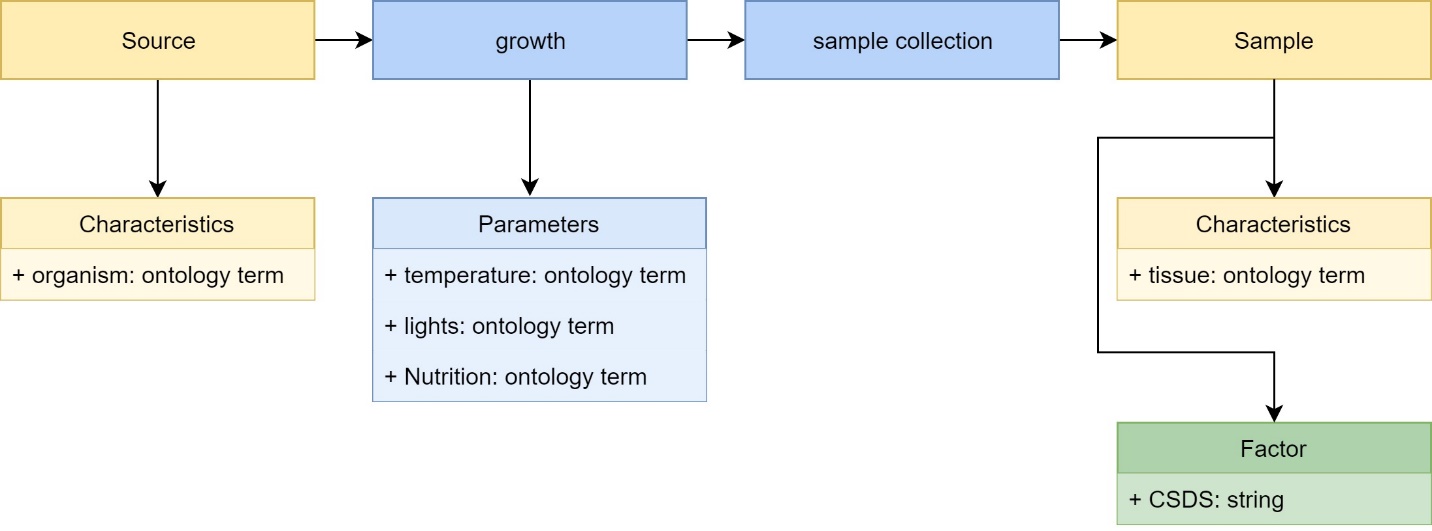
1. Available predefined identifiers related to Mass Spectrometry.

ISA model offers a variety of predefined column identifiers for MS-based Proteomics Assays. *Labeled Extract Name* can be used as an identifier for a material node following a labeling protocol with an additional *Label* tag. There is an addition group of identifiers offered for different types of MS data, including *Raw Spectral Data File*, *Derived Spectral Data File*, *Peptide Assignment File*, *Protein Assignment File* and *Post Translational Modification Assignment File*. These data type identifiers can be followed by *Normalization Name* and *Data Transformation Name* columns. The Table 1 based on the above predefined identifiers can be transformed as it I shown in Table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Name** | **Protocol REF** | **Labeled Extract Name** | **Label** | **Raw Spectral Data File** |
| Sample 1 | iTRAQ | Run 1 | reagent 114 | F1.raw |
| Sample 2 | iTRAQ | Run 1 | reagent 115 | F1.raw |
| Sample 3 | iTRAQ | Run 1 | reagent 116 | F1.raw |
| Sample 4 | iTRAQ | Run 1 | reagent 117 | F1.raw |
| Sample 1 | iTRAQ | Run 2 | reagent 114 | F2.raw |
| Sample 2 | iTRAQ | Run 2 | reagent 115 | F2.raw |
| Sample 3 | iTRAQ | Run 2 | reagent 116 | F2.raw |
| Sample 4 | iTRAQ | Run 2 | reagent 117 | F2.raw |
| Sample 1 | iTRAQ | Run 2 | reagent 114 | F3.raw |
| Sample 2 | iTRAQ | Run 2 | reagent 115 | F3.raw |
| Sample 3 | iTRAQ | Run 2 | reagent 116 | F3.raw |
| Sample 4 | iTRAQ | Run 2 | reagent 117 | F3.raw |

**Table 2:** Predefined Proteomics identifiers for the same Assay table applied on the iTRAQ labeling protocol described in Table 1.

To evaluate the modeling capabilities of ISA framework on Proteomics metadata, two public PRIDE datasets, one with multiplexing MS runs ([PXD007573](https://www.ebi.ac.uk/pride/archive/projects/PXD007573)) and one label-free study ([PXD004501](https://www.ebi.ac.uk/pride/archive/projects/PXD004501)), were annotated manually in ISA-Tab format, by a txt file editor and TSV sheets. These two studies have been chosen among other candidate datasets specifically for their large size and design complexity. Additionally, to access a complete and fully representative manual annotation for the two studies, their publication and supplementary files were used. The Study and Assay table structures for PXD007573 dataset is illustrated in Figure 3 and Figure 4 respectively. The ISA-Tab format for PXD007573 dataset consists of i\_PXD007573.txt, a\_P15177\_1.txt and s\_P15177\_1.txt. Similarly, for PXD004501 dataset the ISA-Tab files are the following, i\_PXD004501.txt, a\_20160515.txt and s\_20160515.txt.



**Figure 3:** Study file identifiers and structural schematic for PXD007573 dataset. Source material node is followed by two protocols (growth and sample collection) and results to samples with specific characteristics and factors.



**Figure 4:** Assay file identifiers and structural schematic for PXD007573 dataset. Study table provides the sample names to Assay table. The samples are first processes thought a sample preparation stage, where proteins are first extracted and then by ISA-Tab combining ability are pooled to create replicates. Proteins are digested and proceed to an iTRAQ protocol, where combining is applied again as in Table 2. The resulted raw files proceed to protein identification and quantification protocols assisted by specialized software and result the corresponding peptide and protein data.

As it is shown in the figures, additional to the good adaptation capability of ISA-Tab format to MS-based Proteomics study designs, it can also offer a detailed mapping of processes to specific parameters and components (including the components parameters). Since, a variety of instruments and software are used during a typical Proteomics experiment, ISA-Tab can be used for complete annotation of every step in an experimental workflow.

**ISA creator and ISA configurator**

After the manual annotation of the two datasets, [ISA creator](https://github.com/ISA-tools/ISAcreator/releases) tool was used to validate the files according to ISA-Tab format. ISA creator tool is not only able to read and validate ISA-Tab files, but also is a very easy to use tool for manual annotation, since it has an integrated ontology term search section connected to 100s of ontologies. ISA-Tab can be created in two ways, manually or by an investigation design wizard. For Assay file design, ISA creator has already predefined mapping templates that covers a variety of experimental designs among different omics. For MS-based Proteomics experiments it offers two different Assay templates, “protein expression profiling” and “protein identification”. All these mapping templates are imported to ISA creator by a default [configuration](https://github.com/ISA-tools/Configuration-Files) in XML format. The Assay templates do not only define the Assay tables (column identifiers) but also have a structural role in the Investigation file, since they determine the STUDY ASSAYS section by predefined terms. Both templates for MS-based Proteomics experiments are annotated by “mass spectrometry” in Study Assay Technology Type field. However, that term is very generic for a study focusing and aiming to annotate Proteomics datasets. Thus, the ISA-Tab files created for the two datasets were not able to be validated by the default configuration.

To overcome that problem and expand the predefined technology type terms, a new configuration specifically for MS-based Proteomics was generated by [ISA configurator](https://github.com/ISA-tools/ISAconfigurator/releases) tool. The new configuration is just a brief example to examine the characteristic and perspectives of ISA configurations. It creates generic Proteomics Investigation and Study files and contains templates for specific types of Proteomics assays, including iTRAQ, Label-free, MS1 label-based, MS2 tag-based, SILAC, SNR and TMT analysis. Additionally, the columns for the Assay table, contain all the predefined MS related identifiers offered by ISA framework. By loading this specific configuration to ISA creator tool, the manually generated annotation files for the two datasets are able to be validated and read by the ISA creator (configuration folder name: isaconfig-2018926).

ISA framework is also implemented by a second format, the ISA-JSON format. ISA tools additionally offer a ISA-Tab to ISA-JSON [converter](https://github.com/ISA-tools/ISAvalidator-ISAconverter-BIImanager/releases).

**Ontologies**

In ISA-Tab format, ontology terms can be used for all required annotation fields. The table with suggested ontologies per topic in session’s minutes was used as an annotation guide. That happened for two reasons, to review some of the suggested ontologies and to relate the topics described in the table with the sections of ISA-Tab format.

Organism, Tissue and Disease topics can be assigned to material characteristics or factors in the ISA-Tab format. EFO suggested ontology covers these topics efficiently and for the two annotated datasets exclusively EFO terms were used.

Sample treatment and Sample preparation topics can be related to study design and in different protocols used in Study and Assay files. Most of the terms used for these topics are from the suggested ontologies (EFO and SEP), however, there are some generic terms like “protein extraction” and “depletion” that do not exist in EFO or SEP.

Software, Software settings and Instrument settings topics can be assigned to protocol components and protocol parameters. The suggested MS ontology provided almost all the needed terms for these topics during the manual annotation. Especially for Software settings values it has a very large collection of tools, however not for parameter names which is necessary in ISA-Tab format. For parameter names PRIDE ontology covered a lot of the missing cases. Additionally, since the software used for the two studies were commercial, the terms are not available in bio.tools yet (Proteome Discoverer, MASCOT), instead MS ontology terms were used.

Analysis workflow operations finally can be assigned to protocols in Assay file. The suggested EDAM ontology contained all the needed ontology terms.

**Summary**

ISA-Tab format appears to be *suitable* for MS-based Proteomics dataset annotation. Its dynamic framework can adapt any design of Proteomics experiments, however, during the manual annotation process each dataset was annotated based on the corresponding study specifications and not by a defined and solid annotation model for Proteomics metadata. The design of such a model is necessary to limit the generic concepts of ISA framework, something that will improve the dataset annotation. That can be done by configuration design in ISA configurator. Additionally, the use of Proteomics-based configurations in ISA creator can transform the tool to a very useful tool for manual annotators with the least investment involved.

The Study table is a flow of processes necessary for the sample extraction from specific material, and the Assay table uses these samples to describe the processes involved for the quantitative measurement extraction. This is how the two tables are used in ISA framework. During manual annotation, the multiplexing processes increased the size and complexity of the Assay file. To reduce these factors, proper tools for user-friendly annotation and simplified visualization will be necessary. There should be a clear definition of all experimental relationships, either in Assay file or in Study file.

ISA is an investigation-centered framework. During the manual annotation, PRIDE accession numbers were used as unique investigation identifiers, nevertheless PRIDE datasets are assay-centered. The use of PRIDE accession number for investigations is not valid when more than one PRIDE dataset derives from the same project and thus they could be used in the same investigation. An example on that is the PXD004501 dataset, since [PXD002213](https://www.ebi.ac.uk/pride/archive/projects/PXD002213) and [PXD003351](https://www.ebi.ac.uk/pride/archive/projects/PXD003351/files) datasets are part of the same investigation too (this was accessed through the supplementary files). Thus, a proper way for unique investigation identifiers selection should be determined.