New Tools for Building BioCompute Objects on DNAnexus, Seven Bridges, and Galaxy for deposit into an access-controlled BioComputeDB

September 13th, 2022

Jonathon Keeney, Ph.D.
George Washington University

Introduction Agenda

- Brief introduction to BioCompute
- Previous work
- Updates
- Today’s workshop
  - BioComputeDB
- Introduction of other speakers

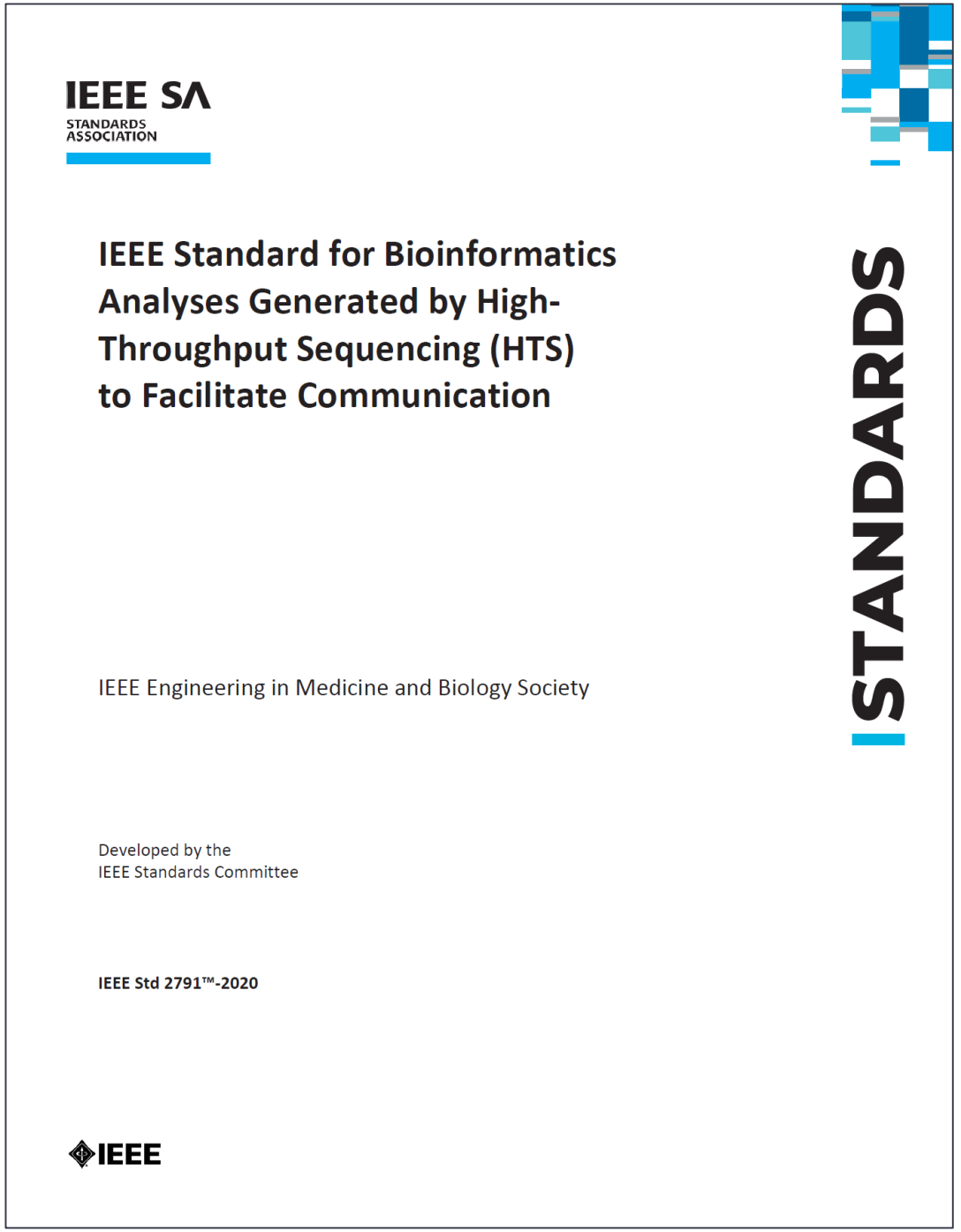
Need for Guidelines

- Lack of standards or guidelines for documentation and reporting of workflows
- Rich metadata not captured in workflow languages

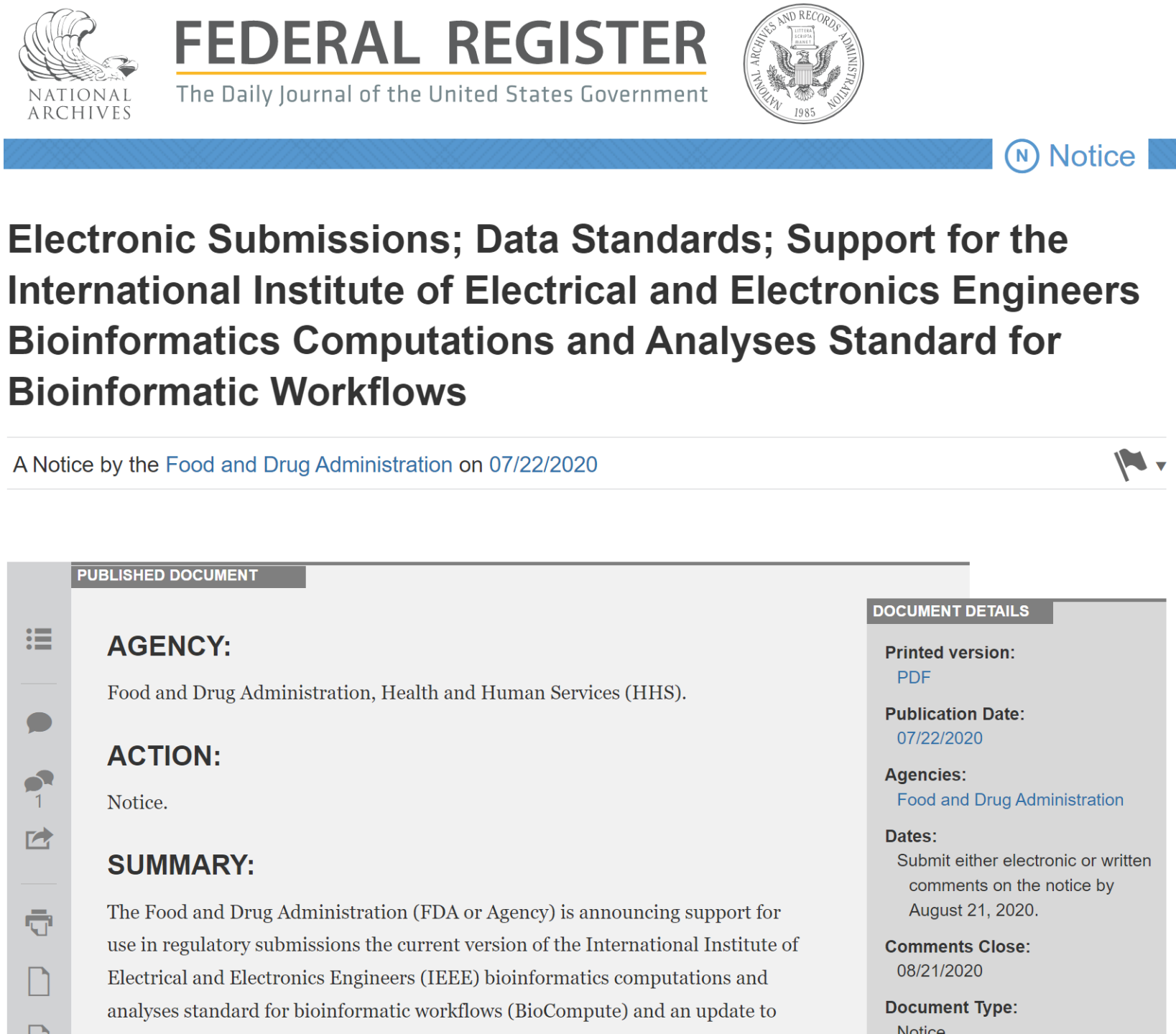
Purpose

- What kind of data needs to be present in order to understand a computational analysis?
- How does that data need to be represented?

- Dozens of working groups
- 3 workshops
- Multiple draft specifications and schemas



Accepted at CBER, CDER, and CFSAN for most drug applications



NATIONAL ARCHIVES

FEDERAL REGISTER

The Daily Journal of the United States Government

Notice

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

PUBLISHED DOCUMENT

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 Electrical and Electronics Engineers (IEEE) bioinformatics computations and analyses standard for bioinformatic workflows (BioCompute) and an update to

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:
08/21/2020

Document Type:
Notice

Domain-based organization

- Who contributed to the work?
- What was their role?

Provenance Domain

User attribution and role in the work

Execution Domain

Description Domain

Parametric Domain

Usability Domain

IO Domain

Error Domain

Extension Domain

Object Information

Domain-based organization

- What was the execution environment?
  - e.g. HIVE, Galaxy, Seven Bridges, DNAnexus, command line, etc.
- Environmental variables?
- Software prerequisites?

Provenance Domain

Execution Domain

Execution environment needed to run the analysis

Description Domain

Parametric Domain

Usability Domain

IO Domain

Error Domain

Extension Domain

Object Information

Domain-based organization

- Keywords
- Description space for each step
- Input/output to describe what each step is doing

Provenance Domain

Execution Domain

Description Domain

Description of each step, including dependencies and IO

Parametric Domain

Usability Domain

IO Domain

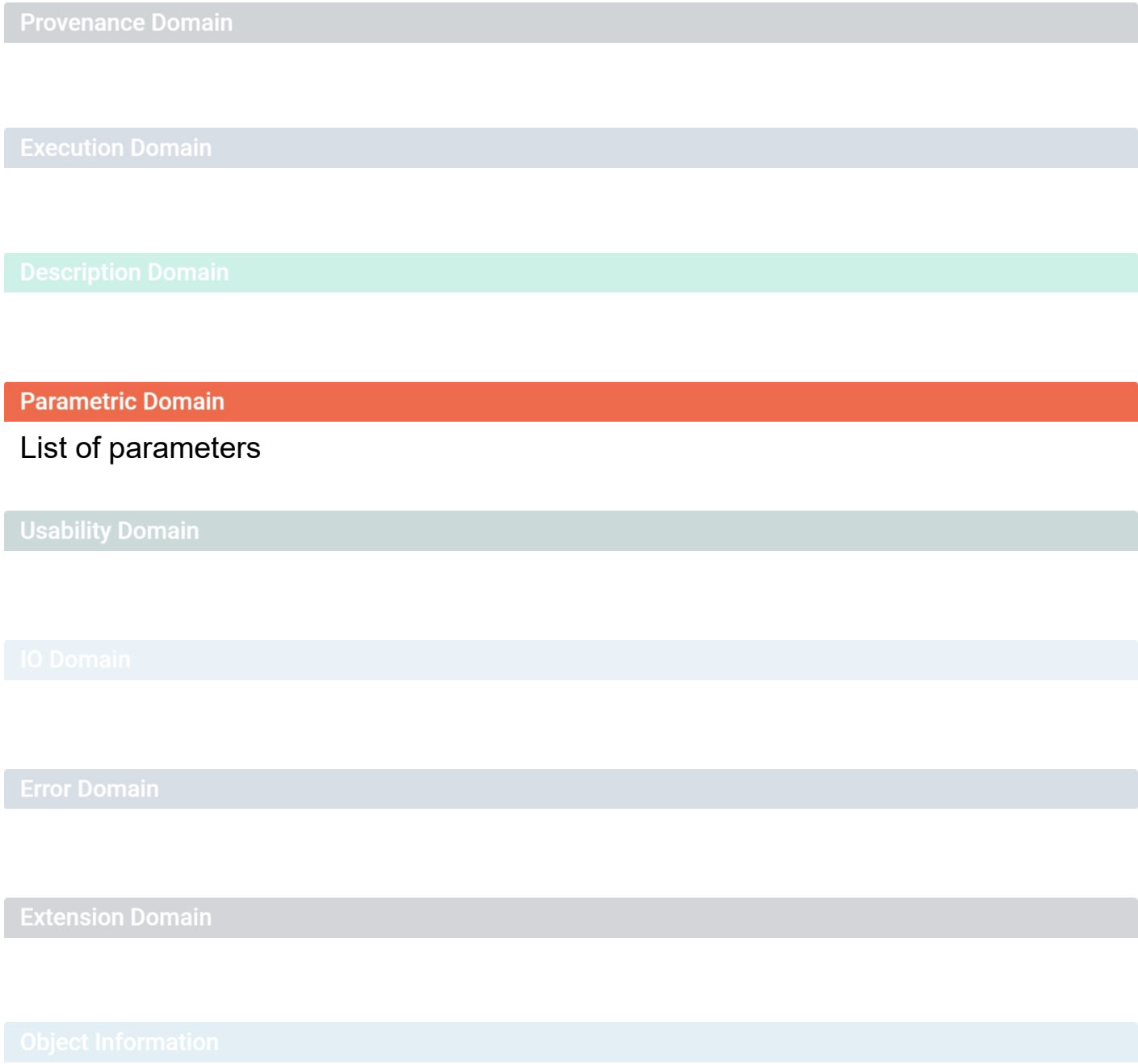
Error Domain

Extension Domain

Object Information

Domain-based organization

- List of all parameters
  - Each parameter associated with a step



Domain-based organization

- Overview of pipeline
  - Purpose, goals, outcomes, context
  - Any other relevant information or author comments

Provenance Domain

Execution Domain

Description Domain

Parametric Domain

Usability Domain

Free text description

IO Domain

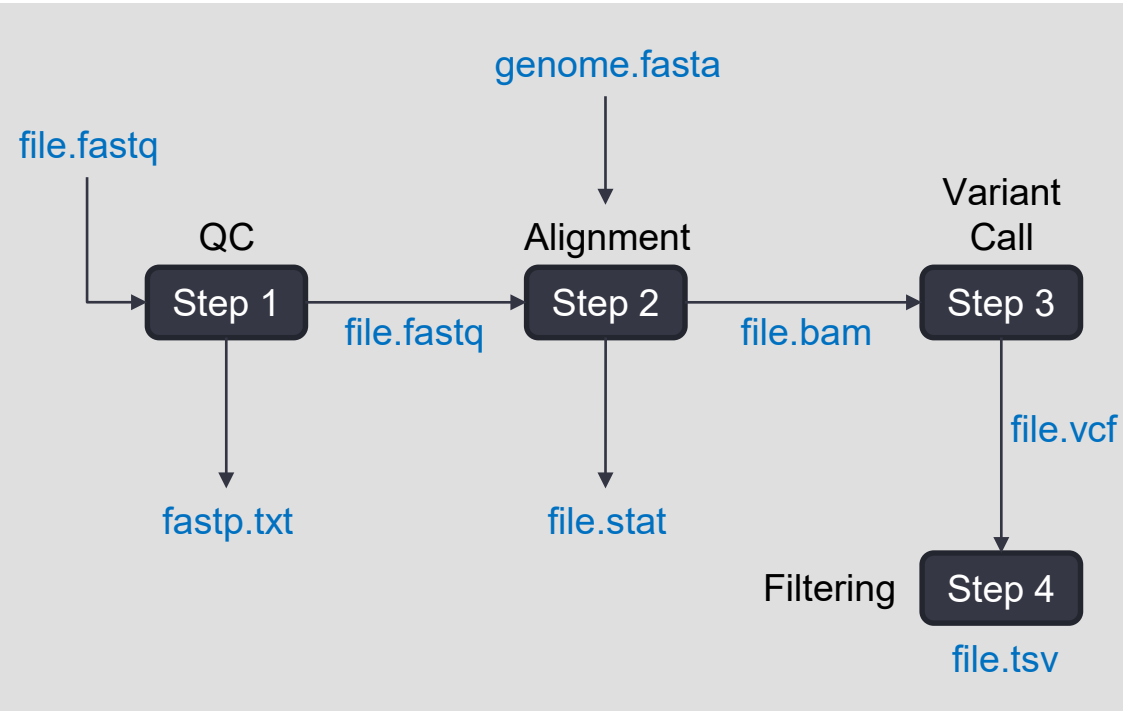
Error Domain

Extension Domain

Object Information

Domain-based organization

- Global input/output
  - Overview of what the pipeline needs and what it produces



Provenance Domain

Execution Domain

Description Domain

Parametric Domain

Usability Domain

IO Domain

Global inputs/outputs to pipeline

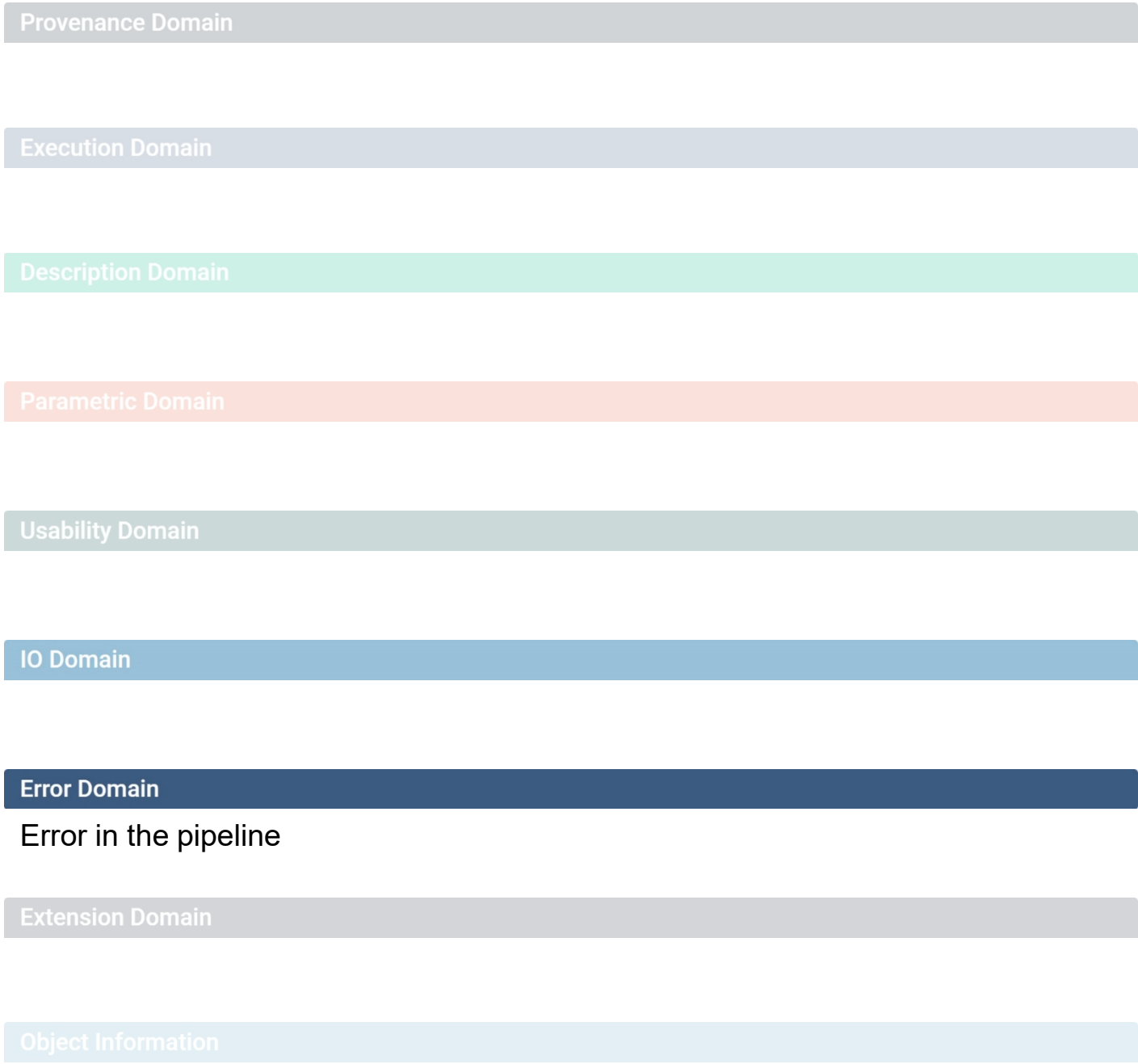
Error Domain

Extension Domain

Object Information

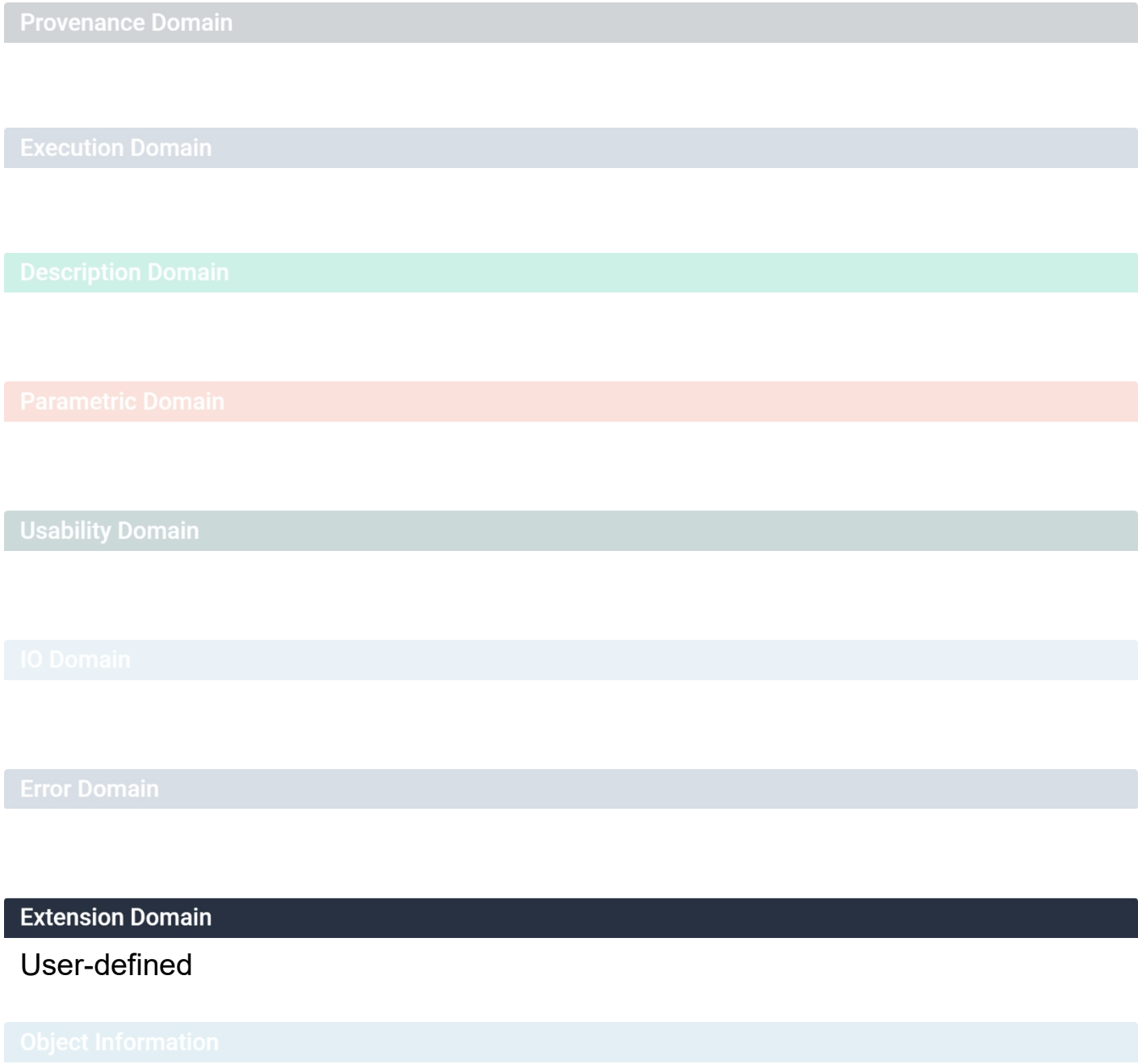
Domain-based organization

- Limits of detection
  - Empirical
  - Algorthmic



Domain-based organization

- User-defined Domain
- Can extend beyond the base schema
- Requires a schema to validate



Domain-based organization

- Object ID
- Specification Version
- eTag

Provenance Domain

Execution Domain

Description Domain

Parametric Domain

Usability Domain

IO Domain

Error Domain

Extension Domain

Object Information

Metadata about the pipeline

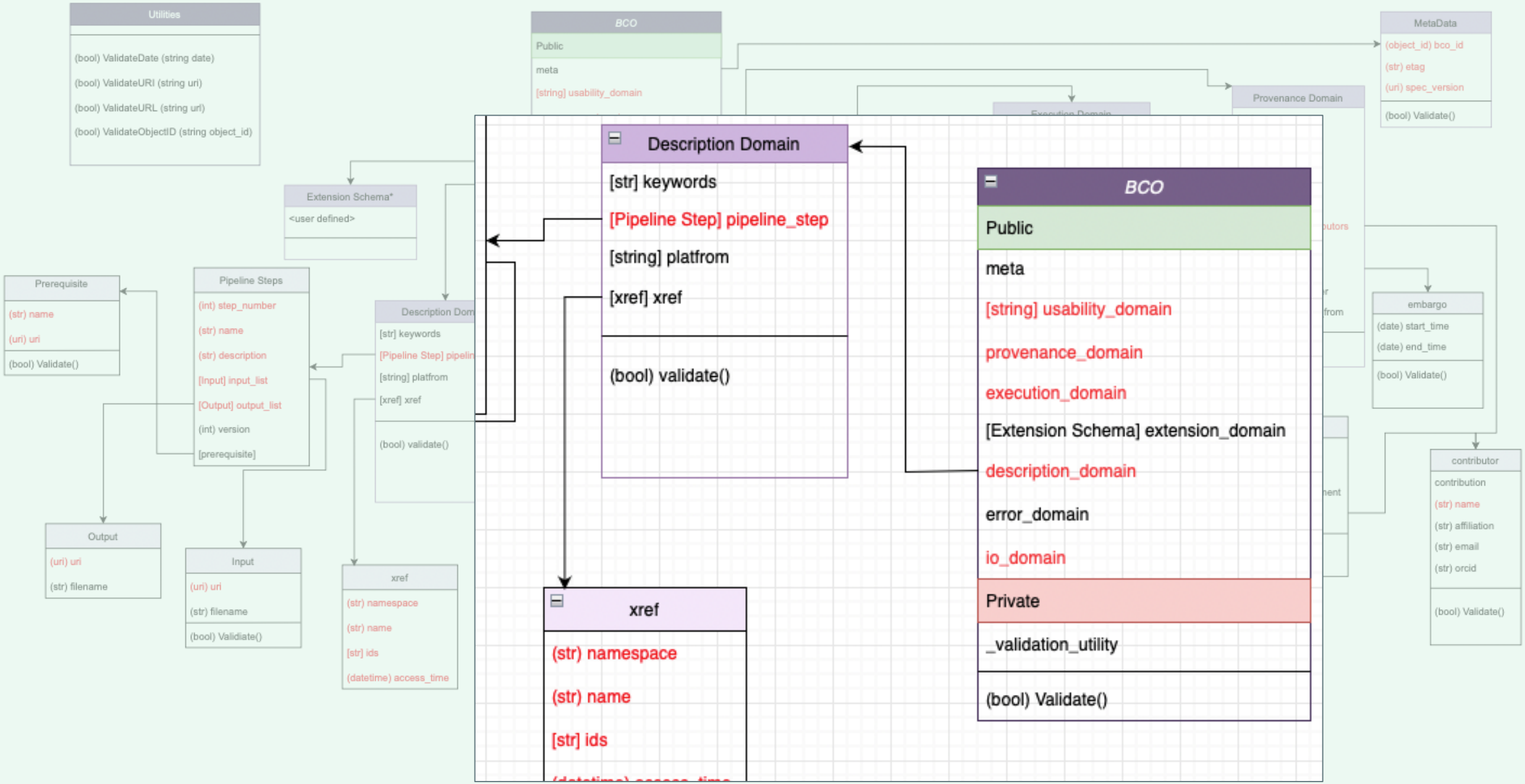
Previously…

- BCO exporter on HIVE
- Form-based editor for making BCOs
- Improved Documentation
- Combination BCOs with other projects
  - Research Objects/Nextflow
  - Common Workflow Language
- Reviewer’s perspective on using BCOs
  - King et al.

Updates: BCO Python Library

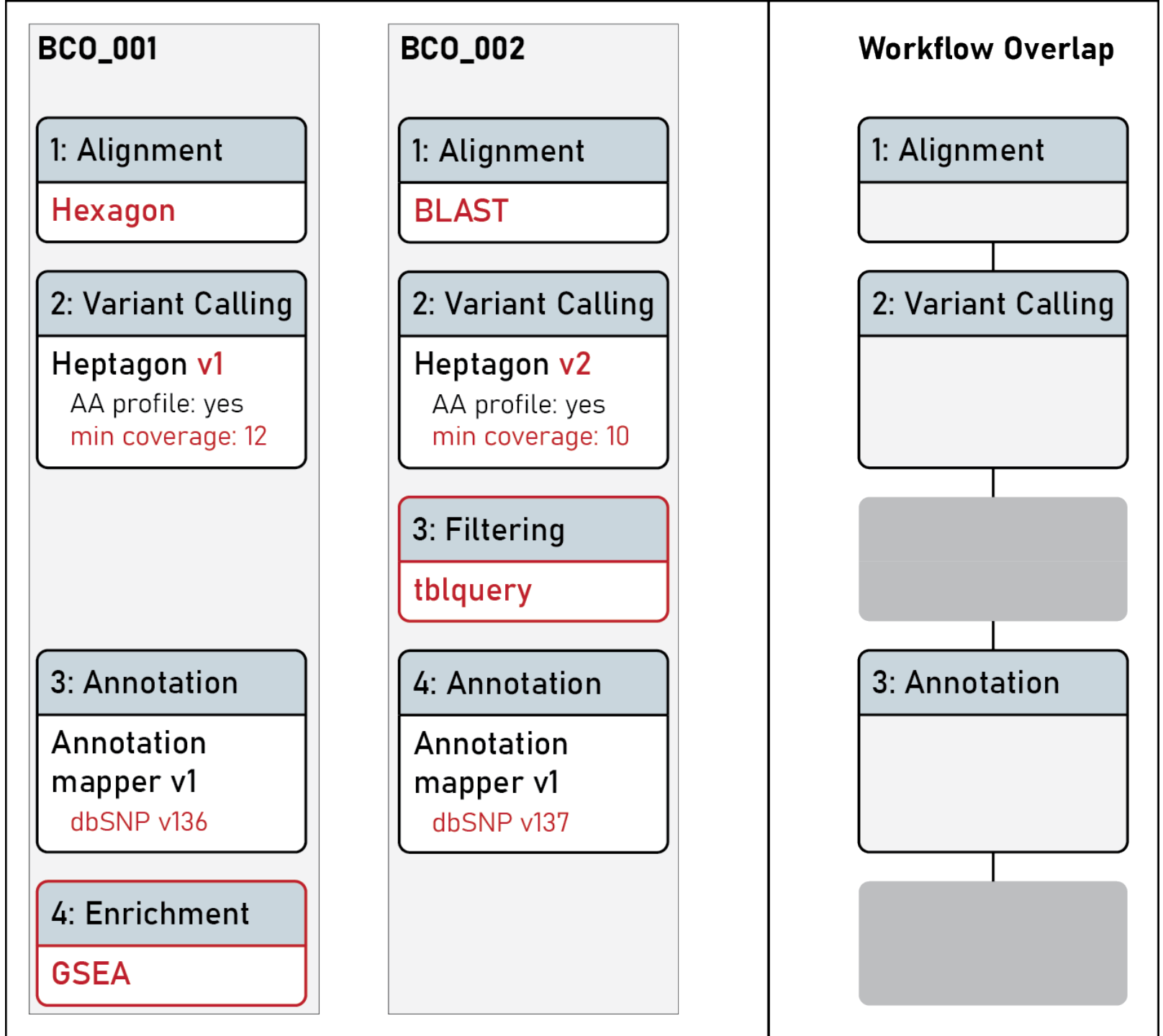
- Importable package for Python
- Programmatically work with BCOs
  - Creates a Python Class Object from a BCO
  - Includes tools for working with the Objects

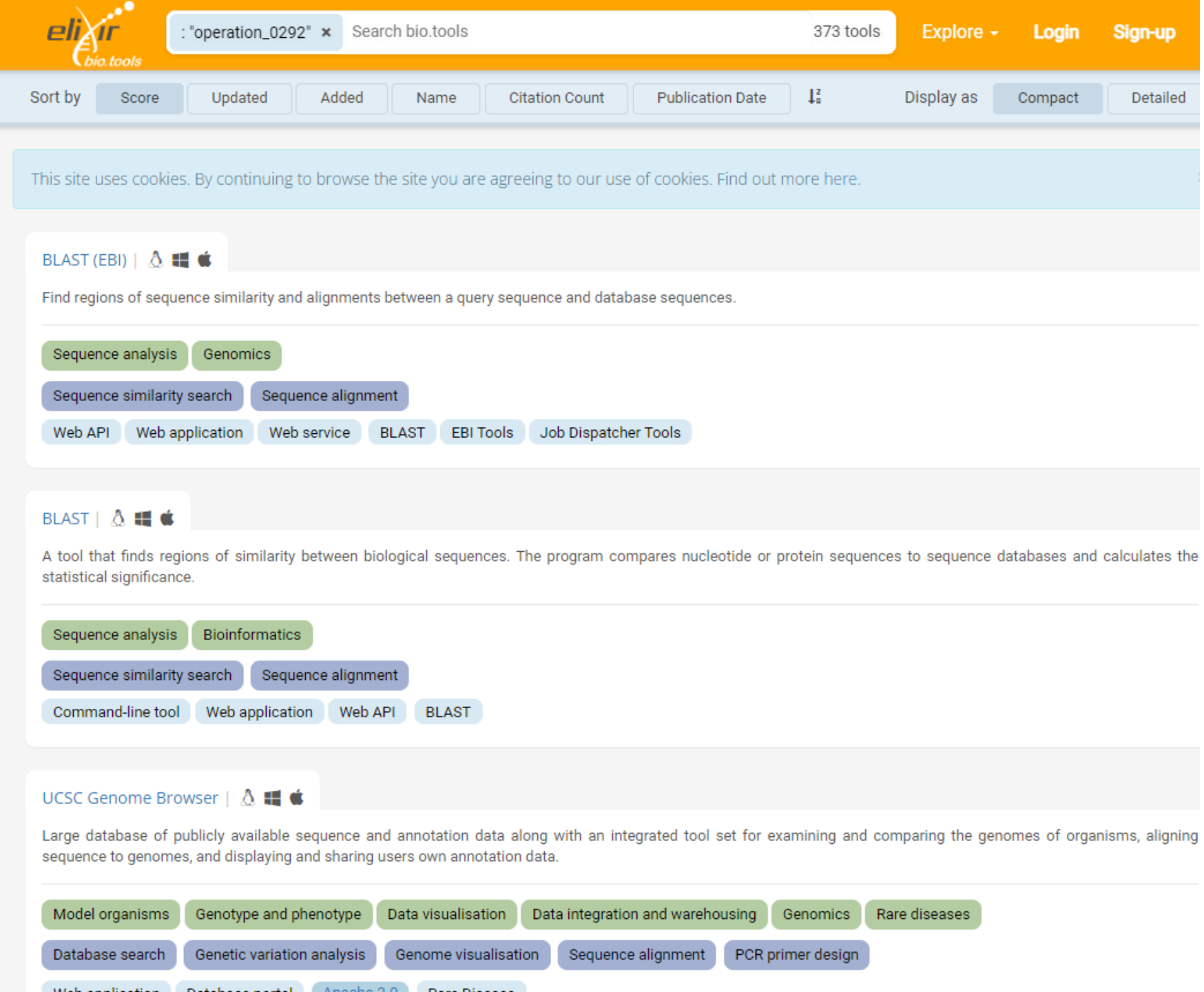
```
>>> import bco_
```


Utilities
(bool) ValidateDate (string date)
(bool) ValidateURI (string uri)
(bool) ValidateURL (string url)
(bool) ValidateObjectID (string object_id)
BCO
Public
meta
[string] usability_domain
MetaData
(object_id) bco_id
(str) etag
(uri) spec_version
(bool) Validate()
Provenance Domain
Extension Schema*
<user defined>
Description Domain
[str] keywords
[Pipeline Step] pipeline_step
[string] platfrom
[xref] xref
(bool) validate()
BCO
Public
meta
[string] usability_domain
provenance_domain
execution_domain
[Extension Schema] extension_domain
description_domain
error_domain
io_domain
Private
_validation_utility
(bool) Validate()
Prerequisite
(str) name
(uri) uri
(bool) Validate()
Pipeline Steps
(int) step_number
(str) name
(str) description
[Input] input_list
[Output] output_list
(int) version
[prerequisite]
Description Dom
[str] keywords
[Pipeline Step] pipelin
[string] platfrom
[xref] xref
(bool) validate()
embargo
(date) start_time
(date) end_time
(bool) Validate()
contributor
contribution
(str) name
(str) affiliation
(str) email
(str) orcid
(bool) Validate()
Output
(uri) uri
(str) filename
Input
(uri) uri
(str) filename
(bool) Validiate()
xref
(str) namespace
(str) name
[str] ids
(datetime) access_time
xref
(str) namespace
(str) name
[str] ids

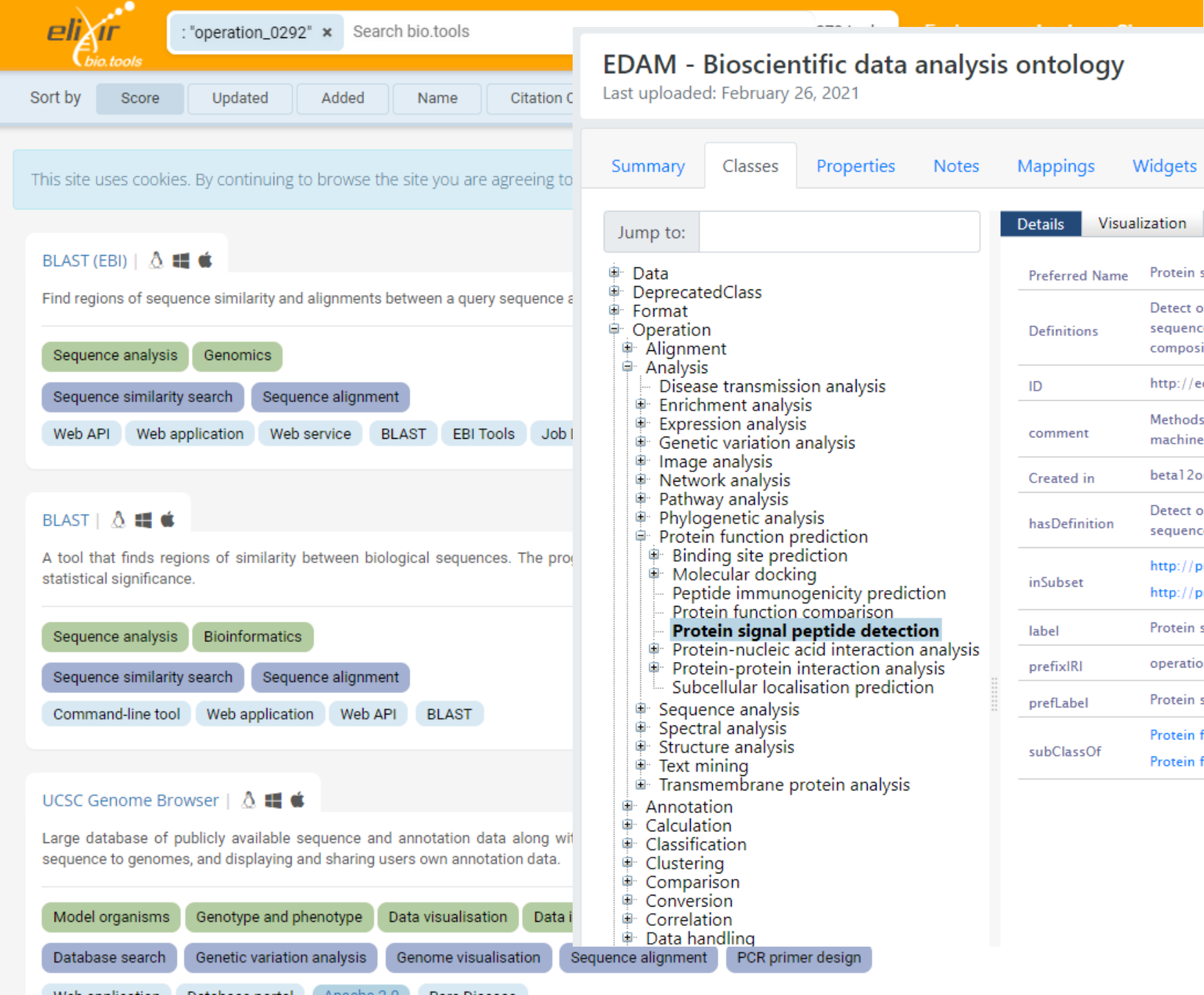
Example usage

- Retrieve publications for specific date range from all authors in Provenance Domain
- Sort studies by number of samples
- Check your environment for compatibility with a particular workflow
- Workflow validator: do file types line up correctly?
- Compare two workflows


BCO_001
1: Alignment
Hexagon
2: Variant Calling
Heptagon v1
AA profile: yes
min coverage: 12
3: Annotation
Annotation mapper v1
dbSNP v136
4: Enrichment
GSEA
BCO_002
1: Alignment
BLAST
2: Variant Calling
Heptagon v2
AA profile: yes
min coverage: 10
3: Filtering
tblquery
4: Annotation
Annotation mapper v1
dbSNP v137
Workflow Overlap
1: Alignment
2: Variant Calling
3: Annotation



Source for Tool Categories



EDAM - Bioscientific data analysis ontology

Last uploaded: February 26, 2021

Summary | Classes | Properties | Notes | Mappings | Widgets

Jump to:

- Data
- DeprecatedClass
- Format
- Operation
  - Alignment
  - Analysis
    - Disease transmission analysis
    - Enrichment analysis
    - Expression analysis
    - Genetic variation analysis
    - Image analysis
    - Network analysis
    - Pathway analysis
    - Phylogenetic analysis
    - Protein function prediction
      - Binding site prediction
      - Molecular docking
      - Peptide immunogenicity prediction
      - Protein function comparison
      - **Protein signal peptide detection**
      - Protein-nucleic acid interaction analysis
      - Protein-protein interaction analysis
      - Subcellular localisation prediction
    - Sequence analysis
    - Spectral analysis
    - Structure analysis
    - Text mining
    - Transmembrane protein analysis
  - Annotation
  - Calculation
  - Classification
  - Clustering
  - Comparison
  - Conversion
  - Correlation
  - Data handling

Details | Visualization | Notes (0) | Class Mappings (0)

Preferred Name	Protein signal peptide detection
Definitions	Detect or predict signal peptides and signal peptide cleavage sites in protein sequences. Methods might use sequence motifs and features, amino acid composition, profiles, machine-learned classifiers, etc.
ID	http://edamontology.org/operation_0418
comment	Methods might use sequence motifs and features, amino acid composition, profiles, machine-learned classifiers, etc.
Created in	beta12orEarlier
hasDefinition	Detect or predict signal peptides and signal peptide cleavage sites in protein sequences.
inSubset	http://purl.obolibrary.org/obo/edam#operations http://purl.obolibrary.org/obo/edam#edam
label	Protein signal peptide detection
prefixIRI	operation_0418
prefLabel	Protein signal peptide detection
subClassOf	Protein function prediction Protein feature detection

biocompute-objects / bcotool Public

Edit Pins | Unwatch 4 | Fork 4 | Star 2

Code | Issues 5 | Pull requests 1 | Discussions | Actions | Projects | Wiki | Security | Insights | Settings

dev | 4 branches | 3 tags | Go to file | Add file | Code

This branch is 11 commits ahead, 3 commits behind main. Contribute

skeeney01 Merge pull request #20 from biocompute-objects/patch-2 ... f903f01 4 days ago 57 commits

bcotool	Merge branch 'dev' into patch-2	4 days ago
data_tests	Fix links to cwl scripts	5 months ago
.gitignore	Update for conversion function	6 months ago
LICENSE	Update license	2 years ago
README.md	Update for conversion function	6 months ago
requirements.txt	adding requirements.txt	15 months ago
tst.json	Updates to API function	2 years ago

README.md

BCO-TOOL

https://github.com/biocompute-objects/bcotool/tree/dev

This is a Command Line Tool that allows for the manipulation of BioCompute Objects. Serveral functionalites are provided (detailed more below in supported modes).

To install:

Run the Git Clone command in the location you would like the repostory:

About

No description, website, or topics provided.

Readme
MIT license
2 stars
4 watching
4 forks

Releases 3

1.2.0 Latest
on Mar 3

+ 2 releases

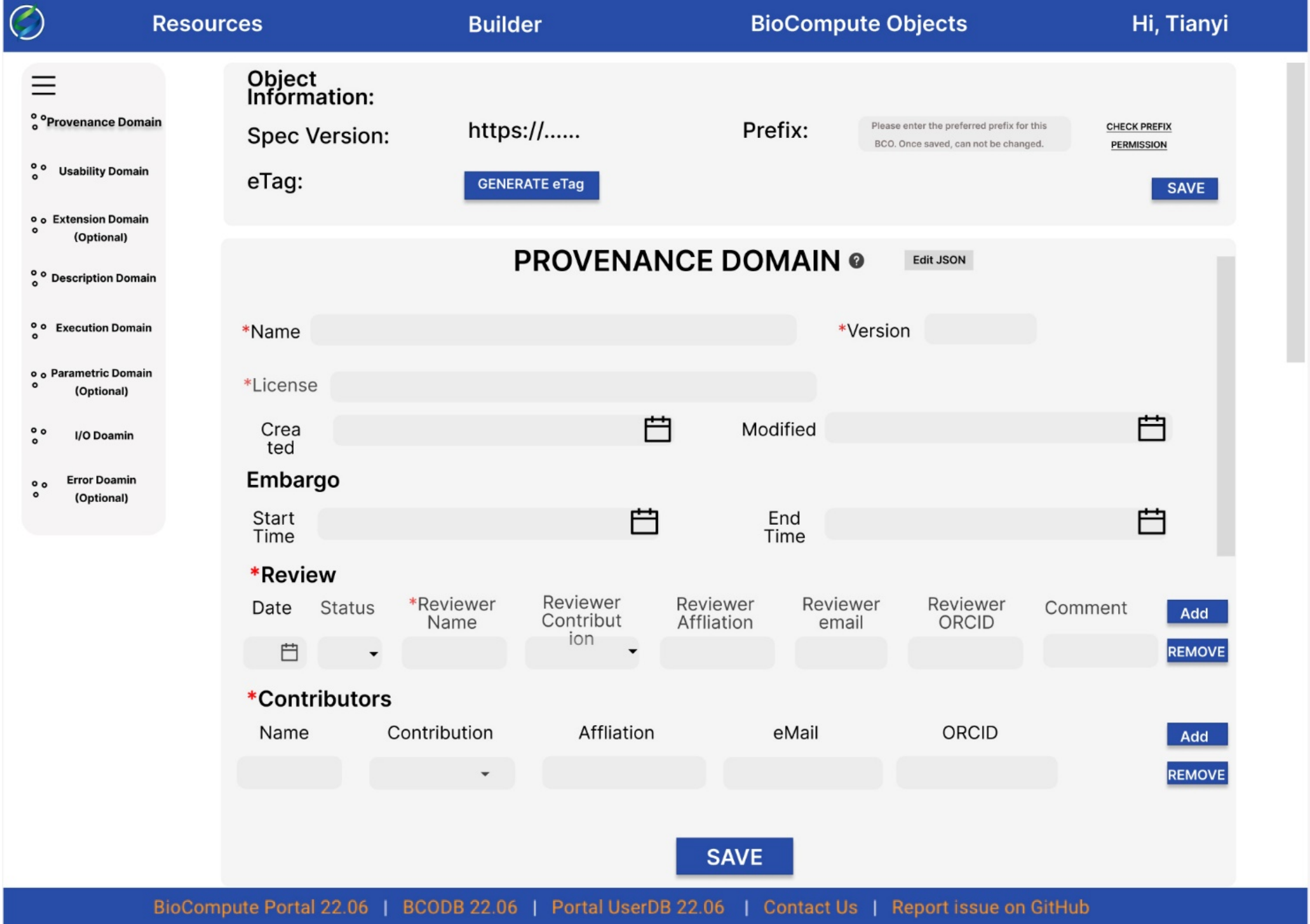
Packages

No packages published
Publish your first package

Contributors 4

HadleyKing Hadley King
acoleman2000 Alex Coleman
skeeney01
rajamazumder

Portal Update



BioCompute Documentation

External site

User Guide
Best Practices
SOP
Tutorials

IEEE 2791-2020

IEEE Standard for Bioinformatics Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication

IEEE Advancing Technology for Humanity

BCO TSC

The Technical Steering Committee of the BioCompute Partnership (TSC) is a body of experienced professionals with BioCompute standard subject matter expertise. See here for the Meeting notes and agenda for all past and the upcomming meetings.

https://biocomputeobject.org

Cloud-based tools for BioCompute

See our resources page for additional tools and services.

HIVE powered by aws

AWS instance of HIVE is temporarily down. Check back later for access.

Access AWS HIVE, the High-Performance Integrated Virtual Environment, on AWS

BioCompute Builder

Use the BioCompute Builder or view objects in the database.
The BioCompute Builder is a platform-

Galaxy PROJECT powered by aws

BioCompute has been merged into the main Galaxy repository. This BioCompute enabled instance of Galaxy on AWS is therefore no longer operational. Thank

News and Events

FDA Notice on BioCompute

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

Tweets from @BioComputeObj

BioCompute Retweeted

Hadley King
@HadleyKingIV · Aug 18

Tools from @SevenBridges @dnanexus and @galaxyproject will be featured! #BioCompute #BioComputeObjects

3

BioCompute
@BioComputeObj · Aug 18

A new BioCompute workshop will be held on September 13th via WebEx! Learn how to create BCO records from your #workflows directly on popular #bioinformatics platforms, and how to deposit them into an access-controlled database.
Please register here:

BioCompute Documentation

External site

User Guide
Best Practices
SOP
Tutorials

IEEE 2791-2020

IEEE Standard for Bioinformatics Analyses Generated by High-Throughput Sequencing (HTS) to Facilitate Communication

IEEE Advancing Technology for Humanity

BCO TSC

The Technical Steering Committee of the BioCompute Partnership (TSC) is a body of experienced professionals with BioCompute standard subject matter expertise. See here for the Meeting notes and agenda for all past and the upcomming meetings.

News and Events

FDA Notice on BioCompute

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

object.biocompute@gmail.com

Cloud-based tools for BioCompute

See our resources page for additional tools and services.

HIVE powered by aws

AWS instance of HIVE is temporarily down. Check back later for access.

Access AWS HIVE, the High-Performance Integrated Virtual Environment, on AWS

BioCompute Builder

Use the BioCompute Builder or view objects in the database.
The BioCompute Builder is a platform-

Galaxy PROJECT powered by aws

BioCompute has been merged into the main Galaxy repository. This BioCompute enabled instance of Galaxy on AWS is therefore no longer operational. Thank

Tweets from @BioComputeObj

BioCompute Retweeted

Hadley King
@HadleyKingIV · Aug 18

Tools from @SevenBridges @dnanexus and @galaxyproject will be featured! #BioCompute #BioComputeObjects

3

BioCompute
@BioComputeObj · Aug 18

A new BioCompute workshop will be held on September 13th via WebEx! Learn how to create BCO records from your #workflows directly on popular #bioinformatics platforms, and how to deposit them into an access-controlled database.
Please register here:

Updates: ARGOSDB

- Database of regulatory-grade genomes of infectious diseases
- Knowledgebase BCOs
- Document curation process

Updates: FDA-ARGOS

- Embleema and George Washington University's bioinformatics research collaboration to expand and develop database for high-quality sequences that can help researchers fight against infectious disease outbreaks

- This project will expand datasets publicly available in FDA-ARGOS, improve quality control by developing quality matrix tools and scoring approaches that will allow the mining of public sequence databases, and identify high-quality sequences for upload to the FDA-ARGOS database as regulatory-grade sequences.

- Building on expansions during the COVID-19 pandemic, this project aims to further improve the utility of the FDA-ARGOS database as a key tool for medical countermeasure development and validation.

FDA-ARGOS Project Outcomes

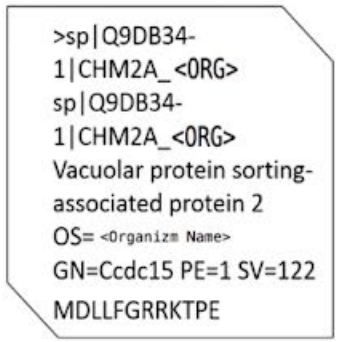
- Identify genomes of microbial species of high clinical relevance qualified as regulatory-grade sequences from public resources and generate annotation data model
- Develop more comprehensive and reliable quality control (QC) assessments of sequence representations
- Prepare NCBI submission packages and deposit regulatory-grade sequences to the FDA-ARGOS database and provide documentation, outreach, and training to FDA personnel

Search by: BCOID, dataset file name, title, description or categories

27 results found.

reviewed protein dataset ARGOS_000033 in FASTA format. [Salmonella typhimurium]

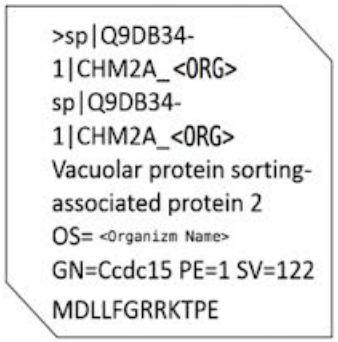
Salmonella typhimurium reference proteome sequences



Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720) reference proteome fasta sequences.

reviewed protein dataset ARGOS_000002 in FASTA format. [Influenza A]

Influenza A reference proteome sequences



Influenza A (A/Puerto Rico/8/1934 H1N1) reference proteome fasta sequences.

reviewed protein dataset ARGOS_000006 in CSV format. [SARS-CoV-2]

SARS-CoV-2 reference proteins list

uniprotkb_ac...	entry_name...
P0DTC4...	VEMP_SARS2...
P0DTD1...	R1AB_SARS2...

SARS-CoV-2 (Wuhan-Hu-1) reference protein accessions and summary annotations.

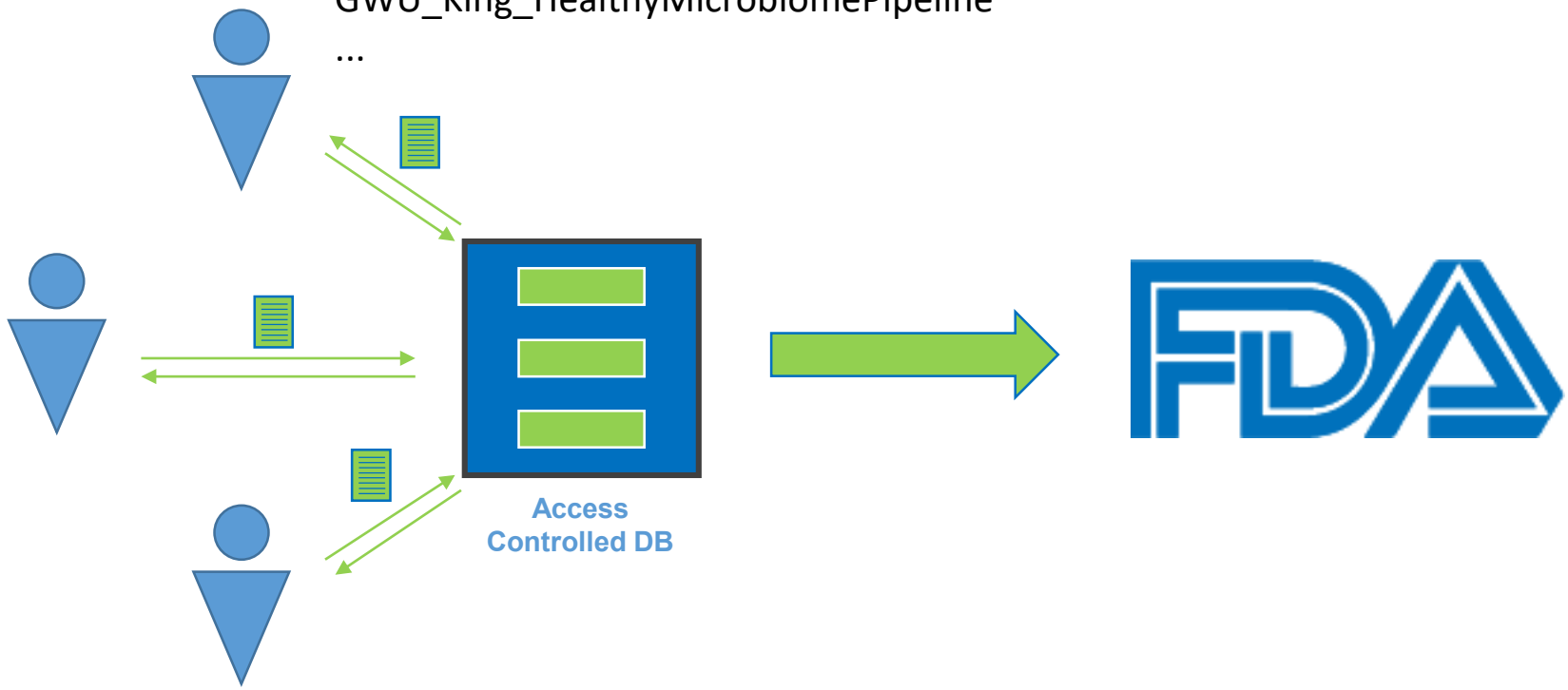
View Details

BioComputeDB

- Centralized registry of workflow Objects
- User-controlled based on Prefix
- Existing DBs at GW and FDA

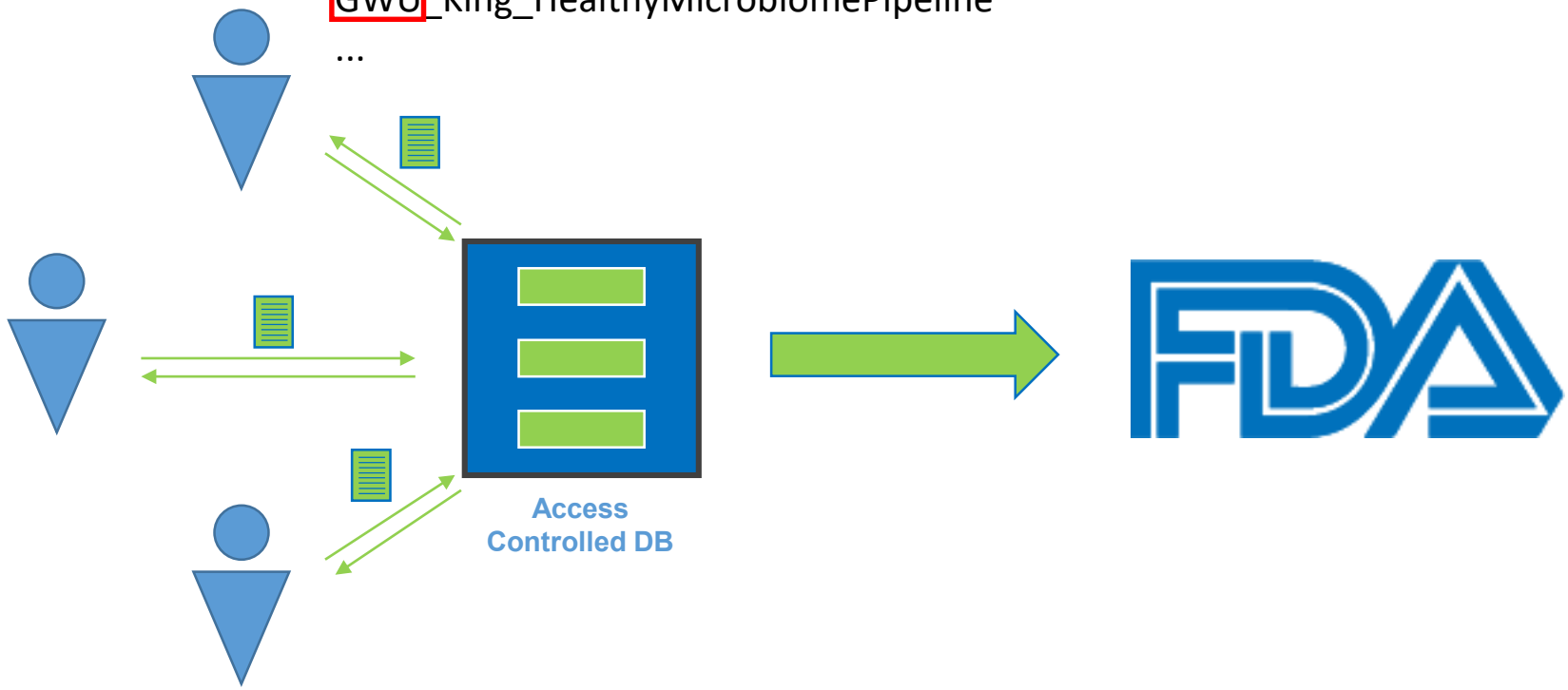
BCOs:
GWU_Vaccine_Keeney01
GWU_Vaccine_Keeney02
GWU_King_HealthyMicrobiomePipeline
...

BCOs:
CBER_HIVE_AAJ001.1
CBER_HIVE_ABC021.4
CBER_HIVE_FFX017.7
...



BCOs:
NCI_Miller_Schuez_EtAl
NCI_Astling_And Smith
NCI_Erikson_EtAl
...

BCOs:
GWU_Vaccine_Keeney01
GWU_Vaccine_Keeney02
GWU_King_HealthyMicrobiomePipeline
...



BCOs:
CBER_HIVE_AAJ001.1
CBER_HIVE_ABC021.4
CBER_HIVE_FFX017.7
...

BCOs:
NCI_Miller_Schuez_EtAl
NCI_Astling_And Smith
NCI_Erikson_EtAl
...

Acknowledgements

BCO Founding Members:

Raja Mazumder Vahan Simonyan

- Konstantinos Karagiannis
- Mark Walderhaug
- Anton Golikov
- Hadley King
- Tianyi Wang
- Rohan Panigrahi
- Sean Keeney
- Lam Phuc
- John Torcivia-Rodriguez
- Alhanouf Altuwayjiri
- Michael Crusoe
- Stian Soiland-Reyes

COMMON WORKFLOW LANGUAGE

MANCHESTER 1824 The University of Manchester

THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC

embleema

GEORGE MASON UNIVERSITY

COLORADO SCHOOL OF MINES OFFICIAL SEAL 1874

clarifai

FDA

BAA: 75F40119C10136

Workshop Schedule

10:30 - 12:00PM	**BioCompute and Galaxy**
	Introduction **Updates** **Questions** **BioComputeDB** Concept, demo functionality with Portal, API, prefixes **Questions** **Galaxy Demo** BCO builds on the platform **Questions**
12:00 - 12:30PM	Lunch Break
12:30 - 2:00PM	**DNAnexus**
	Demo BCOnexus interface to present overview, in-memory editor, schema/standard driven, WDL, DNAnexus integration **Questions** **Demo use cases** Build from scratch, start with a JSON file BCO metadata template and layer in workflow metadata from WDL and DNAnexus; export BCO to DNAnexus and run the workflow **Questions**
2:00 - 2:15PM	Break
2:15 - 3:45PM	**Seven Bridges**
	Demo Seven Bridges tool interface **Questions** **Demo use cases** Generate BCOs from platform output **Questions**
3:45 - 4:00PM	**Open Discussion** Future plans and feedback