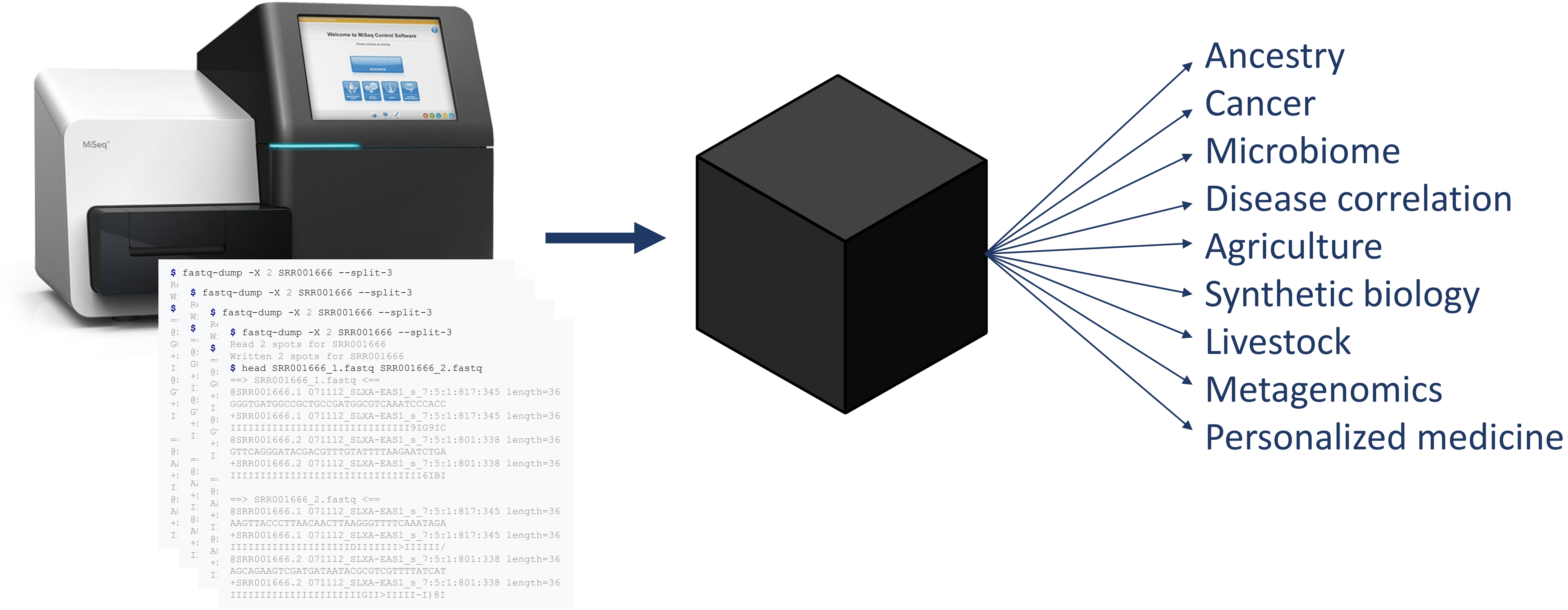
Progress Towards the BioCompute Database

FDA Scientific Computing Board
5 November, 2020

Jonathon Keeney, Ph.D.
George Washington University

BioCompute Objects

HTS Data Flows



```
$ fastq-dump -X 2 SRR001666 --split-3
$ fastq-dump -X 2 SRR001666 --split-3
$ fastq-dump -X 2 SRR001666 --split-3
```

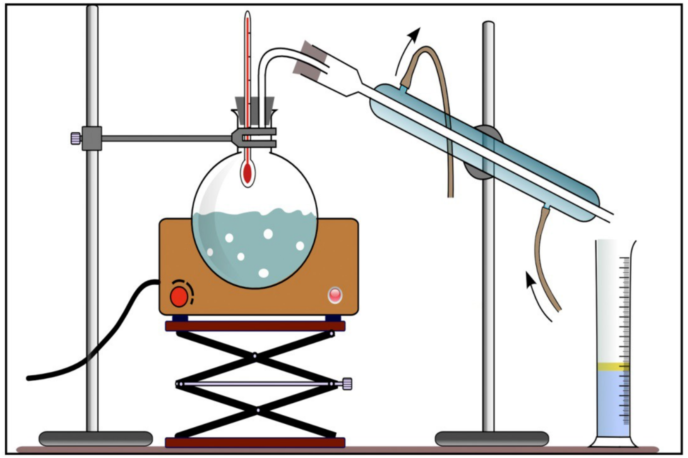
```
$ fastq-dump -X 2 SRR001666 --split-3
Read 2 spots for SRR001666
Written 2 spots for SRR001666
$ head SRR001666_1.fastq SRR001666_2.fastq
==> SRR001666_1.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI

==> SRR001666_2.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
AAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
AGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
IIIIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

Ancestry
Cancer
Microbiome
Disease correlation
Agriculture
Synthetic biology
Livestock
Metagenomics
Personalized medicine

Challenge: Workflow Communication

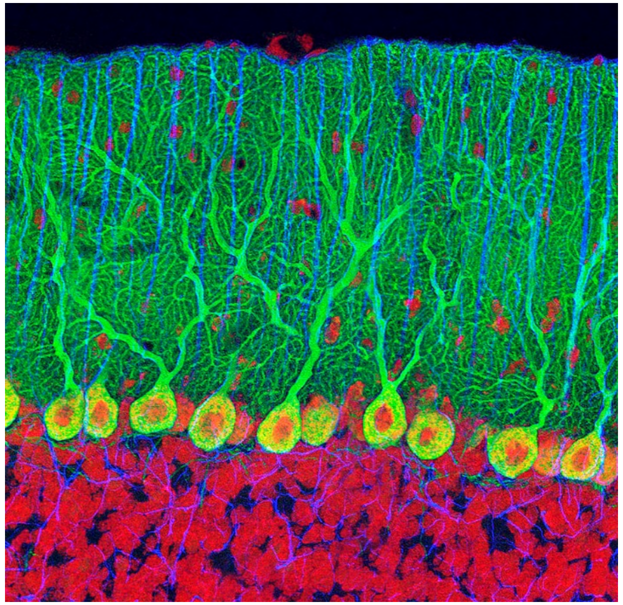
Input

Strain of animal
Tissue type
Protein Lot Number
Humidity
Microscope
…



conditions

Output

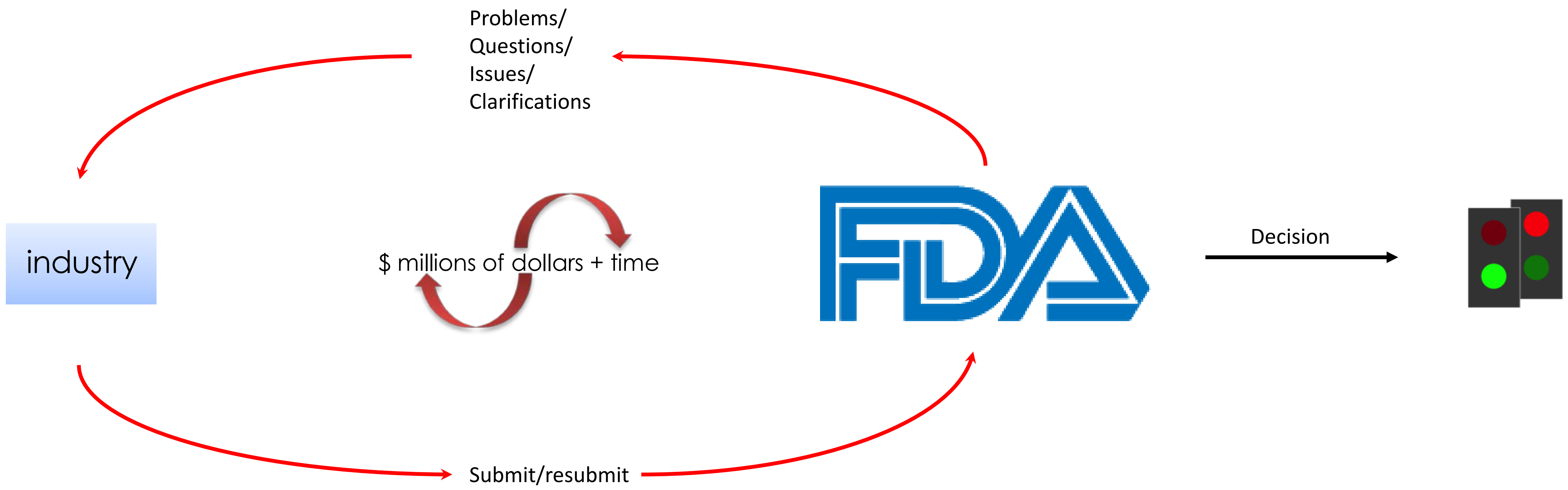


generalized experiment	inputs	parameters	outputs

```
> my_program -i input_file1 -parameter1 value1 -parameter2 value2 -o out_file
```

Analogy: wet lab experiments

Wasted Time and Money



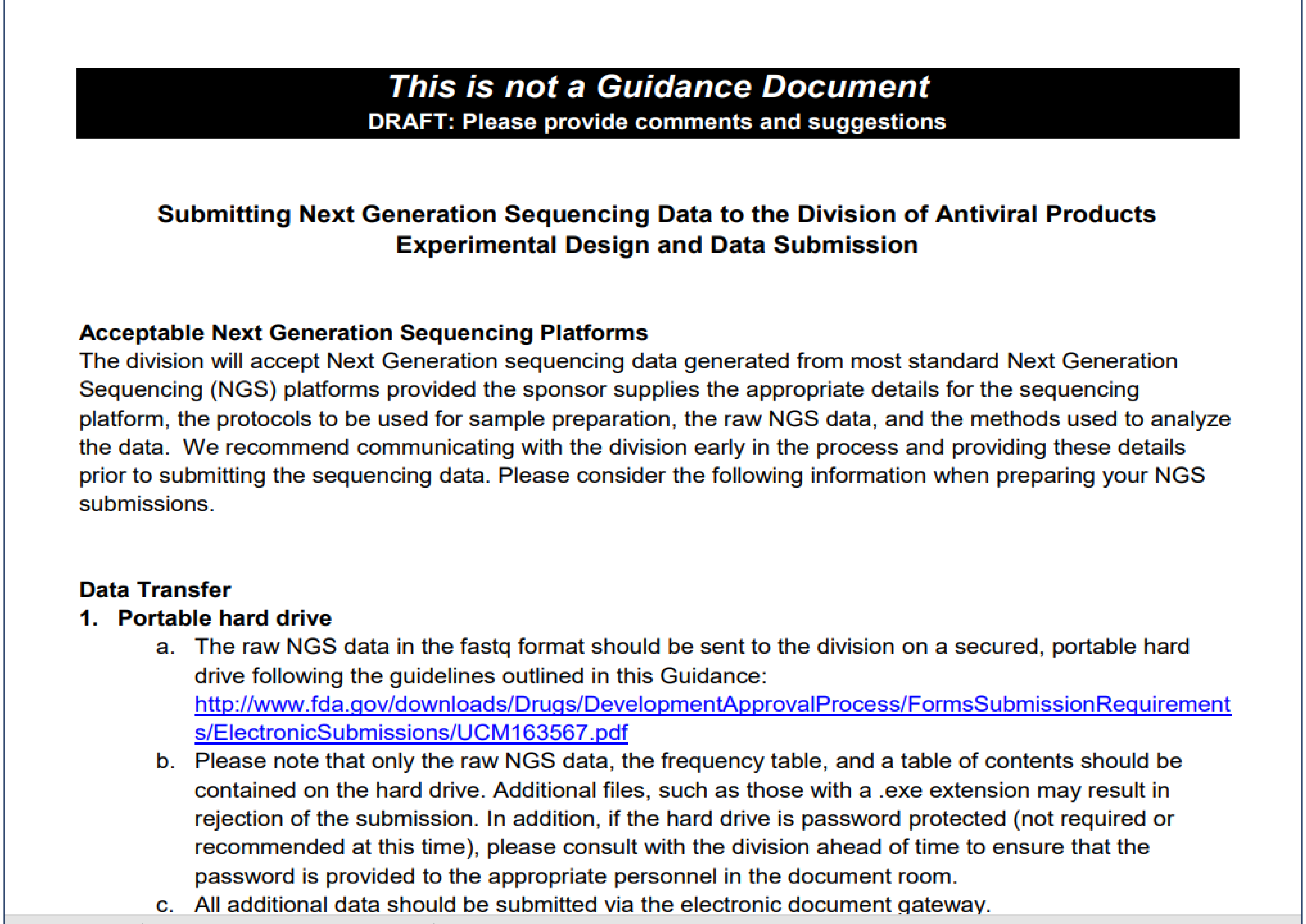
Problems/
Questions/
Issues/
Clarifications

industry

$ millions of dollars + time

Decision

Submit/resubmit



This is not a Guidance Document

DRAFT: Please provide comments and suggestions

Submitting Next Generation Sequencing Data to the Division of Antiviral Products Experimental Design and Data Submission

Acceptable Next Generation Sequencing Platforms

The division will accept Next Generation sequencing data generated from most standard Next Generation Sequencing (NGS) platforms provided the sponsor supplies the appropriate details for the sequencing platform, the protocols to be used for sample preparation, the raw NGS data, and the methods used to analyze the data. We recommend communicating with the division early in the process and providing these details prior to submitting the sequencing data. Please consider the following information when preparing your NGS submissions.

Data Transfer

1. **Portable hard drive**
    a. The raw NGS data in the fastq format should be sent to the division on a secured, portable hard drive following the guidelines outlined in this Guidance: http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM163567.pdf
    b. Please note that only the raw NGS data, the frequency table, and a table of contents should be contained on the hard drive. Additional files, such as those with a .exe extension may result in rejection of the submission. In addition, if the hard drive is password protected (not required or recommended at this time), please consult with the division ahead of time to ensure that the password is provided to the appropriate personnel in the document room.
    c. All additional data should be submitted via the electronic document gateway.

A solution should…

- Be human readable: like a GenBank sequence record
- Be machine readable: structured information with predefined fields and associated meanings of values
- Contain enough information to understand the computational pipelines, interpret information, maintain records, and reproduce experiments
- Be immutable: ensure information has not been altered

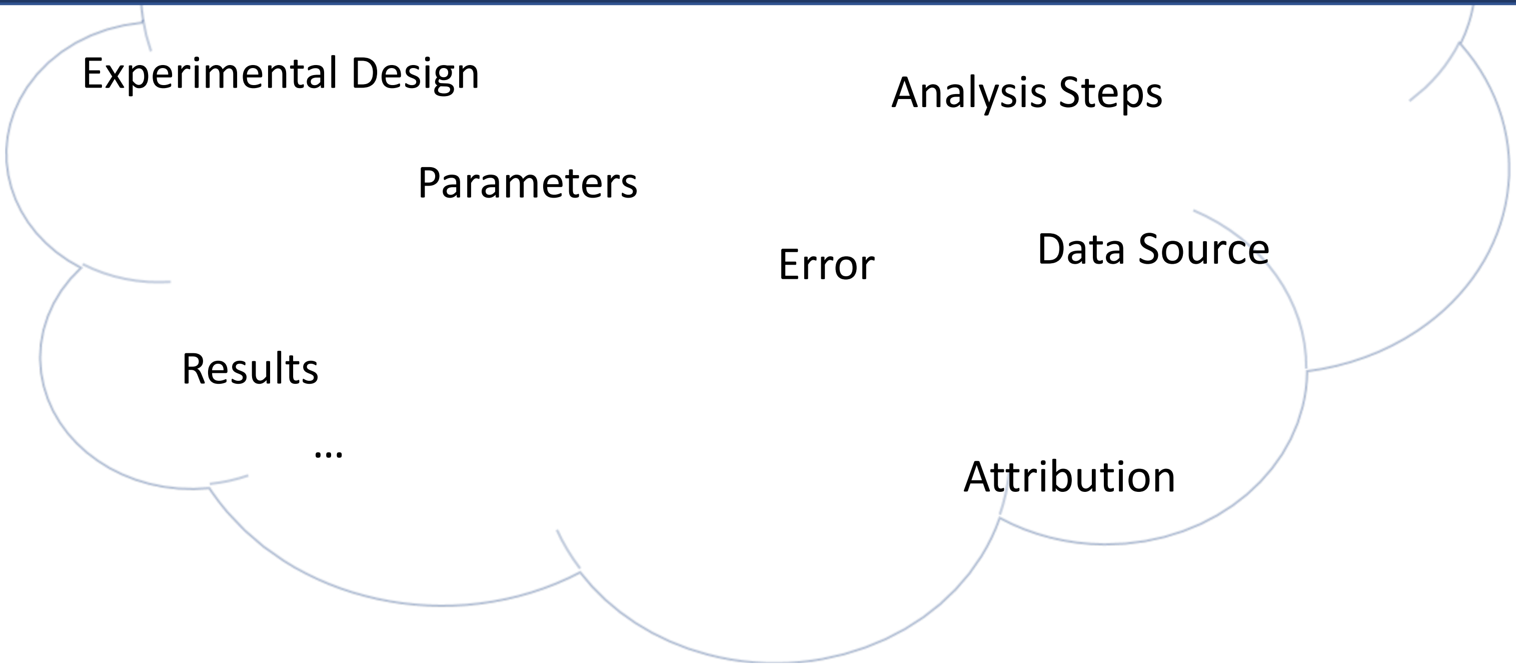
Solution: BioCompute

IEEE approved standard for communicating bioinformatic analysis workflows

- Acts like an envelope for entire pipeline
  - Can incorporate other standards
- Human and machine readable
  - Written in JSON
- Categorized by domains
- Adheres to and encourages F.A.I.R. principles
  - Fully open source
- Adaptable
  - e.g. to other schemas
- Preserves data provenance
- Unique IDs for versioning

BioCompute Objects

Solution: BioCompute



Experimental Design

Analysis Steps

Parameters

Error

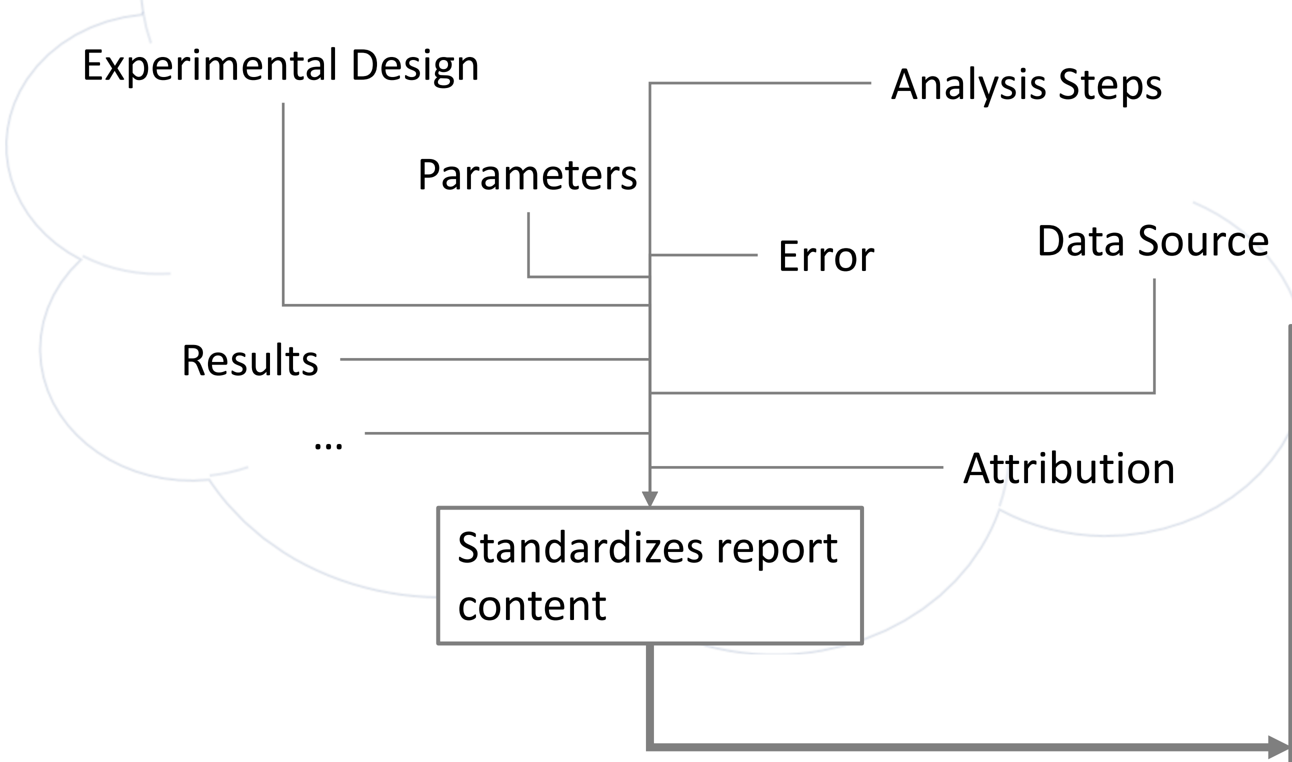
Data Source

Results

...

Attribution

Solution: BioCompute



Experimental Design

Analysis Steps

Parameters

Error

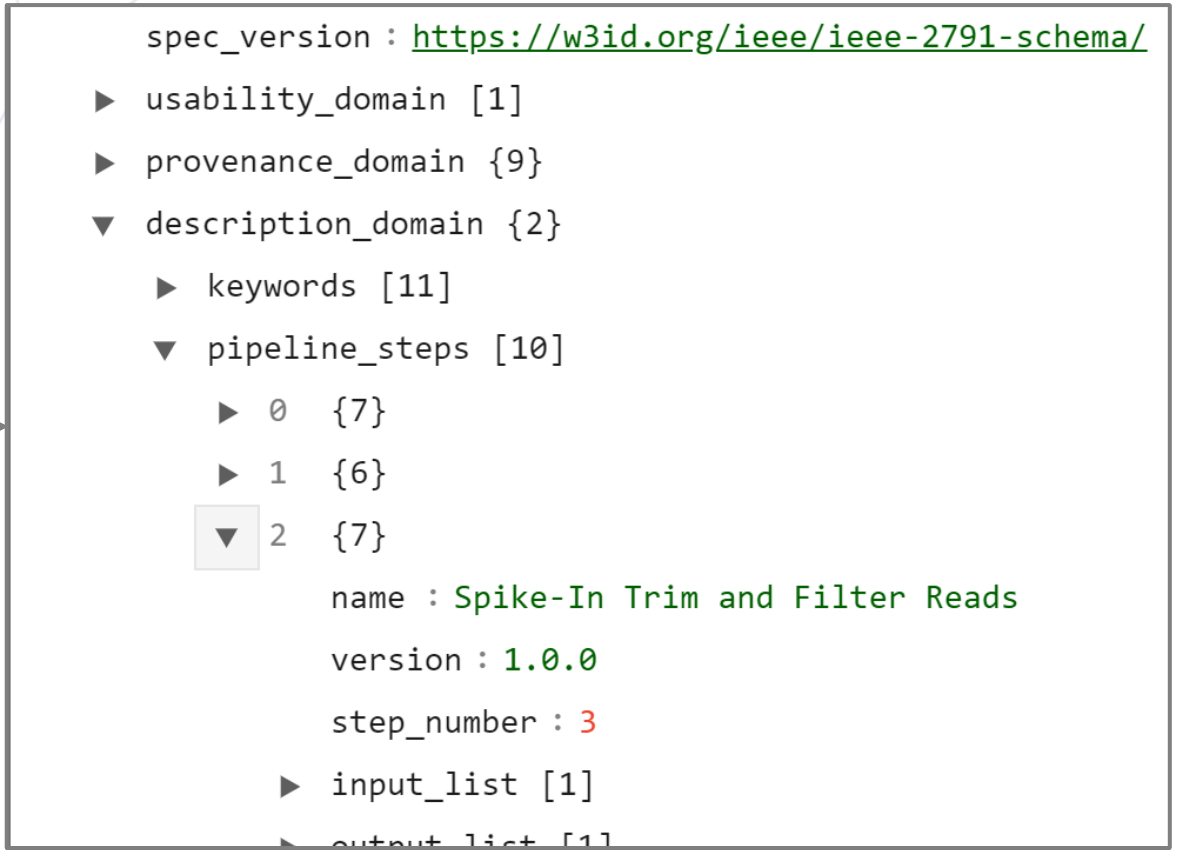
Data Source

Results

...

Attribution

Standardizes report content



```
    spec_version : https://w3id.org/ieee/ieee-2791-schema/
▶ usability_domain [1]
▶ provenance_domain {9}
▼ description_domain {2}
    ▶ keywords [11]
    ▼ pipeline_steps [10]
        ▶ 0  {7}
        ▶ 1  {6}
        ▼ 2  {7}
              name : Spike-In Trim and Filter Reads
              version : 1.0.0
              step_number : 3
            ▶ input_list [1]
            ▶ output_list [1]
```

BioCompute streamlines reporting without enforcing any tool, platform, or workflow strategy.

Machine readability enables customized views

BioCompute Objects

Metadata

```
object_id : https://beta.portal.aws.biochemistry.gwu.edu/bco/BCO_00016916
spec_version : https://w3id.org/ieee/ieee-2791-schema/
etag : fea7e938e6bdf9a2cfcba7fa02f5a5fc3973dccb0b03a64319e1ee29966a5b6b
```

Provenance Domain

```
 provenance_domain :
    embargo :
    created : 2020-08-04T23:50:56.016Z
    modified : 2020-08-04T23:50:56.016Z
    name : Human Healthy Bulk RNA-seq Expression (Bgee)
    version : v-1.0
    obsolete_after : 2020-04-22T23:57:00.000Z
    contributors :
        contribution :
          createdBy
        name : Amanda Bell
        email : amandab2140@gwu.edu
        affiliation : GW HIVE-Lab
        orcid : http://orcid.org/0000-0002-9920-565X
    license : Attribution 4.0 International CC BY 4.0
```

Description Domain

```
 description_domain :
    keywords :
      Gene Expression
      Gene Expression Regulation
      Tissue specificity
    xref :
        namespace : ensembl
        name : Ensembl Genome Browser
        ids :
          Ensembl gene ID
        access_time : 2020-04-22T14:03:00.000Z
    platform :
      OncoMX
    pipeline_steps :
        step_number : 1
        name : oncomx server
        prerequisite :
            uri :
        description : Process data
        input_list :
```

Error Domain

```
error_domain :
  empirical_error:
    D168Y: percentage: 0.56, calls: 0.5615, STDEV.P: 0.00075
  algorithmic_error:
    SCORE_threshold: 0.5, QUALITY: 25, COVERAGE: 5000
```

Parametric Domain

```
parametric_domain :
     param : grep
     value : -r
     step : 1
```

Execution Domain

```
 execution_domain :
    environment_variables :
        key : EDITOR
        value : vim
        key : HOSTTYPE
        value : x86_64-linux
    external_data_endpoints :
        url : https://data.oncomx.org/ONCOMXDS000012
        name : Human Healthy Bulk RNA-seq Expression (Bgee)
    script :
        uri :
          filename : make-dataset.py
          uri : http://data.oncomx.org/ln2wwwdata/software/pipeline/integrator/make-
dataset.py
          access_time : 2020-04-22T14:28:00.000Z
    software_prerequisites :
        uri :
          filename : shell
          uri : https://www.python.org/download/releases/2.7.5
          access_time : 2020-04-22T14:30:00.000Z
        name : Python
        version : 2.7.5
    script_driver : Python
```

IO Domain

```
 io_domain :
    input_subdomain :
        uri :
          filename : Homo_sapiens_UBERON:0000066
          uri :
http://data.oncomx.org/ln2wwwdata/downloads/bgee/current/Homo_sapiens_UBERON:0000066
_AFFYMETRIX_RNA_SEQ.tsv
          access_time : 2020-04-22T20:44:00.000Z
    output_subdomain :
        uri :
          filename : human_normal_expression.csv
          uri : https://data.oncomx.org/ONCOMXDS000012
          access_time : 2020-04-22T20:50:00.000Z
        mediatype : TEXT/CSV
```

Extension Domain

```
 extension_domain :
      dataset_categories :
          category_value : Homo sapiens
          category_name : species
          category_value : normal
          category_name : disease_status
      extension_schema : https://data.oncomx.org/ONCOMXDS000012
```

Usability Domain

```
 usability_domain :
          List of human  taxid:9606  genes with healthy RNA-Seq and Affymetrix expression data
         in Bgee; additional documentation available at
         (https://github.com/BgeeDB/bgee_pipeline/tree/develop/pipeline/collaboration/onco
         MX#information-about-the-files-generated-for-oncomx) Only the subset of RNA-Seq
         data are used to generate the expression profiles for healthy individuals for
         human used by OncoMX.
```

BioCompute participants

Children's National

precisionFDA

ATTAIN

Global Alliance
for Genomics & Health

DISEASE
ONTOLOGY

EPINOMICS

NIH NATIONAL CANCER INSTITUTE

CUROVERSE

Forum for
Collaborative HIV Research

Google

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

EMBL-EBI

DNAnexus

HARVARD
UNIVERSITY

NATIONAL
LIBRARY OF
MEDICINE

NCBI

FHIR
FHIR Genomics

paradigm4

Oxford
e-Research
Centre

UNIVERSITY OF
OXFORD

Seven
Bridges

NIST

Galaxy
PROJECT

NEW YORK
GENOME CENTER®

GENOME NEXT

CRITICAL PATH
INSTITUTE

NIH

research
object.org

HL7

CWL

FLATIRON

MANCHESTER
1824
The University of Manchester

elixir
UNITED
KINGDOM

CUNY

Standardization

Institute of Electrical and Electronics Engineers Standard

BioCompute P2791-2020 approved January 2020

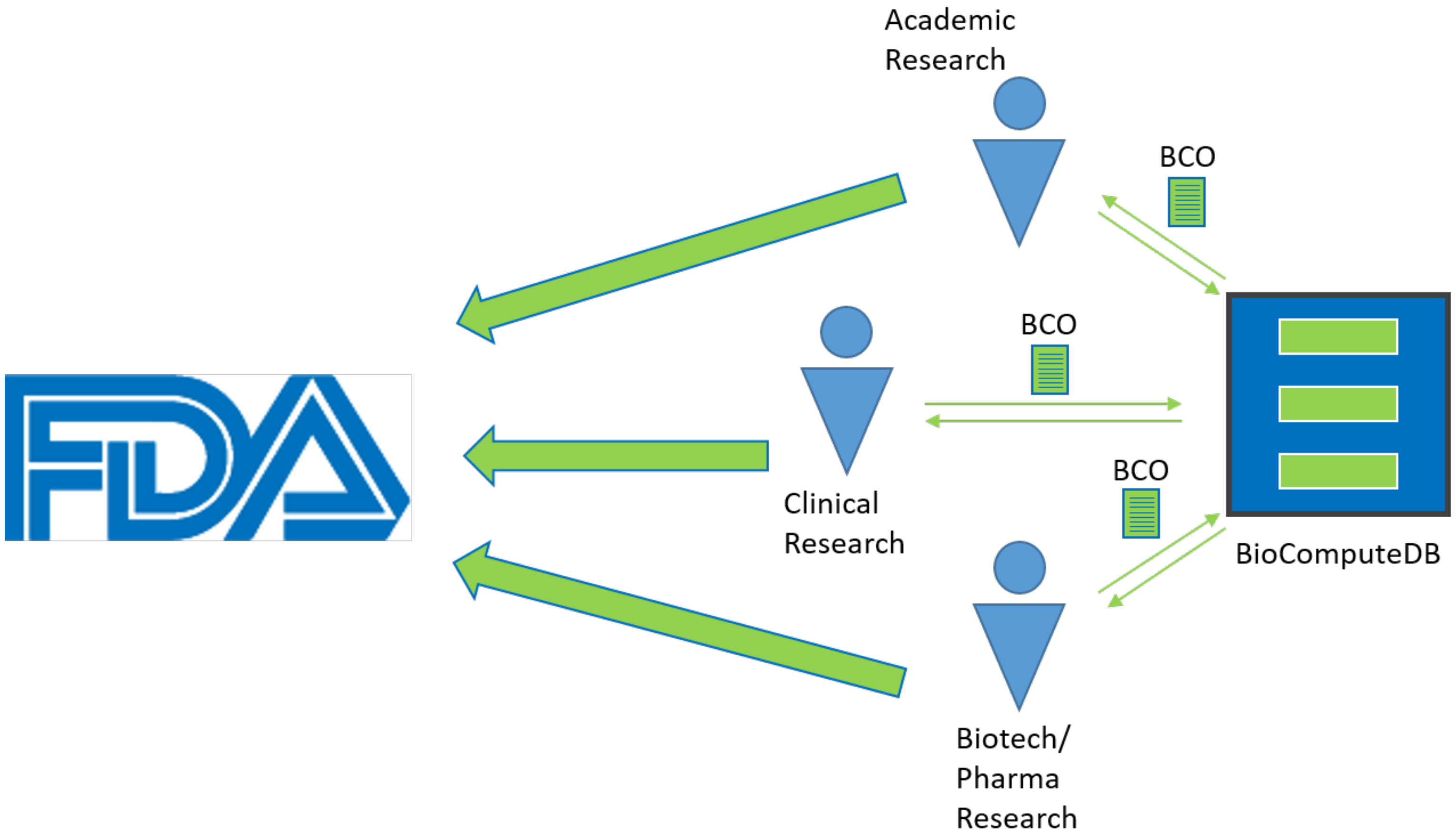
https://standards.ieee.org/content/ieee-standards/en/standard/2791-2020.html

FDA Related Activities

BAA 75F40119C10136

- Training for FDA personnel
- Prototypes
  - BCO-CWL prototypes (portability of execution)
  - BCO-RO prototypes (packaging a BCO with a larger package of resources)
  - Beginning year 3: Seven Bridges Genomics platform integration
  - Beginning year 3: DNANexus platform integration
- Development of an open source database and associated portal

BioCompute DB

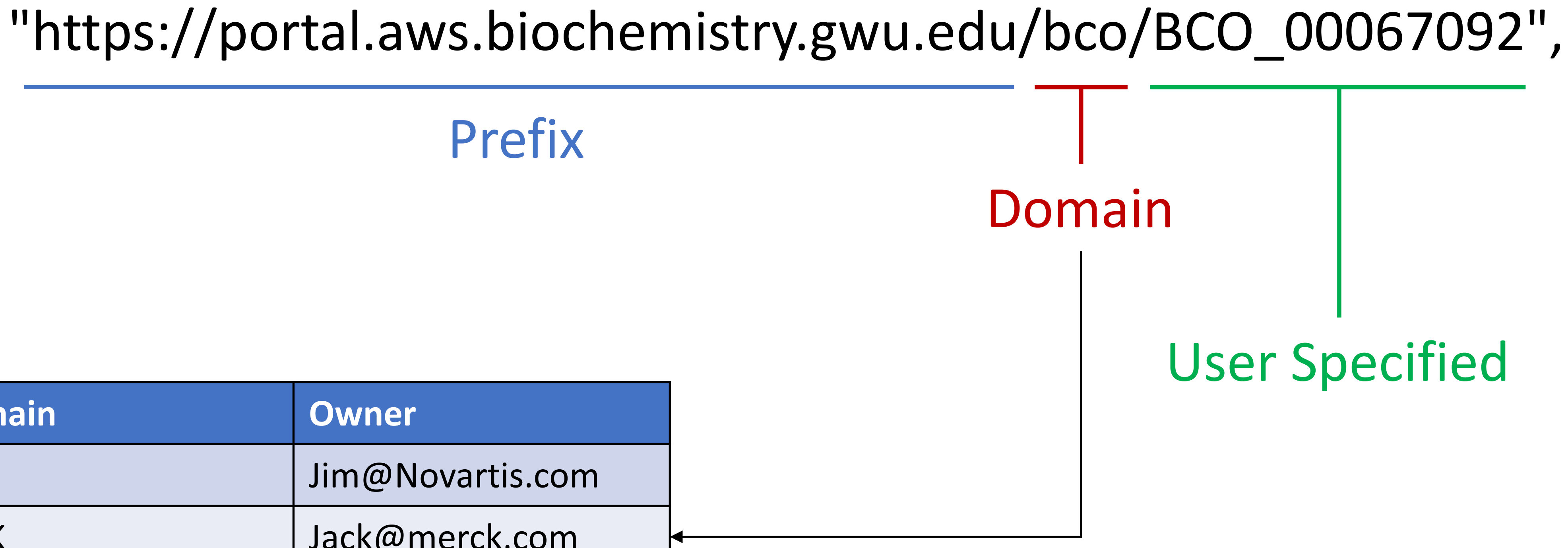


BioCompute DB

- Allow a reviewer to better understand procedures associated with a specific sponsor analysis
  - Facilitate better scientific communication of workflows with little additional communication, outside of initial submission
  - Self education or templates for reviewers
- Central DB for coordination of activities
  - Versioning
  - Checking for name collisions
    - “BCO001.json”
  - Referencing previous BCOs
    - E.g. in “derived_from” field
- Components
  - Database
  - Interface (“Portal”)
    - Registry

BioCompute Objects

BioCompute DB: Registry



"https://portal.aws.biochemistry.gwu.edu/bco/BCO_00067092",

Prefix

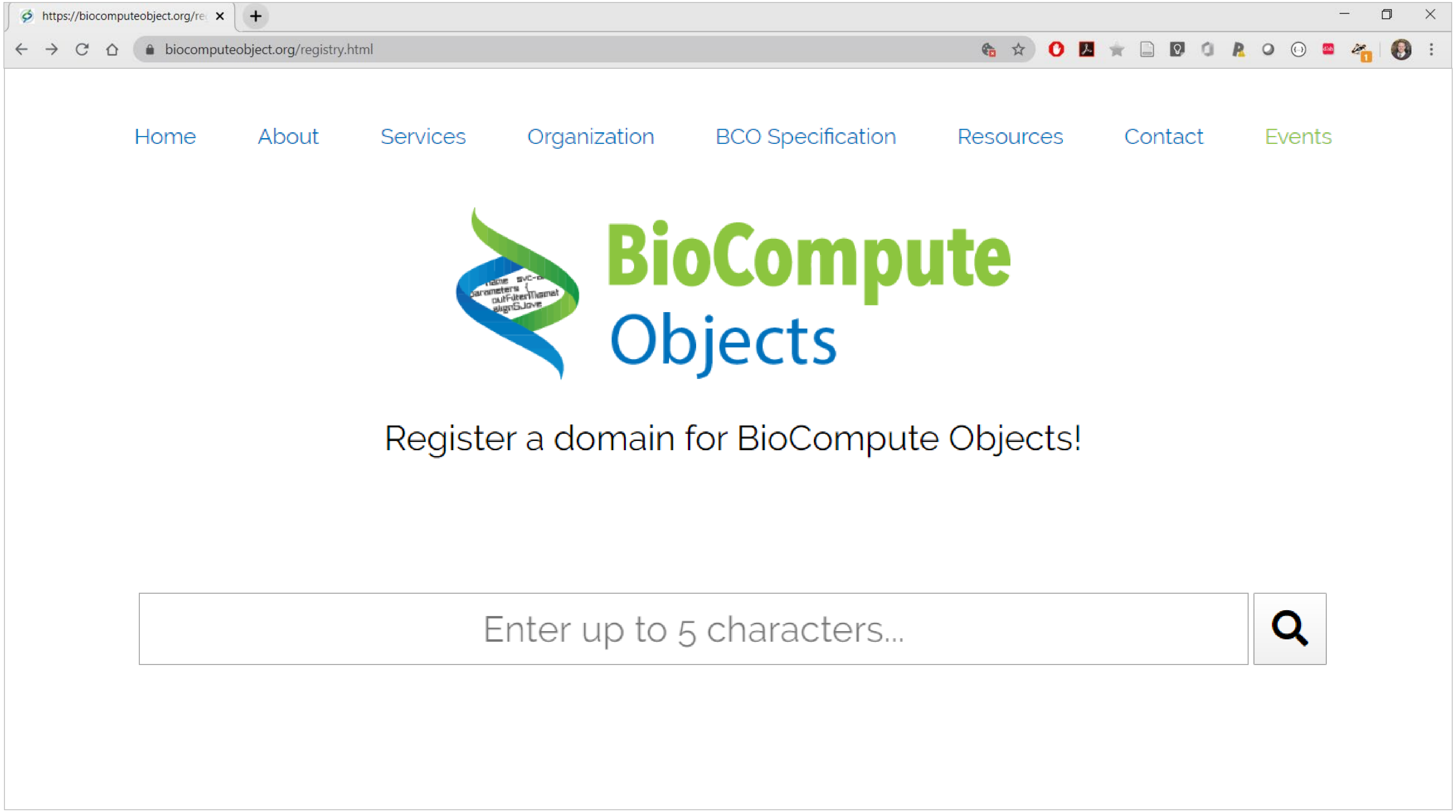
Domain

User Specified

Domain	Owner
NVR	Jim@Novartis.com
MRK	Jack@merck.com
FZR	Susan@Pfizer.com
…	…

BioCompute Registry

BioCompute Registry: Initial Draft



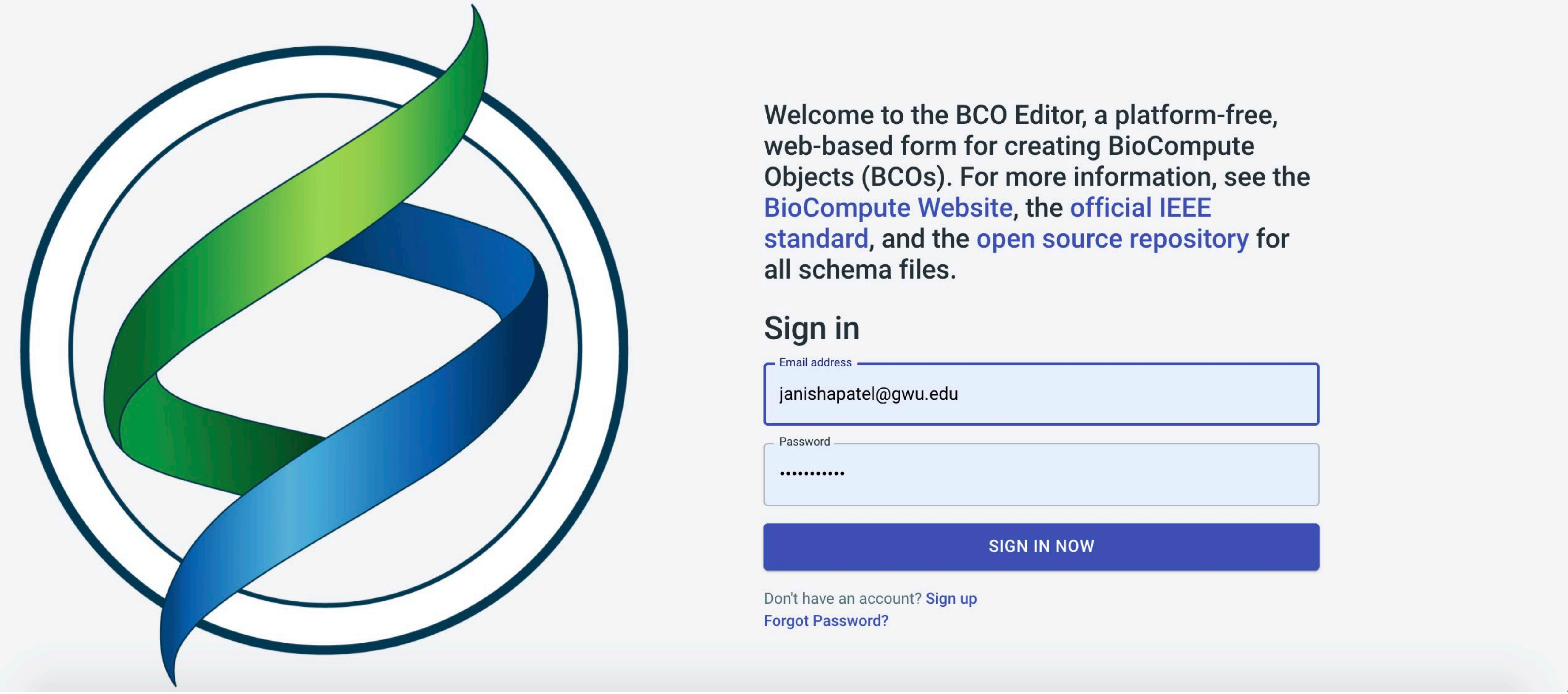
https://biocomputeobject.org/re

biocomputeobject.org/registry.html

Register a domain for BioCompute Objects!

Enter up to 5 characters...

BioCompute Portal: Initial Draft



Welcome to the BCO Editor, a platform-free, web-based form for creating BioCompute Objects (BCOs). For more information, see the BioCompute Website, the official IEEE standard, and the open source repository for all schema files.

Sign in

Email address: janishapatel@gwu.edu

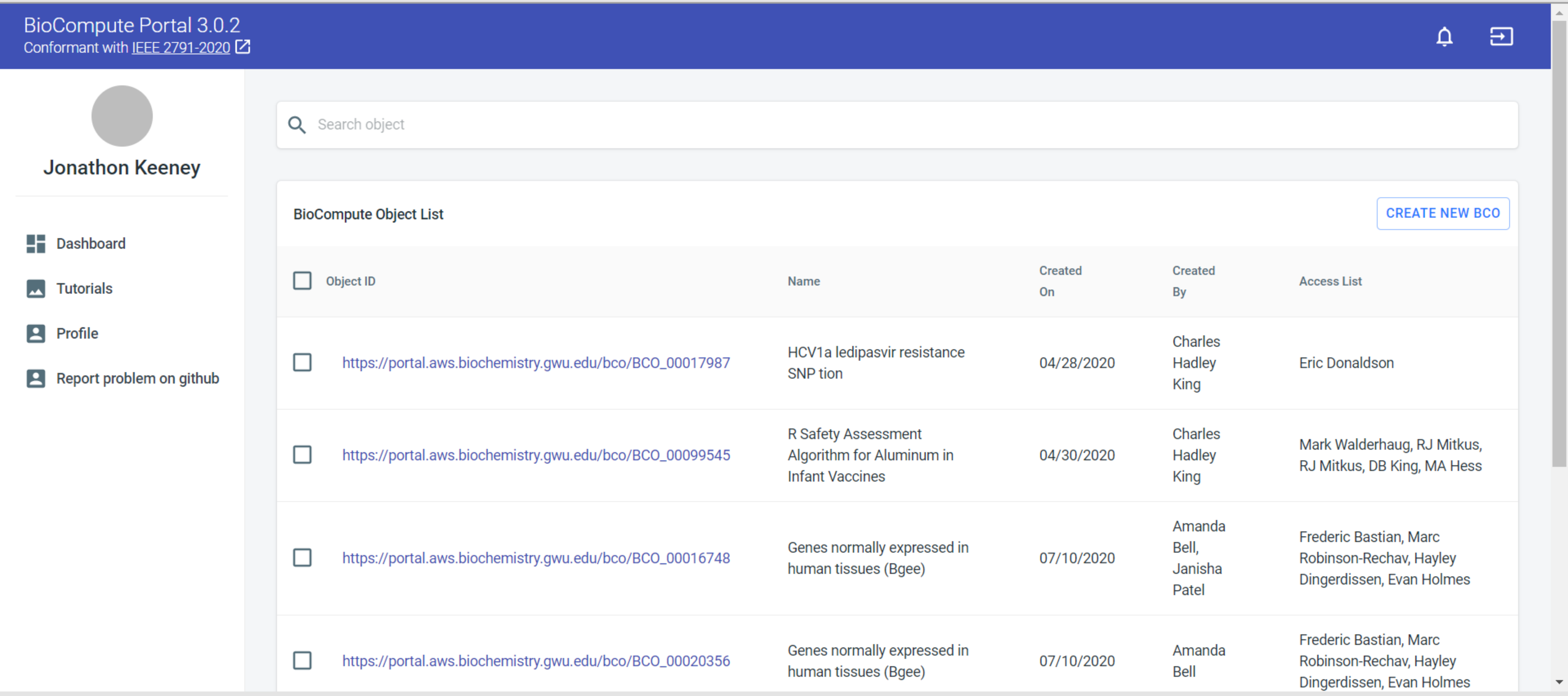
Password: ••••••••••••

Don't have an account? Sign up

Forgot Password?

https://portal.aws.biochemistry.gwu.edu/sign-in

BioCompute Portal: Initial Draft



BioCompute Portal 3.0.2

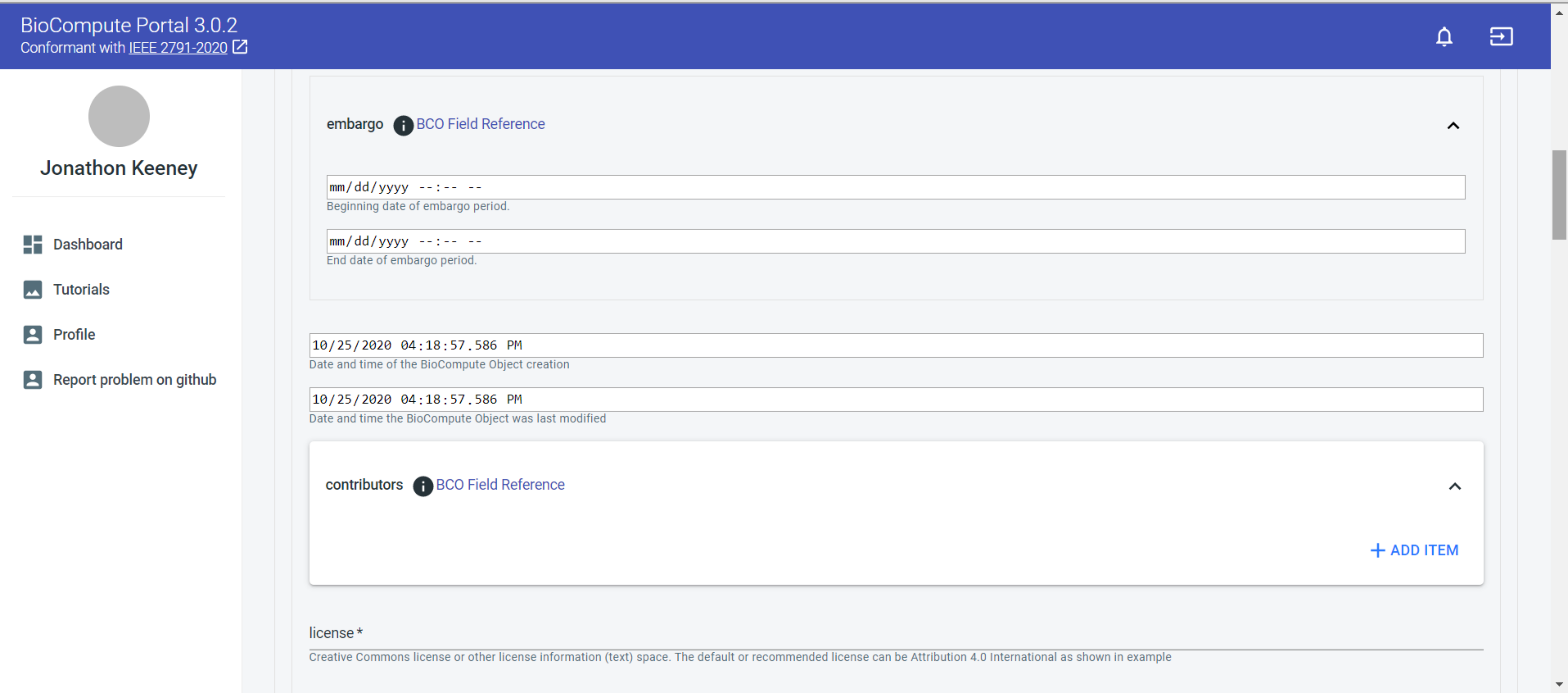
Conformant with IEEE 2791-2020

Jonathon Keeney

BioCompute Object List

□	Object ID	Name	Created On	Created By	Access List
□	https://portal.aws.biochemistry.gwu.edu/bco/BCO_00017987	HCV1a ledipasvir resistance SNP tion	04/28/2020	Charles Hadley King	Eric Donaldson
□	https://portal.aws.biochemistry.gwu.edu/bco/BCO_00099545	R Safety Assessment Algorithm for Aluminum in Infant Vaccines	04/30/2020	Charles Hadley King	Mark Walderhaug, RJ Mitkus, RJ Mitkus, DB King, MA Hess
□	https://portal.aws.biochemistry.gwu.edu/bco/BCO_00016748	Genes normally expressed in human tissues (Bgee)	07/10/2020	Amanda Bell, Janisha Patel	Frederic Bastian, Marc Robinson-Rechav, Hayley Dingerdissen, Evan Holmes
□	https://portal.aws.biochemistry.gwu.edu/bco/BCO_00020356	Genes normally expressed in human tissues (Bgee)	07/10/2020	Amanda Bell	Frederic Bastian, Marc Robinson-Rechav, Hayley Dingerdissen, Evan Holmes

BioCompute Portal: Initial Draft



BioCompute Portal 3.0.2
Conformant with IEEE 2791-2020

Jonathon Keeney

embargo BCO Field Reference

mm/dd/yyyy --:-- --
Beginning date of embargo period.

mm/dd/yyyy --:-- --
End date of embargo period.

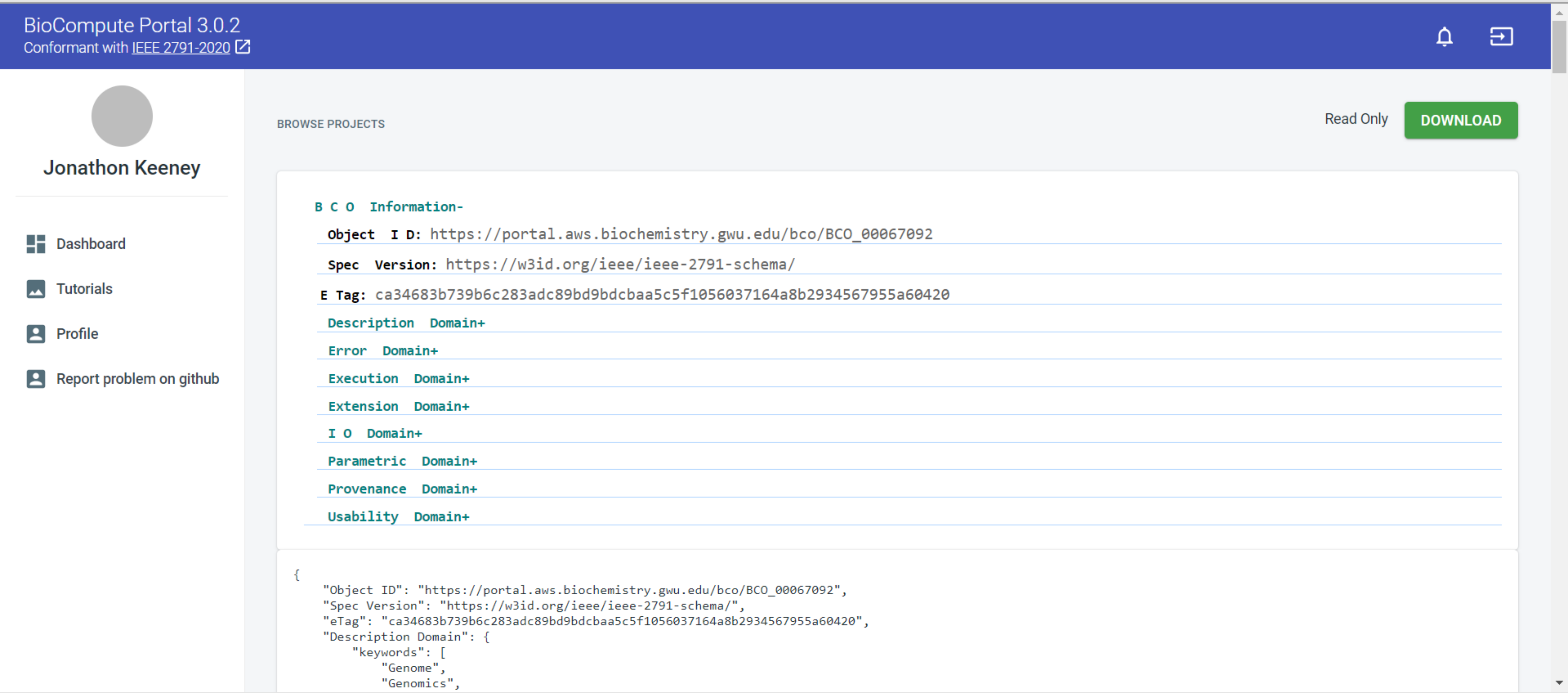
10/25/2020 04:18:57.586 PM
Date and time of the BioCompute Object creation

10/25/2020 04:18:57.586 PM
Date and time the BioCompute Object was last modified

contributors BCO Field Reference

license *
Creative Commons license or other license information (text) space. The default or recommended license can be Attribution 4.0 International as shown in example

BioCompute Portal: Initial Draft



BCO Information-

Object ID: https://portal.aws.biochemistry.gwu.edu/bco/BCO_00067092

Spec Version: https://w3id.org/ieee/ieee-2791-schema/

ETag: ca34683b739b6c283adc89bd9bdcbaa5c5f1056037164a8b2934567955a60420

Description Domain+

Error Domain+

Execution Domain+

Extension Domain+

IO Domain+

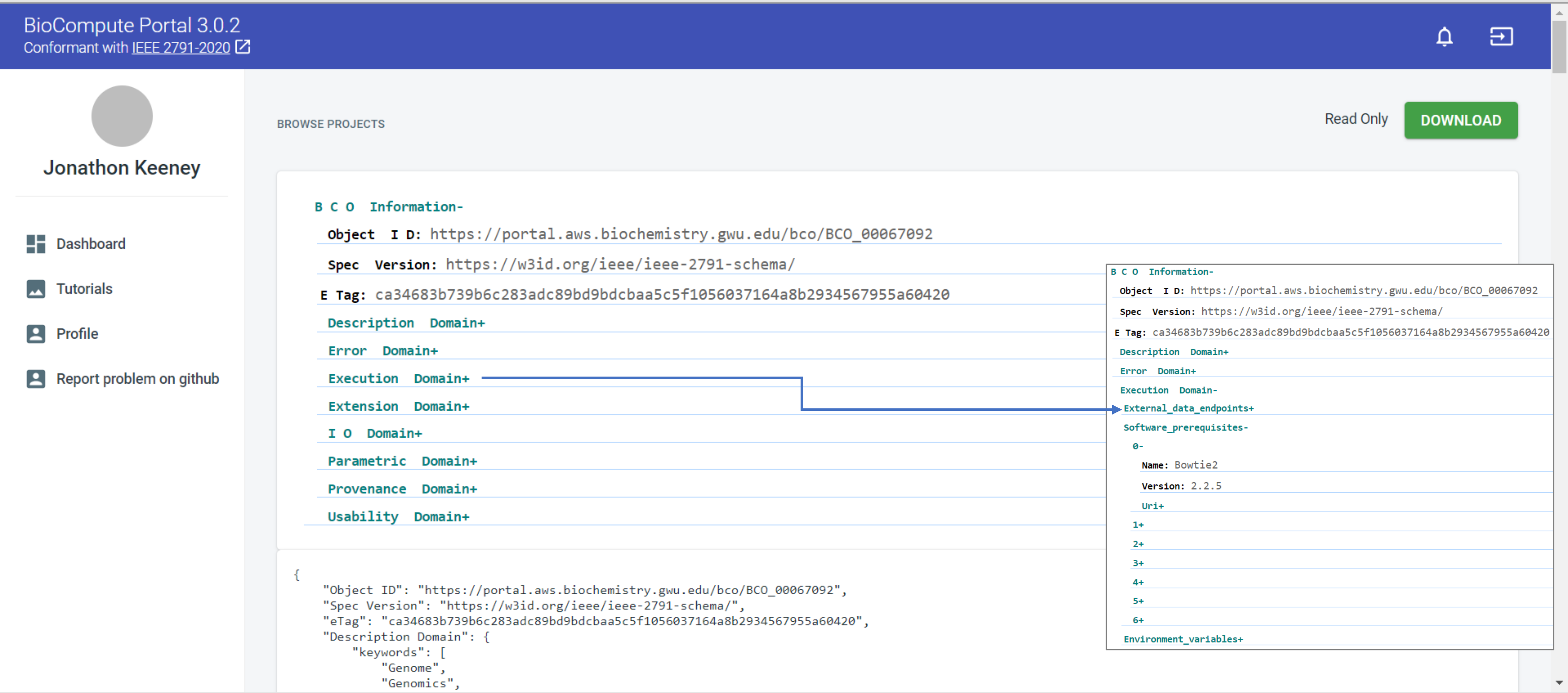
Parametric Domain+

Provenance Domain+

Usability Domain+

```
{
    "Object ID": "https://portal.aws.biochemistry.gwu.edu/bco/BCO_00067092",
    "Spec Version": "https://w3id.org/ieee/ieee-2791-schema/",
    "eTag": "ca34683b739b6c283adc89bd9bdcbaa5c5f1056037164a8b2934567955a60420",
    "Description Domain": {
        "keywords": [
            "Genome",
            "Genomics",
```

BioCompute Portal: Initial Draft



BioCompute Portal 3.0.2
Conformant with IEEE 2791-2020

Jonathon Keeney

BCO Information-

Object ID: https://portal.aws.biochemistry.gwu.edu/bco/BCO_00067092

Spec Version: https://w3id.org/ieee/ieee-2791-schema/

E Tag: ca34683b739b6c283adc89bd9bdcbaa5c5f1056037164a8b2934567955a60420

Description Domain+

Error Domain+

Execution Domain+

Extension Domain+

I O Domain+

Parametric Domain+

Provenance Domain+

Usability Domain+

```
{
    "Object ID": "https://portal.aws.biochemistry.gwu.edu/bco/BCO_00067092",
    "Spec Version": "https://w3id.org/ieee/ieee-2791-schema/",
    "eTag": "ca34683b739b6c283adc89bd9bdcbaa5c5f1056037164a8b2934567955a60420",
    "Description Domain": {
        "keywords": [
            "Genome",
            "Genomics",
```

BCO Information-

Object ID: https://portal.aws.biochemistry.gwu.edu/bco/BCO_00067092

Spec Version: https://w3id.org/ieee/ieee-2791-schema/

E Tag: ca34683b739b6c283adc89bd9bdcbaa5c5f1056037164a8b2934567955a60420

Description Domain+

Error Domain+

Execution Domain-

External_data_endpoints+

Software_prerequisites-

0-

Name: Bowtie2

Version: 2.2.5

Uri+

1+

2+

3+

4+

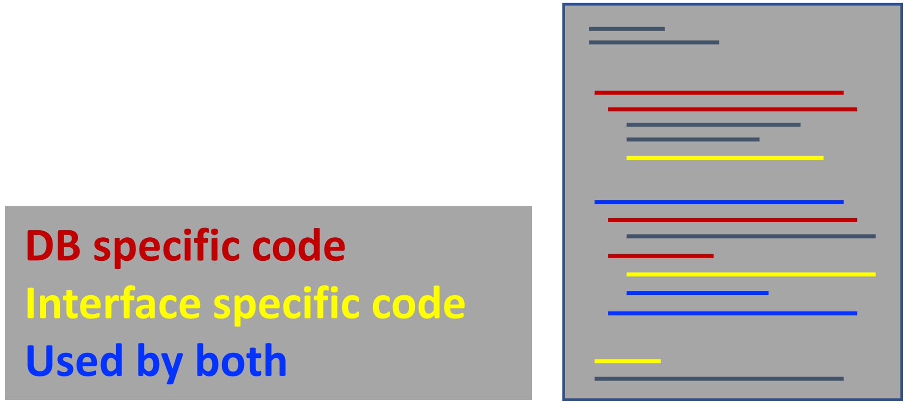
5+

6+

Environment_variables+

BioCompute DB: Original Design Refinements

- Functional, with ideas for improvement
- Monolithic design
  - Interface and database are pieces of the same thing
  - Portal and database joined in source code
  - Full source code must be implemented for instantiation
  - Lack of modularization makes it difficult to change database or Portal without affecting the other
    - E.g. new API features require changing code that would potentially affect the entire project
  - Most fixes/features *ad hoc* code patches



DB specific code

Interface specific code

Used by both

BioCompute DB: Original Design Refinements

- Other ideas
  - Better support for BCOs built outside of Portal
    - Support for CLI submission
    - Methods for external resources (eg HIVE or Galaxy) to deposit BCOs
  - Need better object manipulation tools
    - “Edit Mode” for BCOs in progress and not published
    - Need better support for advanced permissions (view/download/edit/share)
    - Advanced searching
    - Links to external resources or supporting data

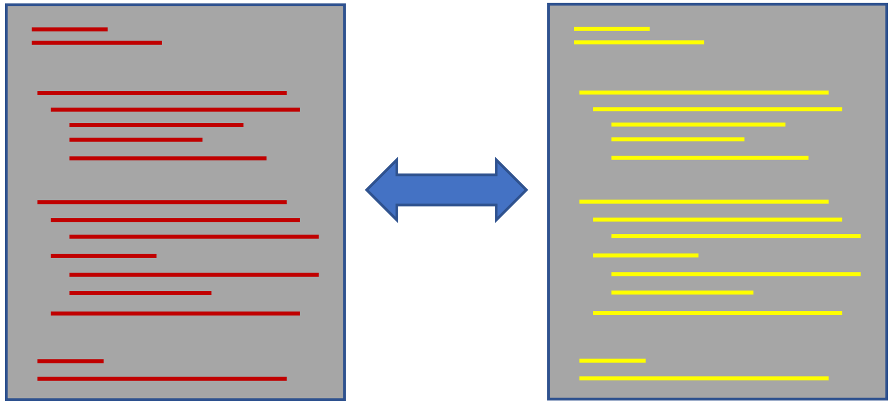
BioCompute DB: New Design

- Fully separated code bases
  - Portal and database
  - Greater flexibility now and in the future
    - E.g. repository-based system like NCBI

DB specific code

Interface specific code

Used by both



BioCompute DB: New Design

- API-driven architecture
  - Easy to interface with
  - Partitioned into “classes”
    - Reusable by other programmers
    - Less effort to expand
  - API follows CRUD paradigm
    - Create, Read, Update, Delete; corresponding to POST, GET, PATCH, DELETE
    - Allows interaction with API to be characterized by operation type
  - Template-based request system
    - Each type of CRUD operation has a defined set of templates allowed for that type
    - Reduces the number of requests that have to be made

BioCompute DB: New Design

- Examples of request types:

POST	Create new BCO; Convert existing object between schemas
GET	Validate JSON object against a schema; request available schemas from a server; search for objects based on fields
PATCH	Modify an existing BioCompute Object based on fields
DELETE	Delete an object based on fields

- GET will likely be most important for cross-organization interaction
  - Users external to an organization can retrieve information about objects created by other users based on characteristics like authors, pipeline steps, time created, etc.

BioCompute DB: New Design Progress

☑ API architecture is in beta

Currently undergoing testing for validating BCOs against IEEE schema

☑ Code modularized

Can now be engaged by command line or browser

☑ Template system implemented

Expanded API functionality

□ Plan to establish first public facing repository in 2 months

House all created objects

□ Convert API into python package

Rapid deployment and configuration

□ Wrap API in OpenAPI/Swagger framework

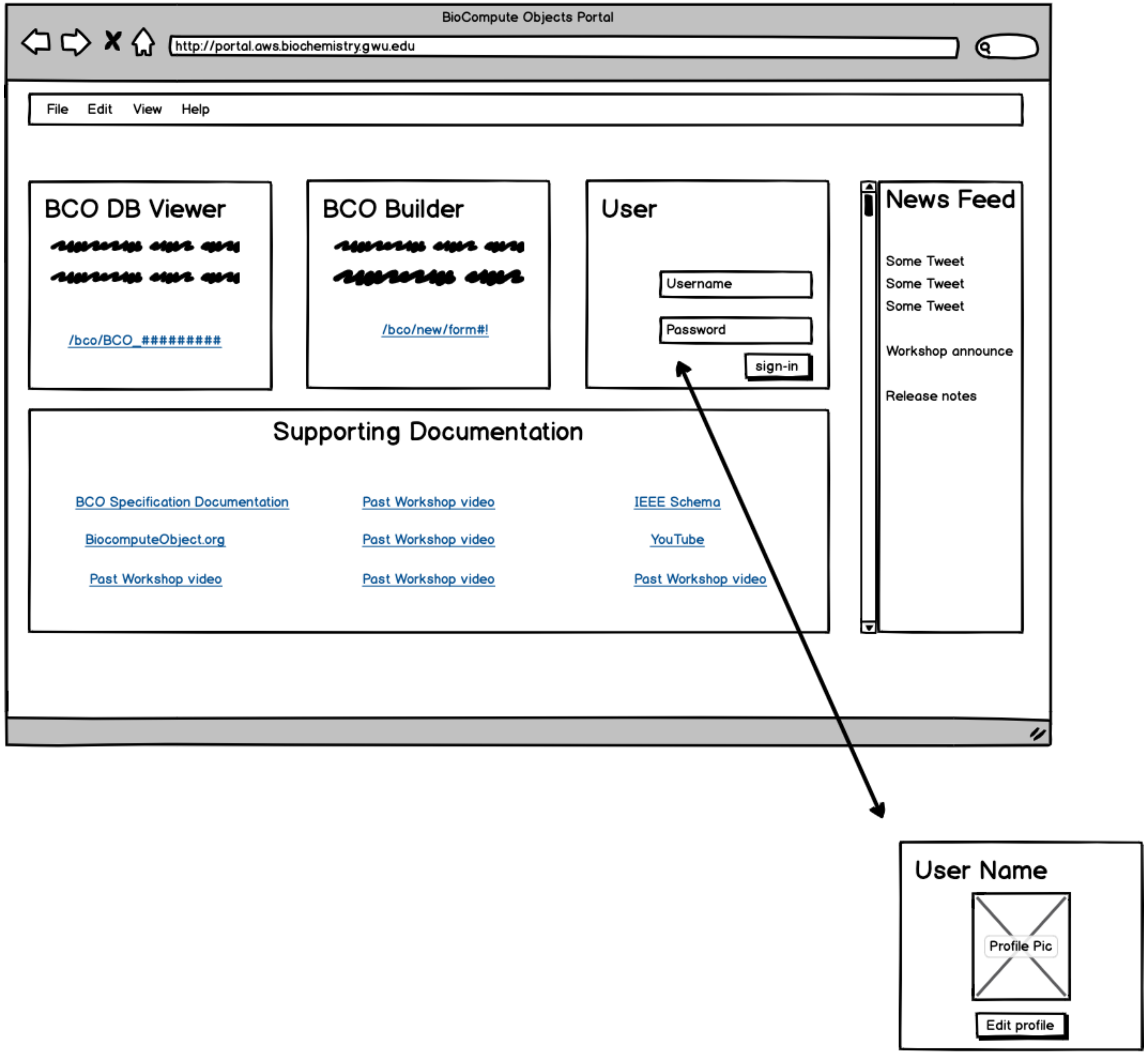
Enables standardization of results

□ Strengthen security of API

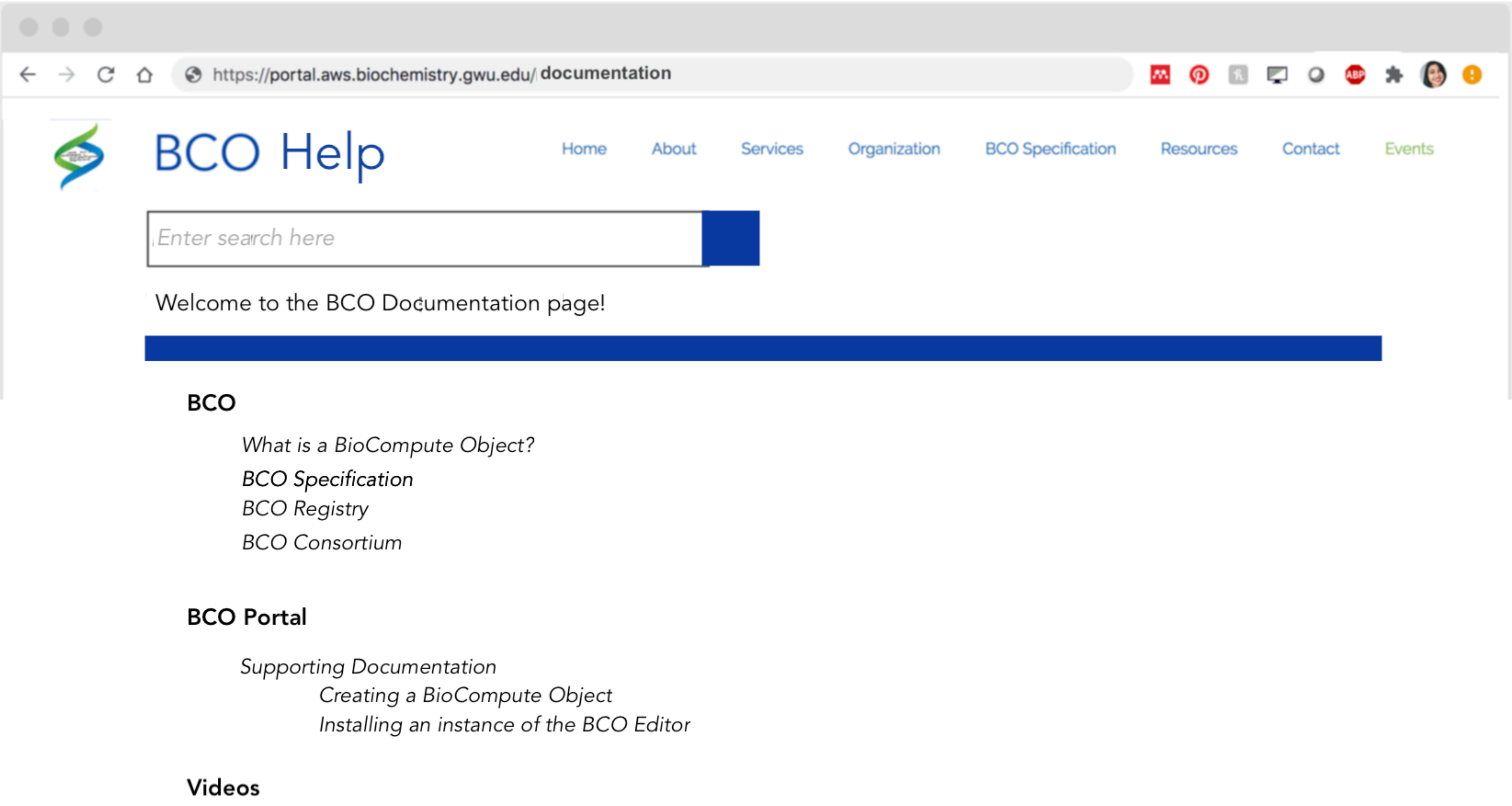
Token-based security authentication, public-key cryptography

BioCompute Objects

BioCompute DB Mockups: Main Page



BioCompute DB Mockups: Documentation Page



Welcome to the BCO Documentation page!

BCO

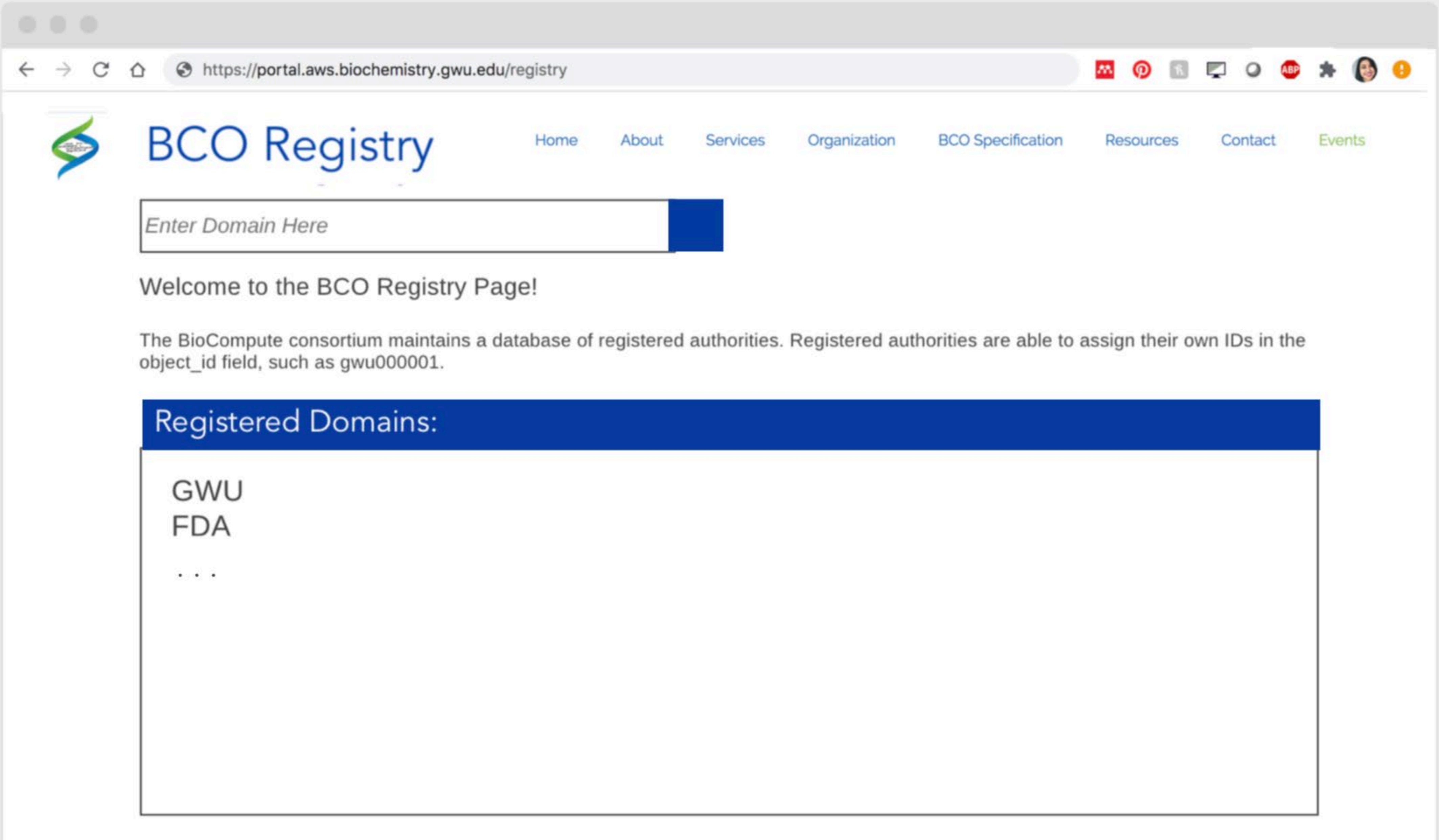
What is a BioCompute Object?
BCO Specification
BCO Registry
BCO Consortium

BCO Portal

Supporting Documentation
Creating a BioCompute Object
Installing an instance of the BCO Editor

Videos

BioCompute DB Mockups: Registry Page



BCO Registry

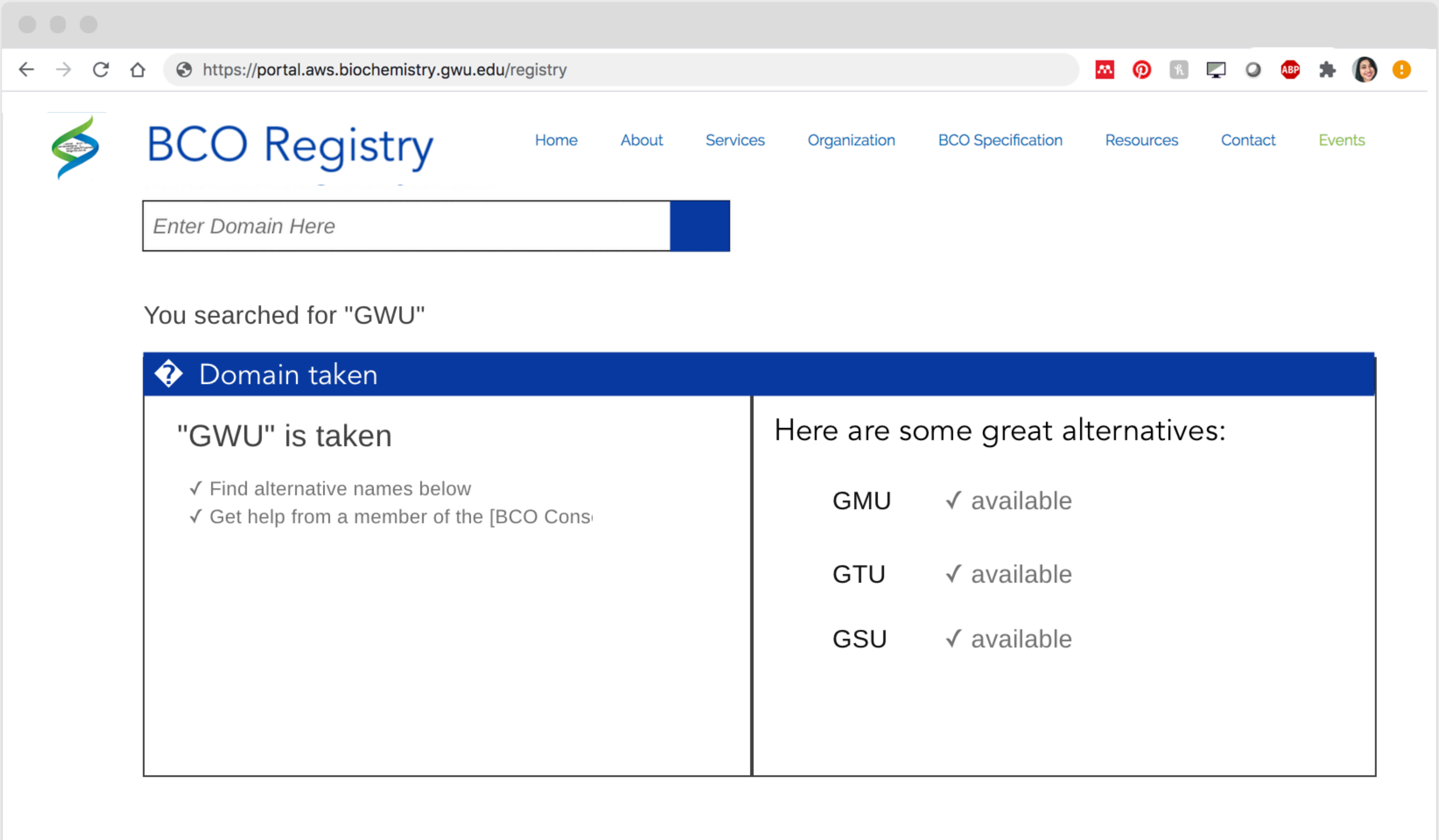
Welcome to the BCO Registry Page!

The BioCompute consortium maintains a database of registered authorities. Registered authorities are able to assign their own IDs in the object_id field, such as gwu000001.

Registered Domains:

GWU
FDA
. . .

BioCompute DB Mockups: Registry Page



Enter Domain Here

You searched for "GWU"

Domain taken

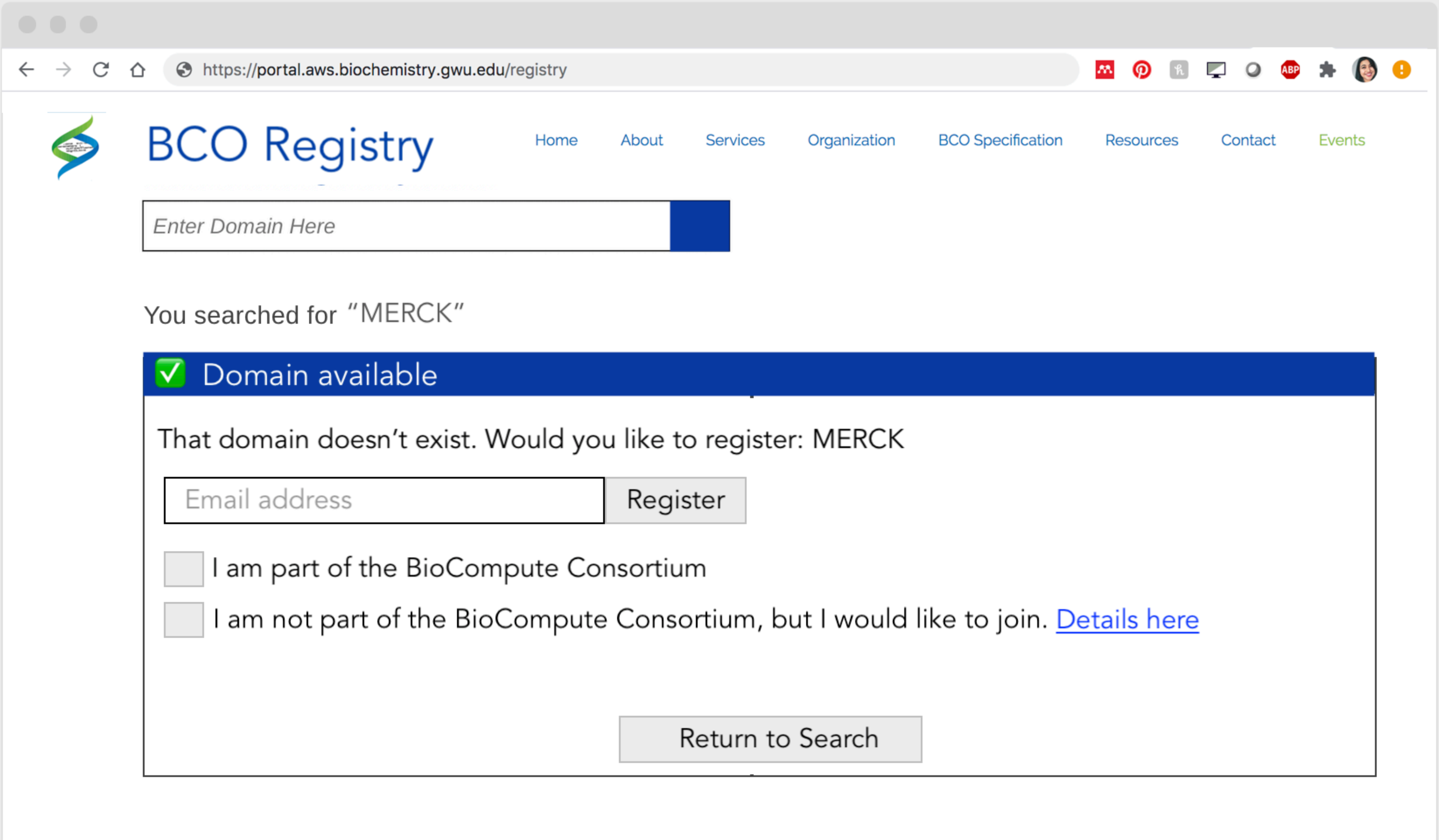
"GWU" is taken

- ✓ Find alternative names below
- ✓ Get help from a member of the [BCO Cons

Here are some great alternatives:

GMU	✓ available
GTU	✓ available
GSU	✓ available

BioCompute DB Mockups: Registry Page



You searched for “MERCK”

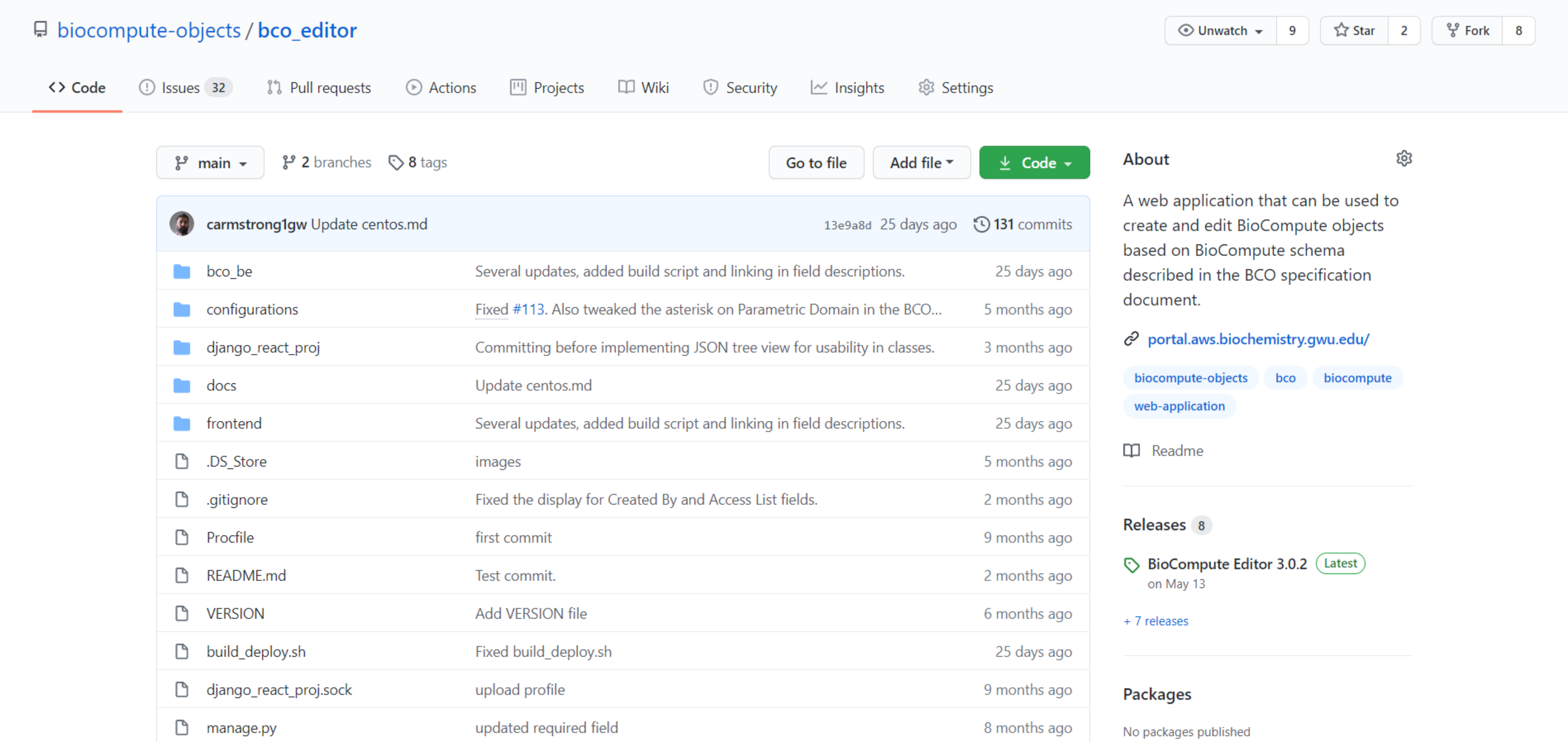
✅ Domain available

That domain doesn’t exist. Would you like to register: MERCK

□ I am part of the BioCompute Consortium

□ I am not part of the BioCompute Consortium, but I would like to join. Details here

https://github.com/biocompute-objects/bco_editor



carmstrong1gw Update centos.md 13e9a8d 25 days ago 131 commits

bco_be	Several updates, added build script and linking in field descriptions.	25 days ago
configurations	Fixed #113. Also tweaked the asterisk on Parametric Domain in the BCO...	5 months ago
django_react_proj	Committing before implementing JSON tree view for usability in classes.	3 months ago
docs	Update centos.md	25 days ago
frontend	Several updates, added build script and linking in field descriptions.	25 days ago
.DS_Store	images	5 months ago
.gitignore	Fixed the display for Created By and Access List fields.	2 months ago
Procfile	first commit	9 months ago
README.md	Test commit.	2 months ago
VERSION	Add VERSION file	6 months ago
build_deploy.sh	Fixed build_deploy.sh	25 days ago
django_react_proj.sock	upload profile	9 months ago
manage.py	updated required field	8 months ago

About

A web application that can be used to create and edit BioCompute objects based on BioCompute schema described in the BCO specification document.

portal.aws.biochemistry.gwu.edu/

biocompute-objects bco biocompute web-application

Readme

Releases 8

BioCompute Editor 3.0.2 Latest
on May 13

+ 7 releases

Packages

No packages published

Next Steps

- Finish beta testing API architecture
- Test template system
- Establish public facing repository
- Convert API into python package
- Wrap API in OpenAPI/Swagger framework
- Strengthen security of API
- Beta test entire system
- Package code for deployment
- Host documentation

Acknowledgements and Contact

Raja Mazumder, Ph.D., PI

Professor

The George Washington University

mazumder@gwu.edu

Jonathon Keeney, Ph.D., Co-I

Assistant Research Professor

The George Washington University

keeneyjg@gwu.edu

Hadley King

Technical Lead

The George Washington University

Chris Armstrong

Development Lead

The George Washington University

Janisha Patel

Outreach Lead

The George Washington University

BioCompute Objects