Feedback from research communities for IAIC Design Workshop

OVERVIEW
In November 2011 IAIC Interim Director Blake Meyers contacted leaders of ten scientific projects or databases to request their advice, and recommendations to help guide the Arabidopsis community effort in developing a new Arabidopsis Information Portal (AIP).

The Arabidopsis community has long maintained strong international collaborative ties and fostered relationships with researchers studying other plants (e.g. maize) and other model organisms (e.g. Drosophila or mouse). Seven of the groups contacted generously responded with insights and advice that were incorporated into materials prepared for the Design Workshop by the IAIC Steering Committee and shared with meeting participants. This valuable feedback and the collection of documents prepared before and during the Design Workshop will, no doubt, continue to be useful to the Arabidopsis community during the AIP-development process and for years to come.

Free text response from Paul Sternberg (Lessons from WormBase for a designed Knowledgebase) follows grid responses.

I. Community and Social Challenges and Approaches

a. How do you get people (researchers, community) to buy in to the project/database? For example, how do you get people to support, utilize, and contribute to your project? Do you have any 'lessons learned'?

Helen Berman Protein Data Bank	It is critically important to actively work with all the stakeholders including the depositors and users of the data. Over the years we have attended professional society meetings, sponsored workshops, created Task Forces, and used electronic media to communicate. This is a long and difficult process but without buy in there is no resource.
Eva Huala TAIR	The most important factor is providing data and tools that people find useful in their work. Other things that help: workshops at meetings and other outreach efforts to make people aware of your resource and teach them how to use it, reciprocal links to other high quality websites to help people find your resource and improve page rank in internet searches, making the site intuitive and easy to use, providing data, tools or services not available elsewhere, for example building custom datasets upon request.
Chuck Gasser, ML Guerinot; TAIR (BoD)	Probably the most critical component for user buy-in is ease of use. I think this is one of TAIR's strengths, having attempted to use databases for other plants and other eukaryotic organisms. TAIR shows very high user numbers relative to other databases, especially when compared to the size of the other communities. Use of the database is one of the best measures of the value of a database. The second most important is support and contribution to the project. This is much more difficult to induce. I believe that TAIR's usability stems in part from the talent and dedication of the staff who have been coordinated to really care about the ease of use and value of the tools they make available. This attitude makes them seek out and rapidly incorporate ideas from the community the serve in general, and from the advisory board on an annual basis. They have been exceedingly responsive to suggestions from the board.
Carolyn Lawrence, Taner Sen; MaizeGDB	We keep communication channels open and respond QUICKLY to provided guidance and feedback. We have learned that failure to respond results in community members who don't provide future guidance and feedback. We publicize new items at MaizeGDB on the front page and via in-person training sessions as well as at the Annual Maize Genetics Conference. Stakeholders' data being advertised on the front page of MaizeGDB as a news item is a huge incentive for researchers to contribute their data to MaizeGDB. The in-person

	training sessions help to convey the message that MaizeGDB is here to serve them, not the other way around.
Janan Eppig; MouseGD	• Initial buy in is achieved when the perceived need is great. For the Mouse Genome Database the initiating need (in 1989, pre-genome), was the growing # of identified genes and the inability to maintain a reasonable map manually – nomenclature, mapping data, and the integration and analysis of those data was beyond the ability of researchers to maintain in an ad-hoc way. The community needs to recognize that without the resource, the building-blocks / infrastructure on which new research is built is unsustainable. As Arabidopsis already is well-positioned as a key model system for many plants and has a completely sequenced genome (with annotation), the argument for need is clearly there. • Continued buy in, including continued community support, utilization and contribution, is achieved through continued proven reliability, accuracy, up-to-date information, and comprehensiveness (in the database's defined scope). One of the big reasons databases fail their communities is when data/information becomes stale, when there are significant gaps in what users anticipate should be there, and when inaccuracies are repeatedly detected and not addressed / resolved well. It is important for a model organism database to be thought of as the "gold standard" for its data content.
Kathy Matthews, FlyBase	Fill common needs with a useful product and focus on high-value data sets to reach critical mass. If 'everyone' is using your database it is in everyone's interest to have their data included. Volunteer efforts for the greater good, despite the best intentions, have not worked well in my experience. Participation needs to result in a direct benefit to the participant's research/teaching/funding/career.
Francis Ouellette; OICR	It is very much a top-down enforcement: there is one way to contribute, and is one place to submit, and really one place to get data from, so that part is "easy".

b. How do you get people to collaborate meaningfully with the database in terms of contributing data sets, adopting standards sets by the database, citing it, etc., when they're not funded directly from a project? Do you have any 'lessons learned'?

Helen Berman Protein Data Bank	The answer to this is the same a 1a. The community must feel it owns the data.
Eva Huala TAIR	Most important: make data submission as simple and intuitive as possible. Also, acknowledge data submissions by displaying the contributor's name attached to their data, provide easy access to a contact person they can interact with to ask questions and get help. Make sure there is community support for a standard before adopting it. Collecting sufficient metadata at the time of submission makes data more useful and allows it to be integrated with other datasets in the future.
Chuck Gasser, ML Guerinot; TAIR	This is really difficult and I don't have any helpful insights.
Carolyn Lawrence, Taner Sen; MaizeGDB	Data are contributed to MaizeGDB and members of the MaizeGDB team seek out new, high-impact datasets. Researchers contribute their data for a few different reasons: because a funding agency requires it, to get others to use the new data, etc. In addition to increasing the visibility of datasets and making those accessible, another 'carrot' that we sometimes provide to data contributors/collaborators is coauthorship on our papers when their contributions to a project are significant. For example, see the author list in the attached POPcorn manuscript, which is currently under review. In addition, when a cooperator (i.e., maize researcher) does a gene review, we put a star in

	his/her page at MaizeGDB (e.g., see http://www.maizegdb.org/cgi-bin/displaypersonrecord.cgi?id=15938). Small incentives such as this help. About standards: we strive to be as transparent as possible to our users. As an example, we only accept submissions to our genome browser if the data are aligned to the current version of the assembly (B73 RefGen_v2). We make sure to put this information on our page for contributions (http://www.maizegdb.org/data_contribution.php).
Janan Eppig; MouseGD	Data will be contributed by researchers if they think it is to their advantage and it must be extremely simple to do. Some things that have been successful for MGD include 1) informing (via email) the researcher when their data are available publicly and providing them accession IDs when appropriate; this includes curation of publications. This serves as an act of engagement; researchers will often then go and review what the database is displaying and are more likely to contribute in future; 2) when presenting talks/posters/demos, point out the "extra exposure" their data receives by its integration into the database and placement in the context of other knowledge; 3) submission forms or instructions have to be super easy – for example, we take text descriptions of phenotypes and do the work of assigning Mammalian Phenotype Ontology terms ourselves and standardizing nomenclature; we will also take many formats for data – spreadsheets, text files, word documents, etc. If the researcher has to either re-format all his/her data or re-enter data, then contributing data becomes a burden -- a stick, not a carrot. As to citing databases, unfortunately this is rarely done in publications, even if the instructions for doing so are online and reviewers rarely notice this oversight.
Kathy Matthews, FlyBase	You have to offer them a meaningful benefit – expert advice on standards, perhaps assistance in applying them; increased visibility for their data; added value to data that has been standardized and integrated with other data; review-proof fulfillment of data sharing requirements, etc. – and you have to make it as painless to participate as possible.
Francis Ouellette; OICR	We have a number of working groups from across the data generating centers; we had (for example) submission standards that were written up, and then ignored, and then tight regulations were put in place to not accept data that didn't fit the agreed-upon standards. The lesson here is to engage data providers, get them to agree of standards, and be very hard at the other end to only accept data that fit these standards. Other lesson is that you make sure you get your standards right the first time: going back to a group and telling them you want another field filled out is not good, and they will eventually ignore you (specially the third time you go back!).

c. What successes have you had to get other groups together to obtain funding across multiple agencies? Are there any individuals at relevant funding agencies (US or elsewhere) that might be worth including in our meetings and activities? Do you have any 'lessons learned'?

Helen Berman Protein Data Bank	We have worked effectively with a consortium of US funding agencies including NIH, NSF and DOE. We have also supported the efforts of our international partners to obtain funding from Wellcome Trust, JST, EU and BBSRC.
Eva Huala TAIR	(Nothing to suggest here.)
Chuck Gasser, ML Guerinot; TAIR (BoD)	(No response given)

Carolyn Lawrence, Taner Sen; MaizeGDB	MaizeGDB has base funding (~$500K) from USDA-ARS via congressional appropriation. Outside funds constitute approximately one-third of the annual MaizeGDB budget. The bulk of outside funds come from the NSF, primarily as subcontracts on researchers' proposals where MaizeGDB will become the long-term repository to house the outcomes of a given project. Other groups who have provided funds to MaizeGDB include USAID (via CIMMYT), the National Corn Growers Association (NCGA), the US Department of Energy, and the Monsanto Company. It is important to note that when MaizeGDB has Working Group meetings, representatives from NSF, USDA, NCGA, CIMMYT, and other database projects have been invited. We never fail to invite USDA, NSF, and the NCGA. It is primarily through our keeping outside groups informed of what we do that opportunities to collaborate and obtain outside funding arise. In addition to serving roles as subcontract recipients on behalf of MaizeGDB, MaizeGDB scientists Lawrence and Sen have served as PI's representing the MaizeGDB project on funded NSF proposals for the PlantGDB project and the POPcorn project (descriptive manuscript attached).
Janan Eppig; MouseGD	Mouse might be an exception to the rule, as there have been several "big" production-scale projects set up trans-NIH (e.g., the Knockout Mouse Project, KOMP) to produce research resources (but not databases). These are still all within NIH, but different institutes. A person from NIH that you might consider including is Dr. Franziska Grieder, director of the Division of Comparative Medicine, NCRR, NIH (GriederF@mail.nih.gov). However, this is still "NIH-only". NIH does make foreign grant awards, but one thing that is of concern is that neither the EU or Wellcome Trust or other European funders do the reverse. There have been a few instances of US researchers with NIH funding and European researchers with EU funding "separate, but collaborative" projects. One must be extremely careful in these instances that the separate projects are doing completely different things, e.g., such as gene model curation versus Gene Ontology development, rather than, for example, sharing curation of gene model annotation work. Funding agencies remain protective of "their own projects".
Kathy Matthews, FlyBase	(No response given)
Francis Ouellette; OICR	Nothing to share here, except that we received funding from the provincial government, a number Federal Canadian agencies, including Canada Foundation for Innovation (an agency that disperses funds for infrastructure: sequencers and computers in our case), and that maybe Canadian groups working internationally could get some funds in a new Genome Canada competition that will be announced soon (early 2012).

d. To what extent is your database (e.g. WormBase, Maize GDB etc.) a portal to access distributed databases, and if it is distributed, how difficult has it been to coordinate the various contributors? Do you have any 'lessons learned'?

Helen Berman Protein Data Bank	Response not given
Eva Huala TAIR	TAIR has done this in a very limited way, we have a GBrowse track (Vista) that is served up from JGI. After a slightly rocky start this has worked pretty well since we impressed upon them that the VISTA server needed to remain up all the time, no extended downtime allowed. I would suggest setting a standard for availability that modules are expected to meet, for example 99% availability (up to 7 hours of downtime per month) , or 99.9 % availability (43 minutes of downtime per month).
Chuck Gasser, ML Guerinot; TAIR (BoD)	TAIR has links to a variety of other databases including insertion mutants, microarray data etc. I don't know how hard it was to set up the links, but they have done a great job and could serve as a model for this.
Carolyn Lawrence,	Please see the attached POPcorn manuscript with the following caveats: In contrast to our understanding of the goals of IAIC, POPcorn addressed portal-function needs

Taner Sen; MaizeGDB	for a narrow slice of datatypes, mainly sequence-indexed information. In addition, the projects to which POPcorn connects are mainly short-term in nature with the final project outcomes anticipated for eventual representation in the MaizeGDB resource directly. A similar incentive can be used for IAIC with a tweak: no researcher would like to see their data disappear after painstakingly producing them. Perhaps IAIC could offer to serve as long-term repository to the projects' outcomes at the end of funding period and request a certain amount to be put into grant proposals to fund data transfer and integration.
Janan Eppig; MouseGD	Worm and Maize do this to a great extent, with distributed databases that are then presented to users with a single interface. MGD instead integrates data from multiple sources upfront. MGD has many links into other resources through coordination of IDs. Rather than distributed database maintenance, there are co-curation activities MGD does with other resources – for example, we co-curate gene-to-sequence relationships with NCBI, and we coordinate gene symbol/name assignments with the Human and Rat nomenclature committees.
Kathy Matthews, FlyBase	(No response given)
Francis Ouellette; OICR	This is turning out to be very difficult, for various reasons. I can talk about it, but prefer not to write about it.

II. Technological Challenges and Approaches

a. What are the most important future needs/challenges that you feel must be addressed for your database (or other major databases) for the next decade? (e.g. what is your list of five top needs/challenges.)

Helen Berman Protein Data Bank	The biggest challenge is to get sustained funding. The funding mechanisms available from the agencies are not tailored to creating and maintaining database resources. In my community it is an expectation that the data deposition and access should be free. The reasoning is that the data are obtained with public funds and should thus be publicly and freely available. Database resources that have charged for data have been severely criticized and not supported by a large user community.
Eva Huala TAIR	1) Funding stability, which is essential to minimize staff turnover, retain institutional memory and carry out long-range planning. 2) Providing good service to two disparate user communities – computational biologists and bench biologists. These groups have very different user interface, data and analysis needs, it may be worth developing two different interfaces to the same database to accommodate this. 3) Handling large datasets from the community – the scientific community is typically not very good at sticking to data format and data quality requirements, which means that the receiving database must run scripts to ensure data quality (i.e. all required fields present) and format correctness before loading. Large amounts of new data also require changes to search and visualization interfaces to keep the user from being overwhelmed by large result sets. There is also a question of whether to accept all such datasets submitted to the resource or to make a judgment on data quality and accept only data that meets quality standards. 4) Transitioning from the era of one reference genome to many sequenced genomes per species. Many questions arise from this, including some that need to be answered by the community as a whole, such as how to provide locus identifiers in different ecotypes, whether a pangenomic gene and protein sequence dataset including functional / wild type versions of all genes found in a species, and if so how to determine which copies are functional. Questions for data providers include how to efficiently store, query and display large quantities of polymorphism data. 5) How to enable and encourage community submissions and updates without

	abandoning data structure or data quality. Wikipedia models don't inherently provide the structure or quality control mechanisms needed to produce consistent, high quality integrated data suitable for use in computational analyses. Lack of community participation is also a big challenge in this area, most community curation tools developed by genome databases are barely used.
Chuck Gasser, Mary Lou Guerinot; TAIR (BoD)	- Continue to curate newly generated gene function and expression data, improved quality of genome sequence etc. in a funding environment that only appears to value things that are "new" and "novel". The community needs these, but also needs timely access to the highest quality data available. - speed at which data can be accessed. The biggest complaint about TAIR is its slowness.
Carolyn Lawrence, Taner Sen; MaizeGDB	i. Long-term funding commensurate with what is needed to meet the articulated informatics requirements of maize researchers. ii. "Big data" – especially large, sequence-indexed diversity data (e.g., SNP) a. Data storage: how? Database or binary? b. Analysis c. Visualization iii. Diversity of available data formats and changing formats iv. Need to accomplish project deliverables while remaining agile: adaptation to emerging technologies is required for both data generation technologies and computational technologies. The "problem" of making data available is neither "solved" nor "solvable" – it is ongoing. In addition, stakeholders use, e.g., Google, Facebook, etc. and do not understand why their information resource is not as responsive. In addition to the very real problem of staying agile, there is a perceived problem from stakeholders who expect large datasets to be quickly analyzed and available. The same scale of data analysis problems exist in, e.g., banking and healthcare, but data consumers in those fields seem to understand that a particular query on a database may take hours or even days to complete. v. Curation is a slow, tedious and time-consuming process. However: without curation of data, no high-quality datasets are made available.
Janan Eppig; MouseGD	1) Continued sustainability in the current funding/economic climate. 2) Rapidity of growth of data (e.g., raw sequence data, genome variation data among individuals and inbred / segregating strains; whole genomes of strains and related species, etc), particularly the implications for annotation streams, curation, and integration processes. 3) Increasing complexity of data. In the next 10 years mouse will have phenotyping data on 20,000 knockouts from large-scale pipeline projects beginning now. In addition to volume, data being captured is increasingly multidimensional, e.g. using high resolution imaging technologies. 4) Expansion of community expectation vs. funding (increasingly large projects "expecting" MGD to incorporate their data post-facto, whether or not those types of data are currently ones supported within MGD. Some large projects have database support for the length of the typical consortium project, usually 5-years; then it "has to be deposited with a relevant resource". 5) Need for new nomenclatures and ontologies to support existing and coming data types. Although many standards, nomenclatures, and ontologies are already in use in MGD, others just do not exist yet or are too immature to be useful for curation, but are needed asap (e.g. there are anatomy ontologies, but no mature cell-type ontology). 6) Storage and bandwidth for maintaining and 'transporting' data. Although I list this, storage capacities have always managed to "keep up" with the march of technology. Bandwidth may be more challenging as data sets increase in size and complexity – but we also may be entering an era where the cloud stores the data and the analyses can be done "there" (i.e. reducing the need to support 'actual' transport of data).
Kathy Matthews,	The volume of high-throughput data is frightening. There are technical challenges, which I have no special insight into, but also curational challenges. We need scalable ways to

FlyBase	distill and summarize data so that the take-home is more accessible and more portable across different domains of expertise.
Francis Ouellette; OICR	i. The correct data model ii. The correct community model iii. Funding iv. Scalability v. Community endorsement/Usefulness (if you build it, and they don't come, it's a waste)

b. What were the technical challenges of integrating data from many web sites, sources, or types, and how do you coordinate data collection across sources? Do you have any 'lessons learned'?

Helen Berman Protein Data Bank	A major challenge with the integration of data from multiple sources is determining and maintaining the provenance of each data source. Many sources archival data sources supplement their "primary" data with information inferred from other resources. It is increasingly difficult to decouple the primary and derived information and this often leads to information cycling between resources. The latter situation makes the correction and update of information among resources very difficult.
Eva Huala TAIR	The biggest challenge in data integration is determining which objects in different datasets are identical. Even for objects with well established identifiers (AGI codes) this is still a significant problem, because not everyone uses the identifier (some may use gene symbols, UniProt IDs, etc), version information (which genome release) is not usually provided, typographical errors are not uncommon, etc. For objects like alleles or germplasms without standardized identifiers the problem is even larger. It's helpful to have data of each type flowing in only one direction – e.g. submitter to module to portal. If two modules take in the same data type and try to exchange and synchronize their data, the process will be more inefficient and error-prone. - Another big challenge is adapting the database schema, data pipelines and user interfaces to new data types, a continuous process for a scientific database. There is a tradeoff to be made between having a flexible schema (slow and awkward for querying, provides fewer data constraints which can make it harder to detect bad data) and a robust schema with better performance and more ways to ensure data quality (but requires more work when the schema needs to be updated to accommodate new data types). - Lesson learned: Planning for future updates is an essential part of the upfront work involved in incorporating a new data type. Things to be considered include how to incorporate corrected submissions from the original submitter and partially overlapping submissions from other submitters. It's important not to underestimate the amount of effort required to build and maintain such pipelines (ETL (Extract Transform Load) technology is helpful here). - Another important point is whether and how to version individual data points, large data releases or the whole database. Versioning allows scientists to correctly reference the dataset they used for an analysis in a publication. Enabling retrieval of older versions of datasets allows researchers to compare a published analysis method with a new method using the same dataset, or verify the results from another group.
Chuck Gasser, Mary Lou Guerinot; TAIR (BoD)	No response given
Carolyn Lawrence, Taner Sen; MaizeGDB	Please see the POPcorn manuscript.

Janan Eppig; MouseGD	The biggest challenge is use of data standards, nomenclatures and accession IDs. While the need for these is much more accepted and appreciated by the community than 10 years ago, it is not a solved problem. Very significant QC and curation effort remains devoted to standardizing data collected from different sources. It helps to work with these other groups early on and hopefully establish consistent data collection and adherence to standards, but, sadly, this is still more the exception than the rule.
Kathy Matthews, FlyBase	(No response given)
Francis Ouellette; OICR	There is really only one route here: either you get the community to endorse one set of standards/formats/ontologies, or you spend your time/energy doing that yourself. Both examples exist in our communities, with the first one being "easier" and the latter more painful.

c. Do you see any gaps in database architecture and technology and how are you going to find solutions to these limitations? What are the current datatypes you're hosting and (if known) indicate how they may differ to those that the Arabidopsis community may need to consider for the next 10-15 years.

Helen Berman Protein Data Bank	The recent emergence of database engines to support text and semi-structured data (e.g. CouchDB & MongoDB) illustrates successful alternatives to traditional relational database technologies employed to address specialized content types. While there have been attempts to create extensions for the relational database systems to handle specialize scientific data types (e.g. chemical graphs), there is now an opportunity to adopt NOSQL type solutions providing detailed access to natural data types common in bioinformatics (e.g. sequence alignments, structure alignments, and annotation domain assignments). Application of domain specific database technologies may result in dramatic advances in query and analysis functionalities in the future.
Eva Huala TAIR	A general comment here on open-source solutions – in many cases there are commercial products to do the same job, and although they are more expensive it may sometimes be more cost effective in the long run to pay for the greater functionality and stability you can get with these products, along with updates that adapt the product to new hardware and operating systems, and product support. Examples include indexing, ETL software, text parsing, relational database software. Regarding data types, how best to deal with a huge number of low-quality genomes for each species needs some careful thought. It will be necessary to integrate, compare, extract useful information and display this analyzed data rather than the raw data of the different genome sequences. Somehow the uncertainty inherent in the lower quality sequences will need to be conveyed.
Chuck Gasser, Mary Lou Guerinot; TAIR (BoD)	No response given
Carolyn Lawrence, Taner Sen; MaizeGDB	Refer back to II.a.2: "Big data" for gaps in database architecture. Datatypes served include (but are not limited to) loci, sequence, genome assemblies, genetic maps, phenotypes (images and descriptions), genetic stock records, allele/variation data, functional information, and references. The only major difference with respect to the Arabidopsis community is the need at MaizeGDB to serve plant breeding data given that maize is a crop.
Janan Eppig; MouseGD	Relational database management systems are not really ontology-friendly, so we do some pre-compute to help. In addition, we are moving to a split system that uses a different back-end and front-end configuration. The back-end

	database (where the data are stored, loaded, edited, QC-ed, etc.) of PostgreSQL, which is a relational database management system and a front-end of Lucene/Solr indices so that effectively much of the PostgreSQL database is de-normalized to allow for faster searches and faster web-page assembly. I would not worry so much about currently available architecture and technology not being up to the task. In general, model organism databases adopt standard, stable technologies and are not on the bleeding-edge of technological change. The bigger challenge is the speed at which publicly-funded resources such as TAIR or MGD can actually transform their technologies as they go along, given that in the best circumstances, budgets, re-training of staff, and re-tooling existing systems that 'work' are hard to justify and fund. And, further, the demands for continued improvement to web features and flexible data access methods for users, and the incorporation of new data types, are far more pressing in priority. <u>Data types incorporated into MGD</u>: Genes, pseudogenes, and gene models including CDS and predicted proteins and non-coding RNAs; cytogenetic markers; genomic and genetic maps; nucleotide and protein sequence associations; spontaneous, induced, and genetically engineered mutant alleles; transgenes; QTL; mutant and conditional phenotypes; mouse models of human disease annotations; Gene Ontology annotations; mouse anatomy, Mammalian Phenotype Ontology, gene product names; Gene nomenclature; Strains; SNPs; CNVs and other variations; protein domains (from InterPro); Mammalian orthologs; literature citations; experimental molecular reagents; functional genomics (gene expression); biochemical pathways; Images of phenotypic mutants and gene expression; links to other tools and other database resources. Significant sections of MGD that have to do with human disease models and mammalian orthology are likely irrelevant for Arabidopsis. For Arabidopsis, or any other primary community database resource, the key is to provide the best infrastructure to support that community's resource needs. Clearly maintaining the canonical catalog of genes / genome features is basic and requires identification of gene models (and other genome features), their naming, location on the genome assembly and some set of information that is available about them (e.g. function, expression, phenotype for mutants). This is not a small undertaking! Arabidopsis is used to model other plants and for basic biological understanding – so these core data types are essential.
Kathy Matthews, FlyBase	(No response given)
Francis Ouellette; OICR	I would suggest, relevant to our discussions here, there are two data types: large infrequently used files (e.g. binary alignment files of sequence reads, the BAM files) and everything else. The BAM files are a problem (if you have lots of them), but they are of interest to a very small group of people: most people in any given community will want the data: the finished sequence itself, along with the variations (phased or not) across that genome: you need not build a system that needs to provide easy access to all data types all the time. I think it is OK for some data types to be harder to get, if they are used less frequently, by a much smaller group of people. Technology is changing too fast right now.

d. What are the challenges of hosting a distributed system (e.g. WormBase) and how do you approach solving them? In a distributed system, how do you maintain updates from other data sources on a regular basis? Do you have any 'lessons learned'?

Helen Berman Protein Data Bank	While web services protocols have made it possible to integrate heterogeneous database resources, achieving the reliability and performance necessary to support interactive access to such systems remains a challenge. Within the Protein Structure Initiative Structural Biology Knowledgebase (SBKB) project we maintain a database of well-defined identifier correspondences to facilitate linking relationships to ~150 database resources. A search of the SBKB uses remote web services in combination with queries of a local correspondence repository to obtain an inventory of related information in the distributed knowledgebase. Linkouts are provided to allow users to drill down to details and services at remote sites. Identifier correspondences are updated on a weekly schedule.
Eva Huala TAIR	See answers to 1D.
Chuck Gasser, Mary Lou Guerinot; TAIR (BoD)	No response given
Carolyn Lawrence, Taner Sen; MaizeGDB	See POPcorn manuscript.
Janan Eppig; MouseGD	This doesn't really apply much to MGD per-se, although we do try to maintain synchronization with other Mouse Genome Informatics Resources. In those cases, generally one database is the 'master' and the other the 'slave' so that data updates flow one way (e.g., MGD gene and allele data flow to the IKMC (International Knockout Mouse Consortium) database; Publications in MGD are used to populate MTB (Mouse Tumor Biology Database)). Although there are no 'lessons learned' here; it is simply practical to declare particular data elements to be authoritative from one source to make synchronization easier.
Kathy Matthews, FlyBase	(No response given)
Francis Ouellette; OICR	Distributed/Federated systems are difficult (less socially and more from a software engineering point of view). The large centralized data warehouse may still be the best way forward for the Arabidopsis community. The analysis needs to be about what datasets will be used and querried against vs what needs to be "served", and how often those requests will be: for example having BAM files on a separate cloud server, and all of the rest of the data on a central (non-cloud) infrastructure may be OK if most people don't need BAM files of the 1000 Arabidopsis genome. The truth of the matter is that although you will want this data type, most users will only want the processed end-point (the differences in genotype), not the alignments (although, obviously a few will want that as well).

Lessons from WormBase for a designed Knowledgebase.

Paul Sternberg, HHMI/Caltech, December 2011. pws@caltech.edu

WormBase is the major database of information about the genome, genetics and biology of Caenorhabditis elegans and other nematodes. It was established by Richard Durbin (Sanger Institute), Lincoln Stein (then Cold Spring Harbor Labs), John Spieth (Genome Sequencing Center of Washington University) and Paul Sternberg (Caltech) in 1999 from the ACeDB database associated with the C. elegans Sequencing Consortium. Lincoln Stein has since moved to the Ontario Institute for Cancer Research (OICR). In 2011, the Cambridge UK group moved to the EBI under leadership of Paul Kersey. Cambridge handles large-scale sequence analysis and database integration. Washington University handles most of the manual sequence curation. Caltech handles biological curation from the literature. OICR handles the website.

I discuss here a few of the lessons I think WormBase has taught us. (Some of these were previously known…)

1. Management of the project should ensure that the project serves all classes of users: individual researchers, educators and students, power users, and non-academic organizations. A serious user, in most cases a biologist, should be in charge or strongly represented in the management team, to ensure that the information resource anticipates growth and direction in the field and serves the community.

2. Multisite projects can work. One advantage is that they prevent insularity. This extends from sociological influences to software. Any one site will not have the best software for everything. It follows that insularity is detrimental. Of course, the downside of multisite projects is decreased or slower communication; modern technology and periodic face-to-face meetings can diminish this issue. Visits by single people from one site to another site can be efficient, rather than a large all-project meeting.

3. Flexibility among personnel and sites is crucial to avoid inefficiency.

4. Flexible curation pipelines allows rapid response to changes in priorities, new data or personnel changes. An example of a rigid pipeline is one in which each paper is analyzed once for all data types that are deemed relevant.

5. While natural language processing (computational linguistics; text-mining) is not perfect, automation can make manual effort more efficient.

6. Be honest about what you can and cannot do. Link to your "competitors" site if they have something you don't have.

7. Be patient. Most bioinformatics funds are not going to the central, longterm information resource but rather to individual short-term projects (3-10 years). Ideally you would work with these projects to incorporate relevant information, but sometimes you just have to wait.

8. Balance heaviness with utility. Don't over-engineer software or processes. For example, don't plan for features that might be needed five years out in the guise of being more efficient. It is more likely that priorities or the landscape will change than your plan will hold. Project management should be effective but not eat up valuable time.

9. Avoid Ego. Databases exist to serve the community.

10. The perfect is the enemy of the good (Voltaire). One should know what the ideal way to do something is (e.g., "ontological purity") so when you choose to deviate you are doing it consciously, not from ignorance or sloppiness.