NCBI BioSample database: capture and storage of sample metadata

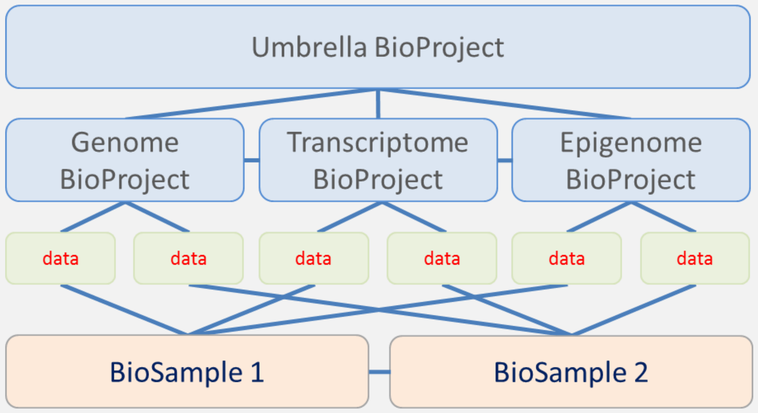
John Anderson, Ilene Mizrachi and Tanya Barrett

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD, USA

The BioSample database (http://www.ncbi.nlm.nih.gov/biosample) was developed to serve as a central location in which to store descriptive information about the biological source materials, or samples, used to generate experimental data in any of NCBI's primary data archives. Typical examples of a BioSample include a cell line, a primary tissue biopsy, an individual organism, or an environmental isolate. A centralized approach for collecting sample information has become necessary due to the expansion in data types and archival databases hosted by NCBI. Data types and databases include traditional nucleotide sequence data (GenBank, dbEST and dbGSS), next-generation sequence data (SRA), expression and epigenomic assays (GEO, Epigenomics), single nucleotide polymorphisms (dbSNP), genotypes and phenotypes (dbGaP) and genomic structural variation data (dbVar). Often, a single sample will be used in several different types of studies. Each of these study types may have a different scope or emphasis, so the sample information collected by each database is often non-uniform. It is also difficult for users to recognize when the data in different studies are derived from the same sample, yet this information could be very useful in biological discovery. In order to connect and unify the source information for these diverse data types, the BioSample database was created. The general aims of this project include:

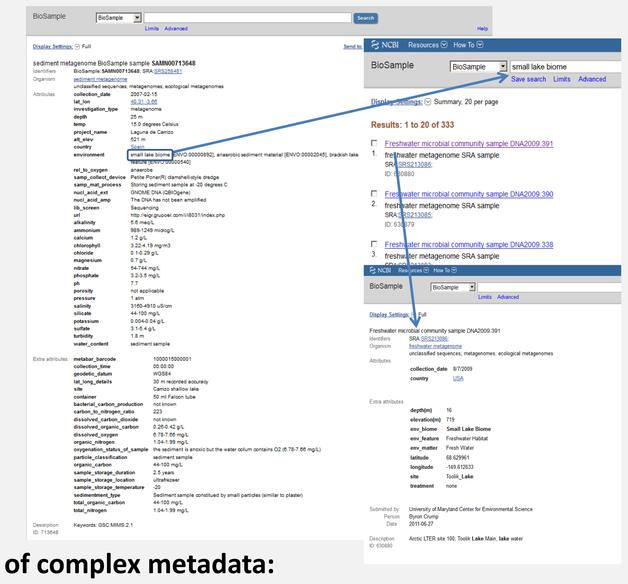
1. to provide a single point of submission for samples that may be referenced as appropriate when making data deposits to archival databases, thereby reducing submitter burden and allowing submitters to explicitly indicate when the same samples have been used across multiple studies;
2. to provide a submission portal that promotes the use of controlled vocabularies for sample attributes, thus helping to harmonize sample descriptions across NCBI databases;
3. to create a searchable resource of sample descriptions indexed in the NCBI Entrez query system;
4. to link samples to corresponding experimental data in multiple archival databases, making it possible for users to aggregate all available data derived from a given sample.

Here we demonstrate the features and user interface of the BioSample database, as well as present the new NCBI Submission Portal through which BioSamples are deposited.



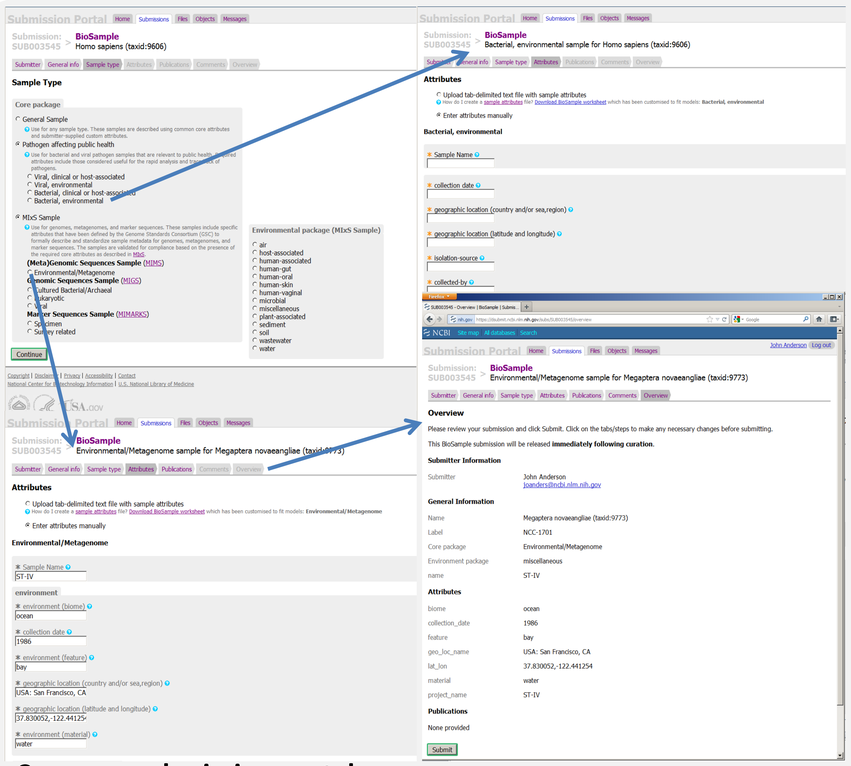
Overview of BioSample integration with other NCBI databases

Schematic depicting how BioSample records are organized and linked with other NCBI objects. This example is composed of one umbrella project that encompasses three subprojects, each of which generated data derived from two BioSample records. Users can query either the BioProject or the BioSample database to retrieve the relevant records, and then navigate through links to the corresponding experimental data which continue to be stored in NCBI's primary data archives, including GenBank, SRA, dbGaP and GEO. This schematic depicts direct links that can be applied between objects; it does not depict links to corresponding records in other NCBI databases, including PubMed, Gene, Genome and Taxonomy.



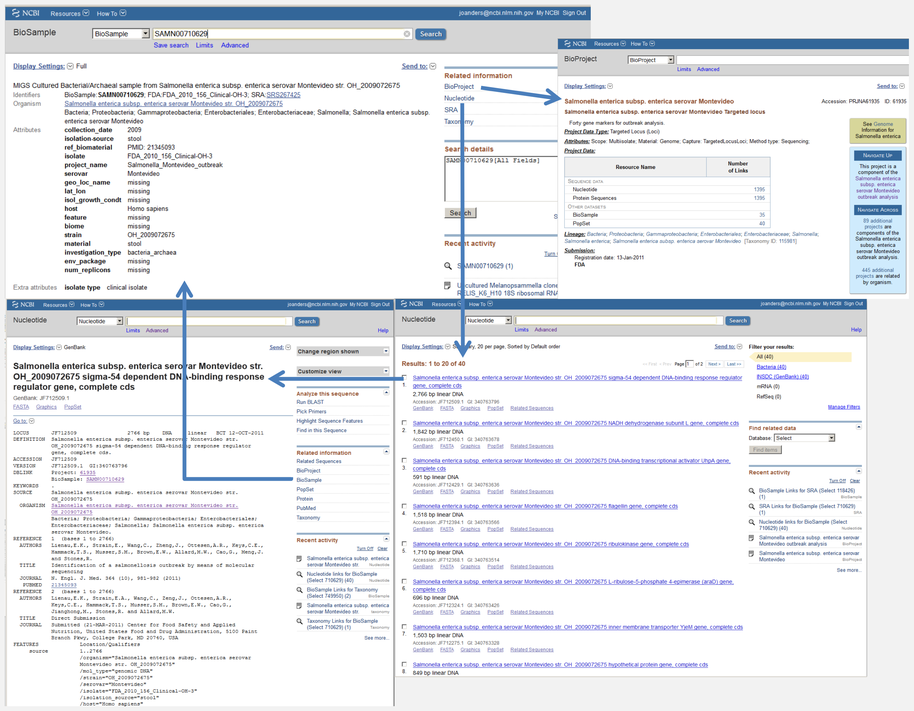
Capture of complex metadata:

A BioSample record is designed to store complete sample information in Label:Value pairs. Flexibility in the data specification allows for both controlled vocabulary and completely customizable attributes.
The attributes are indexed into searchable fields in Entrez. This allows easy clustering of samples by related characteristics.



Common submission portal

An interactive submission wizard guides submitters through the entry of their data.
Controlled vocabularies customized for each sample type allows easy entry of standardized information.
The submission portal also allows custom attributes so that any data related to the sample may be captured.



Linking of records derived from a common sample

All related information is linked through the BioSample record, including BioProjects, taxonomy and primary data records.
In this example, there are 40 GenBank nucleotide records that were derived from this sample. The individual records are linked back to the BioSample, so that it is easy to find the complete information about the source of the DNA.