

## caBIG Workspace Adopter Project Form

Adopters, please complete this form in advance of the caBIG kickoff meeting and return by e-mail to [adamsm@mail.nih.gov](mailto:adamsm@mail.nih.gov). Completed forms will be made available to participants in advance of the meeting to enhance workspace discussions. During our conversations with you, we expressed the aspect of your program that we would like you to develop in the first year of the caBIG pilot; it is this we are asking you to address - here and in your presentation.

### Current Environment:

Cancer research informatics:

#### Staff expertise:

Our informatics development team consists of four full-time personnel plus a faculty advisor with expertise in bioinformatics and clinical informatics. We have the following capabilities:

Database design and development expertise in: ORACLE, MYSQL, MS SQL SERVER

Web application development: J2EE, PERL, ColdFusion, PHP

Systems analysis: functional and technical requirements gathering and documentation, project management, education and training

#### Hardware:

Currently our high performance computing resources include the following systems: two SUN Microsystems SunFire v440 with four 1.062GHz CPU's each web-application/data servers with four 36GB hard disks and 8GB memory each, two SunFire E3800 with 4 CPUs each at 900MHZ/8-MB cache and 4GB of memory; a Sun Fire V880 with 4 900MHZ CPUs, 8 GB memory, and local hard drives of 438 GB in size; and a SUN Enterprise Server E450 with four 480MHz processors, 4GB memory, four 36GB internal drives, and 327GB SUN StorEdge T3WG RAID system. The storage includes 2 Sun StorEdge 3960, each having 2 T3 Arrays with 18 hard-drives of 36GB each.

#### Software available:

Large scale database licenses within OHSU: ORACLE, MS SQLSERVER

Clinical trials software: Surveyor by Permedics

Statistical Software: SAS, Splus, R, Mars, Cart, PASS, Nquery, Statistica, GraphPad

Other Tools: Macromedia, CVS, Power Designer, Crystal Enterprise/Crystal Reports

#### University resources:

The Information Technology Group (ITG) develops, implements and maintains technology-based services and solutions enabling OHSU to effectively manage information to accomplish its missions. ITG supports and administers the hardware and database instances used by the bioinformatics resources of the OHSU Cancer Institute. In addition through the university's High Performance Computing Initiative researchers have at their disposal access to a High Performance Computing grid through a Common User Interface. The High Performance Computing grid is ideal for distributed application development or deployment.

#### Development projects:

##### **Affymetrix Microarray Core Database (AMCDB):**

AMCDB stores in an Enterprise Oracle database microarray data generated on the Affymetrix platform for the institution wide shared resource. AMCDB allows for investigators to securely access and retrieve their own data through a web

interface, dynamically filter and mine the data and retrieve both project and sample level information. This database began development in 2001 and went live in 2002. It contains over 1600 parsed and queryable geneChips data, along with over 5200 parsed and queryable normalized analysis results from these chips. This comprises approximately 130 unique projects, 40% of which are cancer-related and potentially could be shared through the caBIG initiative.

**Spotted Microarray Database (SMCDB):**

This database is in it's beta phase of release. The database is MIAME compliant with project and sample annotation, as well as normalized and raw data.

**Fanconi Anemia Transcriptome Consortium (FACTC) Portal:** A research portal for members of the International Fanconi Anemia Transcriptome Consortium that offers integration of clinical and functional genomic studies. This serves as a model for rare cancers with respect to collaboration and meta-analysis.

**Tools:**

A wide variety of tools have been built for data analysis, filtering, and annotation in Perl, C, R and Java.

**Education and information resources:**

We provide regular workshops, hands on software training and tutorials, and web accessible information for our investigators.

**1. Sponsoring Cancer Center**

Oregon Health and Science University Cancer Institute

**2. Workspace**

Integrative Cancer Research

**3. What "data" are you providing as adopters?**

We are willing to share our microarray data, proteomics data as it comes online, as well as tissue pathology and clinical trials information.

**4. What are the tools you envision would enhance the use and analysis of this "data"?**

If the data is to be shared among members of the grid, we would need appropriate tools to allow export and import of the data. For microarray data, tools based on MAGE-ML can be utilized. However, we will need to develop DTDs to allow sharing of other data types. Visualization tools would greatly facilitate use and analysis of this data, as well as heterogeneous query tools that would allow integration of disparate data types.

**5. What is your ability to evaluate the tools to be adopted?**

Our highly trained staff has experience in both academics and industry and follow best practices for software development and deployment. Through a series of benchmarks and testing, we have first-hand experience with evaluation of performance and stability issues and will apply our expertise and experience to the evaluation of the tools to be adopted. For analysis tools, we collaborate with the biostatistical and computational biology staff and faculty for evaluation of algorithms and protocols. Finally, we have a large membership pool that we utilize for beta-testing, user feedback and feature requests and can collaborate with faculty and staff in Medical Informatics to identify utilization patterns.

**6. How will you provide system integration?**

Naturally, our strategy for integration will depend on the specific tools and applications chosen for adoption during the current planning phase. However, we are willing to adopt NCI/CaBIG standards and protocols with respect to data sharing and development. Specifically, we can map or implement ontologies and data elements, utilize export/import tools based on CaBIG DTDs, and ensure data cleaning measures are taken for data warehousing or data integration. With respect to development, we are willing to develop or utilize standard APIs .

#### **7. How will you provide end-user testing?**

End user testing will be done in a manner consistent with academic and industry standards and based on the content being tested. For example, if the tool is a web interface application industry standards suggest that most bugs and problems can be identified during the course of five passes through a particular process (it is assumed in this projection that the rate of problems diminishes with each pass). Once a clear test plan is established based on the pre-defined validation criteria, end-user training will be conducted recruiting beta-testers from within our available sample population. Because IT staff and computer science students are not good test subjects for end-user software testing, our beta testers will reflect the actual end-user target demographic. We have a diverse base of research and clinical investigators that can be tapped for feedback and testing.

#### **8. How will you provide software validation?**

We found that successful software validation hinges on a thorough and comprehensive requirements specification document usually drafted at the requirements gathering phase- prior to any software development. This document provides the guideline for validation and helps insure that the software met the goals originally outlined in the specification. If multiple sites are working on validation of the same software or system, it will be important that collaboration occurs during the time that validation criteria are established. These evaluation criteria must be in place in order for us to develop and implement our test plan and end-user testing.

#### **9. What are your plans for interacting with the appropriate workspace developers?**

A critical first step is an IP/data sharing policy that will allow each institution to identify what data and products will be shared. This should contain wording about distribution and modification of source code, schemas etc, as well as policies about sharing of clinical data and meeting HIPPA regulations. We find that our most effective strategy for cross-departmental or cross-institutional collaboration is to ensure that clear and direct communication channels have been established. This is critical to make sure that details are not overlooked, participants are well informed and that project scope, costs, time, and quality are adequately managed. From our experience with other collaborative groups, e-mail (in particular discussion threads) are often an excellent way to discuss problems or issues that may arise during our collaboration. As soon as it becomes clear which members of the workspace we will be directly collaborating with and what our timeline is, we will set up regular meetings between staff at both sites. Either telephone or video conferencing can be utilized for these regular meetings. For key stages, on-site visits would occur.