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

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., "Release date" – 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**. A submission must have the **Title** attribute, used to help identify the Submission in the BioStudies DB administrative tools. Use the **Release date** attribute to specify when the submission should become public. 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	E	F	G
1	Submission	!SBM-EPMC3272482					
2	Title	Newly folded substrates inside the molecular cage of the HtrA chaperone DegQ					
3	Release date	20/03/2015					
4							

The following block describes a **Study**. A mandatory attribute for a Study is the **Title** that will be shown to BioStudy 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 (see **Study variable** in the example).

	A	B
4		
5	**Study**	**!S-DIXA001**
6	Title	Transcriptomic fingerprints in human peripheral blood mononuclear cells indicative of g
7	Description	has been successfully applied over the last 30 years to determine early biological effects due to exposure to carcinogens. Despite their success, these early biological effects markers provide limited mechanistic insight, and are unable to detect
8	Study variable	Compound
9	Study variable	Dose
10	Study variable	Tox Class
11	Study variable	Concentration
12	Organism	Homo sapiens
13	Tissue	Human PBMC
14	Compound	Imidazoquinoline
15	Compound	Malonaldehyde
16	Measurement type	transcription profiling
17	Technology type	DNA microarray
18	Technology platform	Agilent
19		

Both Submissions and Studies have identifiers (accession numbers) in the BioStudies Database – in the examples they are "SBM-EPMC3272482" and "S-DIXA001". The exclamation mark tells the BioStudies processing software that these identifiers are globally unique, and there should not be two Submissions or Studies in the database with the same identifier. The identifier pattern for Studies is "S-" followed by 4 uppercase alphabetic characters, followed by a number. The Study accession number is the main identifier of a Study, as exposed through the BioStudies data access interface. The Submission identifier is significant only for a particular submitter, this will be used for tracking Submissions in administrative tools; the pattern is similar, but prefix is "SBM-" instead of "S-". Please contact the BioStudies team for guidance on the use of the 4 letter code and the numeric part of identifiers.

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

4				
5	**Files**	**Type**	**Title**	**Analysis provider**
6	CTCF_6species.zip	GZIP	Peak calls for 6 species	IGHG
7	CTCF_human_5way.gff	GFF	5-way (placental) shared CTCF binding events (hg19)	IGHG
8	CTCF_human_HsaMmuCfa.gff	GFF	Human-mouse-dog shared CTCF binding events (hg19)	IGHG
9	CTCF_mouse_5way.gff	GFF	5-way (placental) shared CTCF binding events (mm9)	IGHG
10				

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

15		
16	File	pgen.0030063.sg001.ppt
17	Type	application(vnd.ms-powerpoint)
18		
19	File	pgen.0030063.sg002.tif
20	Type	image(tiff)
21		
22	File	pgen.0030063.sg003.tif
23	Type	image(tiff)
24		

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 horizontal layout:

11		
12	**Links**	**Title**
13	http://www.ebi.ac.uk/arrayexpress/files/E-GEOD-24891/E-GEOD-24891.sdrf.txt	Link to Sample Information
14	http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-24891/files/	Link to Data Files
15	http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-24891/	Link to ArrayExpress
16		

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		

If you want to describe study authors and their affiliations in a structured manner, use separate blocks for each **Author** and each **Organization**, create arbitrary identifiers for all organizations (o1 and o2 in the example below), and indicate affiliation via **<affiliation>** attributes. Use several lines of the affiliation attributes if the same author has multiple affiliations. Also, please note the use of the 3rd column in this example, to link to the Study that the Authors and Organizations are related to.

21				
22	Author		S-EPMC2924276	
23	Name	Croset V		
24	<affiliation>	o1		
25				
26	Author		S-EPMC2924276	
27	Name	Rytz R		
28	<affiliation>	o1		
29				
30	Author		S-EPMC2924276	
31	Name	Cummins SF		
32	<affiliation>	o2		
33				
34	Organization	o1	S-EPMC2924276	
35	Name	Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland		
36				
37	Organization	o2	S-EPMC2924276	
38	Name	School of Biological Sciences, The University of Queensland, St. Lucia, Queensland, Australia		
39				

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 structure of the Study includes two **Sections**, S1 and S2. Each of them has its own set of attributes, as well as own set of **Links** attached. Again, please note the use of the 3rd column to indicate the parent block of each of the Sections, which is the Study with accession S-BSST1 in this case. It is possible to introduce subsections by referring to the identifiers of their parent sections. In this example we could define the "Raw Data" section as a child of the "Description" section by replacing the second occurrence of "S-BSST1" by "S1".

40			
41	**Section**	**S1**	S-BSST1
42	Title	Description	
43	Description	As published in Science, researchers from Cambridge, Glasgow and Greece have discovered a remarkable amount of plasticity in how transcription factors maintain their function over large evolutionary distances	
44			
45	**Links**		
46	http://www.sciencemag.org/cgi/content/abstract/science.1186176		
47	http://www.ebi.ac.uk/Information/News/pdf/Press09Apr10.pdf		
48			
49	**Section**	**S2**	S-BSST1
50	Title	Raw Data	
51			
52	**Links**	**Title**	
53	http://www.ebi.ac.uk/arrayexpress/experimer	The multi-species CEBPA and HNF4A reads can be found at ArrayExpress with the accesion number E-TABM-722.	
54	http://www.ebi.ac.uk/arrayexpress/experimer	The TC1 expression data can be found at ArrayExpress with the accesion number E-MTAB-178.	
55			

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.