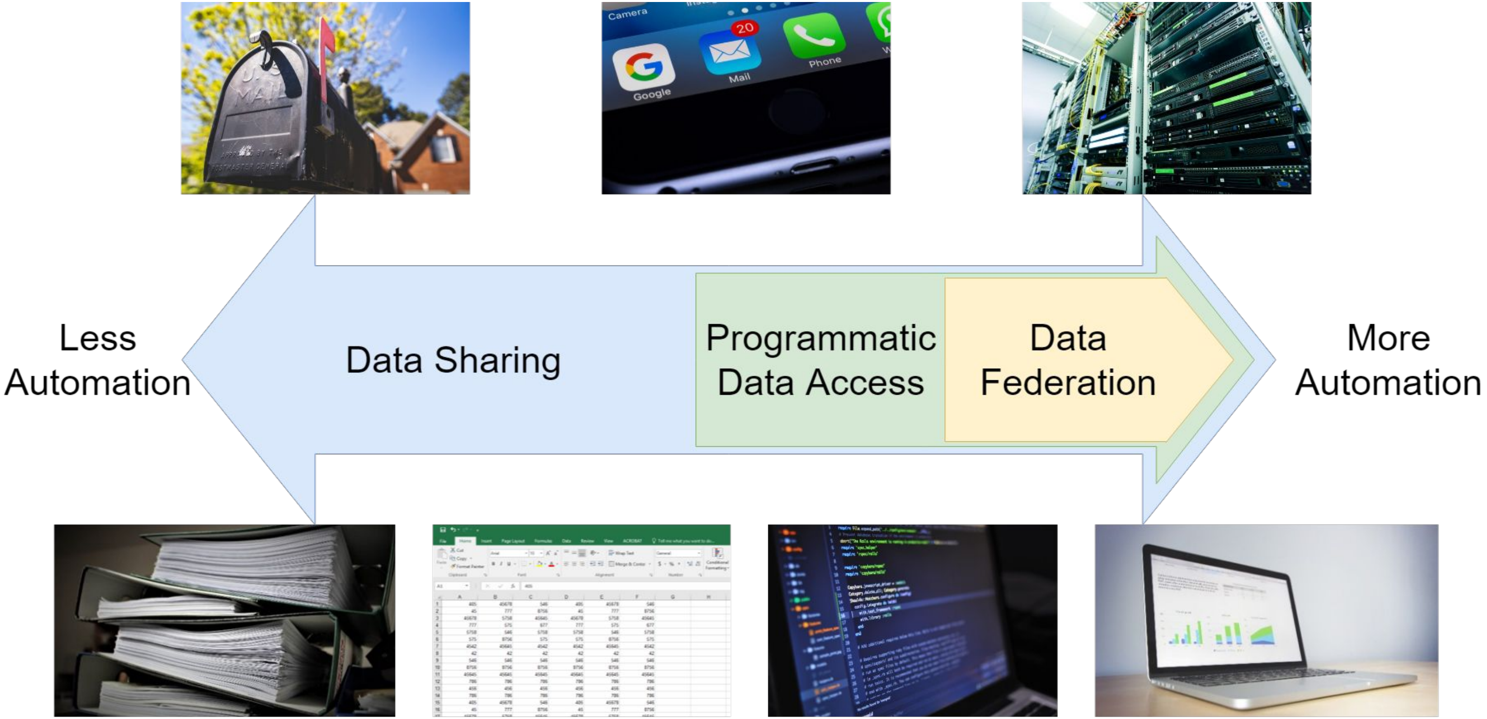
How to Implement Practical Data Federation

Technology Review and Training Material

Previous Results - Defining Data Federation



Less Automation

Data Sharing

Programmatic Data Access

Data Federation

More Automation

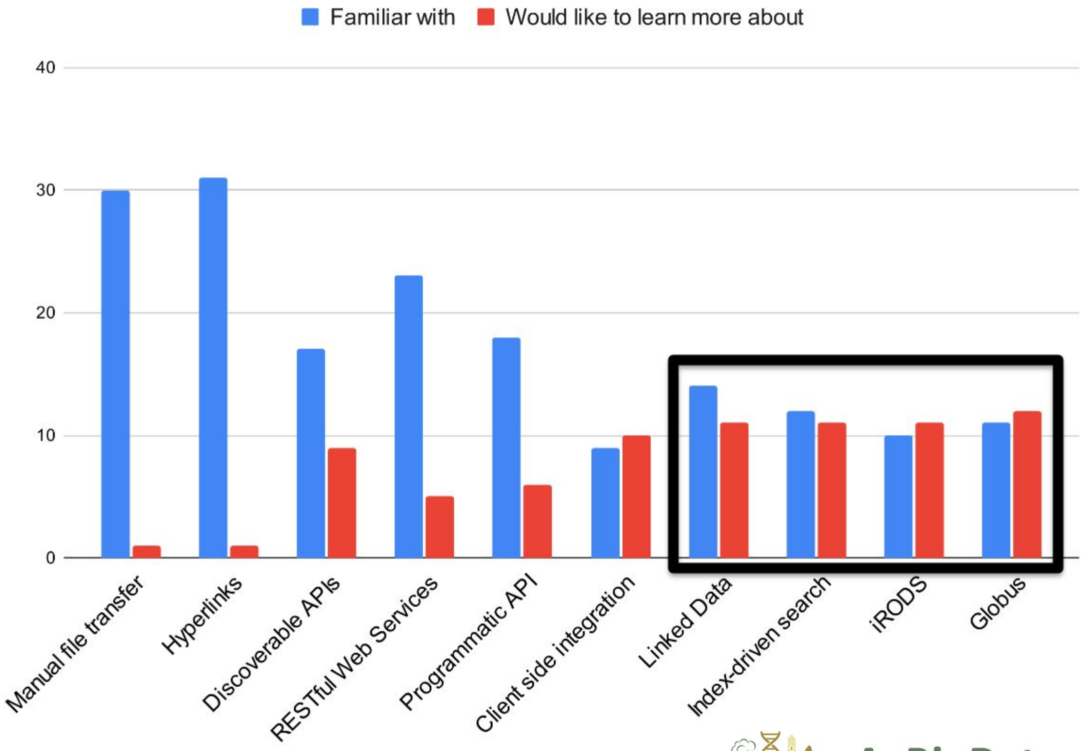
AgBioData

Data Federation Training WG

Previous Results - Technology Awareness

What data sharing technologies ***are you familiar with***?

What data sharing technologies ***would you appreciate learning more about***?



AgBioData

Data Federation Training WG

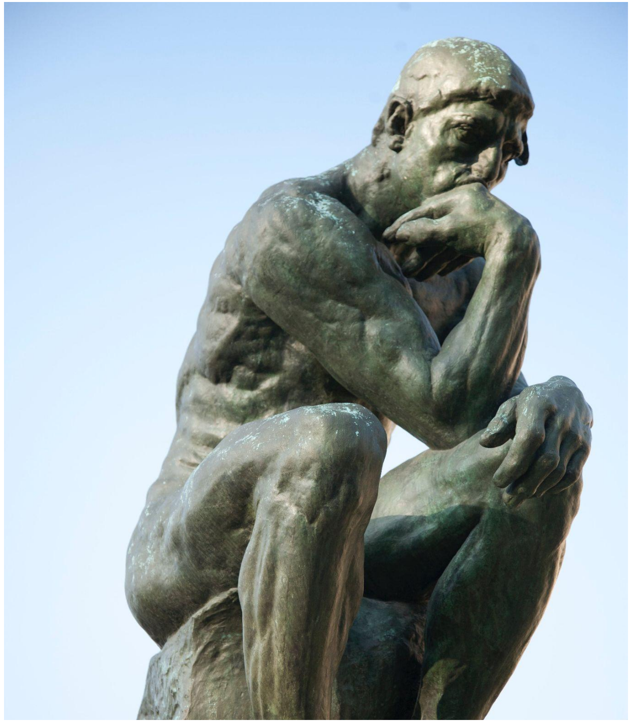
Data Federation Training Working Group

Objectives from the Working Group Proposal

“... This working group will provide training resources on data sharing technologies, either via a collection of existing, vetted training materials; generation of new, written training materials; and/or other materials…”

“...Roughly one third of data federation survey respondents indicated that they would benefit from learning more about Discoverable APIs; Linked-Data; Client-side integration of results from multiple data sources; Index-driven search technologies; Data Management Systems; and Data Sharing via services (e.g. Globus)...”

Getting Started



How do we develop training material for things we are not experts in? Ask the experts!

Brainstorm list of technologies, and find experts in those technologies to teach us.

- Index driven search (feat FAIDARE)
- iRODS
- Globus
- RDF (feat Shallot)
- BrAPI
- GraphQL

AgBioData

Data Federation Training WG

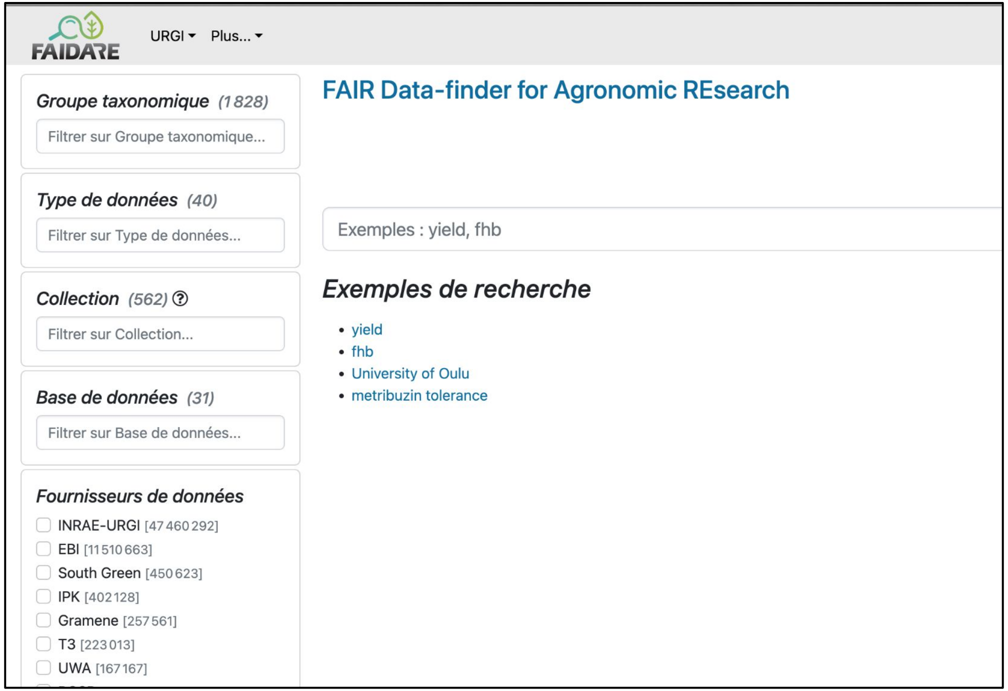
Expert Presentation: Index driven search (feat. FAIDARE)

Cyril Pommier

Use case: Using a shared index to find data from multiple sources through a common interface.

Pros: Greatly increases Findability and Accessibility of data

Cons: Specific solution for a specific use case, not easily generalized



FAIR Data-finder for Agronomic REsearch

Groupe taxonomique (1 828)

Filtrer sur Groupe taxonomique...

Type de données (40)

Filtrer sur Type de données...

Collection (562) ?

Filtrer sur Collection...

Base de données (31)

Filtrer sur Base de données...

Fournisseurs de données

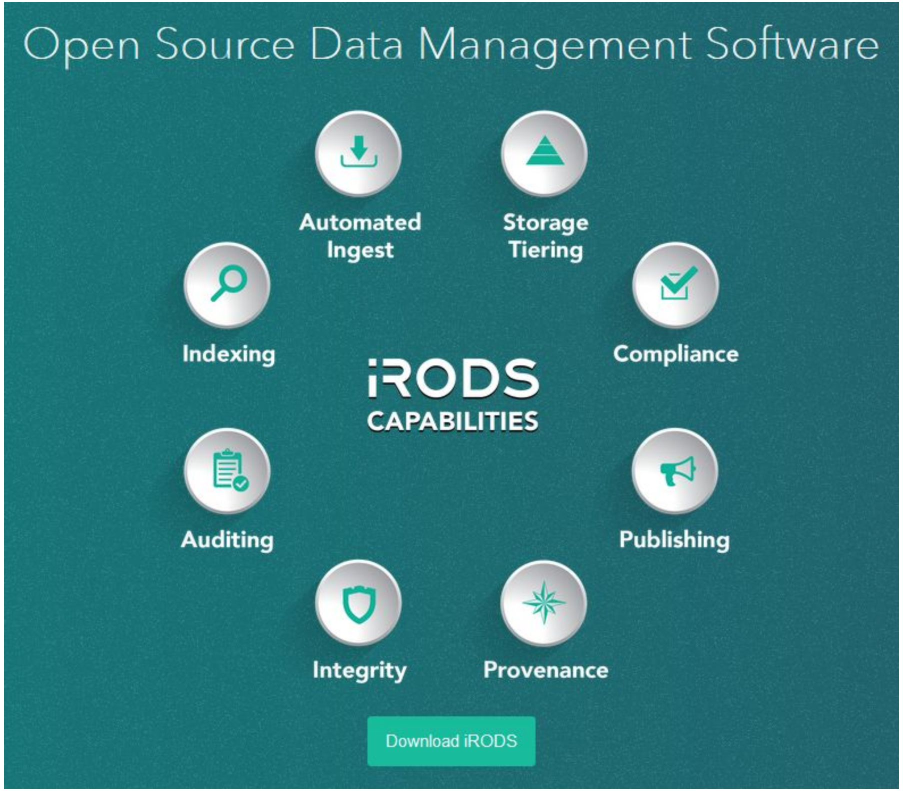
- □ INRAE-URGI [47 460 292]
- □ EBI [11 510 663]
- □ South Green [450 623]
- □ IPK [402 128]
- □ Gramene [257 561]
- □ T3 [223 013]
- □ UWA [167 167]

Exemples : yield, fhb

Exemples de recherche

- yield
- fhb
- University of Oulu
- metribuzin tolerance

Expert Presentation: iRODS



Nirav Merchant

Use case: Raw data access from a shared network of sources, properly annotated shared file system

Pros: Increases Findability and Accessibility of data within a network. Flexible suite of data management tools

Cons: Relies on raw file sharing, without enforced standards or database access. Every node must setup an iRODS system instance.

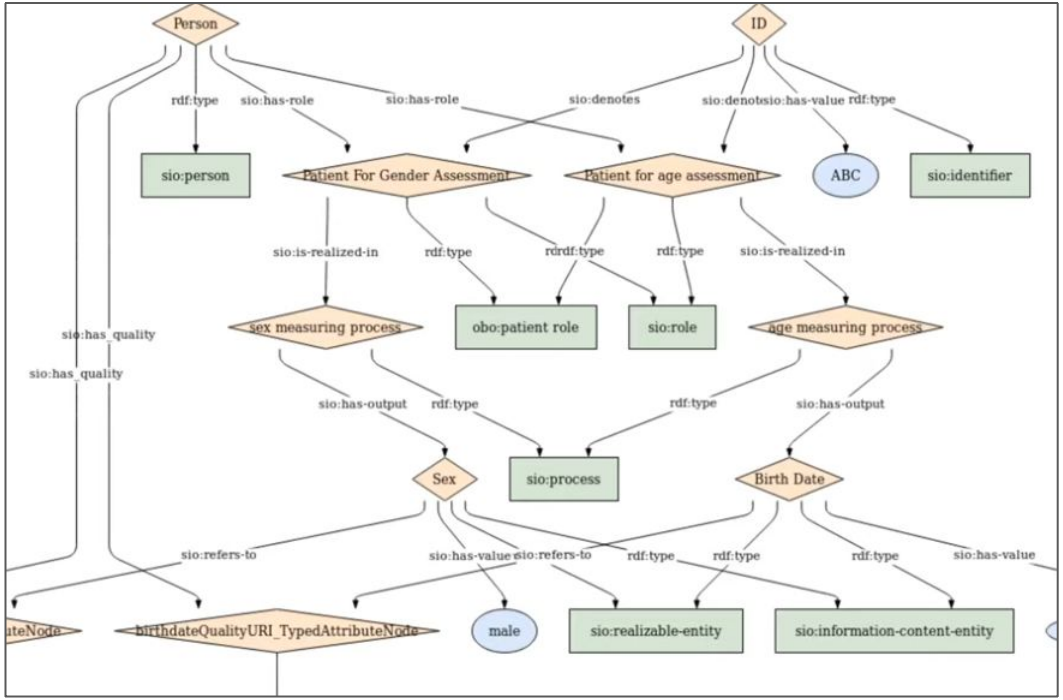
Expert Presentation: RDF (feat. Shallot)

Mark Wilkinson

Use case: Define a shared data model and securely share sensitive data, accessing multiple sources as a single source

Pros: Quickly and securely access common data from many sources with a single query

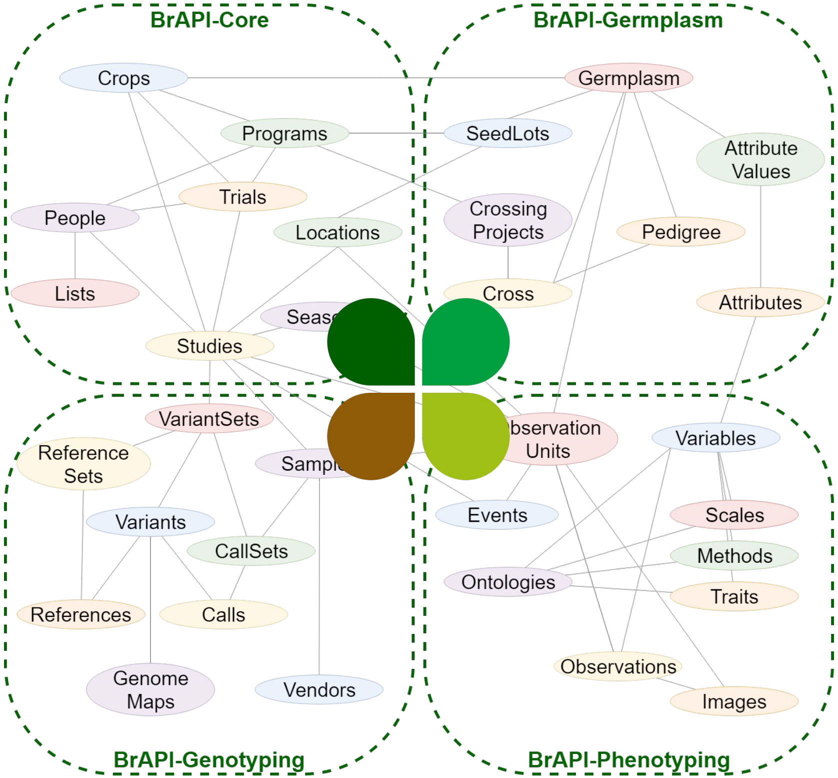
Cons: High cost of setup defining the shared data model, data limited to items every source has in common.



AgBioData

Data Federation Training WG

Expert Presentation: BrAPI



Peter Selby

Use Case: Access specific breeding data from multiple sources using the same standard

Pros: Specific breeding data standard, flexibility to fit many use cases

Cons: Custom implementations can be costly to setup, requires additional technologies to support a network of data sources

AgBioData
Data Federation Training WG

Expert Presentation: GraphQL

Asis Hallab

Use case: Direct query of a data source with a flexible query language

Pros: Lots of flexibility and high speed data access

Cons: High cost to establish a shared data model within a network of data sources

GraphQL

Expert Presentation: Globus

Natasha Pavlovikj

Use case: Efficient storage and sharing of large datasets

Pros: Powerful data sharing functionality, suitable for large datasets

Cons: Subscription based service, closed source, moderate to high learning curve to setup a storage node

Expert Presentation: SOLID and Linked Data

Mark Wilkinson

Use case: Interoperability of datasets when the ownership and control of data is important

Pros: Individuals retain full control of their data, RESTful web services to enable interoperability

Cons: High learning curve and technical knowledge required (for now), specialized tools for interoperability still in development

Data Federation Training Module

Short Term:

- Training module public website
- Expert presentation recordings
- Working group analysis of each tech
- Recommendations for some example use cases

Future Work:

- Additional technologies reviewed and added
- Pilot program to build out an example use case in the AgBioData community

AgBioData

Data Federation Training WG

Data Federation Training Module

https://github.com/AgBioData/DataFederation_WG/wiki/

Members

Abbas Saka	*Sectoral Policies and Institutional Support Manager*
Adediran Daniel Adewole	*Helix Biogen Institute*
Alberto Camara Bellesteros	*CBGP UPM/INIA-CSIC, Madrid, Spain*
Bob Cottingham	*Oak Ridge National Laboratory*
Can Vuran	*University of Nebraska-Lincoln*
Ghulam Sarwar	*Cotton Research Station, AARI, Faisalabad Pakistan*
Jennifer Clarke	*University of Nebraska-Lincoln*
Jinha Jung	*Purdue University*
Marcos Paulo da Silva	*University of Arkansas*
Mark Wilkinson	*CBGP UPM/INIA-CSIC, Madrid, Spain*
Monica Poelchau	*USDA-ARS*
Paola Pesantez	*Washington State University*
Peter Selby	*Cornell University*

AgBioData

Data Federation Training WG