GTL Knowledgebase Community and User Issues

For users, the GTL Knowledgebase (GKB) would be an important tool for accelerating discovery and hypothesis-based fundamental research and rapidly translating this research into critical, practical solutions for global climate change, environmental remediation, energy independence, and alternative fuels. The GKB significantly would influence both basic and applied science, ultimately providing a valuable resource in numerous areas of industrial research. Furthermore, the knowledgebase would lead to transformative technologies, serve as a tool for the overall biological research community, and provide a key methodology for transferring emerging knowledge to industry.

The GKB would uniquely assist the GTL and broader research communities by integrating environmental and biological information into a unified system enabling users to extract existing knowledge, formulate hypotheses, create new data networks, and generate models of complex biological systems. Achieving these envisioned capabilities would require the GKB not only to carry out its data archive mission but also to become a working environment for testing hypotheses using shared data.

Without effective access to information, even the most highly integrated, standardized, complete, and correct set of analytical data is unusable by the broader research community. The knowledgebase project would provide such access by serving as a focal point for data sharing and information exchange within the GTL community (see Appendix 1. Information and Data Sharing Policy, p. 59). The GKB would facilitate these exchanges by supporting a wide variety of data types generated by the general research community and then integrating the data into a common framework linking otherwise disparate systems. Thus a significant challenge for the GKB would be to provide a robust public resource that allows researchers to access GTL data in diverse and flexible ways.

The GTL Knowledgebase also should develop and enforce a data sharing policy that both protects individual researchers' data and ensures the broader scientific community has easy and open access to GKB information. This would require the GKB to actively remove obstacles that typically impede data access and to work vigilantly with scientists to monitor and improve the knowledgebase over time. In addition to initial development of this unprecedented system, the GKB—through outreach, training, and survey-based performance evaluations—would need to continuously assess the critical DOE- and GTL-relevant data needs of researchers and the knowledgebase capabilities for providing them.

Knowledgebase User Community

Findings

An advantage of the GTL Knowledgebase would be having a manageable community of potential users showing early interest in the GKB. The following are general classifications of these target user groups:

- **Data users.** Knowledgebase design should enable users to easily and quickly find data and tools relevant to their research. Such access to GKB data could be achieved through Web interfaces, bulk downloads, and machine use via Web services. GKB features beneficial to researchers include user feedback and ratings showing new users which knowledgebase tools other scientists have found useful. In addition to data, the GKB should provide users with estimates for errors in data and with information about the experiments from which data were derived (e.g., the parameters used on equipment, scientific questions data producers sought to answer with the data, replicate information, and the methods employed). Data users would interact with GKB information in various ways. Some researchers might be interested in qualitative information (e.g., binary data such as presence or absence and protein localization); others might wish to conduct more quantitative analyses (e.g., exploring gene expression levels across different conditions).

- **Data producers.** The GTL Knowledgebase should provide a mechanism enabling data generators to easily deposit their data and metadata into the GKB. This might require the knowledgebase to accept multiple data formats. In addition, data producers may need embargoes allowing them to incorporate their data into the GKB for queries prior to publication, but this feature must be weighed as part of overall operational requirements. Furthermore, producers would like the ability to determine how their data are used. The GKB could provide this service by allowing users to track publications resulting from the use of knowledgebase data. Encouraging GKB participation among producers likely would require offering clear benefits for data deposition, such as GKB graphics and tools for creating publication-ready figures (as in Cytobank) and the ability to use knowledgebase resources and software for analyzing new datasets prior to publication. Moreover, straightforward data analysis tools would serve as incentives for data producers to become GKB users.
- **Software and tool developers.** Developers of software and tools integrated into or connected to the GKB would like to track and evaluate the usefulness of their tools (e.g., via feedback from other users) and the publications resulting from the application of such resources.
- **Data producers and users.** Many GKB users also would be data producers. As they deposit data, models, and tools into the GKB, these users would use resources others have added. Systems biologists, for example, probably would be both producers and users of data.
- **Industry.** Researchers representing the industrial community would take advantage of GKB data and tools to use research results to make significant contributions to DOE mission areas (e.g., climate change research, bioenergy, and carbon cycling).
- **DOE program staff.** DOE program staff likely would be interested in knowing which data and tools were available in the knowledgebase. Participation from this group of users could facilitate evaluation of data-release policies and provide an incentive for data producers and software and tool developers to contribute to the GKB.

Recommendations

- To optimize effectiveness and service to constituents, GKB planners should clearly determine members of the research community who would use the knowledgebase and the science and technology groups who will develop it.
  - Knowledgebase developers could identify the GKB user community by evaluating and pursuing several sources, including current and former GTL grants, publications, conference attendees, DOE phone directories, target institutions (e.g., DOE's Bioenergy Research Centers, JGI, and national laboratories), and Google analytics.

- Contact lists for probable GKB user groups should be developed.
  - These files can be assembled from several sources, including conference abstracts, author lists from publications representing various disciplines, DOE personnel databases, principal investigators awarded DOE research grants, and scientists who contact the GKB website.
- During design and operation phases, the GTL Knowledgebase should survey individuals from each target user group to measure their requirements for new functionality and GKB adequacy.
- User groups associated with large GTL efforts such as centers and confederations should be enlisted to help develop GKB.

[Source: DOE Joint Genome Institute.]

Knowledgebase Interfaces and Portals

Findings

- All knowledgebase constituents—including data producers and users, tool developers, industry representatives, and DOE staff—would access and interact with GKB through several interfaces. These could include the Web, file transfer protocol (FTP) servers, application programming interface (API), Web services, wikis, videos, online tutorials, and software (e.g., MeV, Quackenbush, R/biocurator, and Cytoscape). Knowledgebase interfaces would be nested in various layers such as an entry Web portal leading to GKB subpages with links to analysis tools and FTP servers for data deposition. Organization of these interfaces and portals should be fairly intuitive to meet the needs of both expert and novice users. Furthermore, users accessing GKB data might generate new types of data or refine existing information, facilitating the iterative process of knowledgebase data improvement.

Recommendation

- The GTL Knowledgebase should engage the help of identified user groups, not only to ensure these communities have effective and easy access to the GKB, but also to constantly assess knowledgebase performance in supplying needed and facile services.

Knowledgebase Outreach

Findings

- Outreach activities are excellent not only for fostering knowledgebase awareness, but also for obtaining feedback on the quality and efficacy of GKB resources. Moreover, these activities provide mechanisms for users to be involved in system development. Basic modes of outreach include presentations at scientific conferences, articles in scientific journals, engagement of researchers in individual laboratories, electronically based announcements (e.g., wikis and mailing lists), multiday workshops at conferences, Web-based instruction (e.g., tutorials, webinars, and bulletin boards), and internships for visiting postdoctoral researchers or students. The GKB also would need to target specific areas for focused outreach.

Recommendations

- Significant GKB resources should be dedicated for outreach and mobilization of the scientific community.
- The GKB should engage user groups to ensure effective, easy access and assess knowledgebase performance.
- GKB developers should engage in active communication with the scientific community to obtain input for GKB development and promote widespread and ongoing knowledgebase use. GKB developers can achieve this through the following activities:
  - Seconding staff within large facilities such as DOE's Bioenergy Research Centers and JGI, and using center personnel with expertise in data management.
  - Rotating GKB staff into these centers to develop user expertise.
  - Cross-training graduate students and postdocs by offering GKB training programs in which researchers can interact with developers onsite.
  - Conducting exercises in which knowledgebase developers try to recreate figures from publications using GKB data and methods to anticipate how users might access and analyze data.
  - Providing online and active support for first-time users and new GKB analysis tools and randomly surveying users for detailed feedback.

Knowledgebase Training

Finding

- The sophistication, breadth, depth, and range of GKB capabilities would far exceed previous data resources for probable user communities. Thus users would require training and education to fully understand and exploit GKB tools, resources, and opportunities for advancing their research efforts.

Recommendations

- Pursuant to its full development and operations, the GKB should provide a comprehensive training program to user communities. Training opportunities could include 2- to 4-day modular onsite courses, half- or full-day workshops at relevant conferences, and monthly online sessions and seminars.
- Knowledgebase staff should establish a help-desk service to answer user questions.
- A user-friendly Web resource should be created and include the following features:
  - Extensive documentation of GKB services and capabilities, including quality assurance (QA) methods.
  - Detailed descriptions of standard operating procedures (SOPs) used to generate or analyze data.
  - Information on file format standards and content.
  - Descriptions of data structures.
  - Details on controlled vocabularies used by the research community.
  - Step-by-step directions for using GKB analysis tools.
  - Instructions for downloading data.

Knowledgebase Performance Assessment

Findings

- Continual assessment of GKB performance and definition of system metrics would be essential to the success of the knowledgebase. These tasks could be accomplished by conducting surveys of targeted user groups to gain feedback on GTL community resources. Such feedback would drive further development of GKB resources. Survey responses also would be used to evaluate knowledgebase enhancements as they are released and tested on each target group.
- Moreover, results of user surveys could be used to establish performance measures that would be incorporated into reports for knowledgebase staff, DOE project officers, and other individuals involved in GKB governance. These performance metrics would be the foundations for enhancing many GKB activities. Staff meetings, project plans, and other activities could be optimized to improve metrics over time.
- An iterative process of GKB releases and user feedback would enable the knowledgebase to continue to meet the data and analysis needs of each user community. Furthermore, GKB's Web presence and resources—including a help-desk email address and tools to track system bugs and improvements—would provide ample opportunities for users to offer feedback about the knowledgebase.

Recommendations

- GKB staff should serve as resources and information providers to user communities.
  – Staff presentations and workshops on GKB capabilities would create opportunities for user training and establish further contacts within targeted GKB constituents.
- The GKB user community should be surveyed and the results translated into performance measures and prioritization schemes for ongoing knowledgebase development.

Knowledgebase Data Sharing and Policy Development—Incentives for Depositing Data

Findings

- Universal, straightforward, and productive use of the GKB by the scientific community would require various incentives. In particular, alerting the GKB user community to newly contributed data or tools would provide a significant incentive for participation in the knowledgebase. The GKB could promote such resources by (1) sending users periodic emails about recently submitted data, models, and tools; (2) tracking publications arising from the use of GKB services and emphasizing knowledgebase capabilities and data most cited in scientific literature; (3) showcasing new GKB datasets and tools online on the system's homepage; (4) providing strong graphics and analysis tools to improve and facilitate the publication process; and (5) eliciting user ratings of GKB data and tools (e.g., as Amazon does).
- Researchers also would be encouraged to contribute to the GKB if DOE program managers were able to track data and tools deposited by agency-funded principal investigators. Such a capability would promote knowledgebase participation if, for

example, DOE research grants require release of data and tools to the GKB. As an additional incentive for contributing to the GKB, data and resource providers could be offered early access to newly developed tools (e.g., beta testing), as well as GKB help in maintaining and backing up contributors' data.

Recommendations

- Transparent methods, incentives, requirements, and other inducements should be established to encourage researchers to contribute to and use the GTL Knowledgebase.
- To encourage submission of data and methods into the GKB, knowledgebase designers, along with GTL research programs and their managers, should make this process easy for data producers and tool developers by establishing and supporting standard data formats for different data types.
- To further facilitate knowledgebase use, GKB should develop and provide support for software, processes, and protocols that easily convert data into formats acceptable for knowledgebase entry.

Knowledgebase Working Group

Findings

Envisioned for the GTL Knowledgebase is the complete integration of multiple partners with unique capabilities and science objectives. Successfully integrating these partnerships across DOE mission–focused laboratories, universities, and industry would require a GKB management and governance model combining flexibility and accountability. The GKB management plan likely would focus on organizational structure and reporting mechanisms, operations management, and community assessment and review (see Box 5.1, Elements of the GKB Management Plan, p. 57).

- **Organizational structure and reporting mechanisms.** Critical to GKB success would be a well-defined management team including a director, management staff, and scientific and governance boards. The director's responsibilities would include executing the overall vision of the project and overseeing project direction. The GKB director would be accountable to a governance board for ensuring timely achievement of project goals and milestones. The governance board also would be expected to manage internal and external GKB reviews. In turn, knowledgebase stakeholders in federal agencies such as DOE would oversee and receive reports from the GKB governance board and director. GKB viability also would require a scientific board that would provide vision and recommendations for new directions for the knowledgebase project.
- **Operations management.** The GKB management plan would detail the roles and responsibilities of all senior management staff, which could include a deputy director as well as scientific and technology lead directors. Operational management plans would describe challenges and opportunities associated with coordinating GKB efforts across possible distinctive sites, national laboratories, and academic settings.
- **Community assessment and review.** GKB operating procedures and the system's effectiveness in supporting scientific research would be reviewed every 3 years at minimum. Integral to the review process would be community assessment of the

Box 5.1

Elements of the GKB Management Plan

Governance models define the relationships among various aspects of a program or organization and among key personnel and operations. For the GKB, the governance model should clearly describe the functioning of communications, submission and exchange of knowledgebase information, and establishment of policies, procedures, and personnel responsible for decision making.

- **Roles and responsibilities.** The GKB management plan should define the roles of institutions, programs, and individuals and the commensurate responsibilities of each.
- **Authorities and accountabilities.** The implementation plan should ensure that each role has the requisite authority to bring needed resources to bear and carry out their functions. Accountabilities define responsibilities between parties.
- **Policies, standards, and processes.** The GKB should enlist the input of all GKB stakeholders in defining knowledgebase needs and discussing system solutions for establishing policies and standards.
- **QC and QA protocols.** The GKB's ultimate value to the research community would depend on the degree to which users could rely on the accuracy, fidelity, and completeness of knowledgebase datasets and the tools to use them. Essential to user confidence in GKB resources, therefore, would be the establishment of quality control and quality assurance protocols.
- **Resources and funding.** The GKB would rely on resources and funding from a wide variety of sources. DOE's Genomics:GTL program would provide resources focused on knowledgebase establishment, operations, and maintenance. GTL research programs and centers—each of which would contribute significantly to data management and use—would ensure that research results are rigorously incorporated into the knowledgebase and that GKB-related research directions and priorities are well defined and supported.
- **Program management and staff.** GKB management and staff—in consultation with DOE's Office of Biological and Environmental Research (OBER) and other advisers—would facilitate quality in the research conducted using the knowledgebase. OBER's program management staff would define and approve GKB policies and processes (including the governance model) and would oversee implementation of the GTL Knowledgebase project.

GKB, which would critically depend on the management staff's outreach to the biological community. Also important to knowledgebase success would be project leaders who proactively define emerging needs for GKB-relevant policies on data and information sharing. Developing such polices would require establishing a systematic method for assessing current data and information guidelines, including those adopted by the GTL program. Although all external programmatic reviews would be coordinated with the GKB governance board and director, DOE would have the responsibility of ensuring and formulating the success of such reviews.

Recommendation

- A scientific working group should be formed to work closely with the GKB project throughout the development process. This group should include the following: (1) representatives from all user communities; (2) data producers generating various types of information (e.g., environmental, proteomic, genomic, and transcriptomic); (3) researchers focused on different GTL missions (e.g., bioenergy, carbon cycling, systems biology, and environmental remediation); and (4) experts in computing, informatics, and communications technologies and systems.