ENGINEERING, ARCHITECTURE, INFRASTRUCTURE/INFORMATICS WG

Co-chaired by Matt Vaughn and James Taylor

Collaborative Annotation

Seems to be consensus that community-led curation helps avoid the bottleneck of relying entirely on professional curators

- Not just for genomes. Some model examples:
    - In genomics, SGN is an example. Appointed species chief that oversees annotation and can delegate responsibility. 1500 active curators.
    - ChemSpider - very successful. Funded by RCS. Automatically push data from RCS journals into community curation.
        - TAIR has community annotation as well, and does literature extraction.
    - RFAM uses Wikipedia itself as a platform. This requires an agreement with Wikipedia foundation (iPlant tried this early on and was frustrated)
    - EBI asks for functional annotation from authors who publish papers referring to specific genes.
- Estimate that 500-1000 contributors are required for critical mass
- Buy-in is crucial - some communities that offer collaborative curation have little success.
- Incentivization model for community curation will need careful consideration.

Architectural considerations

- When developing IAP modules, should decide up front whether the content will be community curated and built application to suit
- Immediate idea is to use a Wiki. We're all familiar with it via Wikipedia.
- Wiki content isn't machine readable because it's a mixture of syntactically ambiguous tags and free text
    - Linkages between nodes in a Wiki are fragile because they are written into the content
- Semantic extensions to Wikis have been proposed.
    - Users need to training, platform presents a UI, lose ability to edit Wikitext directly.
    - Still not that machine-readable.
- Minimal standard syntax between Wiki platforms, so IAP would need to settle on a single platform
- So, what's a possible answer?
    - Structured capture into a schema
        - Supplement with tagging interface and the ability to specify free text where appropriate
        - Extensible schema.
            - One reason people move to pushing metadata into free-text is presence of an inflexible schema for information capture.
        - User interface needs careful consideration. Should offer guidance and assistance rather than getting in the way. For an example of the latter, see GEO.

Versioned or Conflicting Information

Fundamental question: How to allow users and their computational delegates to assign authority to information?

- One model is to offer all channels, allow/make users decide when they do a search or perform a computation.
    - But, most don't want to have to decide.
    - Also, for complex queries this can lead to issues due to the combinatorials
- An important role of an annotation resource is to provide a ground truth.
- Assignment of authority: Tools should allow users to override ground truth with their own assertions
    - User assertions must be portable into the future.
        - They are the most important kernels of information because they represent pre-public knowledge.
    - Make it obvious when ground truth is updated
    - Offer tools to map assertions to the updated truth
    - Older versions must be computationally available
        - Information SERVICES need to be versioned not just flat file dumps.
        - Simple RESTful example: http://aip.org/arabidopsis_thaliana/v10/locus/AT4G25530/cds
- Strategies for classes of versioned information
    - Coordinate space
        - Supported via Liftover or similar protocol
            - Liftover is more than file format. UCSC toolkit supports it natively for computing across versions
        - Versioned files and web services
    - Names
        - Versioned entity synonym tables
        - Versioned web services
    - Concepts
        - Changes can be expressed via updated ontologies
        - Need robust strategy for versioning ontologies
        - Web services to allow ontology traversal and query
- Semantic web technology can help reconcile competing or parallel truths
    - Multiple ontologies can apply to a term. Describe relationship among the ontologies themselves.
    - Semantic reasoners can ascertain equivalence.
        - But we will need better support for describing these relationships than Protege

Data storage and presentation as web services

How can we provide high-performance machine-readable interfaces to existing data?

- Many MODs and other data resources start with normalized schemas to fully represent the biological information. Computational query interfaces suffer degraded performance as they assemble complex queries into data views. For an example, see TAIR locus page. This is a bottom-up design.
- The PLAIN data warehouse scheme being developed by TAIR team is query-focused, and is a top-down design
    - Start with use cases, determine views needed, develop schema (can be atop a relational, normalized database) for presenting the view.
    - Pre-compute/index these views. Present to consumers quickly!
    - Can build applications as well as web services to expose views
    - Can provide interface to creating new views and associated query interfaces.
    - TAIR data will be presented, after sunset, via this data warehouse.

    - Not designed for extension - this is to house the current TAIR data
    - Adding a new data type = adding a new view
- Data warehouse approach may be generally useful to presenting legacy data from other sources that has not made it into TAIR by sunset. Once built, they can be maintained with minimal work.
- Since they can present as web services, data warehouses are computable and can be federated.

Federation of data

Federating data prevents any one site from bearing the burden of all storage and retrieval, but creating a robust network of resources is a complex undertaking that doesn't have a clear path forward.

- Current efforts
    - ELIXIR, which envisions a distributed framework with EBI as a hub, has just been funded. Implementation TDB
    - iPlant is currently focused on file-level integration via centralized cloud storage. Works for user data, but not for community-presented structured data.
    - Cross-species efforts such as the proposed PlantHub initiative await funding decisions.
    - Ensembl plants offers a hub for genomics-oriented data for many species.
        - Biomart interface is programmatically accessible
        - But, not the primary repository for some hosted species so there is some duplication of effort
        - Does not encompass other data types (metabolomics, morphometric, etc)
- Joint probability prevents totally distributed data storage. If one runs a query which polls 5 databases, each of which has a 98% uptime, there is a 10% chance it will fail.
    - Argues for presence of some centralized, transparent caching services if we want to do queries in real time.
        - Requires substantial computing and data storage capability. Think of these like the DNS (domain name service) hubs but for larger, more complicated queries.
    - Or, complex queries are run as batch-type operations, with built-in retry on failure. Then, we need to set expectations accordingly.
- We know we want to do this, but everyone seems to be at the same place in terms of implementation.
    - Take-home message is that there isn't a standard to adhere to, or even be attracted towards yet. AIP will need to participate in its development.
- Existing data resources with legacy interface models can be brought in via data warehouse scheme or by adding machine-readable interfaces. But focus should be on future data. At any given time, 90% of your data will be less than 3 years old.
- The Semantic Web – several models for this. Can be quite powerful, but everyone needs to participate. Need to figure out how to fund the additional (usually minor but still non-zero) development needed for SW support.
    - Interop and discoverability interface should probably work at some level WITHOUT semantic technology.

Taxonomic scope of IAP

The AIP will indeed focus on Arabidopsis thaliana, but should offer better support for related species and accessions. Currently, there's a high degree of friction to even compare accessions, let alone sister species.

- Focus on *Arabidopsis thaliana*
    - Try to encompass the species in the data model, rather than just Col-0 accession

        - Representation of accessions/ecotype-specific data in a way that minimizes bias against non-reference accession
        - Present full references where appropriate, present as differential information where not.
        - Ideally, genome sequence represents as a graph where paths correspond to specific accessions.
            - Edge weights can encode certainty/quality information.
            - Conceptually this works, but present generation of tools doesn't function on graphs but on linear contigs.
    - Support cross-species inference.
        - A major value of Arabidopsis will remain as a species that researchers in other systems can turn to in which basic science experiments have been done.
        - Need to facilitate link-out to crop and other model species
        - Homology mapping – perhaps maintain a canonical Arabidopsis -> Green plants homology index.

The Potential Role for iPlant

iPlant is developing a national plant science cyberinfrastructure atop the existing national XSEDE CI. It develops integration and presentation code to make these powerful resources consumable by plant sciences.

- iPlant offers:
    - User-friendly analysis environments including the Discovery Environment (a general purpose bioinformatics analysis platform), the DNA Subway (an educational interface to genomics and phylogenetic applications), PhytoBisque (a high capacity image cataloging and analysis system), each of which consumes the underlying CI and presents a different view to specific consumers.
    - Access to extremely powerful computing resources at University of Texas' Texas Advanced Computing Center, the San Diego Supercomputing Center, and the Pittsburg Supercomputing Center (all three are institutions with strong commitments to the life sciences in high performance computing).
    - A unified, Terascale, high performance, sharing oriented cloud storage solution. The underlying software model supports expansion of individual storage resources as well as multi-site federation and georeplication.
    - User-provisioned virtualization, based on Amazon EC2 functionality (via Eucalyptus) and Openstack. Virtual machines support several use cases including collaborative software development, database and web service hosting, and application of desktop functions to cloud-hosted data.
    - A series of low-level REST APIs for authentication, data management, event monitoring, logging, application discovery, computational job invocation, delegation of authority, user profile discovery, and more. These APIs can be consumed by third parties, such as Galaxy and Bioextract, to construct sophisticated user interfaces that leverage the iPlant CI.
    - A community-extensible framework for publication of functional and interoperable instances of bioinformatics software.
- The iPlant project is in its 4th year and is going into its renewal process starting Jan 2012. The components of the current iPlant CI that are based directly on the national CI will remain stably available regardless of the renewal status of the iPlant project.
- IAIC-developed applications hosted within the iPlant CI will have access to these same resources, and will have the added advantage of co-localization with the physical computing and storage resources.