Engineering, Architecture, and Informatics

Day One: Pointed Questions
# EAI Members

Matt Vaughn  
James Taylor  

Bob Muller  
Lukas Mulller  
Mark Forster  
Paul Kersey  
Francis Oullette  
Klaus Mayer  

Rion Dooley  
Hiren Joshi
What’s the AIP model for data storage and hosting?

Project is responding to:

• Sun-setting of TAIR

• Constellation of un-integrated boutique databases

• Emergence of new, high volume data sets

• Desire for cross-species inquiry

• Desire for dynamic, operational data integration

• Need to enhance discoverability
Some Models

- One central repository
- Geo-replicated central repository
- Distributed federated data sources
- Hierarchically federated data sources
One central repository
Replicated central repositories
Distributed federated data sources
Hierarchical federated data sources
“Cloud” storage

• Will people be willing to store their data sets non-locally?

• Should they? Probably. Let the professionals handle physical data management. But…

• Need to provide explicit QoS agreement

• Need to provide surety than the data can be retrieved in its original format for re-use and as insurance against defunding of the storage resource
What is your model for versioning, attribution, ownership, etc?

- Data releases must have version support at the resource and service level. What questions does that raise? Ensembl is good at this but they run into challenges.

- As data is collected into ever larger aggregates, how is attribution maintained? How is attribution maintained?

- If AIP takes on responsibility for housing primary research products, what questions need to be addressed on ownership, permission, identity management and verification?

- Who holds responsibility for evaluating, developing, and maintaining standards and their support tools?
What is your species-level data model?

- Avoid using Col-0 as an informatic baseline?
- How will the species’ genome structure be defined?
  - As a graph? Are there tools that can work with such a structure?
- For non-genome data, presumably associating metadata with these will address the issue
- What is the definition of an accession or ecotype?
Are you adopting a collaborative model for information curation?

- Avoid the bottlenecks associated with relying upon a small group of professional annotators
- Does the decision need to be at the level of each module or is this a systematic AIP activity?
- What are reasonable user interfaces for community curation activity? Web apps? Do you support machine interfaces as well?
- What is your incentivization process to maintain interest in entering and maintaining community curations?
How are you going to pay for the uninteresting stuff?

• Low-level engineering and development and maintenance that’s not immediately applicable to science. As we move to a science platform, its even more important that the platform itself be maintained.

• Build as little as possible from scratch. New code => technical liability.

• Do you need unorthodox funding schemes, such as microtransactions?

• Fund maintenance unofficially on the margins? <- Not the answer

• Rely on iPlant for low-level elements, since they are likely to be generic to plant science?
What’s the role of iPlant?

• iPlant offers a software and hardware CI aimed at the biological use case

• Capacious, high performance storage. File as well as relational data. Allocatable resource.

• Access to on-demand high performance computing via XSEDE

• An extensible analysis environment plus a programmable API

• Advanced virtualization technology

• Integrated authentication and identity management

• Committed to data interoperability though we are just starting to work on it.
Federation

• BIOMART offers a way to do federation
  • Appears to have a scalability issue with number of federated physical sites
  • Requires conversion of data into Biomart schema
  • May be more oriented towards sharing derived knowledge than primary data
  • Requires adherence to naming scheme