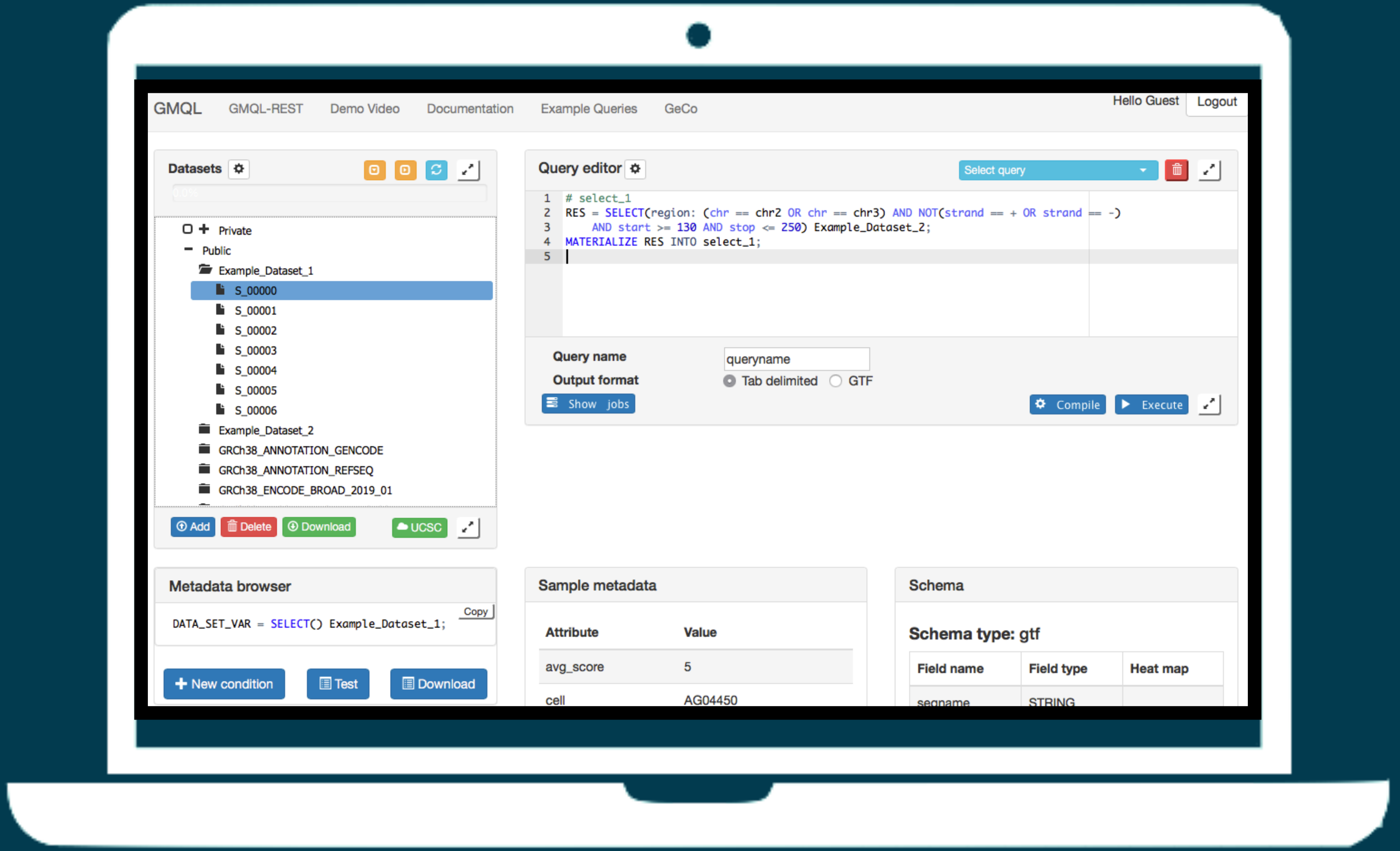
DEIB - POLITECNICO DI MILANO

Federated GMQL

OVERVIEW

GenoMetric Query Language (GMQL)

- Open Source Project
- Enables querying hundreds of datasets and thousands of samples
- Runs on a cloud-computing system based on Apache Spark
- Publicly available through a user-friendly web interface
- Well maintained public repository which integrates genomic datasets from ENCODE, TCGA/GDC and Roadmap Epigenomics



Private
Public
Example_Dataset_1
S_00000
S_00001
S_00002
S_00003
S_00004
S_00005
S_00006
Example_Dataset_2
GRCh38_ANNOTATION_GENCODE
GRCh38_ANNOTATION_REFSEQ
GRCh38_ENCODE_BROAD_2019_01

```
# select_1
RES = SELECT(region: (chr == chr2 OR chr == chr3) AND NOT(strand == + OR strand == -)
    AND start >= 130 AND stop <= 250) Example_Dataset_2;
MATERIALIZE RES INTO select_1;
```

Query name: queryname
Output format: ● Tab delimited ○ GTF

```
DATA_SET_VAR = SELECT() Example_Dataset_1;
```

Attribute	Value
avg_score	5
cell	AG04450

Schema type: gtf

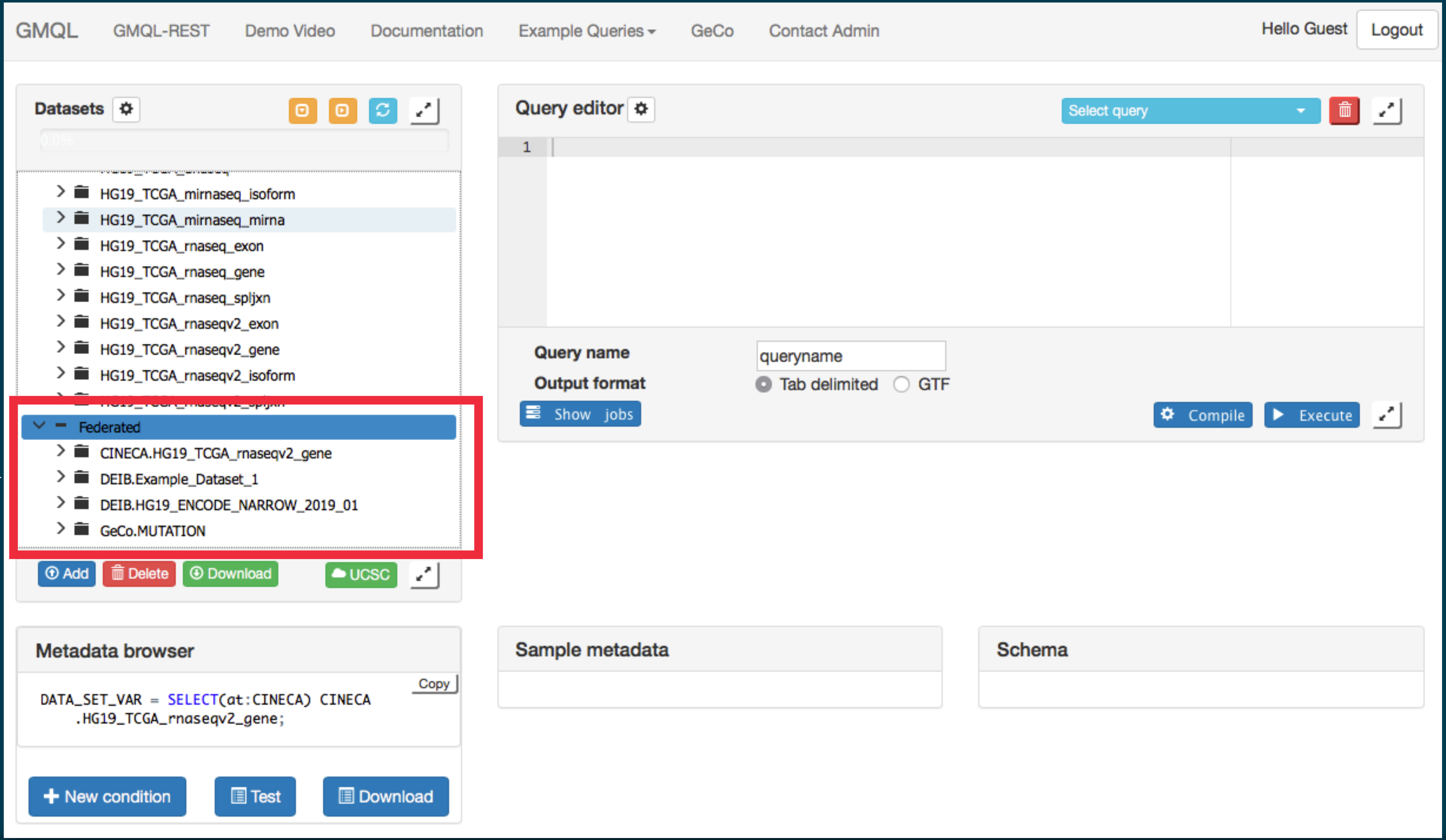
Field name	Field type	Heat map
seqname	STRING	

GECO
USER

GECO
USER
DEIB
ENCODE
CINECA
TCGA

GECO
DEIB
USER
ENCODE
Federation
CINECA
TCGA

Federated Datasets



- HG19_TCGA_mirnaseq_isoform
- HG19_TCGA_mirnaseq_mirna
- HG19_TCGA_rnaseq_exon
- HG19_TCGA_rnaseq_gene
- HG19_TCGA_rnaseq_spljxn
- HG19_TCGA_rnaseqv2_exon
- HG19_TCGA_rnaseqv2_gene
- HG19_TCGA_rnaseqv2_isoform
- HG19_TCGA_rnaseqv2_spljxn
- Federated
  - CINECA.HG19_TCGA_rnaseqv2_gene
  - DEIB.Example_Dataset_1
  - DEIB.HG19_ENCODE_NARROW_2019_01
  - GeCo.MUTATION

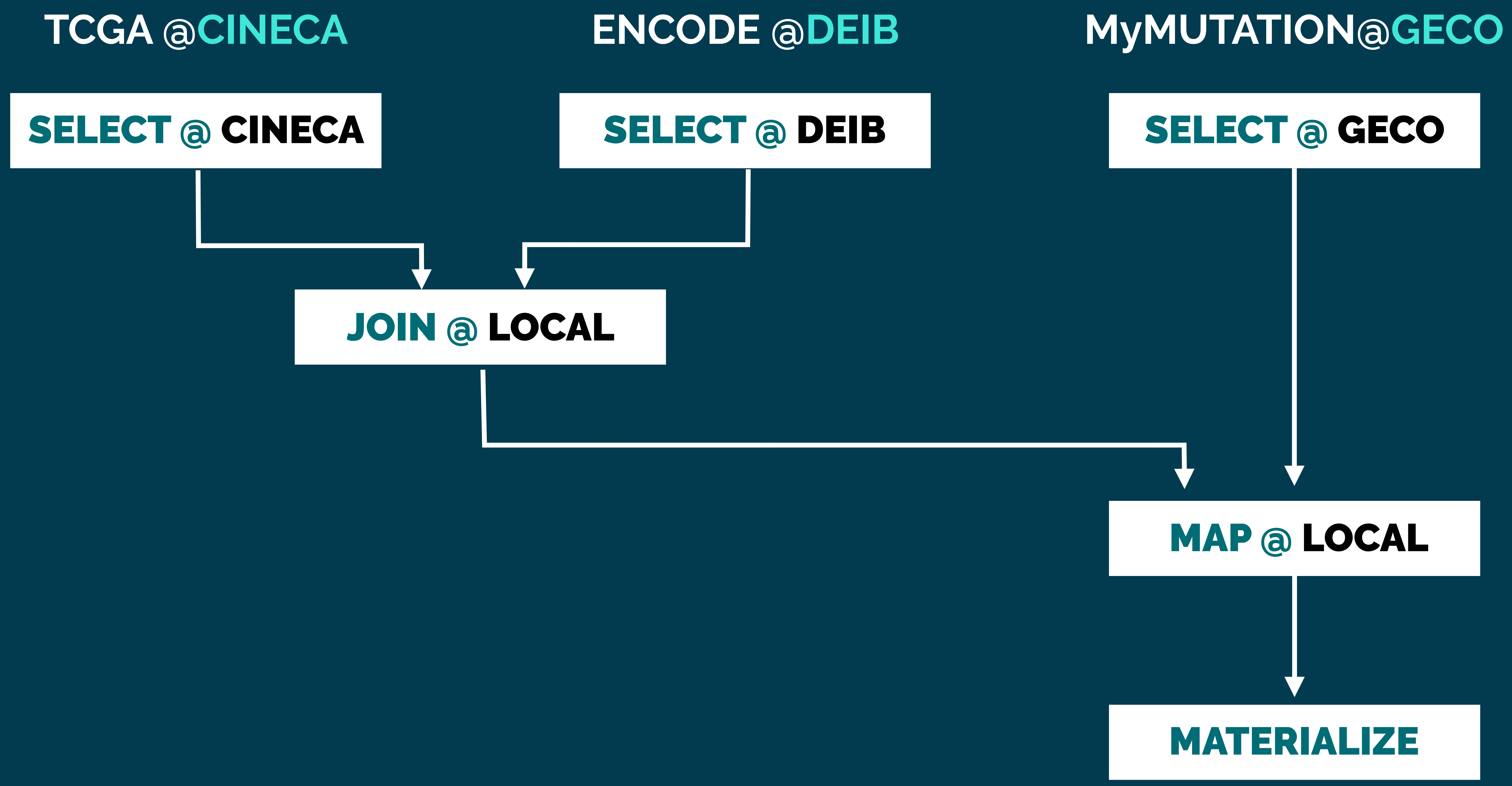
Query name: queryname

Output format: ● Tab delimited ○ GTF

```
1
```

```
DATA_SET_VAR = SELECT(at:CINECA) CINECA
    .HG19_TCGA_rnaseqv2_gene;
```

Federated Query

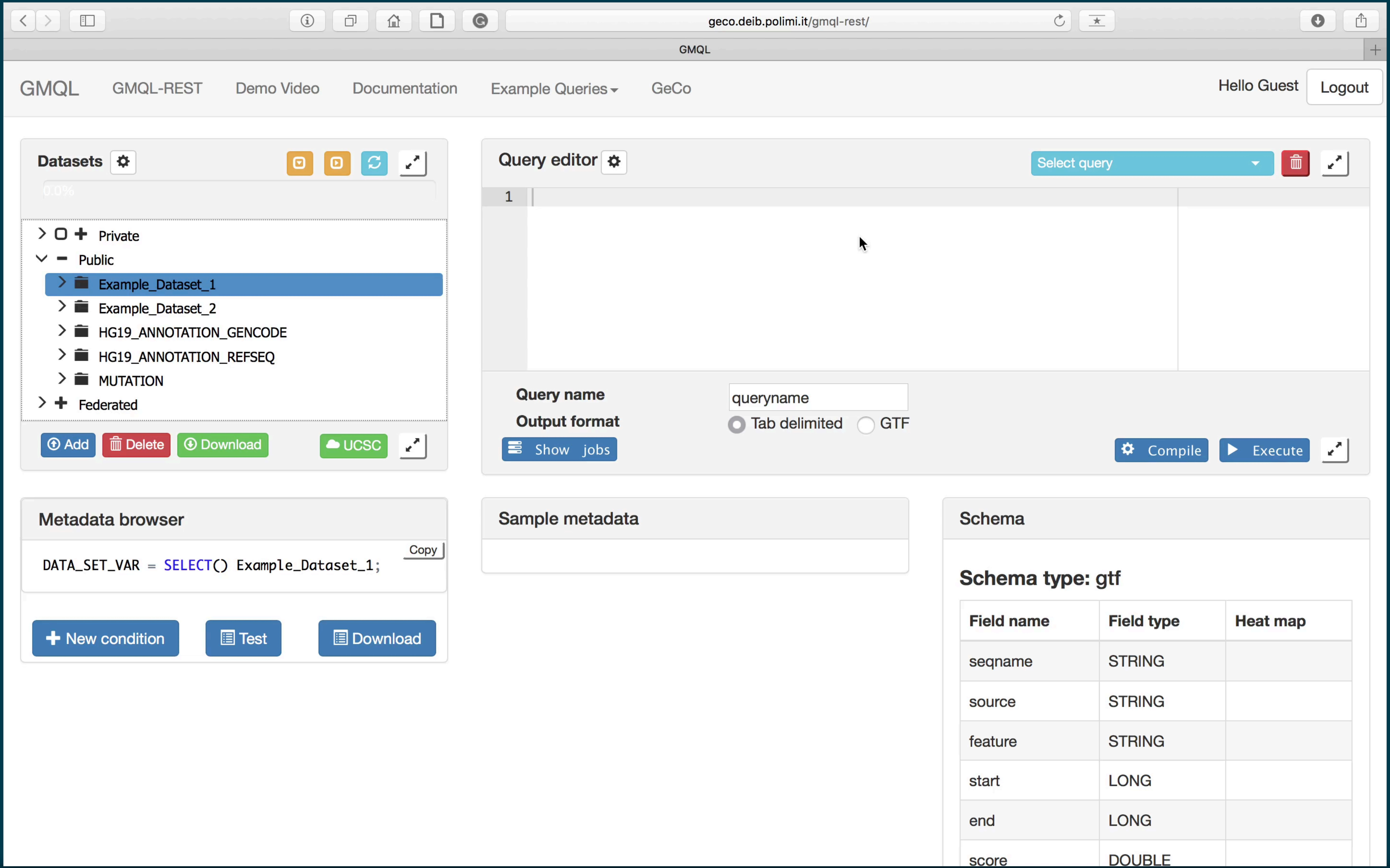


Federated Query

```
# SELECT TCGA AT CINECA
TCGA = SELECT(tumor_tag == "BRCA"; at:CINECA) CINECA.TCGA;

# SELECT ENCODE AT DEIB
ENCODE = SELECT(cell_line == "H1-hESC"; at:DEIB)  DEIB.ENCODE;

# JOIN DS1 AND DS2 AT DEIB
JOINED = JOIN(dist < 0; output: left_distinct; at:LOCAL) TCGA ENCODE;

# SELECT MUTATION AT GeCo
MYMUTATION = SELECT(at:LOCAL) MUTATION;

# Map MUTATION AT GeCo
RES = MAP(count_name: mut_count; at:LOCAL) JOINED MYMUTATION;

# Materialize at GeCo
MATERIALIZE RES INTO ResGenes;
```

Federated Query

```
# SELECT TCGA AT CINECA
TCGA = SELECT(tumor_tag == "BRCA"; at:CINECA) CINECA.TCGA;

# SELECT ENCODE AT DEIB
ENCODE = SELECT(cell_line == "H1-hESC"; at:DEIB) DEIB.ENCODE;

# JOIN DS1 AND DS2 AT DEIB
JOINED = JOIN(dist < 0; output: left_distinct; at:LOCAL) TCGA ENCODE;

# SELECT MUTATION AT GeCo
MYMUTATION = SELECT(at:LOCAL) MUTATION;

# Map MUTATION AT GeCo
RES = MAP(count_name: mut_count; at:LOCAL) JOINED MYMUTATION;

# Materialize at GeCo
MATERIALIZE RES INTO ResGenes;
```



```
DATA_SET_VAR = SELECT() Example_Dataset_1;
```

Query name queryname

Output format ● Tab delimited ○ GTF

Schema type: gtf

Field name	Field type	Heat map
seqname	STRING	
source	STRING	
feature	STRING	
start	LONG	
end	LONG	
score	DOUBLE	

SELECT(cell_line == “H1-hESC”) ENCODE

GECO

DEIB

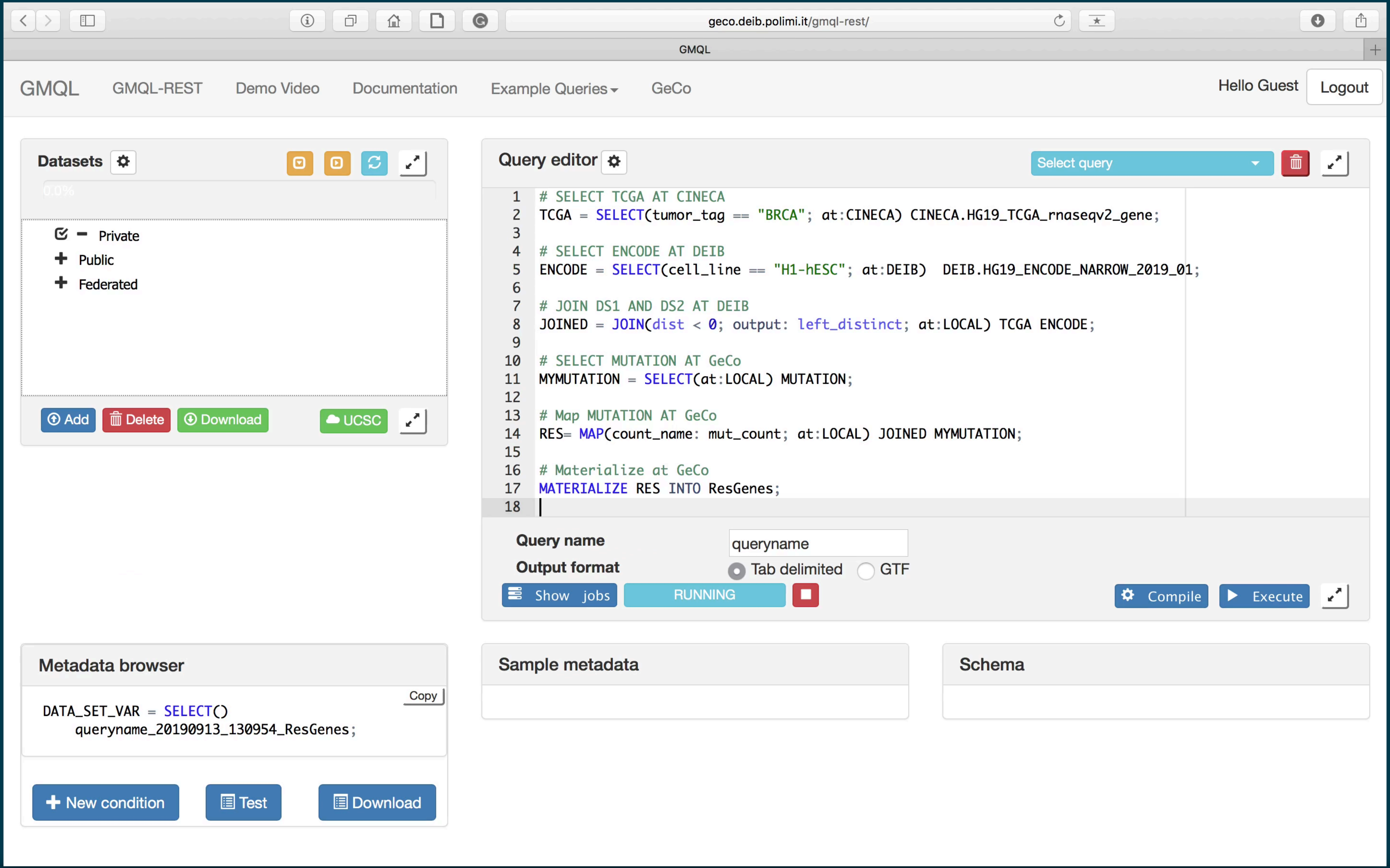
USER

SELECT(tumor_tag == “BRCA”) TCGA

CINECA

JOIN … SELECT … MAP … MATERIALIZE

GECO
DEIB
USER
CINECA



```
# SELECT TCGA AT CINECA
TCGA = SELECT(tumor_tag == "BRCA"; at:CINECA) CINECA.HG19_TCGA_rnaseqv2_gene;

# SELECT ENCODE AT DEIB
ENCODE = SELECT(cell_line == "H1-hESC"; at:DEIB)  DEIB.HG19_ENCODE_NARROW_2019_01;

# JOIN DS1 AND DS2 AT DEIB
JOINED = JOIN(dist < 0; output: left_distinct; at:LOCAL) TCGA ENCODE;

# SELECT MUTATION AT GeCo
MYMUTATION = SELECT(at:LOCAL) MUTATION;

# Map MUTATION AT GeCo
RES= MAP(count_name: mut_count; at:LOCAL) JOINED MYMUTATION;

# Materialize at GeCo
MATERIALIZE RES INTO ResGenes;
```

Query name: queryname

Output format: ● Tab delimited ○ GTF

```
DATA_SET_VAR = SELECT()
    queryname_20190913_130954_ResGenes;
```

DEIB - POLITECNICO DI MILANO

Federated GMQL

OVERVIEW

DEIB - POLITECNICO DI MILANO

Federated GMQL

SETUP

Installing a GMQL Instance

- Using Docker

  Follow the guide at : https://github.com/DEIB-GECO/GMQL-Docker

- Full Installation

  Follow the guide at https://github.com/DEIB-GECO/GMQL-WEB/wiki/installation

Name Server

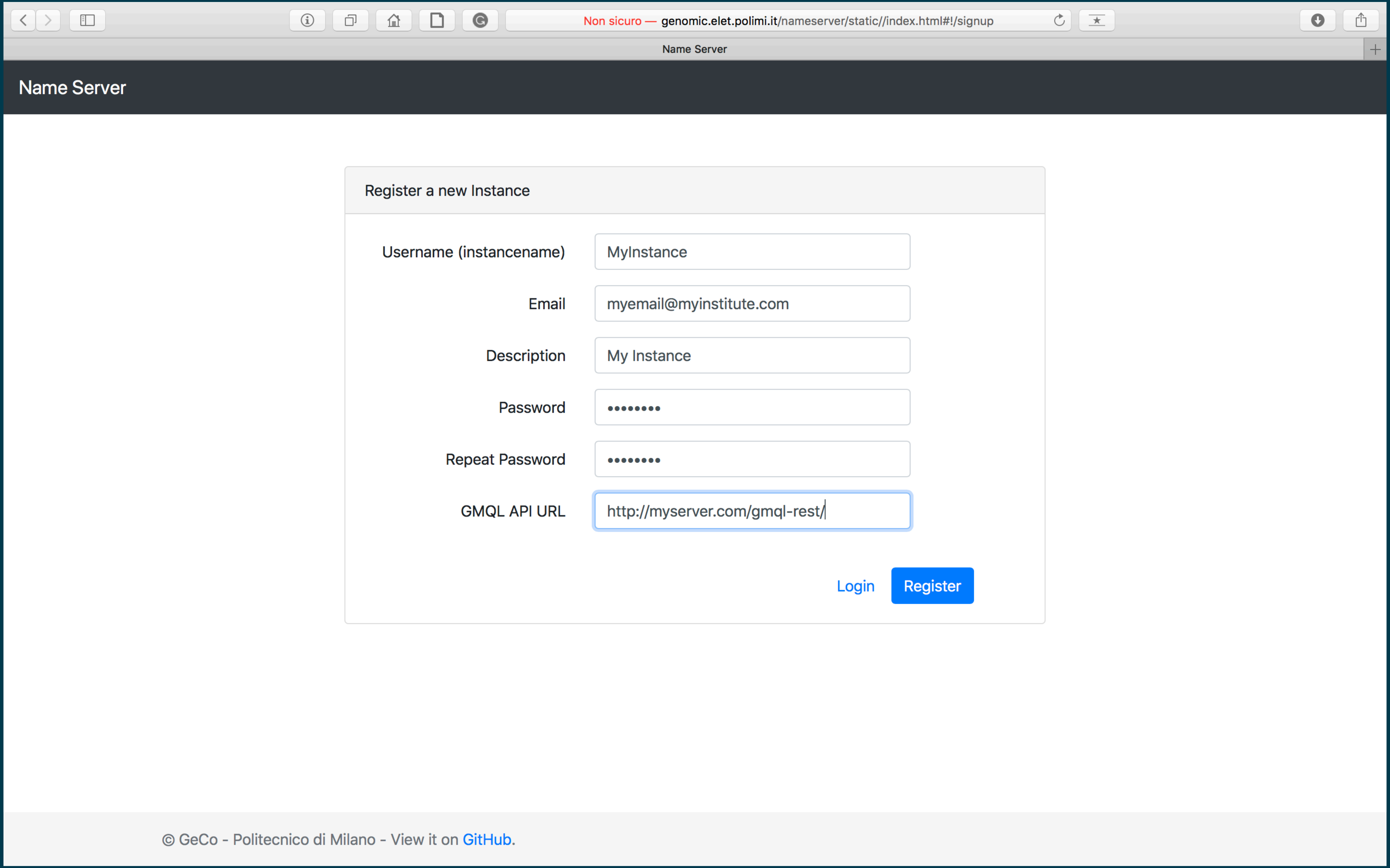
Centralized component of the Federated System that:

- Allows joining the Federation
- Manages:
    - Instances and their authentication mechanisms
    - Federated Datasets and their privacy
    - Groups

Publicly available at : http://genomic.elet.polimi.it/nameserver/

Joining the Federation

- Register your instance on the Name Server
- Login
- Add the required configuration to your repository.xml file



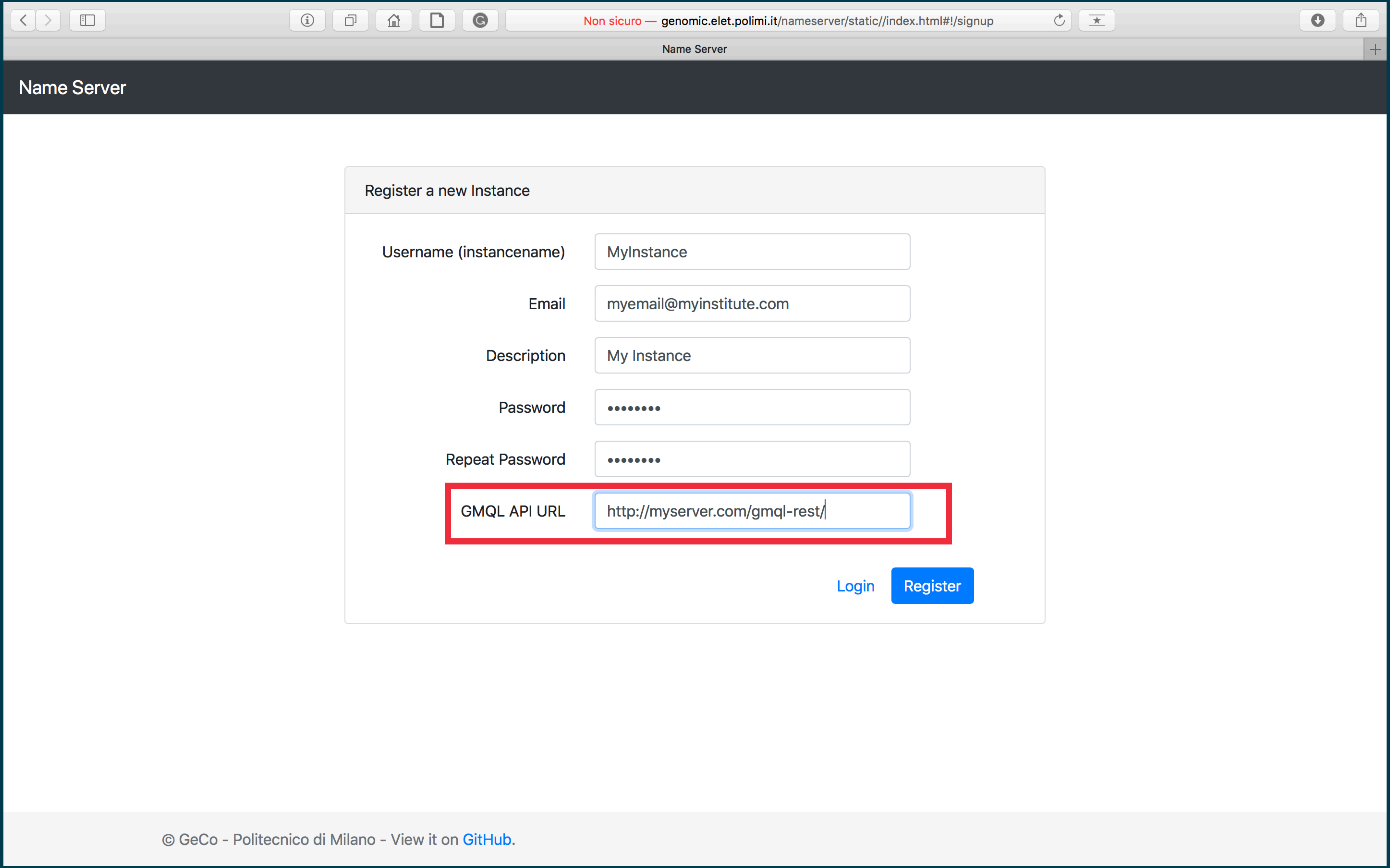
Non sicuro — genomic.elet.polimi.it/nameserver/static//index.html#!/signup
Name Server

Name Server

Register a new Instance

Username (instancename)	MyInstance
Email	myemail@myinstitute.com
Description	My Instance
Password	••••••••
Repeat Password	••••••••
GMQL API URL	http://myserver.com/gmql-rest/

© GeCo - Politecnico di Milano - View it on GitHub.



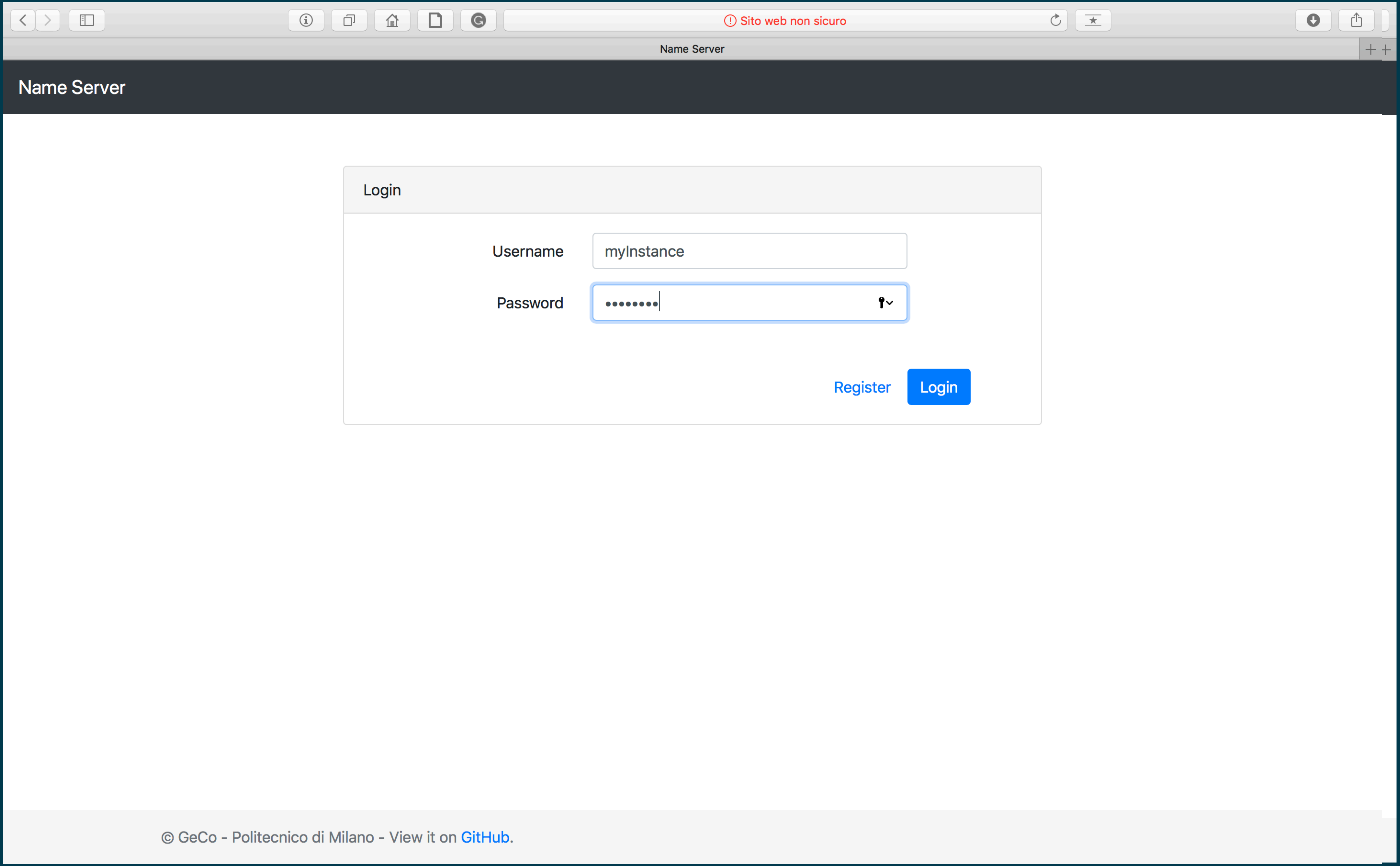
Non sicuro — genomic.elet.polimi.it/nameserver/static//index.html#!/signup
Name Server

Name Server

Register a new Instance

Username (instancename)	MyInstance
Email	myemail@myinstitute.com
Description	My Instance
Password	••••••••
Repeat Password	••••••••
GMQL API URL	http://myserver.com/gmql-rest/

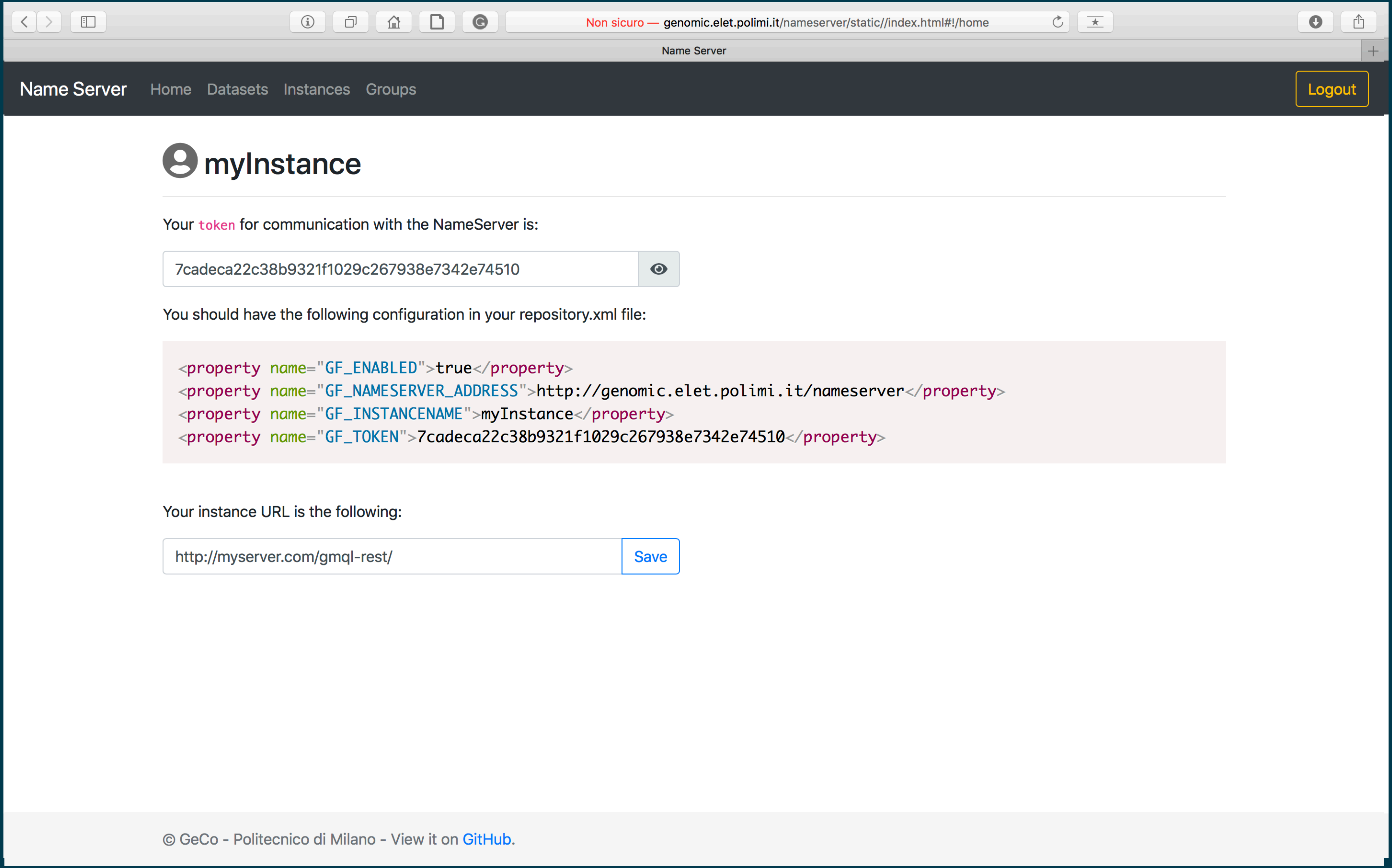
© GeCo - Politecnico di Milano - View it on GitHub.



Username: myInstance

Password: ••••••••

© GeCo - Politecnico di Milano - View it on GitHub.


Non sicuro — genomic.elet.polimi.it/nameserver/static//index.html#!/home
Name Server
Name Server
Home
Datasets
Instances
Groups
Logout

myInstance

Your token for communication with the NameServer is:

7cadeca22c38b9321f1029c267938e7342e74510

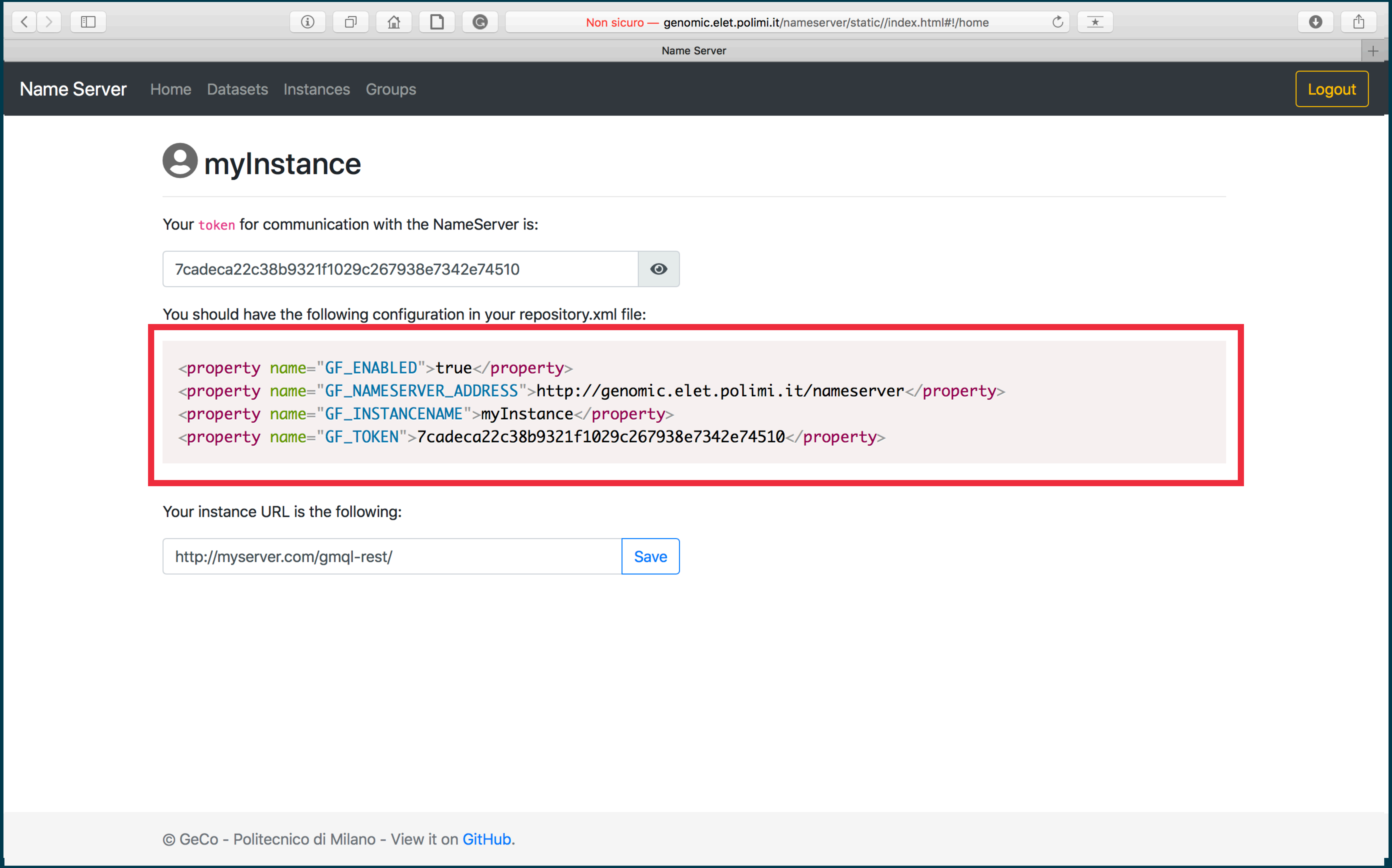
You should have the following configuration in your repository.xml file:

```
<property name="GF_ENABLED">true</property>
<property name="GF_NAMESERVER_ADDRESS">http://genomic.elet.polimi.it/nameserver</property>
<property name="GF_INSTANCENAME">myInstance</property>
<property name="GF_TOKEN">7cadeca22c38b9321f1029c267938e7342e74510</property>
```

Your instance URL is the following:

http://myserver.com/gmql-rest/

© GeCo - Politecnico di Milano - View it on GitHub.


Non sicuro — genomic.elet.polimi.it/nameserver/static//index.html#!/home
Name Server
Name Server
Home
Datasets
Instances
Groups
Logout

myInstance

Your token for communication with the NameServer is:

7cadeca22c38b9321f1029c267938e7342e74510

You should have the following configuration in your repository.xml file:

```
<property name="GF_ENABLED">true</property>
<property name="GF_NAMESERVER_ADDRESS">http://genomic.elet.polimi.it/nameserver</property>
<property name="GF_INSTANCENAME">myInstance</property>
<property name="GF_TOKEN">7cadeca22c38b9321f1029c267938e7342e74510</property>
```

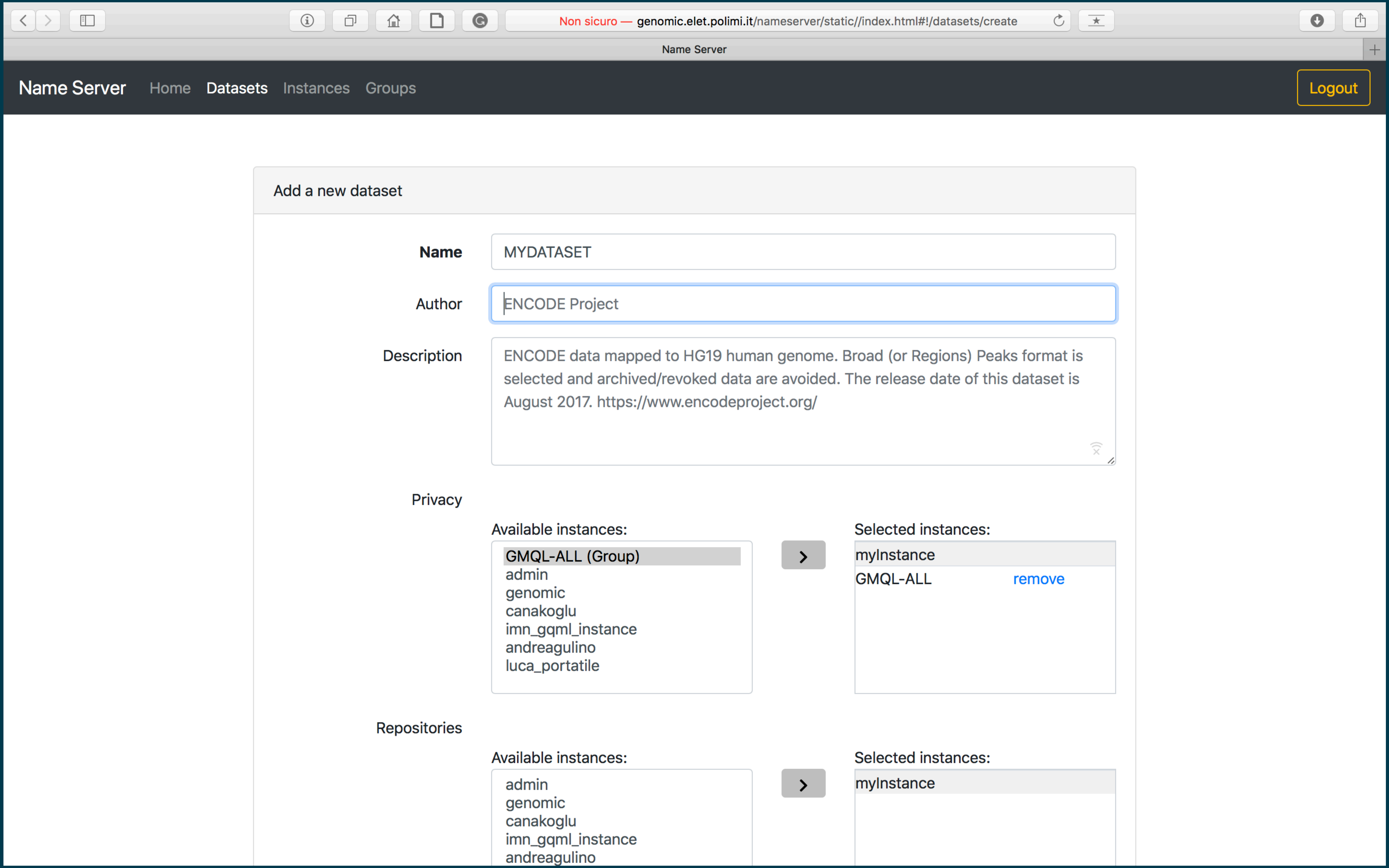
Your instance URL is the following:

http://myserver.com/gmql-rest/

© GeCo - Politecnico di Milano - View it on GitHub.

Federating a Dataset

- Login to the Name Server
- Go to the Datasets section
- Click on Add a new dataset
- Set dataset details and privacy
  - The dataset name must be identical to the name of the dataset as it appears in the public repository of the instance
- Save the new dataset


Non sicuro — genomic.elet.polimi.it/nameserver/static//index.html#!/datasets/create
Name Server
Name Server
Home
Datasets
Instances
Groups
Logout

Add a new dataset

Name MYDATASET

Author ENCODE Project

Description ENCODE data mapped to HG19 human genome. Broad (or Regions) Peaks format is selected and archived/revoked data are avoided. The release date of this dataset is August 2017. https://www.encodeproject.org/

Privacy

Available instances:

GMQL-ALL (Group)
admin
genomic
canakoglu
imn_gqml_instance
andreagulino
luca_portatile

Selected instances:

myInstance
GMQL-ALL remove

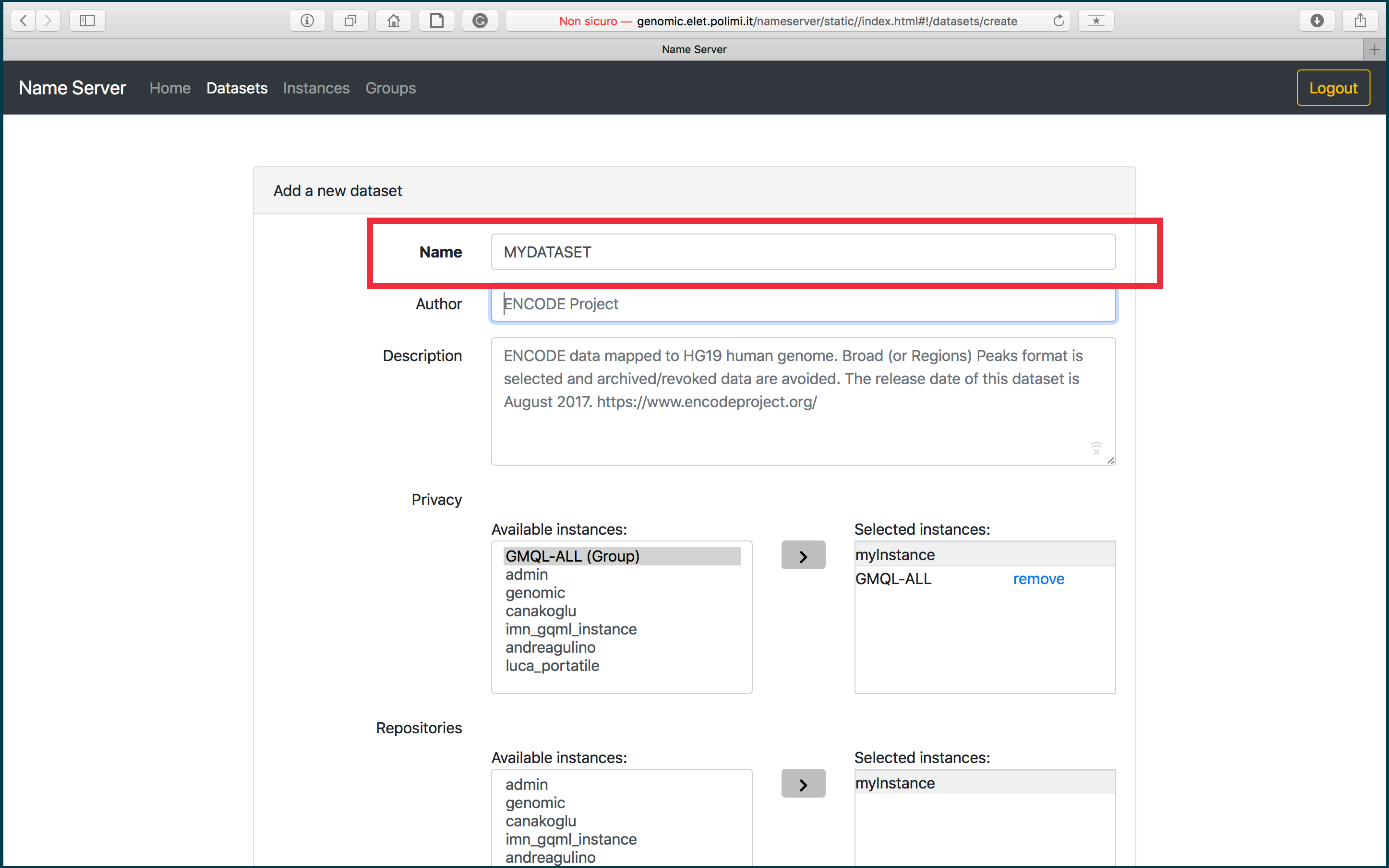
Repositories

Available instances:

admin
genomic
canakoglu
imn_gqml_instance
andreagulino

Selected instances:

myInstance



Non sicuro — genomic.elet.polimi.it/nameserver/static//index.html#!/datasets/create
Name Server
Name Server
Home
Datasets
Instances
Groups
Logout

Add a new dataset

Name MYDATASET

Author ENCODE Project

Description ENCODE data mapped to HG19 human genome. Broad (or Regions) Peaks format is selected and archived/revoked data are avoided. The release date of this dataset is August 2017. https://www.encodeproject.org/

Privacy

Available instances:

GMQL-ALL (Group)
admin
genomic
canakoglu
imn_gqml_instance
andreagulino
luca_portatile

Selected instances:

myInstance
GMQL-ALL remove

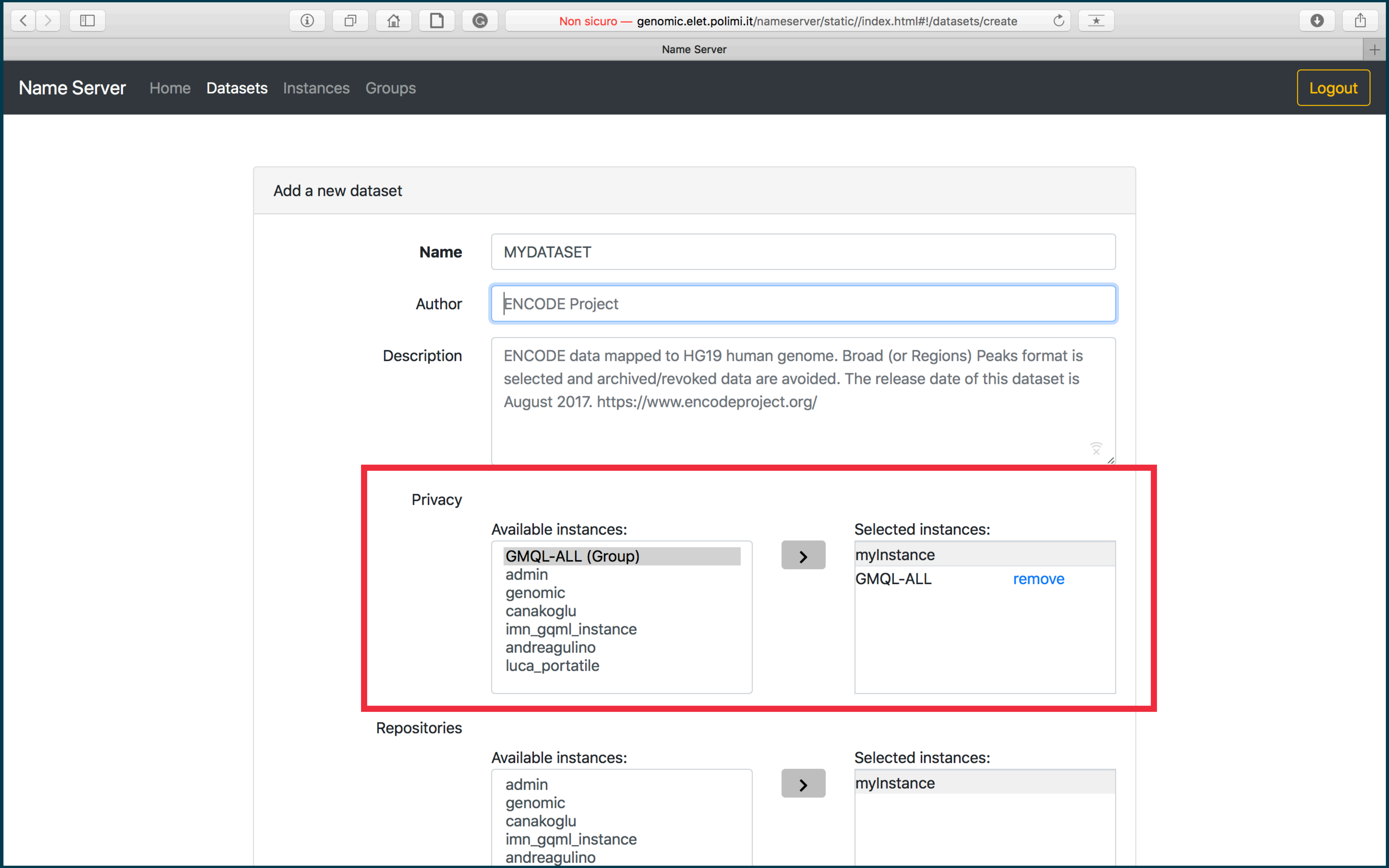
Repositories

Available instances:

admin
genomic
canakoglu
imn_gqml_instance
andreagulino

Selected instances:

myInstance


Non sicuro — genomic.elet.polimi.it/nameserver/static//index.html#!/datasets/create
Name Server
Name Server
Home
Datasets
Instances
Groups
Logout

Add a new dataset

Name MYDATASET

Author ENCODE Project

Description ENCODE data mapped to HG19 human genome. Broad (or Regions) Peaks format is selected and archived/revoked data are avoided. The release date of this dataset is August 2017. https://www.encodeproject.org/

Privacy

Available instances:
GMQL-ALL (Group)
admin
genomic
canakoglu
imn_gqml_instance
andreagulino
luca_portatile

Selected instances:
myInstance
GMQL-ALL remove

Repositories

Available instances:
admin
genomic
canakoglu
imn_gqml_instance
andreagulino

Selected instances:
myInstance

DEIB - POLITECNICO DI MILANO

Federated GMQL

SETUP