BDS User Interface

Monday, February 1, 2016 6:19 PM

This pdf is available from http://djargon.azurewebsites.net circa Feb 2016. Click on 'BDS UI' link at the top.

PURPOSE

The purpose of this document is twofold: First to give an overview of the BDS UI and second to provide the project background as a sequence of short sections (down below).

BDS = Biogeochemistry Data System. Biogeochemistry is the tactical/technical term for Earth System Science; or for the component of ESS particularly concerned with the global carbon cycle if you like.

UI = User Interface. This is particularly concerned with three functions: Inspecting datasets, publishing datasets, and querying datasets.

WHAT YOU NEED

To use the BDS UI you will need a Windows Live ID. Go to http://live.com to establish one if you do not already have one. At some point you will also need to be authorized to use the BDS system by the administrators.

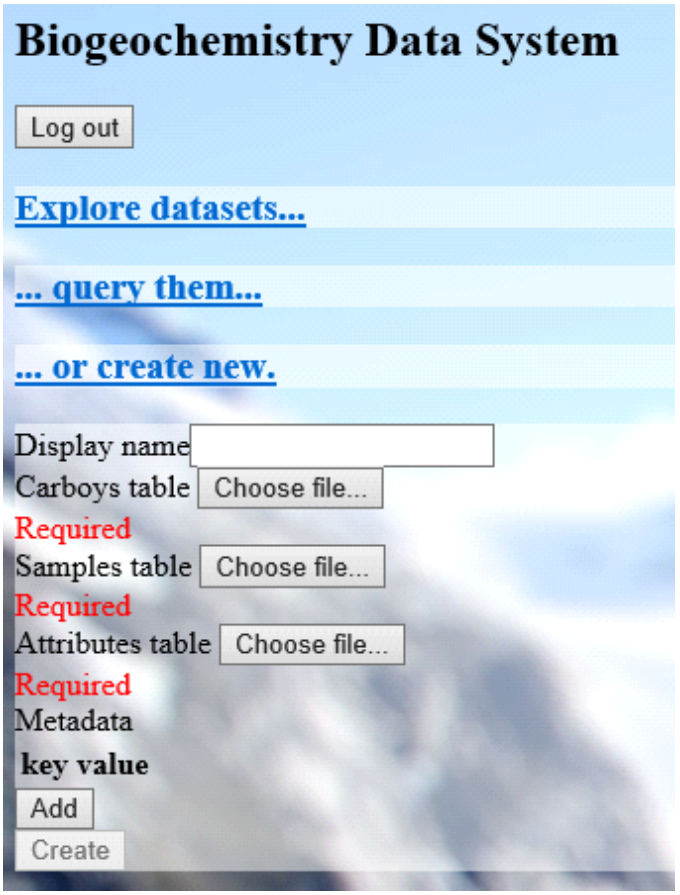
You will also need a working knowledge of dissolved organic matter. The author of this documentation has only a rudimentary understanding of dissolved organic matter, just for reference. What we are talking about here are low-to-intermediate mass molecules that include a lot of carbon and a lot of hydrogen; and generally a fair amount of oxygen with possibly other elements included, particularly nitrogen.

You will need a web browser pointed to the necessary website (see communications from the group) and you may wish to join the group slack account to make it easier to ask questions

Part 1: The BDS UI

Before proceeding we strongly recommend you read through Part 2 below, 'The BDS Background'.

We proceed in four steps: Authenticate, Create, Explore, Query. We work entirely from a Dataset that is part of our test collection; data collected by Aron Stubbins from glaciers in Tibet. Here is the UI starting point (after logging in):



Authenticate

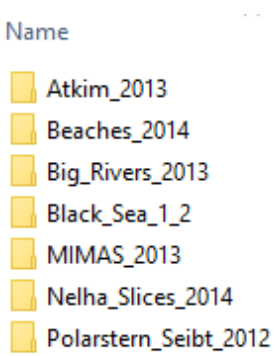
Login/Authentication process at the BDS Web User Interface: Client the authentication button and log in using your Windows Live ID as noted. This type of authentication is called ACS authentication and it happens by means of a pop-up window. If you don't see it you may have a restriction enabled from your browser. This form of authentication involves your (human) interaction. Behind the scenes: When you log in your machine is given a token (sequence of bytes) that enables you to operate the User Interface and see data that you specifically have uploaded: 'Your Data'.

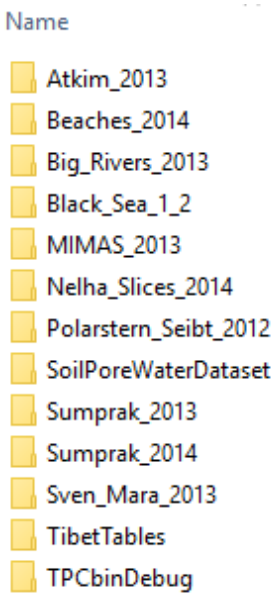
Notice that BDS also has a second authentication mechanism. We call this a Simple ID and it is currently not in play. It is intended for future automation where you want a machine (say a Mass Spec) to automatically publish data to your account without you needing to log in.

Create

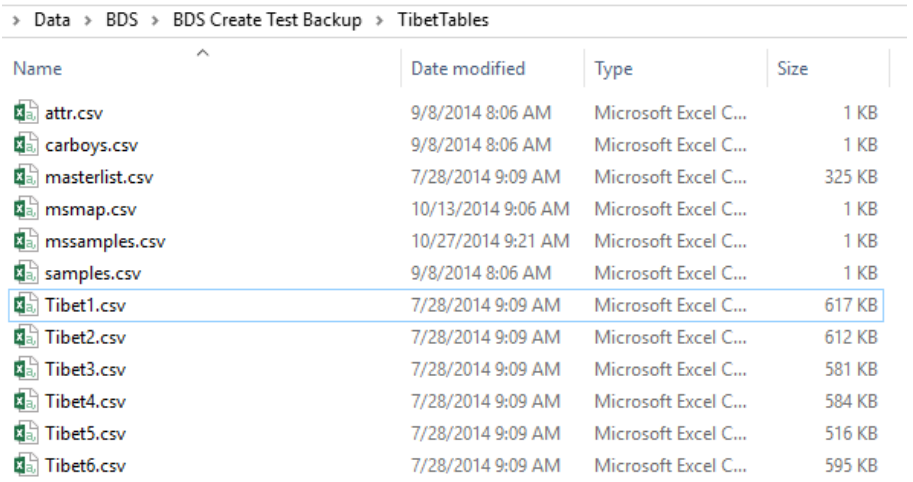
PHYS Type

Set up a folder for your Dataset that contains everything you intend to include in the Dataset upload. Here are the folder contents for our test dataset which in turn resides in the folder TibetTables:

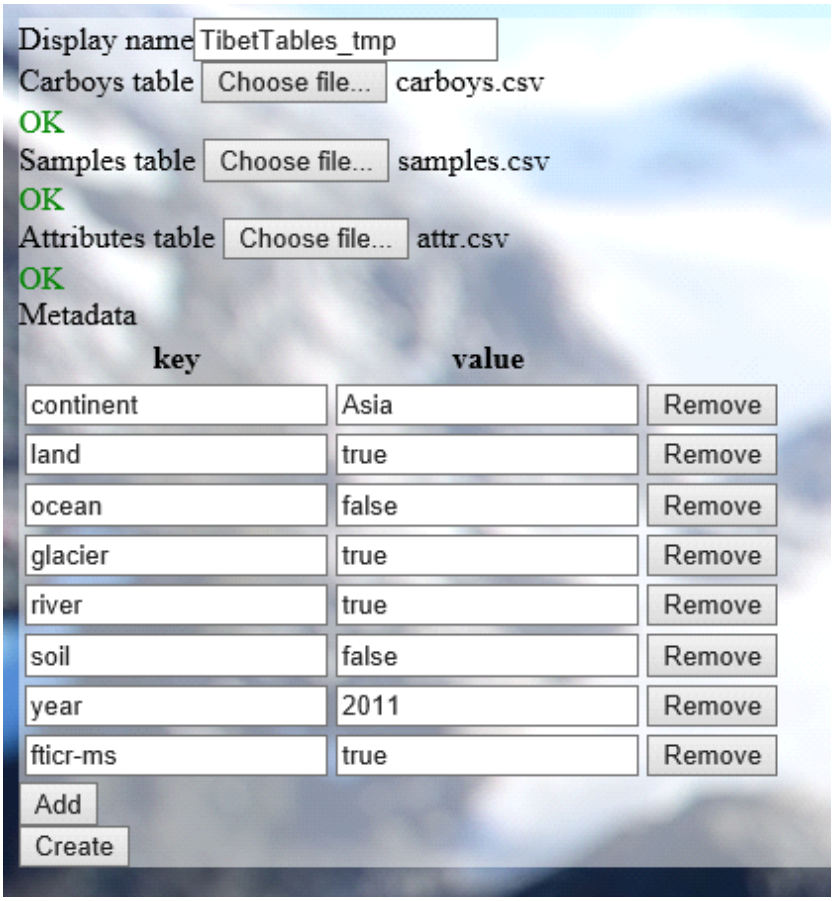




Here are the contents of TibetTables:



Upon populating the three file fields and adding some arbitrary key:value labels my UI looks like this:



Please note that the files are those found in the folder listing above. I now show what is in these three csv files in sequence, noting that it is 'just about as minimal as possible':

carboys.csv:

```
Name,Longitude,Latitude,Elevation,Depth,Timestamp
Carboy 1,84,29,6000,12,2014-01-01T01:02:00
Carboy 2,84.1,29.1,6100,14,2014-01-02T04:02:00
Carboy 3,84.2,29.2,6200,16,2014-01-03T05:02:00
Carboy 4,84.3,29.3,6300,18,2014-01-04T01:02:00
Carboy 5,84.4,29.4,6400,20,2014-01-05T01:35:00
Carboy 6,84.5,29.5,6500,22,2014-01-06T06:02:00
```

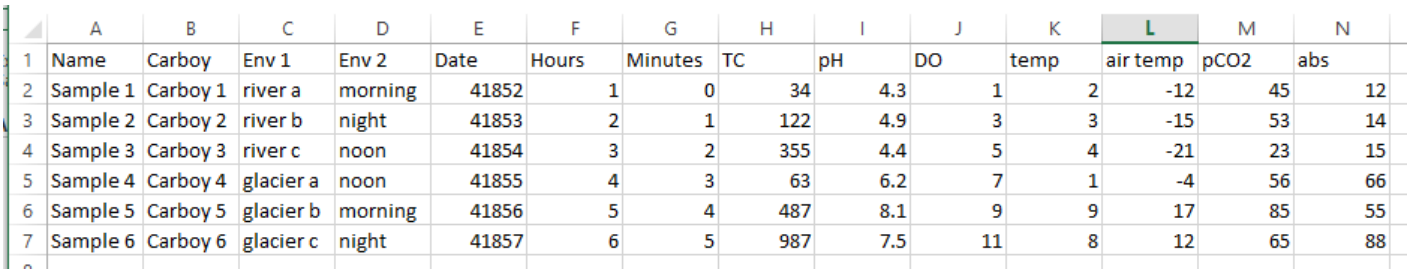
Notice that the Name column is text value names for carboys: Source water masses assumed to be 'in one bottle'. The remaining five columns are not necessary; they are optional.

attr.csv:

	A	B	C	D	E	F	G
1	Name	Method	Unit	Min	Max	Error	Remarks
2	TC	burn	mg/l	0	1000	5	qq
3	pH	yyy		4	10	varies	rr
4	DO	gas	mg/l	0.01	17	0.2	ss
5	temp	therm	deg C	-4	100	0.1	tt
6	air temp	therm	deg C	-100	100	0.1	uu
7	pCO2	xxx	%	0	100	3	vv
8	abs	light	%	0	100	2	ww
9							

Here I have loaded the file into Excel to make it easier to parse the text. Notice that there are 7 rows following the header. These are optional; they could be left blank. Likewise the columns are also optional (although 'Name' may be required). So the entire table is optional. What does it accomplish? It lists attributes that ascribe to each sample. Each sample can have a total carbon value (TC), it can have a pH, a Dissolved Oxygen measurement, a temperature, a light absorbance, etcetera. Now to the samples:

samples.csv:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Name	Carboy	Env 1	Env 2	Date	Hours	Minutes	TC	pH	DO	temp	air temp	pCO2	abs
2	Sample 1	Carboy 1	river a	morning	41852	1	0	34	4.3	1	2	-12	45	12
3	Sample 2	Carboy 2	river b	night	41853	2	1	122	4.9	3	3	-15	53	14
4	Sample 3	Carboy 3	river c	noon	41854	3	2	355	4.4	5	4	-21	23	15
5	Sample 4	Carboy 4	glacier a	noon	41855	4	3	63	6.2	7	1	-4	56	66
6	Sample 5	Carboy 5	glacier b	morning	41856	5	4	487	8.1	9	9	17	85	55
7	Sample 6	Carboy 6	glacier c	night	41857	6	5	987	7.5	11	8	12	65	88

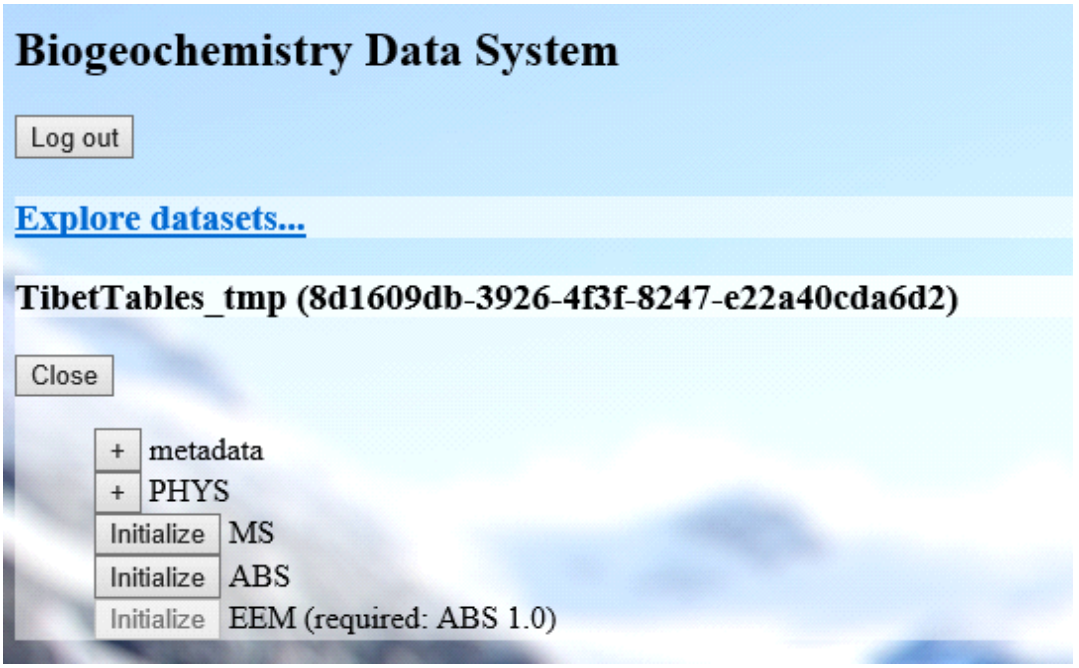
Again using Excel to make the table easy to read. Notice now that we have Sample names in the first column and Carboy names in the second column. These two columns are required. The remaining columns are optional.

The Env1 and Env2 and Date and Hours and Minutes columns are optional; and they are not called out in the attributes file above. It is completely allowed to simply create column attributes inside the Samples file.

The TC, pH, …, abs columns are present in the Samples table and are called out in the attributes table. What this accomplishes is that it allows you to enter a single value for each attribute on the Sample row without having to include all the details of those measurements. Those details are accommodated in the attributes table. So the way to think of it is: The attributes table lets you list out attributes you care about as rows with various details like units and which instrument you used; and then those attributes are corner-turned to become columns in the samples table; where the rows are the Samples and the entries are the simple data values.

Finally notice that we have added a number of key:value pairs, some of which are descriptive and others are boolean. This is a matter of preference and art; whether and how you use this feature.

And then one clicks on Create. (If you are inspecting your table csv files using an application such as Excel the file may be considered 'open' and will fail to load. So close any applications like that and if necessary re-stipulate the files in the UI.)



Upon a successful Create the UI reverts to the Explore view. Notice that I can now proceed to Initialize MS and ABS types but not the EEM type (which depends on ABS existing).

A final note on the Create process: Obviously most of the work is done in the preparation of the three critical tables as csv files. The reason we do carboys to samples is to accommodate the case where a single water mass is divided into multiple samples, for example in an incubation experiment. By accommodating this detail and the variability in method the BDS is designed to lower the barrier to publication and enable considerable detail in the description of water samples should the researcher so desire.

This completes the creation of a PHYS Type dataset. PHYS is one of several supported types but is first among equals: You must go through the PHYS creation process before you can add other types such as ABS or MS.

MS Type

To be written still.

ABS Type

To be written still.

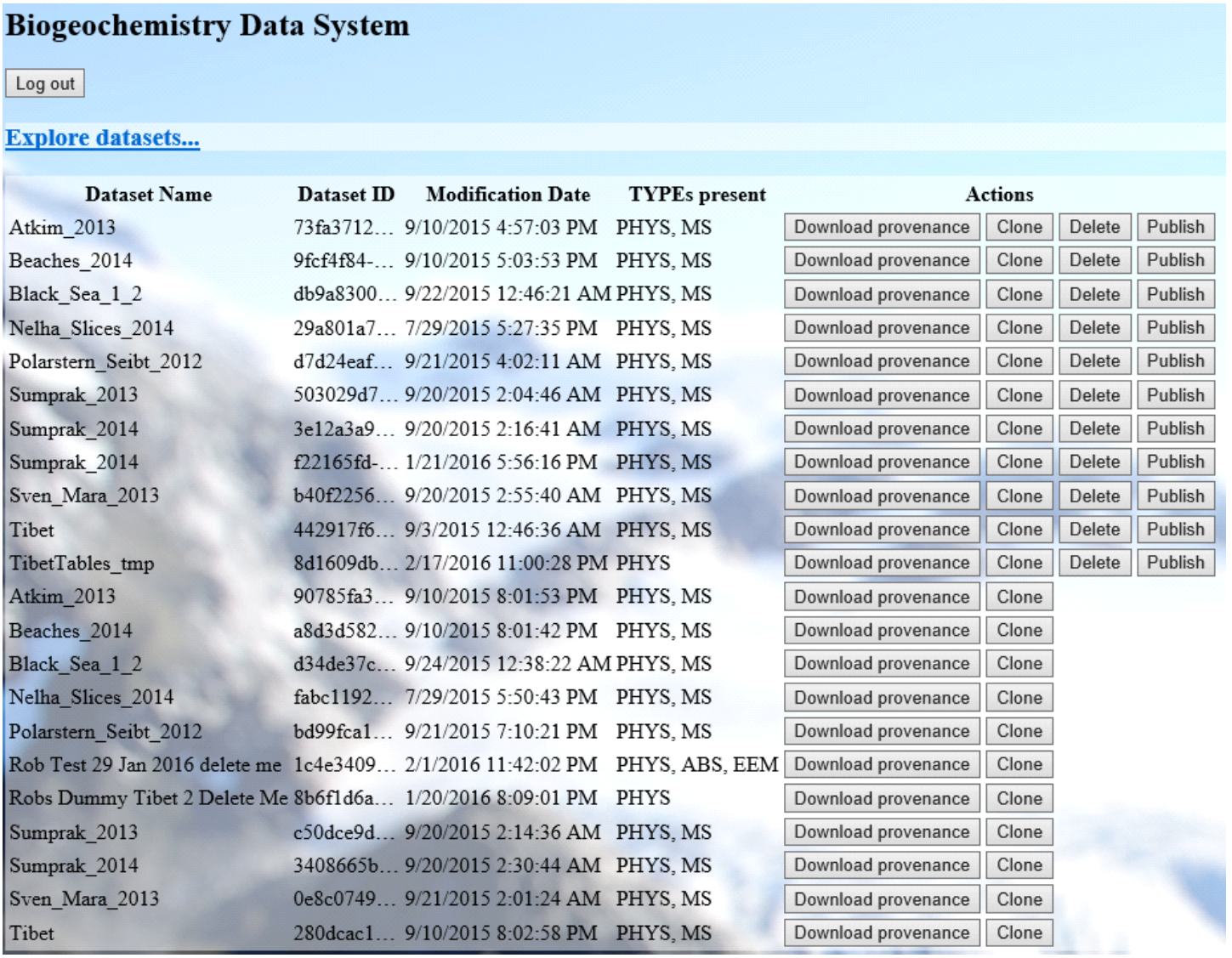
EEM Type

To be written still.

Explore

Above we created a test 'TibetTables_tmp' dataset which appears in the Explore view. Clicking on the metadata and PHYS '+' buttons produces this view:

Note that the key:value information is present here as well as some long string identifiers. These are used programmatically so there is never a need to remember them. By clicking on Close we return to the main Explore view which lists both private and public Datasets:



The upper set of Datasets are private. I can Delete or Publish them; and I can also click on them and add new Types. The lower set of Datasets do not have 'Delete' or 'Publish' buttons. They are static copies of private datasets that have been Published; and they are visible to anyone visiting the system who has authenticated.

The Clone button permits you to make copies of Datasets (Public or Private) that will exist in your Private collection. These can be subsequently modified; and also subsequently published. This provides a mechanism for evolving Public datasets without allowing a researcher to directly modify a public Dataset. The rationale for this is as follows: If a Public dataset can be edited by anyone then it could potentially break another researcher's use case. Instead we work from clones of datasets so that researchers can restrict the impact of changes to their own work.

Query

To be written still…

Part 2: The BDS Background

These Q/A pairs are intended for participating scientists using the BDS.

Q: What do I need to know about?
A: If you'd like to have an 'origins' perspective on this system then we suggest you review -- in sequence -- the following BDS components. Please first read through this entire Q&A section to get a system-level perspective; then learn sufficient details so that BDS makes sense to you. For example it would be a pity to Create 40 datasets only to learn (after the fact) that you could have included metadata with each of them that would make them much more useful. Here are the mini-topics:

1. The research model: Perfunctory versus Exploratory work in data science.
2. The data system model: 'Types x Levels x Datasets' and 'Carboys-to-Samples'
3. The publishing model: How metadata is optional but makes your data more valuable.
4. The automation model: How the BDS will become easier to use.
5. Scaffold model: How BDS can grow out, grow up, and fill in.
6. User Interface A: Dataset viewing / retrieval
7. User Interface B: Creating a new dataset
    a. PHYS type creation
    b. ABS to EEM
    c. MS
    d. Building in Metadata
8. User interface C: Dataset Query
9. Provenance: What is it, how do I use it?
10. Death Star essay

Notice that parts 6, 7 and 8 are provided here as abbreviations of Part 1 above: The BDS UI instructions proper.

Q: What are all these BDS models; and what the hell does BDS stand for in the first place?
A: BDS stands for Biogeochemistry Data System. This is a very generic term for something that has a very specific purpose: The storage, sharing, and querying of data concerning Dissolved Organic Matter (or DOM). DOM is a vast ensemble of medium-sized carbon molecules that permeate the earth's hydrosphere. The term BDS is general, however, because the system itself is expandable to accommodate new types of data. So it begins from a very specific basis and has aspirations to adapt to what scientists need. This leads to the notion of models: The BDS is conceived and built from a set of models which are just 'idea sets' for how to construct a useful data mangement system. Without further ado then here are the models…

Q: What is the BDS Research model?
A: This is the idea that some areas of research involve data processing that is *perfunctory,* in other words it can be automated with not a lot of human intervention and fine-tuning; and then there are other areas where it is all about human intervention and interpretation. Obviously this is a spectrum; and what we hope to accomplish in BDS is the automation of some of the perfunctory processing in order to save researchers time. In the case of DOM we have several spectral analysis methods that are applied to water samples and in particular to the organic molecules dissolved in that water. A machine produces a 'raw-ish' dataset from those samples which is then typically processed further to arrive at a higher-level result. Doing this processing is tedious; so why not try and automate it? We come back to this in further detail below when we describe the Automation model.

Q: What is the BDS Data System model for Types x Levels x Datasets?
A: A 3-dimensional structure with x, y and z axes. The x-axis is data Types, initially PHYS, MS, ABS and EEM. These are described below; but the only required Type is PHYS which is an abbreviation for PHYSICAL, i.e. a physical description of a water sample that presumably contains organic carbon in dissolved form. Other Types (MS, ABS, EEM) need not be present in a Dataset; but PHYS must be present as the means of *defining* a dataset. Second axis, the z-axis, is data Level. For a given Type (say MS) there is a sequence of Levels that represent successive steps in (mostly) perfunctory processing. Level 1.0 is typically ingest data; and 1.1, 1.2, 1.3 and so on represent incremental stages of processing. Level 2.0 is

used to represent a data reduction step; so for example data might be clustered into 5 sets from 50,000 original data values. This would represent a data reduction factor of 10,000 and so the output or result would be assigned a Level of 2.0 or higher. Level 3.0 does not exist yet but is intended to represent a hybridization across Types. Finally the third axis is Datasets which are quantified in terms of source Carboys (masses of water) that are in turn quantified as Samples. This is described in more detail next.

Q: What is the BDS Data System model for Carboys and Samples?
A: A Dataset is defined as a number of Carboys. A Carboy is in turn some number of Samples. It may be Zero or One or Many but both the Carboys and the Samples are given unique names and the Samples are mapped to the Carboys. Most often we expect a 1:1 mapping; so the Carboys would in this case be redundant. However the motivation is very important: We expect a geochemist studying DOM to often enough draw multiple Samples of water from the same source water mass, for example in doing a photodegredation experiment. The Carboy-Sample mapping supports this practice systematically rather than obliging the scientist to try and build it into a Sample naming convention or by some other means.

Q: What is the BDS Publishing model?
A: As a BDS User you have a unique identity or login that uses your Windows Live ID. If you do not have one you can get one for free from Microsoft; and you can use any email address for that; you just have to remember the password you associate with it. This 'gets you in' to the BDS interface. Once there you can publish data; and that will automatically be Private to you. You will see in the User Interface both Public data and your Private data. The latter includes a Publish button; so you can contribute to the Public data pool. Finally as noted you create new Datasets as Carboys mapped to Samples but very little information is needed in order to accomplish this. That is, the barrier to entry for publishing Datasets in BDS is quite low. This is intentional to try and keep the BDS as flexible as possible.

Q: What is the BDS Automation model?
A: Very simply BDS allows you to do Inspect / Publish / Query on Datasets manually through a web User Interface. This is fun to do… once. Twice and it may start to get tedious even if it is saving you hours of time by running PARAFAC for you. So what we want to work towards is a fully automated pipeline that goes from you putting a physical water sample into your instrument to you having the BDS data available at all processing levels *without you having to do anything.* In order to reach this Automation Nirvana we will still need to build some of the bridging software. Automating what can be automated (provided it actually works) will save you time and permit you to do more actual research.

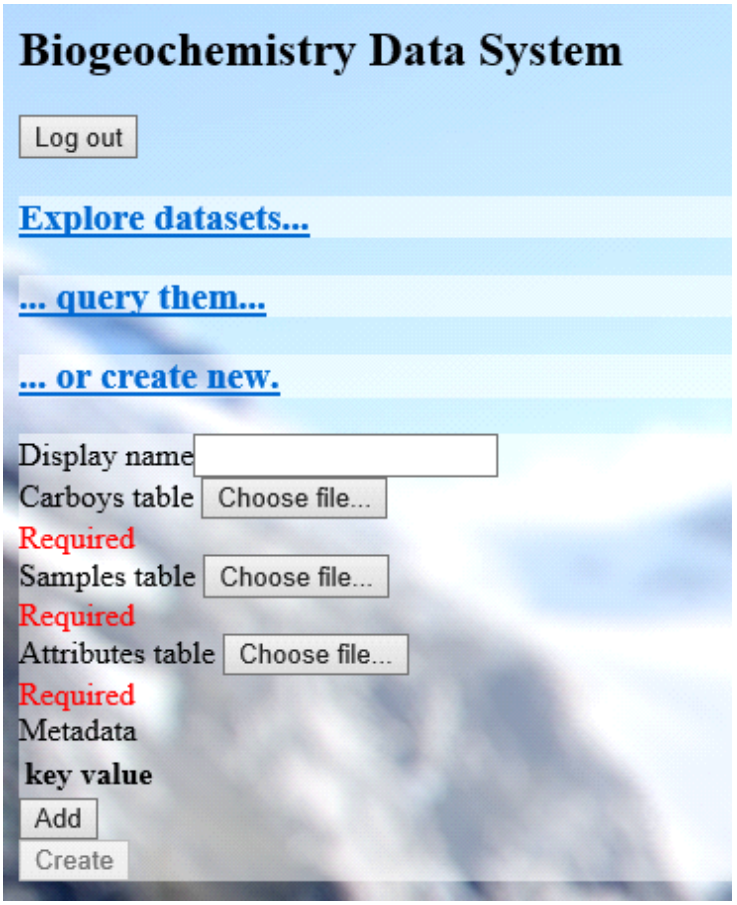
Q: What is the BDS Scaffold model?
A: BDS has certain data Types and Levels defined; with associated construction software built in. However BDS also accommodates new Types and new Levels through an administrative process. This is not documented here; but it is intended to be fairly straightforward and easy to do. This is yet to be demonstrated in practice. The point is that BDS is as much scaffolding to build new machinery as it is some existing machinery. The main consideration is of course that once some new Levels become available: How do existing Datasets come in to play? The answer is not clear; but one approach is to visit each existing Dataset and update its processed status by clicking appropriate buttons in the UI. We will cross these bridges at a later date; the main point here is that the system is a scaffolding that will allow the BDS to grow and evolve.

Q: What is the BDS URL?
A: Sorry; this is redacted until it goes public, alas.

Q: How does the User Interface data viewer work?
A: You click on the top hyperlink **Explore Datasets…** to expand the data view. Here you see both your private datasets and public datasets as a series of rows. You can click on an individual dataset to expand the view of it. You can keep expanding this view into the metadata view, and then Types and Levels. A Download button will exist where there is data to download. If there is no data for a particular type then there will be a button present for you to Initialize that Type. In other words: You can go from inspecting

to expanding upon a particular Dataset. You can also Close the dataset to return to the higher-level view. Public datasets are static: They cannot be expanded in the way that your private data can be expanded.

Q: How do I Create a new Dataset using the User Interface?
A: Click 'Create new' after you have logged in and upload the three requisite files. (You will need to prepare these per the information in Part 1, above.)



Q: How do I specifically Create a PHYS Type Dataset
A: First create a Carboys table as a CSV file. Create a Samples table as a CSV file that maps to the Carboys. Create an attributes table as a CSV file that provisions the attributes you want to assign to your samples. Then return to your Samples table and for every attribute you have added create a new column that uses that attribute name. Enter values for each of your samples and save this table. This is all that is required to configure the PHYS type; and notice that the attributes are optional; you can do the bare minimum of creating a blank table and it will still create the Dataset.

Q: How do I specifically Create a MS Type Dataset?
A: You don't. Rather: You add MS Type data to an existing PHYS Type Dataset. So see above; and then the option to add MS Type to the Dataset will appear in the UI.

Q: How do I add ABS Type and EEM Type data to an existing PHYS Type Dataset?
A: See above for MS Type: Same thing. In fact you must create both the PHYS Type Dataset and then add the ABS Type data prior to being able to add EEM Type data. That is: EEMs depend upon both PHYS and ABS Types; and this dependency is built into the UI.

Q: How do I build metadata into my Dataset?
A: There are two mechanisms for including metadata. First you can add key:value pairs directly through

the User Interface (see screen capture above). These will be stored in association with your Dataset and will be subsequently searchable; so it is advisable to include them if you anticipate doing searches down the road. The second and more detailed form of metadata is associated with samples in the PHYS Type. Here (see above) you can stipulate data attributes in the attributes table and then ascribe these attribute values to the Dataset samples in the samples table. Attributes can be text fields or numerical values; for example pH or site description or latitude or altitude or depth or dissolved oxygen measurement or weather conditions. Unfortunately at this time you cannot include images as metadata unless you manage to translate them to a text format.

Q: How do I conduct a Dataset Query using the User interface?
A: Using dropdown menus and using the Query Grammar. The dropdowns allow you to stipulate where in the BDS scaffolding to look for data with three levels of granularity: Datasets, tables and rows. A Dataset query will return the IDs of one or more successful Datasets. That is: Datasets that match the high level query criteria. These results are not going to be particularly useful as such; but they can be used for finer-grained queries. The second level of granularity is tables within datasets, again identified by ID. The third and finest level of granularity is a row query; and so this provides you with rows from tables from Datasets. Herein are the actual data values that we typically find useful.

The Query grammar was developed to be analogous to how database queries work; so you can create logical and arithmetic conditions. The example we like is from FTICR-MS: 'Give me all the rows where the molecular formula includes exactly 15 carbon atoms'. This will return rows from tables from Datasets with C15 in the molecular formula. Hence you will need to know in advance that Level 1.3 Tables contain these formulas in order for your query to make sense.

Finally note that you can query from your own data or from the public data pool.

Q: How do the Provenance files work once I download them?
A: This topic (kilroy) is To Be Addressed. Requires installation (?) of MEWorkViewer I believe. But in short the provenance files open in your browser and give you a view of the workflow for a particular Dataset. They are built on an application called Modeling Environment (ME) that has come out of Moscow State University and Microsoft Research Cambridge. The nice thing about the provenance is that it automatically produces certain graphics and charts that allow you to easily look into the inner workings of the Dataset processing flow.

Q: Now that I've done all that what is the 'Death Star Essay'?
A: This is a reference to the Second Death Star which was operational despite being only partially completed structurally. BDS is operational today but not finished; and we intend to keep it operational as we build new components. The difficult challenge is to do this without making past efforts obsolete. The primary emphasis once we are past the basic nuts and bolts is to *create suggested metadata standards so* that new published datasets have value, as noted above. And we have no plans to use the BDS against the Rebel Alliance; just let me make that clear.