BioStudies data submission format

13/06/2017

There are two ways to submit data to BioStudies – via submission format (described here), or by using a web-based submission tool.

The BioStudies Database uses a simple tab-delimited file format that can be created in any spreadsheet editing tool such as Microsoft Excel. We call this format PageTab (**Page** layout **Tab**ulation format). The main idea behind this format is that it provides means to describe a study, its attributes, the associated files, and links; all this information will be presented to users in a way that closely resembles the input provided. The BioStudies project will evolve and work with data providers to define community-specific constraints on how the data should be described, and to build specialized ways to visually represent certain data structures; however, the baseline functionality will still enable rapid data publishing for cases where such conventions have not yet been made.

Information in PageTab is organized in blocks, where each block spans multiple non-empty lines. Blocks are separated by one or more empty lines. Each row represents a property of the item being described in the block. Each property has a name (e.g., "ReleaseDate" – see the example below) and a value (e.g., "20/03/2015"). Column number three is used occasionally – this will be explained below.

The first block should always be a **Submission**. In the example below, "S-EXMP1" is the identifier (also known as the accession number) of the study that is being submitted. The identifier pattern for Studies is "S-" followed by 4 uppercase alphabetic characters (that identify the source of the Study), usually followed by a number. The Study accession number is the main identifier of a Study, as exposed through the BioStudies data access interface. If the data submitter does not know the exact accession number of the Study being submitted, they can specify only the prefix of the Study, e.g. **!{S-EXMP}** . In this case the BioStudies system will find the first unused accession number having the defined prefix, e.g., S-EXMP2. Please contact the BioStudies team for guidance on the use of the 4 letter code and the numeric part of identifiers.

The **RootPath** attribute helps the BioStudies system locate the data files (more detail below). The **AttachTo** attribute is used to attach your submission to one of the previously defined projects within BioStudies – use this if advised to do so by the BioStudies staff. Use the **ReleaseDate** attribute to specify when the submission should become public – please make sure that Excel understands this value as a date. If the date is in the past, the submission will be made public as soon as it is loaded into the database.

	A	B	C	D
1	Submission	S-EXMP1		
2	RootPath	directoryName		
3	AttachTo	Some Project		
4	ReleaseDate	13/03/2017		
5				

Alternatively, if you want to make the submission publicly accessible and do not want to define a release date, add the "Public" access tag in the 3rd column of the first line.

	A	B	C	D
1	Submission	S-EXMP1	Public	
2	RootPath	directoryName		
3	AttachTo	Some Project		
4				

The following block describes a **Study**. A desirable attribute for a Study is the **Title** that will be shown to BioStudies users when browsing Studies or looking at search results. Use of all other attributes depends on the domain, and it is up to the submitter to use keys and values that give a suitable overview of the Study and facilitate keyword search in the database. In case of multiple values for the same key, repeat the line as many times as necessary.

5					
6	Study				
7	Title	A genomic Multiprocess survey of machineries th			
8	Description	Understanding cells as integrated systems requir			
9	Study type	high content screen			
10	[Ontology]	EFO			
11	[TermId]	EFO_0007550			
12	Number of screens	1			
13					

In this example, [Ontology] and [TermId] are **attribute value qualifiers**, providing additional information about the attribute value ("high content screen"). BioStudies user interface knows how to render these qualifier types; unknown qualifiers will be displayed only on mouse-over.

The subsequent blocks in the submission file can define data **Files** attached to the Study. You may describe the **Types** of the Files, their **Descriptions**, as well as any other attributes that help describe files – like **Format** in this example.

151				
152	Files	Type	Format	Description
153	lib.txt	library file	tab-delimited text	Bioneer haploid deletion library v.2 r
154	proc.txt	processed data	tab-delimited text	This file contains information about t
155				

This layout is useful if there are many files to describe. Alternatively, use a layout similar to that of the Study block, where a single block describes a single file:

155				
156	File	lib.txt		
157	Type	library file		
158	Format	tab-delimited text		
159	Description	Bioneer haploid deletion library v.2 modified		
160				
161	File	proc.txt		
162	Type	processed data		
163	Format	tab-delimited text		
164	Description	This file contains information about the phenc		
165				

Use the **Links** section if you want to include and describe arbitrary hyperlinks. Similarly as for Files, use a horizontal or a vertical layout. This illustrates the vertical layout:

165							
166	Link	http://idr-demo.openmicroscopy.org/webclient/?show=screen-3					
167	Type	Raw data					
168	Data format	Evotec/PerkinElmer Opera Flex					
169	Image organization	The genome wide screen was repeated twice and then a follow up re					
170							

There is an alternative way to define Links that should be used when the Study refers to records in bioinformatics resources that the BioStudies database knows about – see <here> for a list of types of identifiers that can be used. The user interface will generate clickable hyperlinks for each of Links described in this manner. In the example below, the Study refers to the European Nucleotide Archive and to reference SNPs in dbSNP:

16			
17	**Links**	**Type**	
18	AC192820	ENA	
19	rs2250341	refSNP	
20	rs733107	refSNP	
21			

Please describe study authors and their affiliations in a structured manner, using separate blocks for each **Author** and each **Organization**: create arbitrary (but unique within the Study) identifiers for all organizations (o1 in the example below), and indicate affiliation via **<affiliation>** attributes. Use several lines of the affiliation attributes if the same author has multiple affiliations.

16		
17	Author	
18	Name	Rafael Carazo Salas
19	Email	r.carazo-salas@gen.cam.ac.uk
20	Role	submitter
21	<affiliation>	o1
22		
23	Organization	o1
24	Name	Department of Genetics, University of Cambridge, D
25		

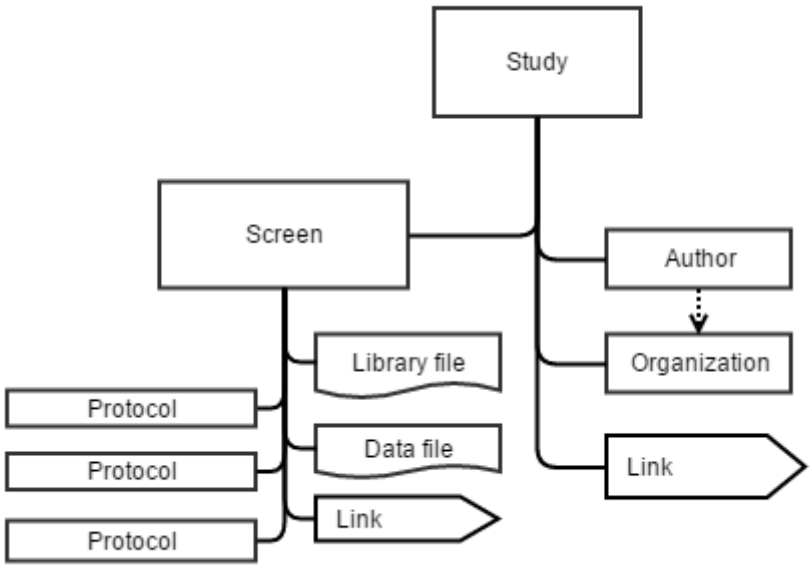
The mechanism for describing Authors and Organizations is a special case of a more general mechanism for creating hierarchical Study descriptions, and attaching Files and Links at the appropriate places in this hierarchy. See the example below. Here, the Study includes a **Section** identified as "idr0001-graml-sysgro/screenA"; the exact value of this identifier is not important since it will not be shown in the BioStudies user interface, but it is used when referring to this Section in the PageTab document (see below). The Section type is "Screen" – unlike "Author" and "Organization" sections, BioStudies will not attempt to interpret this section in any special way, and will simply display it in the user interface. If definitions of Files and Links follow a particular Section of a Study, BioStudies will attach them to this section, rather than the top-level object, the Study.

25		
26	Screen	idr0001-graml-sysgro/screenA
27	Description	Primary screen of fission yeast knock out mutan
28	Technology type	gene deletion screen
29	Type	primary screen
30	Screen size	Plates: 192
31	Screen size	5D Images: 109728

Sections by default are attached to the Study object. It is also possible to introduce deeper hierarchies; in this case, use the third column to indicate the parent block of each of the Sections. In the example below, we introduce a "Protocol" Section, and attach it to the above-defined "Screen" Section "idr0001-graml-sysgro/screenA".

65			
66	Protocol	screenA : growth protocol	idr0001-graml-sysgro/screenA
67	Type	growth protocol	
68	[Ontology]	EFO	
69	[TermId]	EFO_0003789	
70	Description	KO mutants were grown exponentially for >48 hr and imaged in	
71			

The diagram below shows the overall structure of this example Study. It has been loaded here: https://wwwdev.ebi.ac.uk/biostudies/studies/S-EXMP1 . The PageTab file is available here: https://wwwdev.ebi.ac.uk/biostudies/files/S-EXMP1/S-EXMP1.pagetab.tsv .



General notes on file formatting:

- We can accept files created in popular spreadsheet programs, with extensions such as .xls, .xlsx, .ods
- If the first character on a line is #, then that line is ignored by the software processing PageTab files – use for including comments helpful during the submission creation and/or modification stage, e.g., if the files are prepared by more than one individual, or, if a certain project/ community creates a PageTab template to guide data submissions. If it is necessary to use # as the first symbol of the line, precede it by \ - as an example, instead of using "Number of screens" attribute, use "\# of screens".