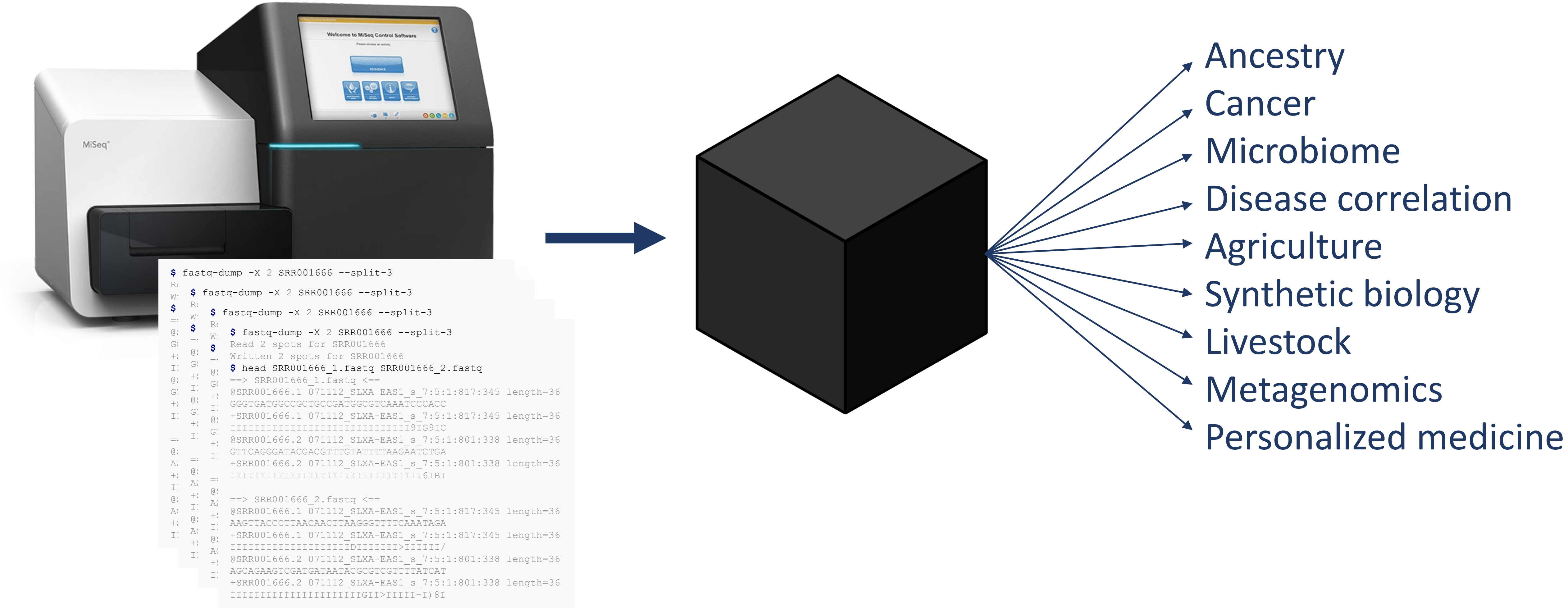
BioCompute

BioCompute
Objects

HTS Data Flows



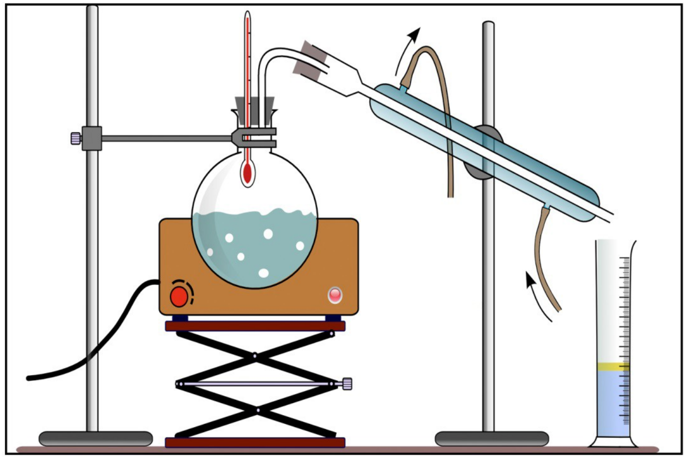
```
$ fastq-dump -X 2 SRR001666 --split-3
Read 2 spots for SRR001666
Written 2 spots for SRR001666
$ head SRR001666_1.fastq SRR001666_2.fastq
==> SRR001666_1.fastq <==
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 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
IIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

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

Challenge: Workflow Communication

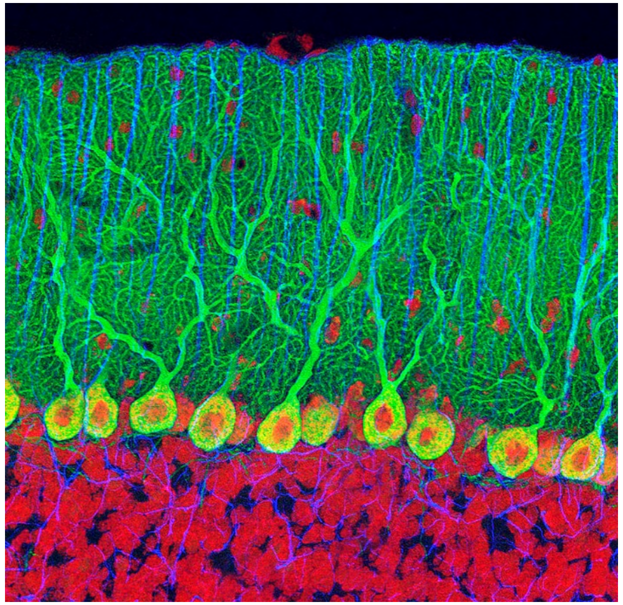
Input

Strain of animal
Tissue type
Protein Lot Number
Humidity
Microscope
…



conditions

Output

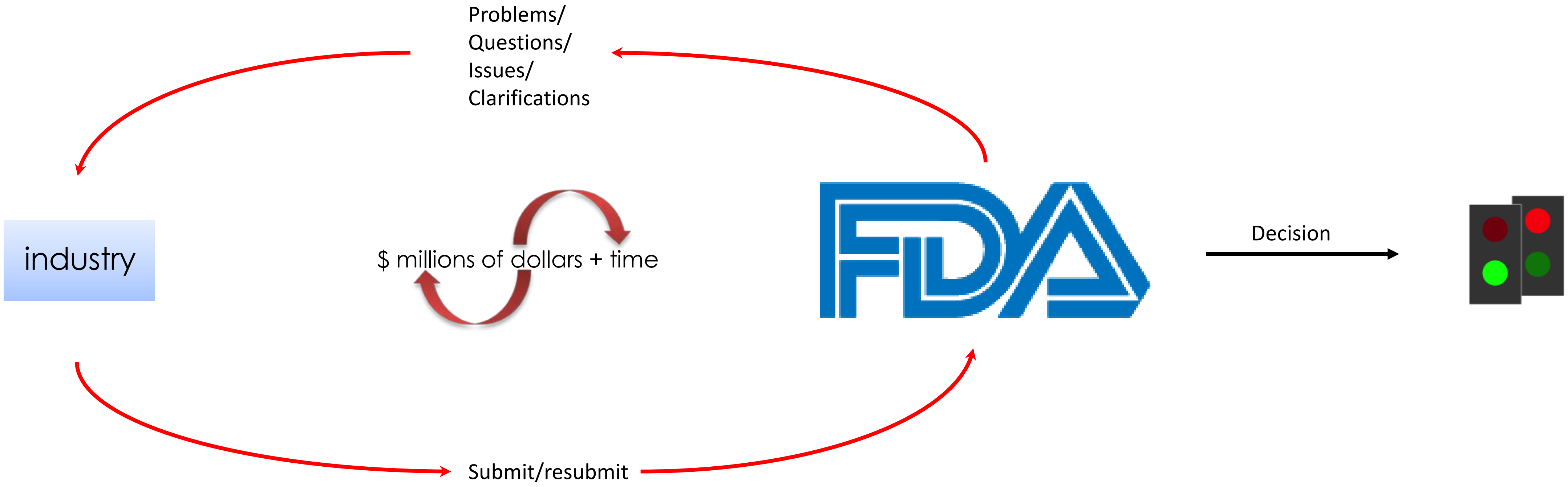


generalized experiment **inputs** **parameters** **outputs**

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

Analogy: wet lab experiments

Wasted Time and Money



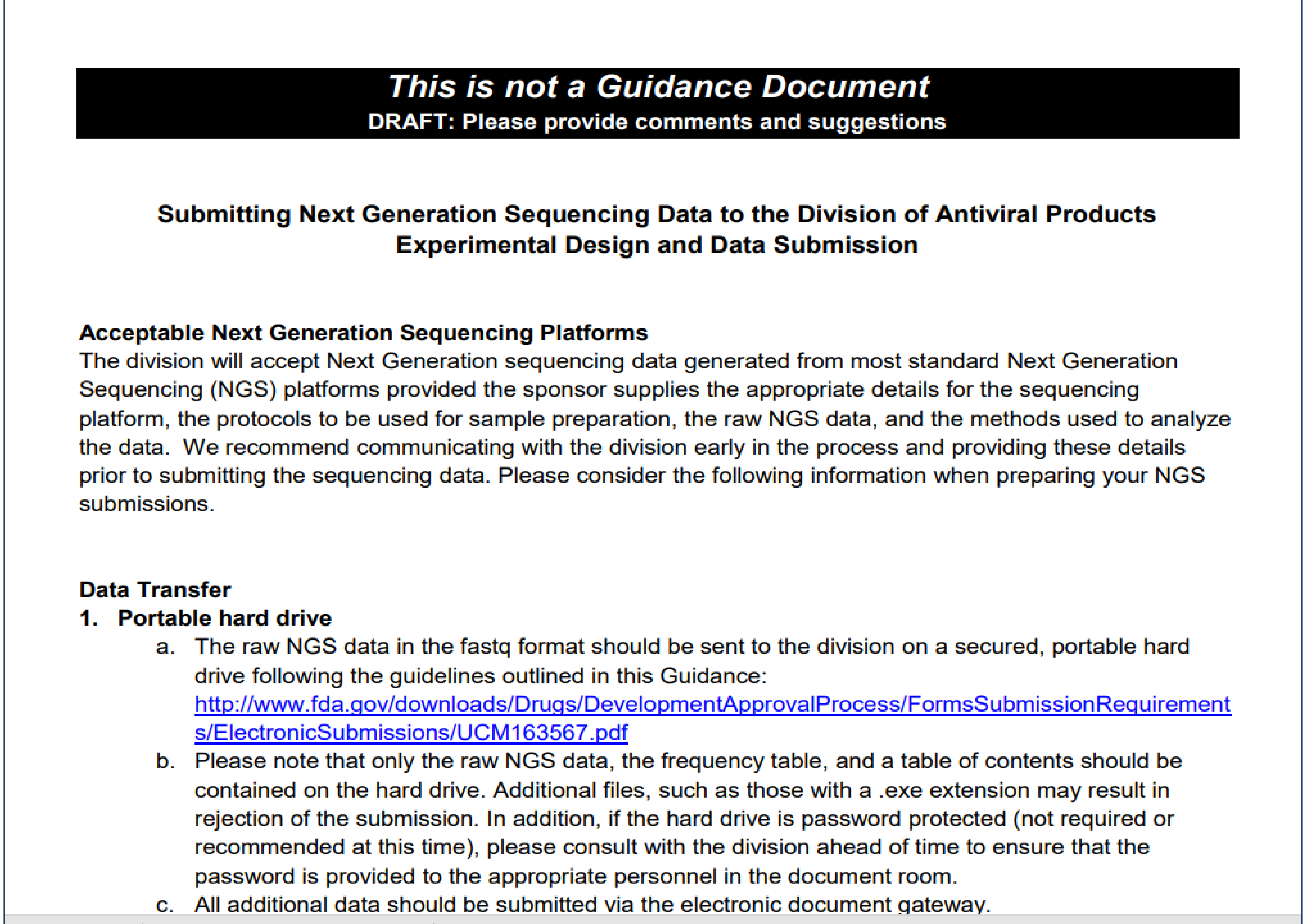
Problems/
Questions/
Issues/
Clarifications

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

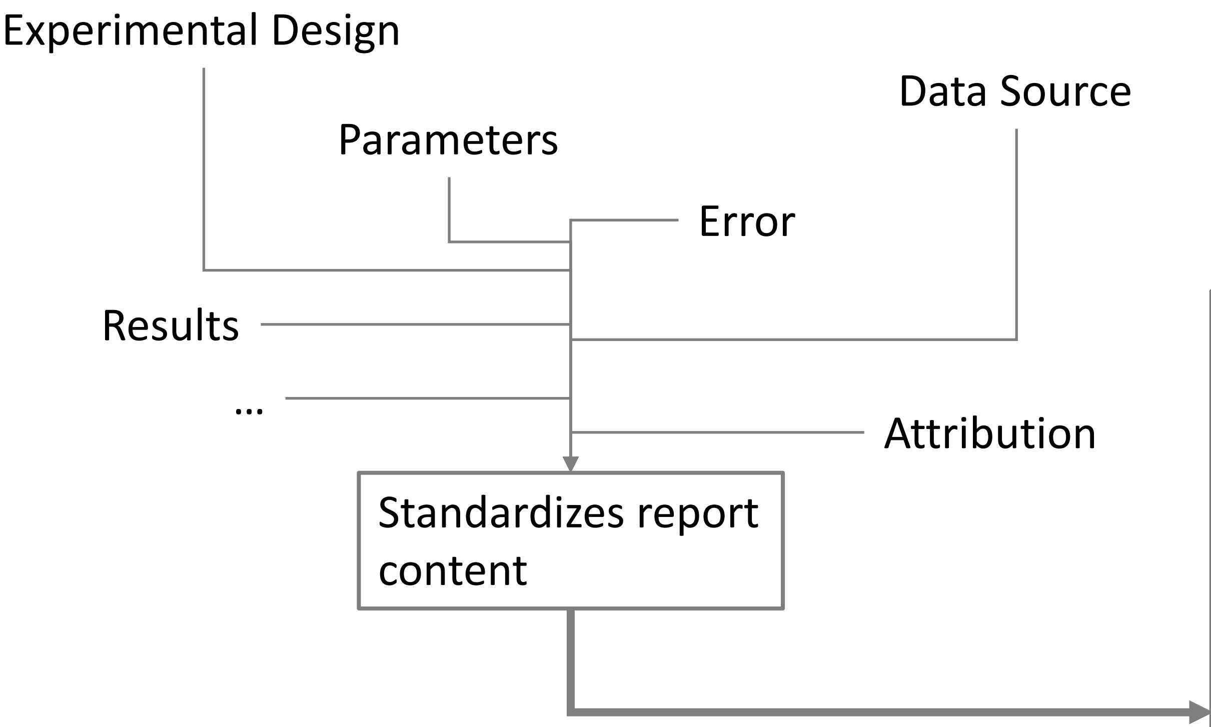
Solution: BioCompute

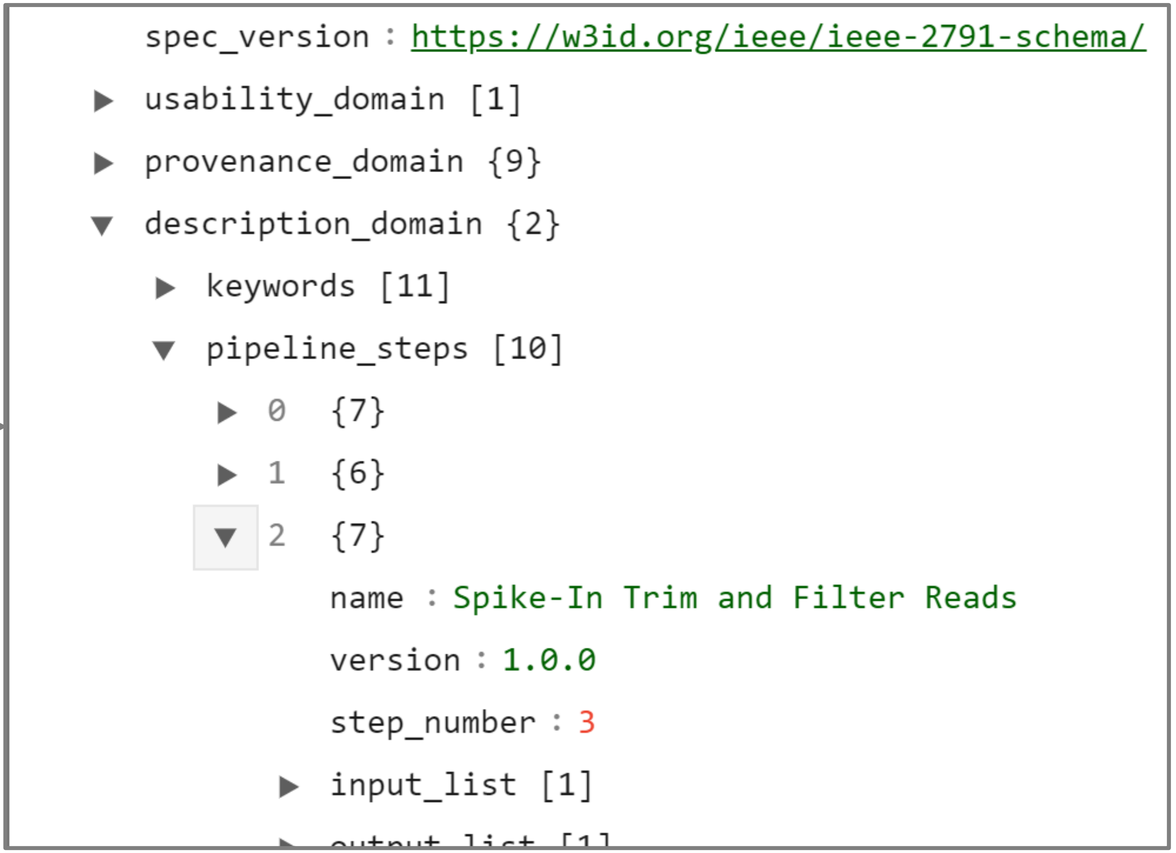
IEEE approved standard for communicating bioinformatic analysis workflows

- 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





```
    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]
```

Machine readability enables customized views

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

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 : None
```

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 Objects

BioCompute participants

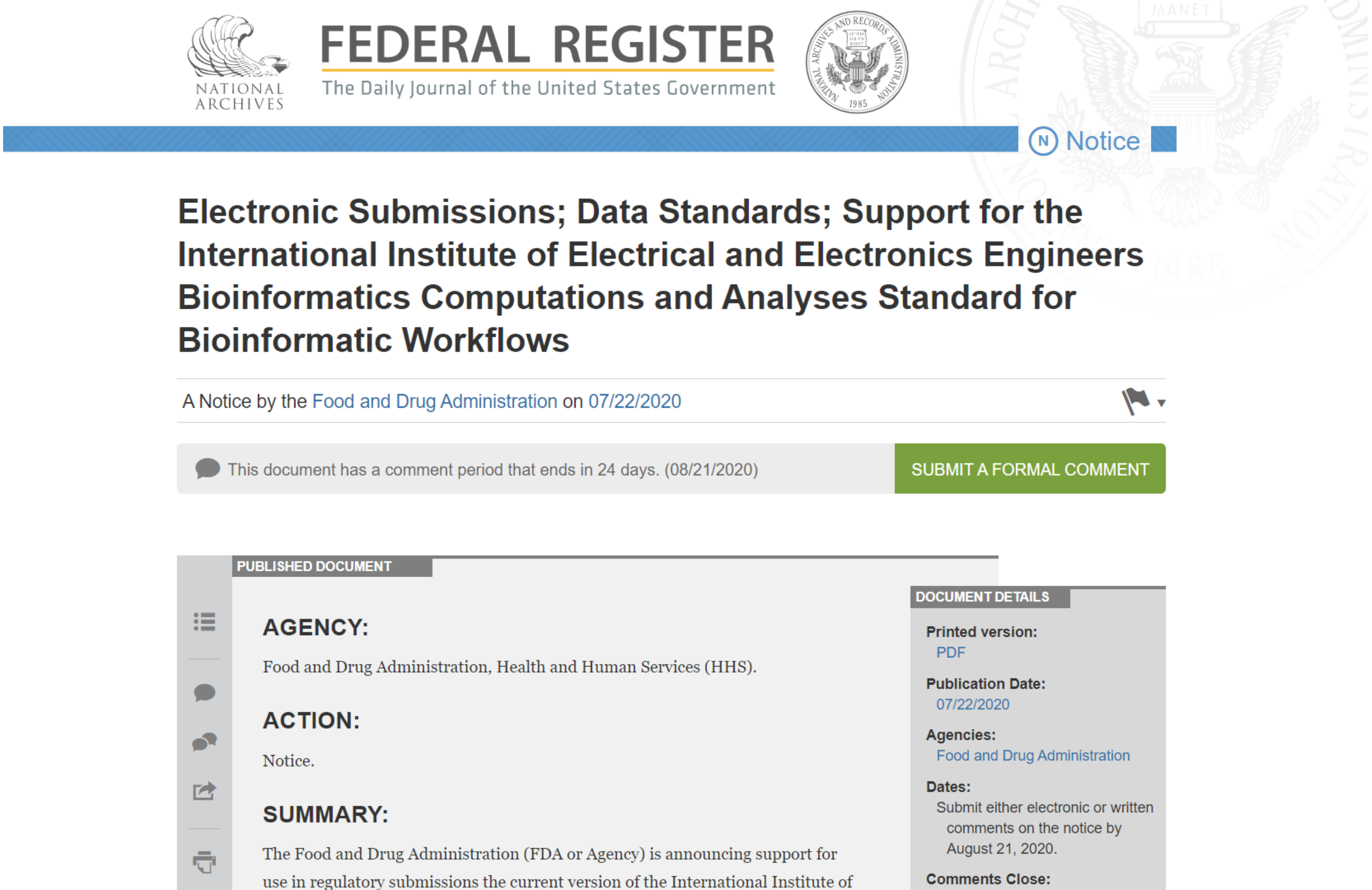
Children's National
precisionFDA
ATTAIN
Global Alliance for Genomics & Health
DISEASE ONTOLOGY
EPINOMICS
NIH NATIONAL CANCER INSTITUTE
CUROVERSE
Forum for Collaborative HIV Research
DNAnexus
HARVARD UNIVERSITY
NATIONAL LIBRARY OF MEDICINE
Google
CDC CENTERS FOR DISEASE CONTROL AND PREVENTION
EMBL-EBI
FHIR
FHIR Genomics
paradigm4
Oxford e-Research Centre
UNIVERSITY OF OXFORD
NCBI
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
CU NY

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



Electronic Submissions; Data Standards; Support for the International Institute of Electrical and Electronics Engineers Bioinformatics Computations and Analyses Standard for Bioinformatic Workflows

A Notice by the Food and Drug Administration on 07/22/2020

This document has a comment period that ends in 24 days. (08/21/2020)

AGENCY:

Food and Drug Administration, Health and Human Services (HHS).

ACTION:

Notice.

SUMMARY:

The Food and Drug Administration (FDA or Agency) is announcing support for use in regulatory submissions the current version of the International Institute of

DOCUMENT DETAILS

Printed version:
PDF

Publication Date:
07/22/2020

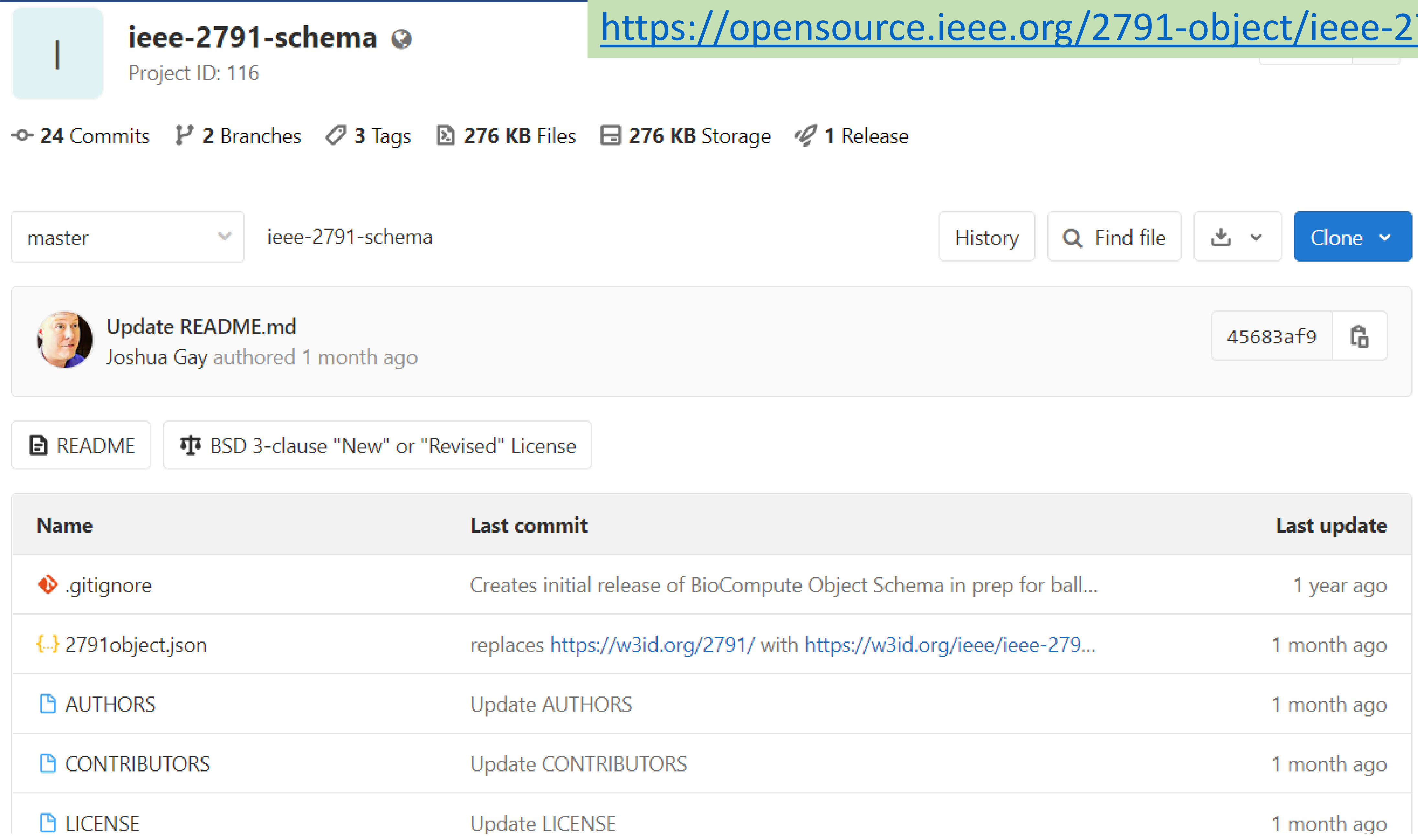
Agencies:
Food and Drug Administration

Dates:
Submit either electronic or written comments on the notice by August 21, 2020.

Comments Close:

BioCompute Schema Files

https://opensource.ieee.org/2791-object/ieee-2791-schema/



ieee-2791-schema

Project ID: 116

24 Commits · 2 Branches · 3 Tags · 276 KB Files · 276 KB Storage · 1 Release

Update README.md

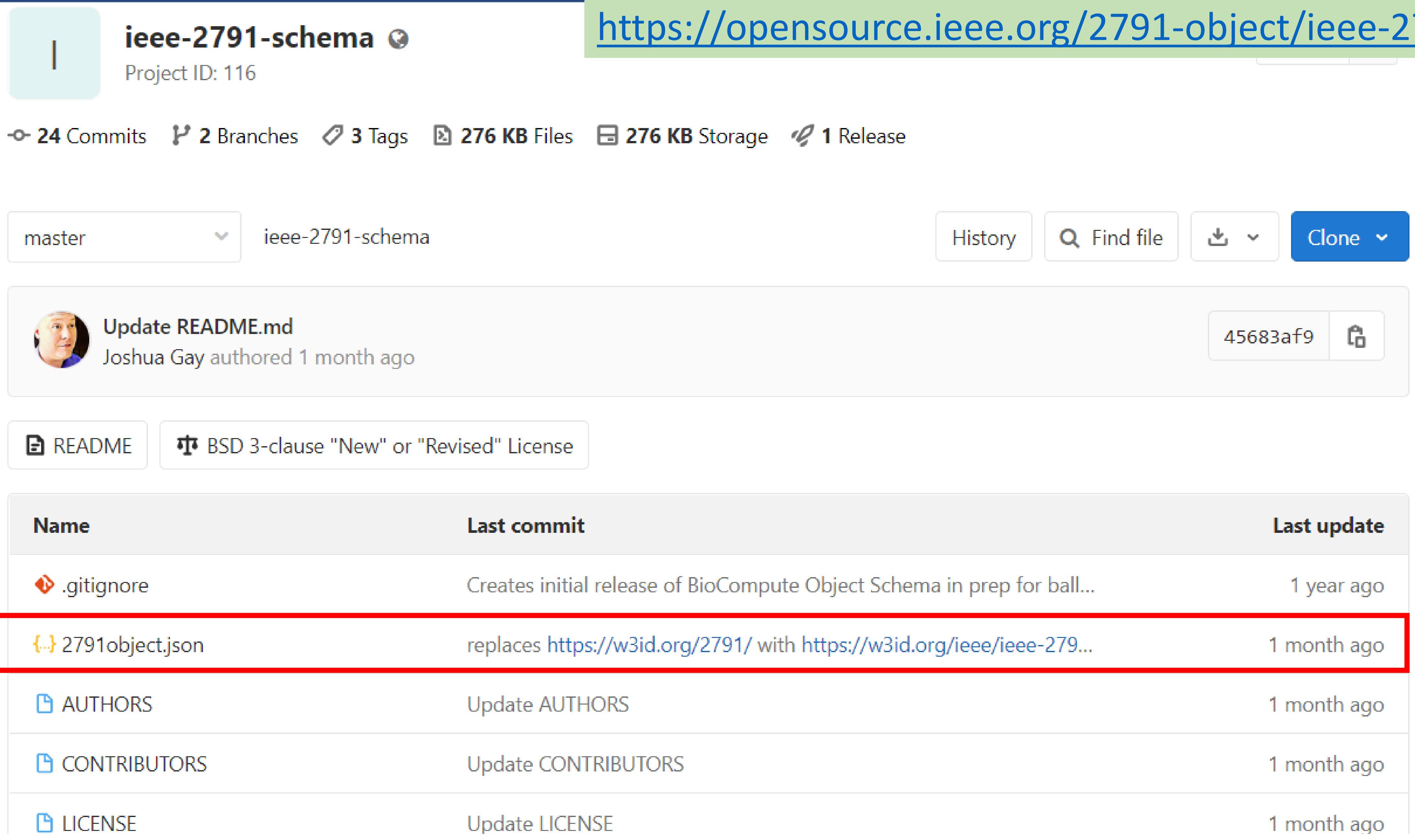
Joshua Gay authored 1 month ago

45683af9

Name	Last commit	Last update
.gitignore	Creates initial release of BioCompute Object Schema in prep for ball...	1 year ago
2791object.json	replaces https://w3id.org/2791/ with https://w3id.org/ieee/ieee-279...	1 month ago
AUTHORS	Update AUTHORS	1 month ago
CONTRIBUTORS	Update CONTRIBUTORS	1 month ago
LICENSE	Update LICENSE	1 month ago

BioCompute Schema Files

https://opensource.ieee.org/2791-object/ieee-2791-schema/



I

ieee-2791-schema

Project ID: 116

24 Commits **2** Branches **3** Tags **276 KB** Files **276 KB** Storage **1** Release

Update README.md

Joshua Gay authored 1 month ago

45683af9

Name	Last commit	Last update
.gitignore	Creates initial release of BioCompute Object Schema in prep for ball...	1 year ago
{..} 2791object.json	replaces https://w3id.org/2791/ with https://w3id.org/ieee/ieee-279...	1 month ago
AUTHORS	Update AUTHORS	1 month ago
CONTRIBUTORS	Update CONTRIBUTORS	1 month ago
LICENSE	Update LICENSE	1 month ago

Key Features of a BCO

- Abstract away workflow based on commonalities
  - Platform/tool/protocol independent
- Usability Domain
  - Free text description
- Data provenance
  - Data manifest, track files from beginning to end
  - Track user attribution (“authoredBy,” “contributedBy,” “reviewedBy,” etc.)
- Verification Kit
  - Error Domain + IO Domain
  - Sanity check: given the input files and the inherent error, is the output this analysis claims to have gotten valid?
- Extensible
  - Extension Domain
  - Open source repository
- Embargo Domain
  - Prevent others from viewing a BCO for any amount of time

BioCompute Objects

BCOs for Biocuration

- Workflow is abstracted

Within environment:

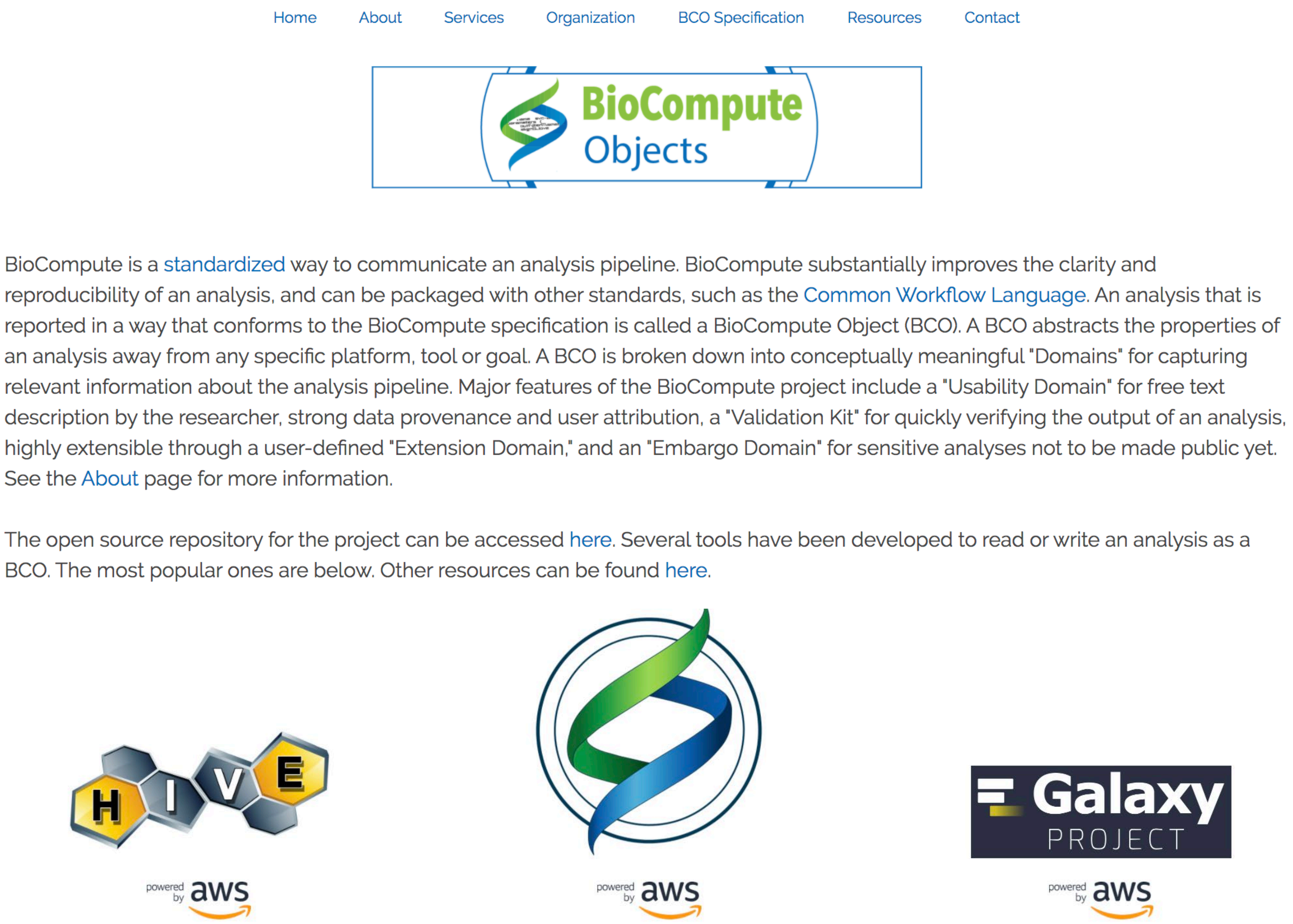
[Input → transformation steps/parameters → output]

+ Relevant annotation

- Strong provenance and user attribution
  - Features are native to BCO
- Extensible
  - Unique features of datasets can be captured without losing the benefits of standardization

Advantages

- Data can be worked with programmatically
  - Know exactly what kind of data to expect and in exactly what format
- Standardization of data curation for teams
  - OncoMX consists of multiple geographically distributed individuals
- Flexibility
  - BCOs standardize a workflow description while preserving the ability to describe all of the unique features of curation



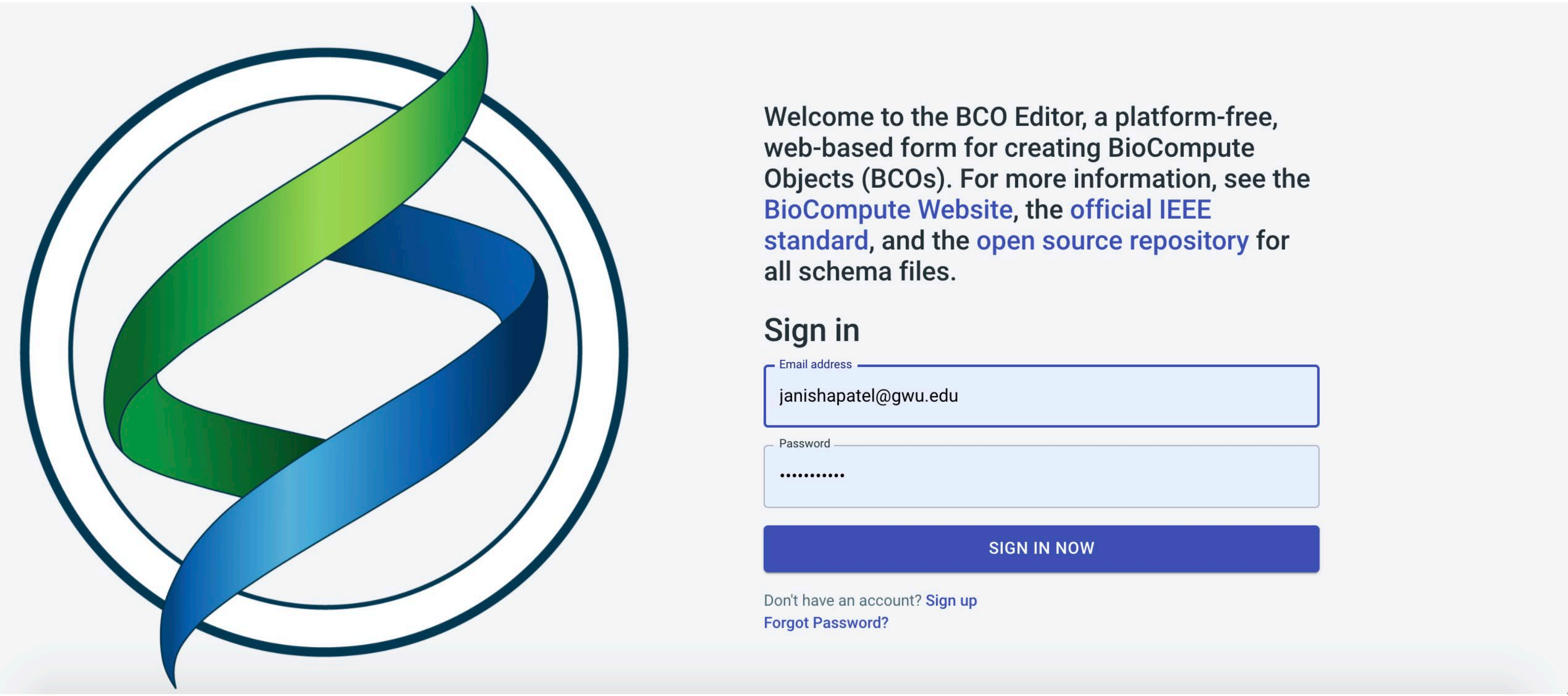
Home About Services Organization BCO Specification Resources Contact
BioCompute Objects

BioCompute is a standardized way to communicate an analysis pipeline. BioCompute substantially improves the clarity and reproducibility of an analysis, and can be packaged with other standards, such as the Common Workflow Language. An analysis that is reported in a way that conforms to the BioCompute specification is called a BioCompute Object (BCO). A BCO abstracts the properties of an analysis away from any specific platform, tool or goal. A BCO is broken down into conceptually meaningful "Domains" for capturing relevant information about the analysis pipeline. Major features of the BioCompute project include a "Usability Domain" for free text description by the researcher, strong data provenance and user attribution, a "Validation Kit" for quickly verifying the output of an analysis, highly extensible through a user-defined "Extension Domain," and an "Embargo Domain" for sensitive analyses not to be made public yet. See the About page for more information.

The open source repository for the project can be accessed here. Several tools have been developed to read or write an analysis as a BCO. The most popular ones are below. Other resources can be found here.

https://biocomputeobject.org/

BioCompute Portal



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

Contact

Jonathon Keeney, Ph.D.
Assistant Research Professor
The George Washington University
keeneyjg@gwu.edu

BioCompute Objects