Improving Samples Descriptions in the Software SEEK.

Natalie J Stanford[1*], Norman Morrison[1], Stuart Owen[1].

1. School of Computer Science, University of Manchester, Manchester, UK.

*Corresponding Author: natalie.stanford@manchester.ac.uk.

Motivation

The SEEK platform is a web-based resource for sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes. It preserves associations between them, along with information about the people and organisations involved. The SEEK software has been used to support a large range of systems and synthetic biology projects, featuring as the main data management platform for all projects in the SysMO Consortium, the Virtual Liver Network, and more recently EraSysApp. It is now open to the community for self-managed projects. Since its inception the platform has been built, and maintained, using a network of PALs (Project Area Liaisons), who constantly test and advise on real world usage of the platform.

After the most recent round of funding, SEEK has been adopted as a major platform for the Research Infrastructure FAIRDOM. To achieve the infrastructures goals of ensuring all research outcomes are FAIR (Findable, Accessible, Interoperable, Reusable), the platform has broadened the knowledge networks used in development. One key aspect of this work was to work was to address how fit-for-purpose many aspects of the metadata collection for research recording were. One of the clubs established was based around samples. Initially samples in SEEK had been developed to cater for users within the Virtual Liver Network. As a result, the formatting was not flexible enough to accommodate the complex and changing landscape of interdisciplinary life-sciences. The club was set to work out the next steps for establishing a framework for samples recording in SEEK.

It was clear from the initial meeting of Samples Club, that there needed to be an improvement in how samples were represented in SEEK. In current interdisciplinary life-sciences projects, the breadth of samples can be vast, and therefore can only be approached with a flexible, and user defined reporting.

For more information about the Samples Club, and attendees please visit http://fair-dom.org/communities/samples-club/

If you are an expert in samples, and would like to comment on the document or join upcoming meetings please contact community@fair-dom.org

SEEK Samples Framework

Definitions:

Sample: entity (material or data) that is converted into a new item (material or data) via a process (physical or computational).

> **Material Sample:** refers to a physical sample in the laboratory, examples include cell pellets, extracted biological material (DNA, protein, metabolites), plasmids, etc.
>
> **Data Sample:** refers to digital objects, these can include machine data from analysis of material samples, models, descriptions of plasmids, DNA sequences in digital format, MassSpec Spectra etc.

Process: refers to defined activities (experimental protocols, data analysis, algorithms etc) that can be defined, explained and reported, that convert one sample into another. In essence this is an experiment. A process can only occur if a protocol and/or a standard operating procedure is associated with it.

What processes are possible?

We see a number of ways in which samples can be processed

Data > Data

An example of data to data process would be taking a a spectra from GCMS in a MzXML file and post processing the data to identify which metabolites are present, and their associated peak area.

Data > Material

An example of data to material would be the gene sequence of an enzyme required to modify the metabolism of a cell, then this information being transformed into a plasmid for inclusion in a cell.

Material > Material

An example of a material to material conversion would be taking a sample of a cell culture, and extracting and derivatizing the metabolites from the sample ready for metabolomic GCMS profiling.

Material > Data

An example of material to data conversion would be taking an extracted and derivatised sample of metabolites, and running it on a GCMS machine for profiling. The output from the machine would be the spectra, available in a form such MzXML.

Samples do not necessarily need to have a 1:1 relationship, we also expect there to be a 1:MANY, and a MANY:1 relationship

1:MANY

An example of a 1:MANY relationship would be taking a cell sample and splitting it into 3 different samples: (i) metabolite profiling; (ii) proteome analysis; (iii) genome analysis.

Taking a data description of a plasmid, and this description being used to generate a number of plasmid samples would be an example of a 1:MANY relationship with a transfer data>material.

Taking a raw data set and analysing it using several different computational methods, to produce several different processed datasets would lead to a 1:MANY, data > data relationship.

MANY:1

An example a MANY:1 relationship would be pooling of samples taken from a culture at t0 to ensure there was enough biomass for a particular analysis.

Taking multiple derived ‘omics datasets (transcriptomics, proteomics, metabolomics) and integrating it into a model would be an example of a MANY:1 data>data relationship.

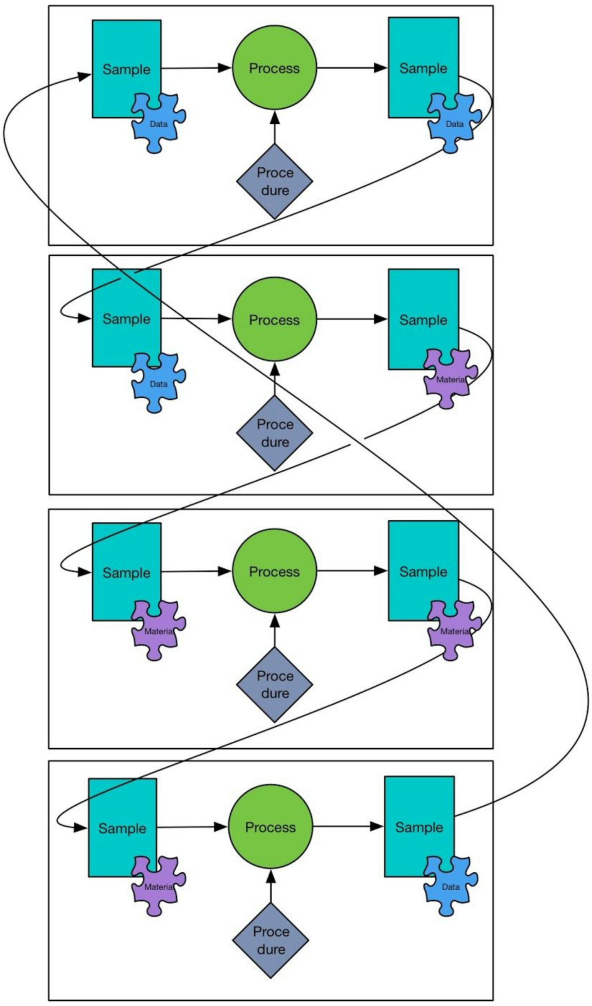

Figure 1: Here we show 4 separate instances of sample processing. Each represents a type that will be implemented: data > data; data > material; material > material; and material > data. The samples are modified via a process. The process is clarified by the information that is attached to it. In these examples a procedure is attached to it. The procedure will be a description of the experiment which can be in many forms including electronic lab notebooks, text descriptions, standard operating procedures, and any other relevant information and/or data sources.

Defining and describing samples in SEEK

An instance of a sample will now be defined and described by its metadata. The metadata will be based on terms available in ontologies (e.g. JERM ontology). The user will be given a common base of information required for the sample, and will be free to choose what extra information is required. The user can choose to follow minimum information guidelines if appropriate.

Defined in the SEEK interface

The user will be able to define the important information that should comprise the description of the sample. Once defined the template can be saved and reused. The user can populate the information using numbers and text directly in SEEK.

Rightfield templates (download)

Where a sample has been defined by its required information in SEEK, a rightfield template can be downloaded. Users can then fill in the template with numbers and text as appropriate. The filled in template will be able to uploaded back into SEEK to generate a sample automatically.

Rightfield templates (upload)

Templates prepared using Rightfield can be uploaded into SEEK, which will use the annotations and information in the spreadsheet to automatically generate a sample in SEEK.

Pulled from openBis

Samples in SEEK will follow the same flexible approach as samples from openBis. Therefore samples defined in openBis will be able to be directly pulled from openBis into SEEK.

How to handle current SEEK biosamples

culture > samples will now be covered by the 1:Many structure of samples, with a material>material conversion. The same categories, and associated metadata can be used, however the interface will change.

Priority: ensure that there is an appropriate template in the new version of SEEK that can replace this functionality.

Processing samples in SEEK

The process function is, in essence, an experiment. In the ISA (Investigation>Study>Assay) framework that SEEK uses, experiments are best represented by Assays. Therefore samples must be able to be linked directly to assays as an input and output. Data files, operating procedures, standard operating procedures, analysis code and other relevant files can also be linked to assays, providing an intuitive and flexible way in which the process can be described.