OHDSI Gold Standard Phenotype Library Working Group

Community Call
Progress Update

Aaron Potvien

April 2, 2019

Gold Standard Phenotype Library (GSPL)

(Talked about why we need the GSPL on January 15th)

Objective:

To enable members of the OHDSI community to **find, evaluate, and utilize community-validated cohort definitions** for research and other activities.

FAIR Principles

- GSPL development is being guided by FAIR Principles

- Reference: The FAIR Guiding Principles for scientific data management and stewardship by Wilkinson *et al.* (2016)

Findable
Accessible
Interoperable
Reusable

FAIR Principles

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Library Architecture Formulation

End User

Authors

Librarians

Validators

“Gold Standard” you say?

- What it **isn’t:**

  - Imposing rules to make sure phenotypes have “good enough” metrics.

- What it **is:**

  - Librarians making sure that certain “**gold standard processes**” are being followed when a phenotype is submitted to the library and when a phenotype is validated.

Gold Standard Processes

Author Data Elements

- Metadata:
  - Title
  - Author(s) and Affiliations
  - Date of Submission
  - Modality (Rule-Based or Computable)
  - Links to implementation/config files on GitHub
- Development:
  - Purpose and Intended Use
  - Development Methodology
  - Flowchart
- Identify CDM Dependencies:
  - Conditions
  - Drug Exposures
  - Labs
  - Measurements
  - Notes NLP
  - Observations
  - Procedures
  - Visits
- Provenance:
  - Other phenotype definitions this phenotype was derived from or inspired by

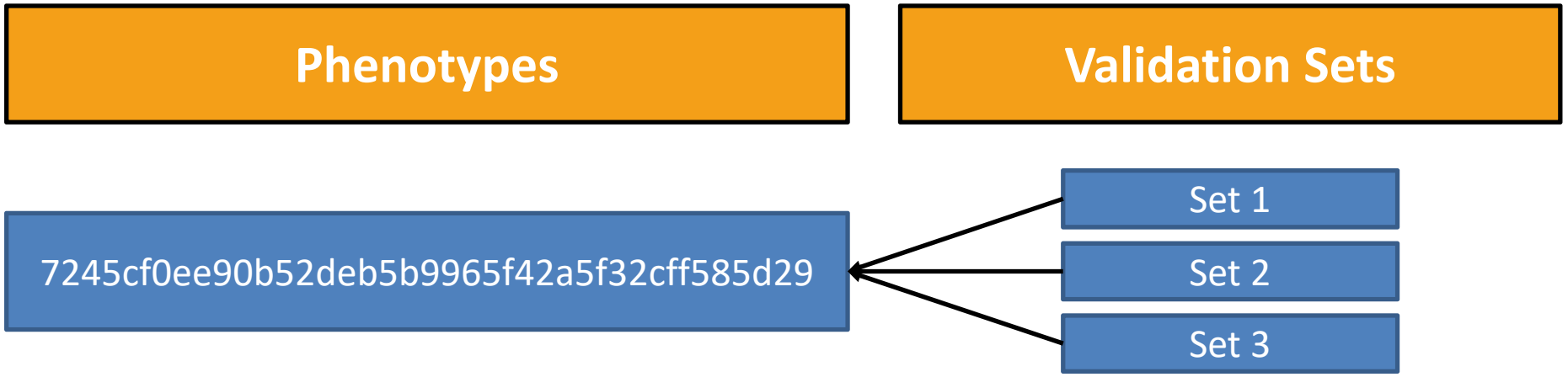
Gold Standard Processes

Validator Data Elements

- Metadata:
  - Title
  - Author(s) and Affiliations
  - Date of Submission
  - Hash of phenotype evaluated
  - Validation procedure
- Metrics:
  - Sample Size
  - True Positives/Negatives
  - False Positives/Negatives
  - Was a THEMIS-certified dataset used?

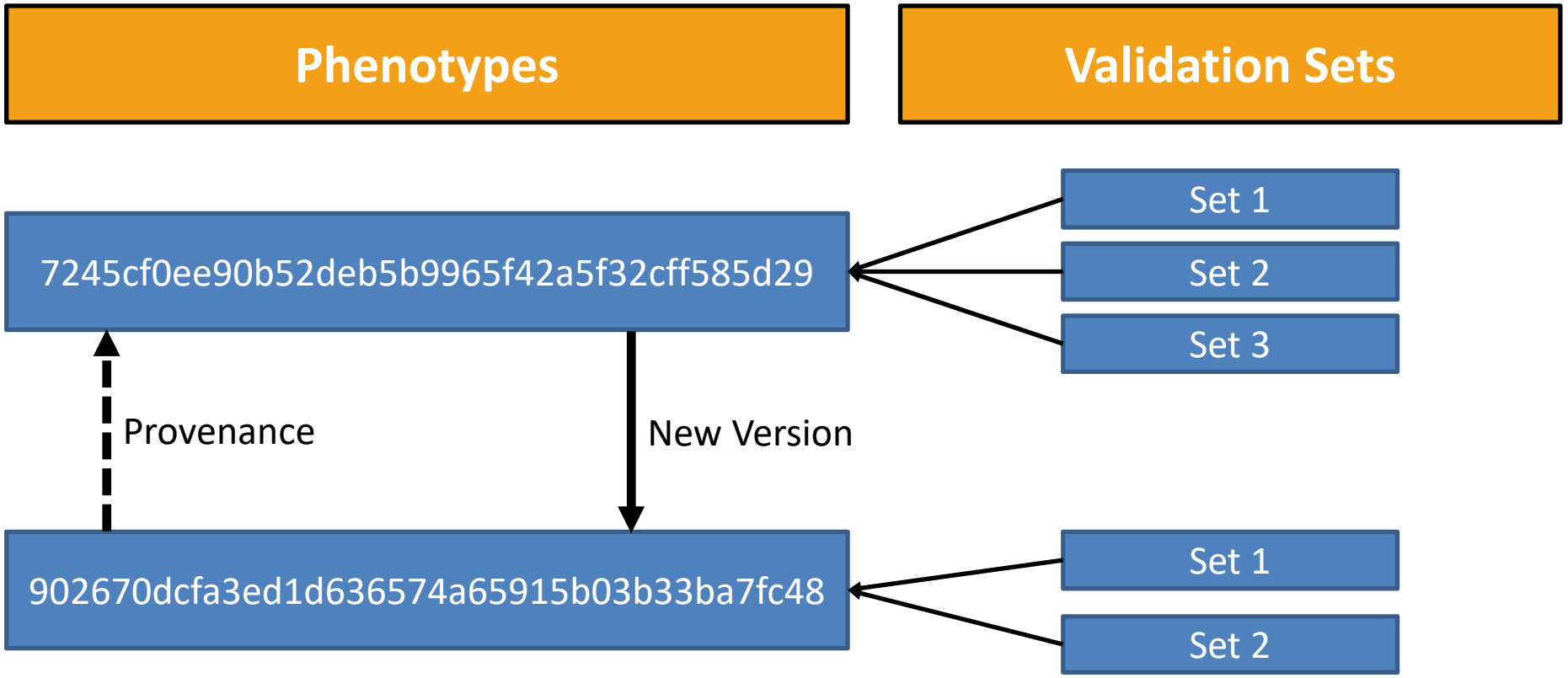
Hash-based Linkage

- A Phenotype is identified by a hash of its implementation file



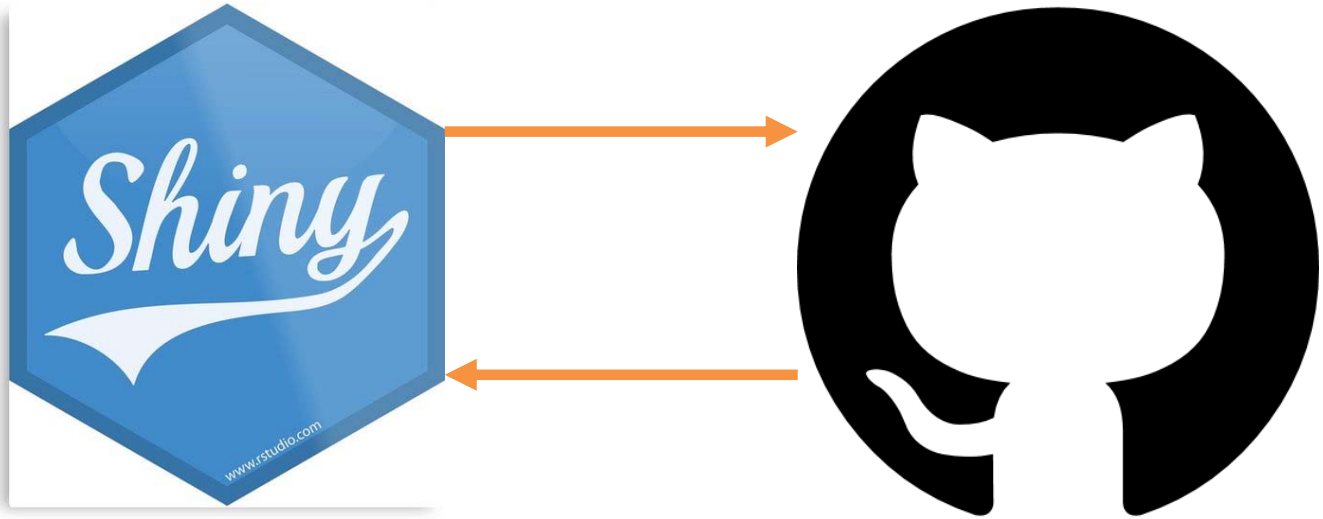
Hash-based Linkage

- A Phenotype is identified by a hash of its implementation file



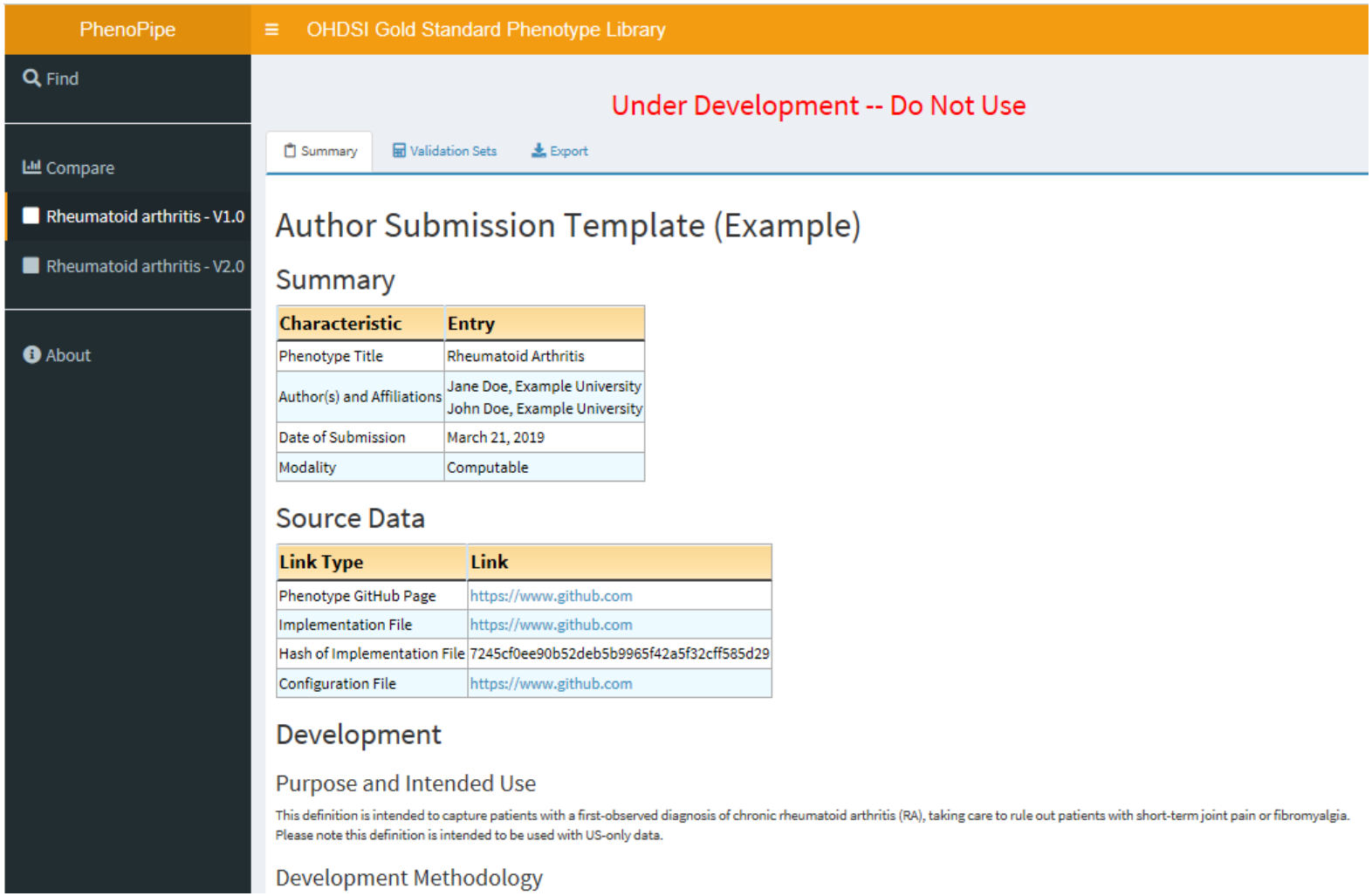
Library Implementation

- Data for the library will be stored on GitHub
- A companion Shiny application will exist to help with searching through this data, compare and contrast phenotypes, etc.



Shiny App Viewer

data.ohdsi.org/PhenotypeLibraryViewer/



Combining OHDSI Toolsets

Aphrodite (Juan Banda)

https://github.com/OHDSI/Aphrodite

- **Can create phenotypes** probabilistically by learning good phenotypes from a set of noisy labels
- Built to interface with the OMOP CDM to automatically create and utilize features using all data in your CDM (or a subset, if you choose)
- Machine learning takes into account more features than what could be considered by hand, and labeling heuristic is less time consuming
- Performs internal validation and is easy to share (config file tracks how it was built; binary object output tracks the definition itself)

Combining OHDSI Toolsets

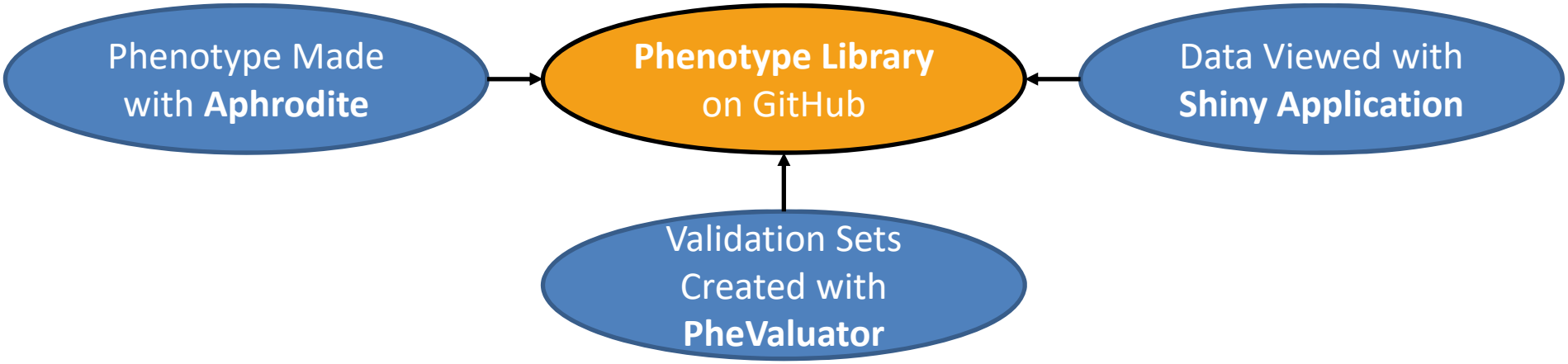
PheValuator (Joel Swerdel)

https://github.com/OHDSI/PheValuator

- **Can evaluate phenotypes** to see how well they perform, offering an alternative to low-powered and time-consuming clinical review
- Uses a diagnostic predictive model to assign a large sample of people a predicted probability of having the condition
- Assess "Truth" based on an extremely specific cohort (xSpec) or extremely sensitive cohort (xSens)
- Produces *all* metrics (not just PPV) for a complete understanding of phenotype definition performance
- Like Aphrodite, will automatically output documentation needed for being a Gold Standard Process.

Combining OHDSI Toolsets

- Combining these tools can help to populate the library.



- **Not required** to be "gold standard" but available to help facilitate the process and avoid pitfalls!

Feedback Welcomed!

Forum:
http://forums.ohdsi.org/t/requirements-development-for-the-ohdsi-gold-standard-phenotype-library/4876

Wiki:
http://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:gold-library-wg

Aphrodite:
https://github.com/OHDSI/Aphrodite/

PheValuator:
https://github.com/OHDSI/PheValuator/

Viewer Application:
http://data.ohdsi.org/PhenotypeLibraryViewer/

My e-mail:
Aaron.Potvien@gtri.gatech.edu

Thanks!