The website Covid 19

Mass spectrometry Coalition

<https://covid19-msc.org/>

Working Setup

In person lab time

10.30 -Thursday

10.30 - Friday

Recurrent meetings

**Konstantinos :**

- [Zoom](https://ucl.zoom.us/j/7219642208?pwd=SjR5WVVFUGg2MU9lVitEdnRlU2ZaZz09)

**Covid-19 msc:**

- [Zoom](https://zoom.us/j/94889193537)

- [Microsoft Teams](https://eur01.safelinks.protection.outlook.com/ap/t-59584e83/?url=https%3A%2F%2Fteams.microsoft.com%2Fl%2Fmeetup-join%2F19%253ameeting_ZjFjMWI3ZTYtNDgzOC00Mzc5LWIxOTUtYjljYmFlYmNjNmEz%2540thread.v2%2F0%3Fcontext%3D%257b%2522Tid%2522%253a%2522c152cb07-614e-4abb-818a-f035cfa91a77%2522%252c%2522Oid%2522%253a%25225e2e30aa-d5aa-47a0-9abe-014dc7cc030f%2522%257d&data=05%7C01%7Ck.thalassinos%40ucl.ac.uk%7C1a417abafda34023ca7008da812005ac%7C1faf88fea9984c5b93c9210a11d9a5c2%7C0%7C0%7C637964271815901591%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=C%2BLeK%2FnucpHcqiqmPZ7h46mEwRzVOw%2ByTeJnx67CFmU%3D&reserved=0)

Drupad:

- [Microsoft Teams](https://teams.microsoft.com/l/meetup-join/19%3ameeting_NWJjYWY2ZTktNTVlMC00YTJlLWI0YTgtZjNmODlmOTU0Yzk1%40thread.v2/0?context=%7b%22Tid%22%3a%22c152cb07-614e-4abb-818a-f035cfa91a77%22%2c%22Oid%22%3a%220849e83c-1bcd-4792-8dee-20ea7a84d6a2%22%7d)

Database Access

**RedCap Manchester** [**https://www.redcap.rss.mhs.man.ac.uk/**](https://www.redcap.rss.mhs.man.ac.uk/)

User ID: **giammarco.ferrari**

password: **Cov19-MSC**

*password recovery question:* What was your childhood nickname?

*answer:* Gian

**PostgreSQL**

local database

master password: **123**

database superuser password: **123**

port number the server should listen: **5432**

database port: 5053 - why it asks me again?

Web Hosting

**tsohost**

<https://www.tsohost.com/>

Database Hosting

**free database hosting for testing**

<https://www.quora.com/Which-free-hosting-allows-external-access-to-MySQL-database-If-there-is-no-such-good-free-hosting-please-recommend-me-a-good-but-cheap-hosting-that-can-do-that>

<https://www.000webhost.com/>

<https://profreehost.com/features/>

<https://bit.io/pricing>

User ID: Covid19MSC

psw: https://somee.com/

<https://somee.com/FreeAspNetHosting.aspx>

User ID: giammafer

psw: Giam\_mafer75

<https://somee.com/doka/DoHelpTopics.aspx?docode=false&thnid=72>

<https://somee.com/doka/DoHelpTopics.aspx?docode=false&thnid=52>

Apparently no external access

[giammafer.retool.com](https://www.freesqldatabase.com/)

**optional**

<https://www.youtube.com/watch?v=TMGHOW8Hzvw>

**bit.io**

<https://docs.bit.io/docs>

free database for testing

The Idea

Take samples with different range of severity.

Collecting data:

proteomics

metabolomics

lipodomics

We must make the repository with all data and link the data with the metadata.

Create a dashboard where is possible to collect again the raw data and reprocess everything.

Work package 3.

Flask for the database in the backend.

We want to connect pride - metabolomic

From <https://covid19-msc.org/action-plan/> I considered the point 5:

5.      Open data repositories

Critical to our collective effort is how we manage the data. We suggest that for omics data you use a data repository (ideally PRIDE for proteomics datasets, as part of ProteomeXchange), and that you tell us the accession number. PRIDE enables the inclusion of the submitted datasets in the EMBL-EBI COVID-19 data portal (<https://www.covid19dataportal.org/>). After deposition we will record these along with the requisite metadata to identify the datasets. We will soon be updating the website with a dedicated page where you can submit this information. The website will also be updated regularly so that many researchers can access the processed data. It is through this that we will gain the most and most quickly.  For native data on recombinant viral antigens get in touch as we will make the data repository and we will curate that data for all to examine and to compare to data from patient samples.

What the best way to connect these data (metadata are necessary)

The data collection activity could be performed in several sites (at the moment Surrey).

The purpose of the database will be to collect metadata from the affiliates and perform reprocessing and comparative analysis.

The main sources of comparison will be samples coming from [PHOSP](https://www.phosp.org/people/) ad [ISARIC](https://isaric4c.net/virtual_site_visit/).

PHOSP is a study patients that survived Covid and followed up.

ISARIC is without follow-up.

The current project that we refer for the definition of metadata is Frimley.

Storyboard

The website should manage two main operation:

* Upload record
* Download record

**Upload a new record**

A researcher just completed the collection of his data.

The researcher should submit his data on one of the repository specific for each omics data type:

|  |  |
| --- | --- |
| **Omics Data Type** | **Repository** |
| Proteomics | PRIDE |
| Metabolomics | MetaboLights |
| Lipodomics | Zenodo |

The **accession code** to these data,whatever is the repository where they have been uploaded, will be part of the patients record uploaded on the Covid19-msc website.

The user should knows if its experimental group of patient has already generated other omics data from other researches. This information is essential for create new group codes or not.

* *YES he knows*
  + In this case the researcher must know the Covid19-msc's accession code for the specific group of patients already present in the Covid19-msc database.
  + **He must use this code to fetch the specific record.**
  + Finally the record must be updated with the accession code related to the specific repository where the omics data have been uploaded.
* *NO he does not know*

In this case the website must generate a new code for a new entry

Download a record

A researcher wants to access the data for .........

Therefore she/he goes to the website and start the search on the search line

It could also refine the search providing additional information like metadata

For which

META-RELATIONSHIP

V1



Input

Raw Data **Proteomics**



Access online database

**PRIDE**



Input

Raw Data **Metabolomics**



Access online database

**MetaboLights**

**Research Subjects**



Input

Raw Data **Lipodomics**



Access online database

**Zenodo**

Input Metadata

Accession Code **Z**

Accession Code **M**

Input Metadata

Input Metadata

**Input Record**

**Cov19-MSC Database**



Accession Code **P**

**Cov19-MSC Website**

V2



Input

Raw Data **Lipodomics**



Access online database

**Zenodo**



Input

Raw Data **Metabolomics**



Access online database

**MetaboLights**



Input

Raw Data **Proteomics**



Access online database

**PRIDE**

Accession Code **Z**

Accession Code **M**

Accession Code **P**

Input Metadata

Input Metadata

Input Metadata

**Subjects**

**Group**



**Cov19-MSC Database**

**Input Record**

**Cov19-MSC Website**

**Subjects Group Descriptor**

**Projects**

V3



Input

Raw Data **Proteomics**



Access online database

**PRIDE**

**Subjects**

**Group**

Input Metadata

Input Metadata

Input Metadata



Input

Raw Data **Metabolomics**

Access online database

**MetaboLights**



Input

Raw Data **Lipodomics**



Access online database

**Zenodo**

**INPUT**

**SEARCHING**



**Cov19-MSC Database**

**Input Record**

**Cov19-MSC Website**

**Searched Record**

METADATA

**Relationships of Cov19-MSC main identifiers**

The metada should describe the raw file. Then the link should point to a repository. The link must be stored in the single metadata raw. The link represented by a subject to its omics data (the raw data for each omics analysis) must be solved embed by the database schema and represent the subject point of view.

**Subject**

***[omics]*.raw**

However, in this project there are ‘sites’ that generate the data. Because it is considered that the cohorts of subjects are collected from the sites.

These sites could be Surrey or Manchester (I call them ‘physical sites’) or repositories like PHOSP, ISARIC Frimley (so far is like this but in the future it could change).

**Subject**

***[omics]*.raw**

**Site**

**Raw Data**

**Sample Processing**

**Sample Delivery**

Lipodomics

Raw Data

Metabolomics

Raw Data

Proteomics

Raw Data

08122022\_S222\_something\_**proteo**\_serum.raw

**Metabolomics**

**Lipodomics**

**Proteomics**

08122022\_S222\_something\_**metabo**\_serum.raw

Manchester

08122022\_S222\_something\_**lipo**\_serum.raw

Surrey

**S\_222**

S\_222\_serum\_madeupusername\_02112022\_**01**.raw

**Metabolomics**

**Lipodomics**

**Proteomics**

S\_222\_serum\_madeupusername\_02112022\_**02**.raw

Surrey

S\_222\_serum\_madeupusername\_02112022\_**03**.raw

Which representation for REDCap?

At the moment **REDCap** is the main source of metadata eligible for the Cov19-MSC database. It should include all the metadata models that has been suggested or imported from other data sources. This set of metadata must be refined in the future based on the requirements that will be set by the consortium.

**REDCap**

**Frimley**

PHOSP

ISARIC

Surrey

Manchester

*We may have same/similar column names but the way data is encoded may not be the same. We do that before it gets uploaded/recorded on Redcap. Then that way we have one version of variable names and encoded input data no matter where the sample was collected. (Drupad mail 04/11/2022 - 16:06)*

Intersection

REDCap

ISARIC

Frimley

Frimley

ISARIC

**REDCap**

PHOSP

PHOSP

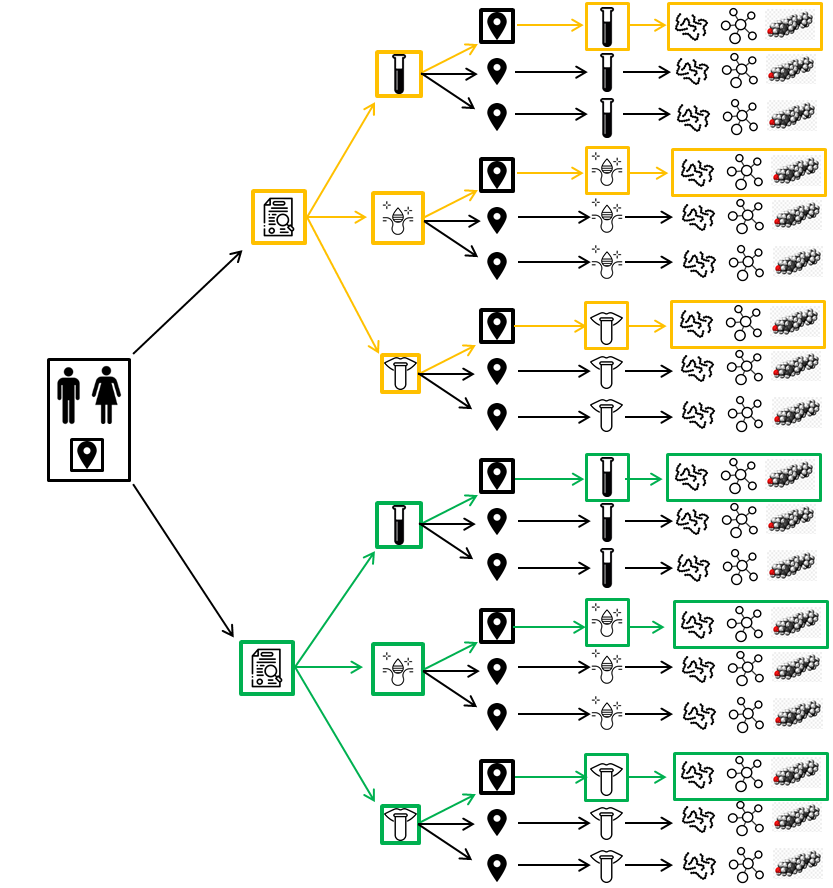
*For now, Redcap will try to capture as much info as we have available from all 3 sampling sites. So it may not be an intersection, but eventually, from a data analysis point of view, we may decide to keep only variables that are meaningful and present in all the data. (Drupad mail 07/11/2022 - 10:46)*

However, the requirement to generate a data catalog is related to the concept of having a link to the specific group of subjects with the desired chacteristics.

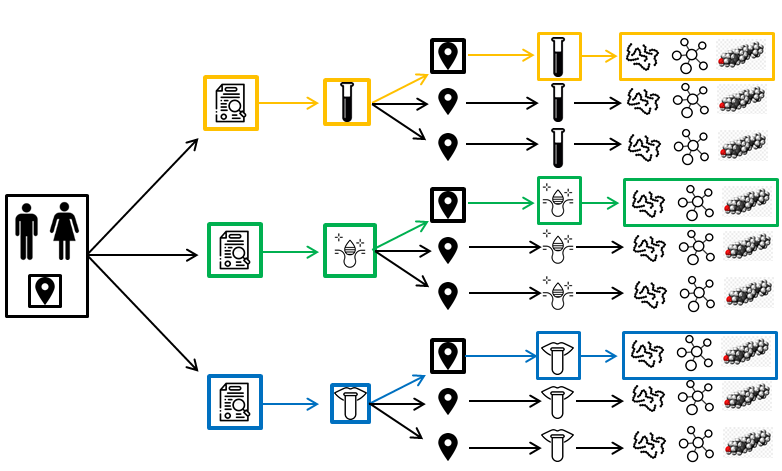
As a consequence, the ‘site’ cannot represent a logical link to these characteristics. Instead, the published ‘study’ where the specific group of subjects/cohort has been investigated could be a meaningful identifier. The criteria used to enrole any subject into the study can justify the subjects group features.

Now the question is on the multiomics point of view.

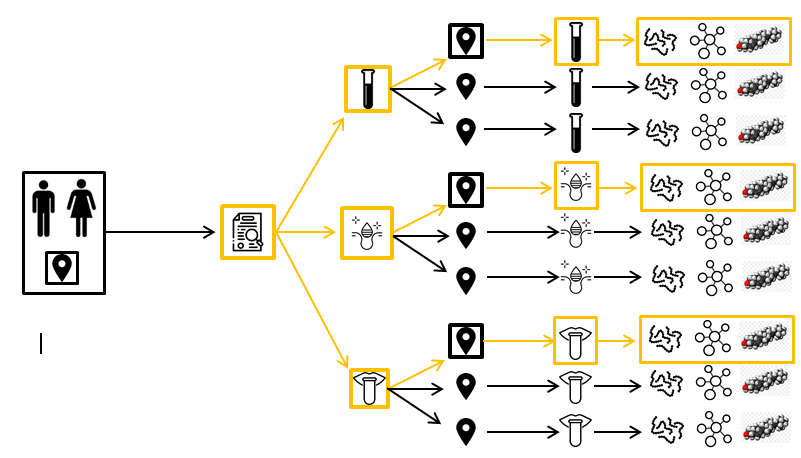
Is it possible that a study investigates one or more omics aspects and another study investigates other omics aspects on the same group of subjects?



Where this case it will be the most likely one:



Or the multiomics scenario on the same group of subjects is always related to a single study?



MAGE-TAB-Proteomics

A valuable reference about the metadata definition is the [MAGE-TAB-Proteomics](https://pubmed.ncbi.nlm.nih.gov/34615866/) that is a project to adapt the MAGE-TAB for transcriptomic for proteomics data management.

<https://github.com/bigbio/proteomics-metadata-standard>

<https://www.youtube.com/watch?v=Gd87sPWSH8U>

*By repurposing and extending the MAGE-TAB for Proteomics, we aim to provide a format for future submissions of multiomics experiments to ProteomeXchange partners and better integration with other omics data.*

MAGE-TAB-Proteomics - Main Descriptors

***Proteomics Experiment properties*** - [organism, organism part, and biological replicate]

***Data File*** - [fractionidentifier, technical replicate accession, label (in the case of labeling methods), and data-file name]

As descriptors, they are made combined different features but they are not entity in the database.

The focus of the MAGE-TAB-Proteomics is on the phenotype expressed by the sample . Indeed, the **IDF** (Investigation Description Format) file format that contains information describing the **study** (authors/submitters, protocols, and publications). While the **SDRF** (Sample and Data Relationship Format) allows the **link of a sample to all its raw files**. Raw files comes from technical replicates over the same sample. The point of view is on the phenotype that is trated as factor. A factor/phenotype is defined by the collection of all the samples that refer to it.

Instead, Cov19-MSC is focused on the subject (investigated through different types of samples). The concept of phenotype it is not appropriate to describe a sample (while the sample type is a good identifier). Finally the dease is not a valid attribute because the desease is always COVID-19. Maybe the stage of the desease is a good descriptor.

Descriptors

Which descriptors for the Cov19-MSC???

So far I have a look of the REDCap columns and I grouped them with my own semantic.

ID relationship

*We have Surrey (also known as Frimley) sample called S\_222. Surrey has metadata where the sample is called S\_222. Surrey performed metabolomics and call their file name S\_222\_serum\_madeupusername\_02112022\_01.raw   
  
Now, Manchester has S\_222 and they do metabolomics. They call their file 08122022\_S222\_something\_metabo\_serum.raw*

|  |  |
| --- | --- |
| **REDCap** Identifier | **Site**  Identifier |
| R001 | S\_222 |
| R002 | S\_223 |
| R003 | PHOSP\_394 |
| R004 | PHOSP\_395 |
| R005 | I\_0239 |
| R006 | I\_0240 |

S\_222

R001

PHOSP\_394

 I\_0239

**REDCap** Identifier

**Surrey**

Identifier

**PHOSP** Identifier

**ISARIC** Identifier

|  |  |  |  |
| --- | --- | --- | --- |
| **REDCap** **Identifier** | **Site name** | **Site Identifier** | **SampleType**  ***Saliva*** |
| R001 | Surrey | S\_222 | type\_3 |
| R002 | Surrey | S\_223 | type\_3 |
| R003 |  | PHOSP\_394 | type\_3 |
| R004 |  | PHOSP\_395 | type\_3 |
| R005 |  | I\_0239 | type\_3 |
| R006 |  | I\_0240 | type\_3 |
| R007 | Manchester | S\_222 | type\_3 |
| R008 | Manchester | S\_223 | type\_3 |

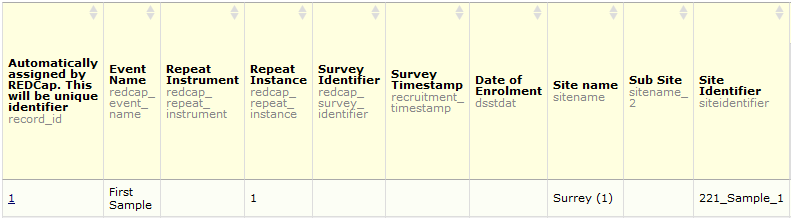
|  |
| --- |
| **Sample Reference** |
| S\_222 |
| S\_223 |

|  |  |  |  |
| --- | --- | --- | --- |
| **REDCap** **Identifier** | **Site name** | **Site Identifier** | **SampleType**  ***Saliva*** |
| R001 | Surrey | S\_222 | type\_3 |
| R002 | Surrey | S\_223 | type\_3 |
| R003 |  | PHOSP\_394 | type\_3 |
| R004 |  | PHOSP\_395 | type\_3 |
| R005 |  | I\_0239 | type\_3 |
| R006 |  | I\_0240 | type\_3 |
| R007 | Manchester | S\_222 | type\_3 |
| R008 | Manchester | S\_223 | type\_3 |

Same Cohort Different Aliquots

223

222

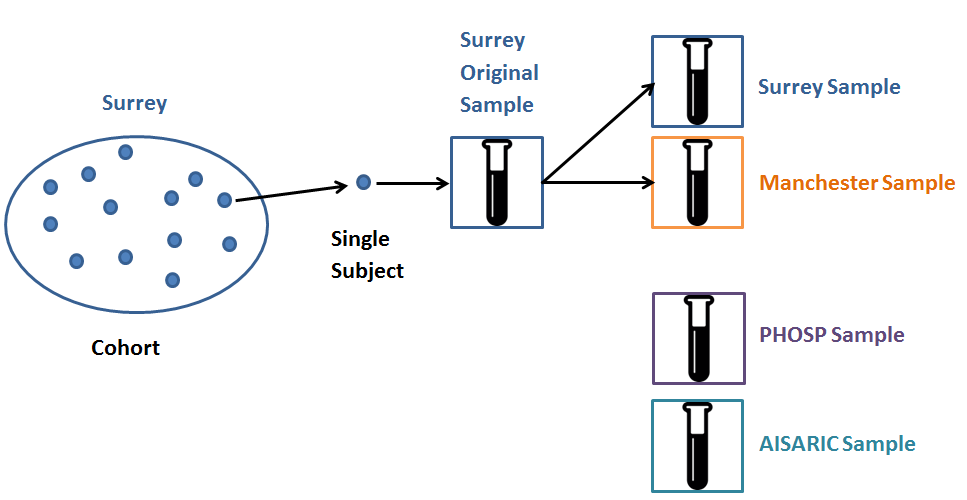


.

.

.





**Collecting all samples in REDCap**

|  |  |  |  |
| --- | --- | --- | --- |
| **REDCap** **Identifier** | **Site name** | **Site Identifier** | **Omics** |
| R001 | Surrey | S\_222 | metabolomics |
| R002 | Surrey | S\_223 | metabolomics |
| R003 |  | PHOSP\_394 | metabolomics |
| R004 |  | PHOSP\_395 | metabolomics |
| R005 |  | I\_0239 | metabolomics |
| R006 |  | I\_0240 | metabolomics |
| R007 | Manchester | S\_222 | metabolomics |
| R008 | Manchester | S\_223 | metabolomics |

**Same Cohort**

**Dfifferent Aliquots**

**Same Subject**

**Dfifferent Aliquots**

|  |  |  |  |
| --- | --- | --- | --- |
| **REDCap** **Identifier** | **Site name** | **Site Identifier** | **Omics** |
| R003 | PHOSP | PHOSP\_394 | metabolomics |
| R004 | PHOSP | PHOSP\_395 | metabolomics |
| R005 | ISARIC | I\_0239 | metabolomics |
| R006 | ISARIC | I\_0240 | metabolomics |

***One correction is - we will never have site name of the place where sample is analysed. Each analytical lab is doing specific analysis so ‘type of analysis’ column is enough to tell us who analysed that sample. Site name will only have Surrey/PHOSP/ISARIC. When the project started, we were going to have another set of samples from Manchester, but that hasn’t happened.***

**DO NOT CONSIDER MANCHESTER ANYMORE**

The table below try to answer two questions:

* It is possible to find all the rows that belong to the same subject?
* It is possible to understand for each single row from which site they come?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Covid-MSC Subject Identifier** | **REDCap Record Identifier** | **Event Name** | **Site name** | **Site identifier** | **Sample Type** | **Analysis Type to be undertaken** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | serum | **metabolomics** |
| C19-001 | 2 | First Sample | Surrey | 001\_Sample\_1 | serum | **proteomics** |
| C19-001 | 3 | First Sample | Surrey | 001\_Sample\_1 | serum | **lipodomics** |
| C19-001 | 4 | First Sample | Surrey | 001\_Sample\_1 | sebum | **metabolomics** |
| C19-001 | 5 | First Sample | Surrey | 001\_Sample\_1 | sebum | **proteomics** |
| C19-001 | 6 | First Sample | Surrey | 001\_Sample\_1 | sebum | **lipodomics** |
| C19-001 | 7 | First Sample | Surrey | 001\_Sample\_1 | saliva | **metabolomics** |
| C19-001 | 8 | First Sample | Surrey | 001\_Sample\_1 | saliva | **proteomics** |
| C19-001 | 9 | First Sample | Surrey | 001\_Sample\_1 | saliva | **lipodomics** |
| C19-001 | 10 | Second Sample | Surrey | 001\_Sample\_2 | serum | **metabolomics** |
| C19-001 | 11 | Second Sample | Surrey | 001\_Sample\_2 | serum | **proteomics** |
| C19-001 | 12 | Second Sample | Surrey | 001\_Sample\_2 | serum | **lipodomics** |
| C19-001 | 13 | Second Sample | Surrey | 001\_Sample\_2 | sebum | **metabolomics** |
| C19-001 | 14 | Second Sample | Surrey | 001\_Sample\_2 | sebum | **proteomics** |
| C19-001 | 15 | Second Sample | Surrey | 001\_Sample\_2 | sebum | **lipodomics** |
| C19-001 | 16 | Second Sample | Surrey | 001\_Sample\_2 | saliva | **metabolomics** |
| C19-001 | 17 | Second Sample | Surrey | 001\_Sample\_2 | saliva | **proteomics** |
| C19-001 | 18 | Second Sample | Surrey | 001\_Sample\_2 | saliva | **lipodomics** |

**This dummy database rows try to catch the case of a single subject that has been sampled and resampled in Surrey. Moreover, for both samples have been taken all the three types of samples allowed in REDCap and for each of them have been performed all the analysis types. But now only for Surrey site.**

**Why there are redundant data?**

Because I used the REDCap data present at the 10/11/2022 and they could change in the future.

**Why I would like to add the column 'Covid-MSC Subject Identifier':**

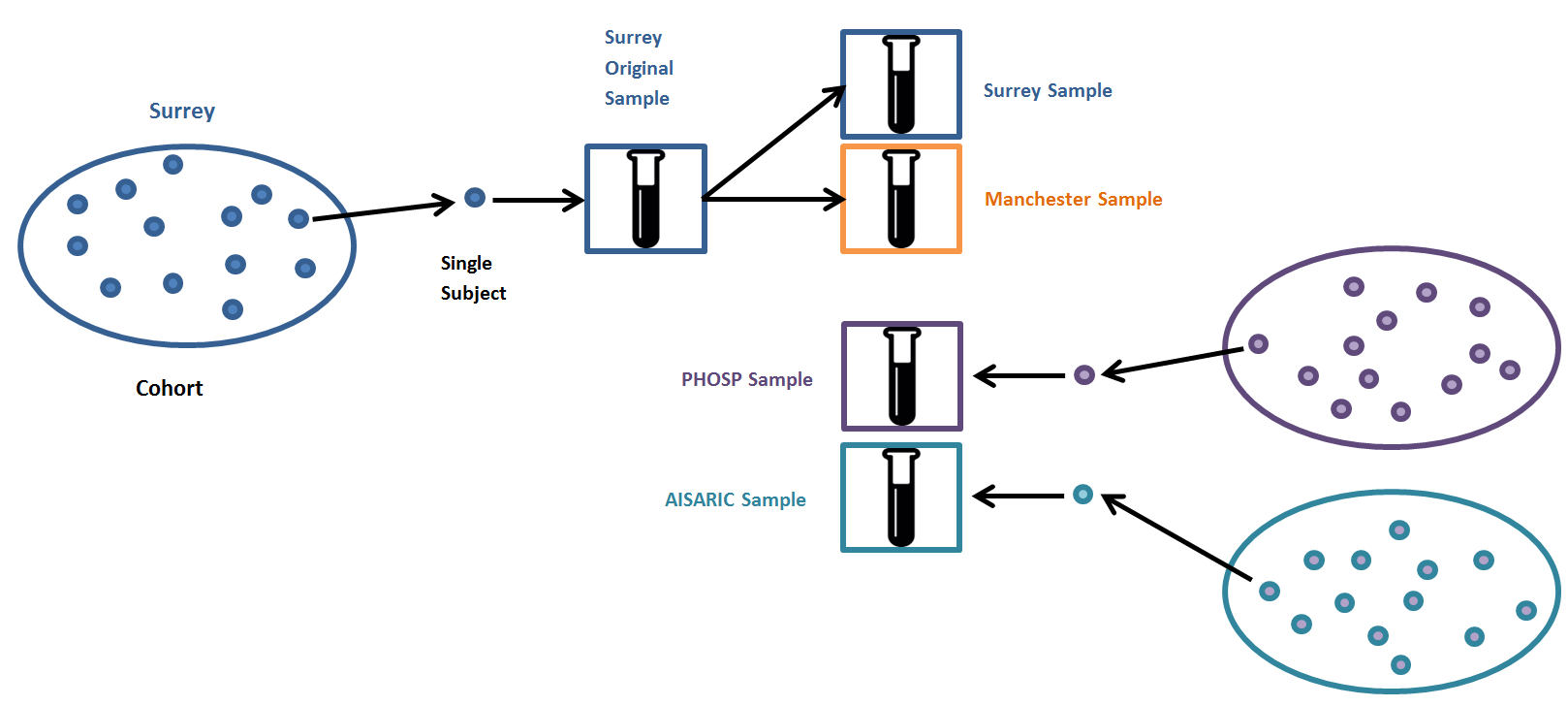
Because the subject is the entity from which are generated all the samples in each omics analyses. I need an ID that allow to collect all the metadata that refer to a specific subject.

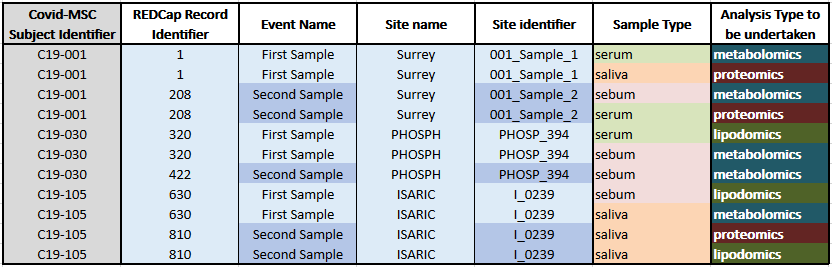
Looking at the data on REDCap, it seems that this identifier is already present in the column **'Site Identifier**'.

However, in this column is present also the value of the 'Event Name' column. For instance, the rows for the Site Identifier = **001\_Sample\_1**, it reports data for the subject **001\_** and the **Sample\_1** took from this subject.

By the way, we have several options to generate a subject ID that span from generating an ID from the scratch, to pre-process data in REDCap in order to obtain a unique identifier that collects all the samples refferred to a specific subject.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Covid-MSC Subject Identifier** | **REDCap Record Identifier** | **Event Name** | **Site name** | **Site identifier** | **Sample Type** | **Analysis Type to be undertaken** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | serum | **metabolomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | serum | **proteomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | serum | **lipodomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | sebum | **metabolomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | sebum | **proteomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | sebum | **lipodomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | saliva | **metabolomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | saliva | **proteomics** |
| C19-001 | 1 | First Sample | Surrey | 001\_Sample\_1 | saliva | **lipodomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | serum | **metabolomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | serum | **proteomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | serum | **lipodomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | sebum | **metabolomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | sebum | **proteomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | sebum | **lipodomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | saliva | **metabolomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | saliva | **proteomics** |
| C19-001 | 208 | Second Sample | Surrey | 001\_Sample\_2 | saliva | **lipodomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | serum | **metabolomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | serum | **proteomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | serum | **lipodomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | sebum | **metabolomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | sebum | **proteomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | sebum | **lipodomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | saliva | **metabolomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | saliva | **proteomics** |
| C19-030 | 320 | First Sample | PHOSPH | PHOSP\_394 | saliva | **lipodomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | serum | **metabolomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | serum | **proteomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | serum | **lipodomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | sebum | **metabolomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | sebum | **proteomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | sebum | **lipodomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | saliva | **metabolomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | saliva | **proteomics** |
| C19-030 | 422 | Second Sample | PHOSPH | PHOSP\_394 | saliva | **lipodomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | serum | **metabolomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | serum | **proteomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | serum | **lipodomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | sebum | **metabolomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | sebum | **proteomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | sebum | **lipodomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | saliva | **metabolomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | saliva | **proteomics** |
| C19-105 | 630 | First Sample | ISARIC | I\_0239 | saliva | **lipodomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | serum | **metabolomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | serum | **proteomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | serum | **lipodomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | sebum | **metabolomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | sebum | **proteomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | sebum | **lipodomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | saliva | **metabolomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | saliva | **proteomics** |
| C19-105 | 810 | Second Sample | ISARIC | I\_0239 | saliva | **lipodomics** |





**Collecting all samples in REDCap**

g@

INPUT - OTPUT

single omics data type

<https://www.breezetree.com/articles/excel-flowchart-shapes/>

**LOADING DATA v1**



EMBL-EBI COVID-19 Data Portal



Access online databases

PRIDE



Input

Raw Data

Accession Code

Input Metadata

**Cov19-MSC Database**



**Input Record**

**Cov19-MSC Website**

**LOADING DATA v2**

**Create Input Record**

**Generate Accession Code**

**START**

**Search for Existing Code**

**New Entry?**

**yes**

**no**

**Cov19-msc**

**DB**

**STOP**

**Update Exsisting Record**

**START**

**Does the data match the database schema?**

**no**

**Cov19-msc**

**DB**

**Create Input Records**

**yes**

**STOP**

**Processing Data**

**no**

**Input data provide need new static data?**

**yes**

**Update Static Data**

**DOWNLOADING DATA v1**

Searching



Access online databases

PRIDE

**Cov19-MSC Database**



Metadata

**Accession Code**

Accession Code

**Cov19-MSC Website**



Output

Raw Data

**DOWNLOADING DATA v2**

**Cov19-msc**

**DB**

**Input Search Criteria**

**START**

**Return list of**

**file links**

**STOP**

**Cov19-MSC Metadata**

**STOP**

**PRIDE : Accession Code**

**MetaboLight : Accession Code**

**Zenodo : Accession Code**

**Input Search Criteria**

**Cov19-msc**

**DB**

**START**

Visualization Record

The most important fields to visualise in the table

meeting 18/10/2022

The subject group point of view is confirmed. In term of searching operation, what will be more relevant between subject group and Project ?

*Sample Delivery* - from which research center the data are coming

*Sample Processing* -

Proteomics

Metabolomics

Lipodomics

SWATH proteomics

*Data Processing* - Which software or method has been used

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group ID | Project Title | Researcher Name | Institution | Description | SampleProcessing | Sample Delivery | Protocol | Data Processing | Repositories | Links |

accession code

project title

researcher

sample delivery

description

sample processing

protocol

data processing

repository name

**Project**

group ID

group size

group description

**Subject Group**

0..3

1..\*

researcher name

researcher institution

**Researchers**

1..\*

Institution name

**Sample Delivery**

1..\*

type

**Sample Processing**

1..\*

type

**Protocol**

1..\*

type

**Data Processing**

1..\*

repos accession code

repository name

accession code

**Repositories**

1..\*

type

**Data Processing**

1..\*

type

**Protocol**

1..\*

type

**Sample Processing**

1..\*

Institution name

**Sample Delivery**

1..\*

1..\*

researcher name

researcher institution

**Researchers**

1..\*

repos accession code

repository name

accession code

**Repositories**

1..\*

project\_ID

project title

researcher

sample delivery

description

sample processing

protocol

data processing

repository name

**Project**

group ID

group size

group description

**Subject Group**

21/10/2022

Drupad gave his suggestion

**Project Descriptor**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Project Title | Researcher Name | Institution | Description | SampleProcessing | Sample Delivery | Protocol | Data Processing | Repositories | Links |
| Group ID |  | | | | | | | | | |
| Size |
| Age  **Group Descriptor** |
| Gender |
| oxygen |
| drugs |
| survival rate |

group ID

group size

age

institution

survival rate

drugs

oxygen

**Subject Group**

group ID

group size

**Subject Group**

gender type

survival rate

**Gender**

age min

age max

mean age

**Age**

**Oxygen**

**Drug**

28/11/2022

It is worth to say that the database will be query by the users while the DB manager will be in charge of the input data operations. For these reasons, it is possible to:

**Reduce the number of simple tables**

This means that some of the tables created so far could be removed from the database. The conditions to do so are:

* The table contains ID + identifier.
* The identifiers list is short (3 or 5 elements).
* It is rarely need to be updated.

The solution could be to hardcode options menu in the HTML or a vector used later on in the page.

However, most of the foreign keys to these tables will be represented by indexes. Because it is supposed to be just a few options for each table, the '**double**' data type it looks like the best option to implement these indexes.

Sample\_Types

Analysis\_Types

Events

Must became vectors in models.py

Fisical schema tool

<https://dbdiffo.com/dbdiffo.php>

Page Layouts - Omics

Navigation Bar

Data Base Rows / Search Results

Footer

Logo

**Search**

SEARCH RESULT

Navigation Bar

Footer

Logo

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Project Title | Researcher Name | Description | SampleProcessing | Sample Delivery | Protocol | Data Processing |
| *#MSC001* |  |  |  | P - M - L |  |  |  |
| *#MSC002* |  |  |  | P - M - L |  |  |  |
| *#MSC003* |  |  |  | P - L |  |  |  |
| *#MSC004* |  |  |  | P |  |  |  |
| *#MSC005* |  |  |  | M - L |  |  |  |
| *#MSC006* |  |  |  | P - M - L |  |  |  |
| *#MSC007* |  |  |  | L |  |  |  |
| *#MSC008* |  |  |  | M |  |  |  |
| *#MSC009* |  |  |  | P - M - L |  |  |  |

RECORD DETAILS

- Sample Processing Level

Navigation Bar

|  |  |  |
| --- | --- | --- |
| **Group ID** #MSC001 | | |
|  | | |
| *Proteomics* | Publication 1 | PRIDE |
| *Metabolomics* | Publication 2 | MetaboLight |
| *Lipodomics* | Publication 3 | Zenodo |

Footer

Logo

RECORD DETAILS

- Sample Processing Level

- Raw Data Links Level

Navigation Bar

|  |  |  |
| --- | --- | --- |
| **Pride Accession Code** #PXD036103 | | |
| *File Type 1* | link 1 | file size 1 |
| *File Type 2* | link 2 | file size 2 |
| *File Type 3* | link 3 | file size 3 |
| *File Type 4* | link 4 | file size 4 |
| *File Type 5* | link 5 | file size 5 |

Footer

Logo

**Search**

Page Layouts - Group Descriptor

Navigation Bar

Searching Form

Footer

Logo

SEARCH RESULT

Navigation Bar

Footer

Logo

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Group ID* | Group Size | Female | Male | Survival Rate | Oxygen | Drug | Sample Delivery | Sample Processing |
| *#MSC001* | 1234 | 678 | 123 | 90% |  |  | University of Oxford | P - M - L |
| *#MSC002* | 2355 | 135 | 234 | 95% |  |  | University College Hospital | P - M - L |
| *#MSC003* | 2956 | 754 | 234 | 91% |  |  | University of Cambria | P - L |
| *#MSC004* | 1946 | 123 | 658 | 90.5% |  |  | Charing Cross Hospital | P |
| *#MSC005* | 1612 | 345 | 446 | 92.7% |  |  | University of Edinburgh | M - L |
| *#MSC006* | 367 | 244 | 123 | 93% |  |  | St Thomas Hospital | SP - P - M - L |
| *#MSC007* | 1532 | 354 | 243 | 96% |  |  | University of Cambridge | L |
| *#MSC008* | 3528 | 234 | 768 | 91.2% |  |  | Queen Mary Hospital | M |
| *#MSC009* | 1946 | 123 | 658 | 90.5% |  |  | University of Sussex | P |
| *#MSC010* | 3528 | 234 | 768 | 91.2% |  |  | Queen Charlottes and Chelsea Hospital | M |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

RECORD DETAILS

- Sample Processing Level

Navigation Bar

Footer

Logo

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group ID** #MSC006 | | | | | | | | |
| Group Size | | Survival Rate | | Oxygen | Drug |  | | |
| 367 | | 93% | |  |  |
| Female | Male | Female | Male |  |  |
| 244 | 123 | 60% | 40% |  |  |
| Sample Processing | | Sample Delivery | | Data Processing | Protocol | Raw data files | Project Title | Researcher |
| Proteomics | | University of Sussex | | software 1 |  | 5 |  |  |
| SWATH Proteomics | | University of Oxford | | software 2 |  | 8 |  |  |
| Metabolomics | | University of Sunderland | | software 3 |  | 6 |  |  |
| Lipodomics | | University of Suffolk | | software 4 |  | 9 |  |  |

Navigation Bar

Footer

Logo

|  |  |  |
| --- | --- | --- |
| **Group ID** #MSC006 | | |
| **PRIDE Accession Code** | #PXD036103 | |
| **Sample Processing** | Proteomics | |
| **Raw Data Files** | 5 | |
| File name 1 | 1.5 Gb | link 1 |
| File name 2 | 1 Gb | link 2 |
| File name 3 | 1.3Gb | link 3 |
| File name 4 | 1 Gb | link 4 |
| File name 5 | 1.2 Gb | link 5 |

Database

Oracle Database Instance

<https://docs.oracle.com/cd/E11882_01/server.112/e40540/startup.htm#CNCPT005>

[AWS Postgres inserts with SQLALchemy more than 1000 times slower than local SQLite](https://stackoverflow.com/questions/56353333/aws-postgres-inserts-with-sqlalchemy-more-than-1000-times-slower-than-local-sqli)

<https://stackoverflow.com/questions/56353333/aws-postgres-inserts-with-sqlalchemy-more-than-1000-times-slower-than-local-sqli>

[varchar(n) size?](https://dba.stackexchange.com/questions/271722/varcharn-size)

<https://dba.stackexchange.com/questions/271722/varcharn-size>

PostgreSQL

8.1. Numeric Types

<https://www.postgresql.org/docs/current/datatype-numeric.html#DATATYPE-INT>

Metadata

Metada are generated in order to link different types of project datafiles. Maybe!!! However this is not enough!!!

FAIR

Metadata must follow FAIR <https://www.go-fair.org/fair-principles/f2-data-described-rich-metadata/>

PRIDE

In PRIDE there is a standard for metadata, the Proteomics Sample Metadata Format.

*The Proteomics Sample Metadata Project aims to standardize the way ProteomeXchange partners and the proteomics community capture the relation between the* samples *and the* data *generated within a PX submission.*

The standard derives from the [MAGE-TAB](https://www.fged.org/projects/mage-tab/) extended and repurposed for Proteomics data.

<https://github.com/bigbio/proteomics-metadata-standard>

<https://www.fged.org/projects/mage-tab/>

API

PRIDE API

<https://www.ebi.ac.uk/pride/ws/archive/v2/swagger-ui.html#/>

Flask API

<https://anderfernandez.com/en/blog/how-to-create-api-python/>

Python API

<https://realpython.com/api-integration-in-python/#rest-architecture>

**ENDPOINTS - DOWNLOADING DATA**

Apparently the most important endpoint for query PRIDE are:

[**/projects/{accession}**](https://www.ebi.ac.uk/pride/ws/archive/v2/swagger-ui.html#!/projects/getProject)

<https://www.ebi.ac.uk/pride/ws/archive/v2/swagger-ui.html#!/projects/getProject>

[**/projects/{accession}/files**](https://www.ebi.ac.uk/pride/ws/archive/v2/swagger-ui.html#!/projects/getFilesByProject)

<https://www.ebi.ac.uk/pride/ws/archive/v2/swagger-ui.html#!/projects/getFilesByProject>

Web Development

Bootstrap

How to Edit, Customize, and Override Bootstrap CSS to Suit Your Brand

<https://blog.hubspot.com/website/how-to-override-bootstrap-css>

Adding CSS styling to your website

https://pythonhow.com/python-tutorial/flask/Adding-CSS-styling-to-your-website/

SVG

SVG Tutorial: How to Code SVG Icons by Hand

<https://www.aleksandrhovhannisyan.com/blog/svg-tutorial-how-to-code-svg-icons-by-hand/>

Retool

<https://retool.com/>

SQLalchemy

DB supported: SQLite, Postgresql, MySQL, Oracle, MS-SQL, Firebird, Sybase and others, most of which support multiple DBAPIs

Postgres, SQLAlchemy, and Alembic

<https://www.learndatasci.com/tutorials/using-databases-python-postgres-sqlalchemy-and-alembic/>

Using PostgreSQL through SQLAlchemy

<https://www.compose.com/articles/using-postgresql-through-sqlalchemy/>

Data Management With Python, SQLite, and SQLAlchemy

https://realpython.com/python-sqlite-sqlalchemy/#working-with-sqlalchemy-and-python-objects

SQLite (for database connection)

https://docs.sqlalchemy.org/en/14/dialects/sqlite.html

SQLAlchemy data types

Chapter 4. SQLAlchemy Type Engines

<https://www.oreilly.com/library/view/essential-sqlalchemy/9780596516147/ch04.html>

PostgreSQL Data Types and Custom SQL Constructs

<https://docs.sqlalchemy.org/en/14/dialects/postgresql.html>

[How to create integer of a certain length with Sqlalchemy?](https://stackoverflow.com/questions/36942506/how-to-create-integer-of-a-certain-length-with-sqlalchemy)

<https://stackoverflow.com/questions/36942506/how-to-create-integer-of-a-certain-length-with-sqlalchemy>

Comparison Postgre SQLAlchemy

https://towardsdatascience.com/here-is-the-reason-why-sqlalchemy-is-so-popular-43b489d3fb00

sqlite3

https://docs.python.org/3/library/sqlite3.html

TKinter -SQLalchemy

<https://github.com/EmaSMach/tkinter-form>

<https://github.com/EmaSMach/tkinter-form/commit/6c3400f47f5bb01e654177cba01db3ece49674d3#diff-b01eeed0ec6fc4f4799e44c7b4084e649d6eee4f1aca52ac935553ff82108398>

Ajax

<https://en.wikipedia.org/wiki/Ajax_%28programming%29>

AJAX with jQuery

https://flask.palletsprojects.com/en/1.1.x/patterns/jquery/

[Html auto refresh table](https://stackoverflow.com/questions/6521968/html-auto-refresh-table)

<https://stackoverflow.com/questions/6521968/html-auto-refresh-table>

[How to GET data in Flask from AJAX post](https://stackoverflow.com/questions/37631388/how-to-get-data-in-flask-from-ajax-post)

<https://stackoverflow.com/questions/37631388/how-to-get-data-in-flask-from-ajax-post>

[Get selected item value from Bootstrap DropDown with specific ID](https://stackoverflow.com/questions/24620741/get-selected-item-value-from-bootstrap-dropdown-with-specific-id)

https://stackoverflow.com/questions/24620741/get-selected-item-value-from-bootstrap-dropdown-with-specific-id

Interesting search: *ajax return 404 on flask*

<https://stackoverflow.com/questions/47166043/flask-404-response-from-ajax-request>

<https://stackoverflow.com/questions/43272910/flask-jquery-get-404>

<https://stackoverflow.com/questions/72908106/error-404-on-a-post-request-from-ajax-to-flask>

<https://stackoverflow.com/questions/40831915/flask-app-with-ajax-post-gives-404-not-found>

Site Deploiement

[Flask request without port in url](https://stackoverflow.com/questions/19627195/flask-request-without-port-in-url)

<https://stackoverflow.com/questions/19627195/flask-request-without-port-in-url>

Deploying to Production

<https://flask.palletsprojects.com/en/2.2.x/deploying/>

References

Messner et al., **Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection**, July 22, 2020, *Cell Systems 11*, 11–24, doi: <https://doi.org/10.1016/j.cels.2020.05.012>

Perez-Riverol et al., **The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences**

7 January 2022, Nucleic Acids Research, Volume 50, Issue D1, Pages D543–D552, doi: <https://doi.org/10.1093/nar/gkab1038>

Io, ad esmpio, sto utilizzando PostgreSQL come database.

Attualmente sono in una fase in cui mi hanno risposto quasi definitivamente sugli aspetti più importanti dei dati che vorrebbero modellare nel database. Quindi ho in mente un elenco di tabelle ed una struttura del db che vorrei implementare (fino ad ora ho fatto solo delle prove con sette tabelle per vedere come funzionava SQLAlchemy).

A questo punto però sto facendo un'analisi dei dati cercando di definire il range che li rapprersenta. Alla fine, nelle colonne delle tabelle, mi aspetto di trovare per la maggior parte dei numeri, qualche stringa e dei booleans.

Per la parte delle colonne di tipo numerico, ci saranno casi in cui il range andrà magari da 1 a 3. Questo perchè nel front-end farò dei dropdown menu che prendono le loro opzioni di scelta da delle liste statiche che verranno definite in programmazione. Ma potrebbe essererci anche il caso in cui la mia variabile ha un range 25 - 110.

Quindi ho provato a vedere in SQLAlchemy se era possibile limitare la dimensione delle variabili. Ad esempio per un intero da 1 a 3 mi chiedevo se ci fosse un tipo di intero small o addirittura la possibilità di dichiarare proprio il range da 1 a 3, trovandosi poi una dimensione del dato che passa magari da 4 a 2 bytes.

Paper Digest

Writefull

Penelope.ai

Litmaps

Bio-render

Otlet.io ----- share biological samples

Notion -----> Instapaper -----> Readwise

<https://www.youtube.com/watch?v=mii3RO0SVfo>

Evernote (Endnote)