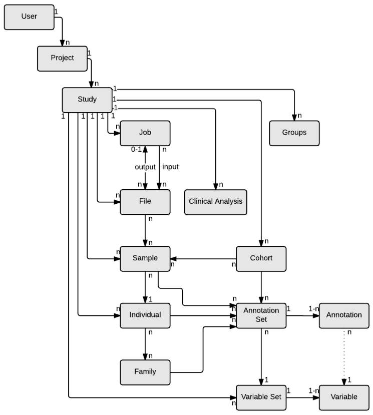
Data Management

OpenCGA Catalog allows researchers to store really diverse meta-data. Data models can be found in the right. Remember that OpenCGA *Catalog* provides an **authenticated environment** to manage data, you can create *users* and *groups*. Users will define *projects* and *studies* to organise all data and set permissions.

Table of Contents:

Remember Catalog Data Models:



Users

A *User* is generally a person who will be using OpenCGA. The idea in OpenCGA is that every single person have its corresponding user created in OpenCGA. Every user should be authenticated (see Authentication section) to be able to perform any action. However, to be able to perform any actions, users will need to be granted some specific permissions or to have a specific category within the *Study* (see Sharing and Permissions section).

There are defined two different type of users (referenced in Catalog as the user accounts):

- **full**: these users have permission to create *projects* and *studies.*
- **guest**: users that will not have the possibility to create their own *projects* and *studies.* Despite this, these users will still be able to collaborate (view, write...) in other user's studies as long as they have been granted the proper permissions.

Groups

You can create *group* of users, this will simplify data permission management. Groups are defined at *study* level, i.e. each *study* contains different groups. *Groups* can only be created by the study *owner* or the study *admins.* A *Group* of users will generally bring together users that have something in common. Groups are strongly related to permissions in Catalog (see Sharing and Permissions section). For example, let's imagine that we have 5 different departments in our institution and each department requires different permissions to the data. In that case, we could think of creating as many groups of users as different departments we have in our institution and give the specific permissions to those groups (not to the users) that have been created in OpenCGA. Doing it this way have lots of benefits:

- A user belonging to different departments (groups) will have the permissions from all the groups he/she belongs to.
- If one user leaves the department, we would just need to remove that user from the corresponding group. That user will automatically lose the permissions the group has*.
- If one user starts in the department, we would just need to add that user to the corresponding group. That user will automatically gain the permissions the group has.

* Unless the user had some or all the permissions granted to the group defined in a different group he /she might still belong to or assigned directly to the user.

All *studies* have always two administrative groups that cannot be deleted or renamed, these are *admins* and *members*.

Projects

Any ***full*** user can create any number of *projects* (and *studies).* A *Project* is *a piece of planned work or an activity that is finished over a period of time and intended to achieve a particular purpose* (Cambridge dictionary definition). A *Project* in Catalog is understood as a scientific project for one concrete species. Any project in Catalog will contain at least a name, an alias (project identifier) and the species organism. But it can also contain the organisation and a description of the project.

Projects are used as the central piece for variants storage.

Studies

Projects are composed by a set of studies. A *Study* is *the activity of examining a subject in detail in order to discover new information* (Cambridge dictionary definition) Any project owner can create as many studies as necessary. Most of the Catalog data models, except for *User* and *Project* belong to a particular *Study,* so it can be seen as the central piece in OpenCGA Catalog. A *Study* contains, similarly to *Project*, a name and an alias (study identifier). Optionally, it can have a description as well.

Groups

Despite the explanation of *Groups* from a previous section, *Groups* are actually defined within a *Study*. Different studies can have different groups of users that are basically defined by the *Study* owner or administrators. By default, every *Study* is created with two reserved groups (*admins* and *members*). The roles of these two groups is described in Sharing and Permissions section.

Variable Sets and Annotation Sets

One of the most outstanding features of OpenCGA Catalog is the ability to not only store any type of data in the database, but also filter and query by any of the values populated by the researchers. This can be achieved with what we have called *Variable Sets* and *Annotation Sets.*

A *Variable Set* is a set of *Variables*, understanding as a *Variable* the complete definition of a field that need to be populated. In other words, a *Variable Set* could be seen as a template of a form that is given to the patient containing the points the patient should fill in. A *Variable Set* will look similar to the table shown above. That *Variable Set* is composed of four well described *Variables*:

	Name	Type	Required	Allowed values
Variable	Sex	Categorical	Yes	MALE, FEMALE, UNKNOWN
Variable	Age	Integer	Yes	NA
Variable	Mother name	Text	Yes	NA
Variable	Affected	Boolean	Yes	NA

Every *Study* can have as many different *Variable Set* definitions as necessary.

The values defined for each of the *Variables* are called *Annotations*, and the population of a whole *Variable Set* is called *Annotation Set.* This means that an *Annotation Set* only makes sense and is always related to one concrete *Variable Set.*

There are four *Annotable* data models: *Sample, Individual, Family* and *Cohort*. Each entry from these data models can have *Annotation Sets* as can be seen in the diagram in the right margin. An *Annotation Set* will look to something similar to:

	Key	Value
Annotation	Sex	MALE
Annotation	Age	60
Annotation	Mother name	Jane
Annotation	Affected	Yes

OpenCGA allows querying by any of these key-value pairs.

Files

OpenCGA Catalog keeps track of all the files and folders containing the relevant data. Every *File* registry contains the physical path where the files/folders are stored in the file system (uri). Besides this, Catalog creates a virtual file structure so no matter what the real location of the files are, users can organise and work with those files differently. Everything related to *Files* can be found in the File Management section.

Individuals and Families

We understand an *Individual* as a subject (typically a person) for which some analysis will be made. A group of *Individuals* with any parental or blood relationship is called *Family*. Any of these two data models can have *Annotation Sets* defined.

Samples and Cohorts

A *Sample* is any biological material, normally extracted from an *Individual*, that is used for a particular analysis. *Cohorts* contains groups of samples sharing some particular conditions such as "healthy" vs "infected". Any of these two data models can also have *Annotation Sets* defined.

Clinical Analysis

A *Clinical Analysis* contains all the information of the *Individuals* and *Samples* involved to perform a real clinical analysis. It also allows storing the interpretations derived from the results.

Job

OpenCGA Catalog allows running different tools. This tools can be any of the ones built in OpenCGA, but also any external tool the user might need to use. Every time the user calls to a analysis webservice to run anything, a new *Job* is created. This jobs contain the essential information of the task that needs to be run. A daemon is in charge of checking whether there are any prepared, queued or running jobs and update the information.

OpenCGA supports SGE (Sun Grid Engine) that accepts, schedules, dispatches, and manages the remote and distributed execution of large numbers of standalone or parallel jobs.