

# BIOMEDICAL INFORMATICS WITHOUT BORDERS: ENABLING COLLABORATION TO STRENGTHEN RESEARCH AND CARE

## 1. Requirements for Systems that Optimize how Biomedical Scientists Choose Collaborators

The increasingly collaborative nature of biomedical science and the growing complexity of today's research projects challenge researchers to choose the most appropriate and best-qualified collaborators. To date, most collaborations are still established through traditional means, such as personal contact and searching the literature. However, this approach does not scale well in the face of large and growing pools of potential collaborators. Electronic systems, called expertise locating systems can potentially ameliorate this situation, but to date have played only a minor role in helping scientists do so. This study used a comprehensive literature review and contextual inquiries and semistructured interviews with 30 medical scientists to develop a preliminary set of requirements for electronic systems designed to help optimize how biomedical researchers choose collaborators. The requirements include aspects such as comprehensive, complete and up-to-date online profiles that are easy to create and maintain; the ability to exploit social networks when searching for collaborators; information to help gauge the compatibility of personalities and work styles; modeling various aspects of "proximity," which influences the success of collaborations; and recommendations for effective searching and making "non-intuitive" connections between researchers. The assumption of scarcity on which many existing expertise locating systems seem to be built is not supported by our results. Researchers primarily seem to be concerned with identifying the most promising collaborations, not with not being able to find collaborators. Clearly, one challenge for designing expertise locating systems is that seeking, evaluating and choosing scientific collaborators is a complex decision-making process that is poorly understood. Therefore, future work should validate the preliminary requirements we developed, further describe and analyze the dynamics of collaboration decisions among biomedical scientists, and articulate implications for system design.

### AUTHORS AND AFFILIATIONS:

Titus Schleyer, DMD, PhD, School of Dental Medicine, University of Pittsburgh; Heiko Spallek, DMD, PhD, School of Dental Medicine, University of Pittsburgh; Brian Butler, MS, PhD, Joseph M. Katz, Graduate School of Business, University of Pittsburgh; Sushmita Subramanian, MHCI, Intel Corporation; Daniel Weiss, MHCI, The MITRE Corporation; M. Louisa Poythress, MHCI, Brulant, Inc.; Phijarana Rattanathikun, MHCI, Adobe Systems, Inc.; Gregory Mueller, MHCI, DeepLocal, Inc.

## 2. Developing caBIG™ Compliant Applications An Internal Case Study

The development and deployment effort behind caArray 2.0 presented many opportunities to learn how to move from concept to an application that meets the acceptance criteria into the caBIG pipeline of tools. Built from inside NCI-CBIIT, many lessons were gained from a procedural and technical perspective. We offer these lessons to help cancer centers and the larger community adapt, adopt and ultimately integrate their applications and data into the caBIG world. Areas of focus include end-user engagement; infrastructure adoption and adaptation; environment automation; interdependent feedback; and techniques for compliance. Future challenges of integration and co-development will be covered.

### AUTHORS AND AFFILIATIONS:

Brent Gendleman, 5AM Solutions; Eric Tavela, 5AM Solutions; Todd Parnell, 5AM Solutions; Rashmi Srinivasa, 5AM Solutions; Paul Duvall, Stelligent

### 3. Bridging Continents in Clinical Research for Duke University and Beijing Cancer Hospital using caBIG™ Tools: Planning Phase

Connecting cancer research data globally with Cancer Biomedical Informatics Grid (caBIG™) standards and tools not only will reduce costs but also speed up research. In planning this Phase II clinical trial in breast cancer, lead by Duke University Comprehensive Cancer Center and Beijing Cancer Hospital of Peking University, investigators have a primary objective to optimize patient accrual and a secondary objective to pilot the Clinical Trials Management Suite (CTMS) software suite in a global environment. This project will enable Duke to partner with the Beijing Cancer Hospital in global fight against cancer and allow China's adaptation and adoption of CTMS suite to allow data exchange leveraging the use of common data elements (CDE's) and Case Report Form (CRF) Templates in collecting the data and securely sharing it on the caGrid. Duke has significant experience with the caBIG™ platform in local and multi-center trials, making this project the next logical step in their deployment plans. This presentation demonstrates how many challenges are addressed, such as remote data collection, monitoring, reporting and statistical analysis. Technical issues, such as network bandwidth and multi-lingual forms are also addressed. Plans for support across continents and time zones are also discussed.

#### **AUTHORS AND AFFILIATIONS:**

Robert Annechiarico, BS, Duke Comprehensive Cancer Center, Vijaya Chadaram, RN, BSN, Duke Comprehensive Cancer Center, Mohammad Farid, MS, Duke Comprehensive Cancer Center, Kimberly Blackwell, MD, Duke Comprehensive Cancer Center, Jun Ren, MD, Beijing Cancer Hospital of Peking University, Wei Chen, PhD, Duke University Health System

### 4. caBIG™ Terminology Metadata: An Overview

The cancer Bioinformatics Grid (caBIG™) is an information technology program that develops software tools to support cancer research efforts by establishing a common infrastructure that can be used to share data and applications across organizations. Interoperability is a key factor in caBIG™'s infrastructure and is very dependent on terminologies. The growing demand for service-oriented access to terminologies in the caBIG community is anticipated to result in increased publication of respective services on the caGrid. Efficient discovery, administration and query of these resources require definition and use of consistent and reliable metadata that can be queried at the service level. Primary use cases were collected and studied in the categories of Resource Identification, Internationalization, Intended/Allowed Usage, Provenance, Administration and Download and converted into a UML model. A metadata model was defined to satisfy these primary use cases while allowing for future expansion. In this work, we describe the UML terminology metadata model and how it will be used.

#### **AUTHORS AND AFFILIATIONS:**

Thomas Johnson, BS1, Salvatore Mungal, MS2, Brian Davis, PhD3, Frank Hartel, PhD4, George Komatsoulis, PhD4, Hua Min, PhD5, Scott Oster, MS6, Michael Riben, MD7, Denise Warzel BBA4,1 Mayo Clinic, Rochester, MN; 2Duke University, Durham, NC, 33rd Millennium, Inc. Waltham, MA, 4NCI CBIIT, Rockville, MD, 5Fox Chase Cancer Center, Philadelphia, PA, 6Ohio State University, Columbus, OH, 7MD Anderson Cancer Center, Houston TX.

### 5. BioBIKE: Creative Analysis of Bioinformation by Noncomputational Biomedical Researchers

The recent availability of vast amounts of data in the forms of genomic sequences and mass experimental data sets has been the engine of rapid progress in many areas of biomedical research. But progress has tended to be broad rather than deep, because few researchers are able to build specialized computational tools to look at information in new ways. The most fundamental new insights occur most often to those minds primed by experience and fed by novel encounters with data not already filtered by existing models or through the minds of technical assistants. Agility in today's information-rich world demands the ability to work creatively with the large data sets and therefore to program the computer. Yet the great majority of biomedical researchers are ill equipped to do this. Instances of BioBIKE (Biological Integrated Knowledge Environment) have been developed that seek to encompass all available knowledge pertinent to a coherent community of researchers. One distinguishing feature of BioBIKE is that the knowledge contained in the system may be manipulated using a programming environment integrated with the data. The end user is thus spared the need to find data and transform it to meet the requirements of specific applications. The programming environment is graphical,

adopting the conventions of word processors that virtually all researchers will have encountered. It is far more powerful than the usual query-based interfaces. It places in the hands of noncomputational biomedical researchers the power of a general purpose, and programming language adapted to the needs of bioinformatic analysis. Examples will be given of how BioBIKE allows users to combine analysis of mass data (e.g., microarrays) with metabolic information (e.g., genes related to given pathways) and genomic information (e.g. sequences upstream from a set of genes) to facilitate discovery (e.g., of regulatory motifs) and the creative use of information.

#### **AUTHORS AND AFFILIATIONS:**

Jeff Elhai, Center for the Study of Biological Complexity, Virginia Commonwealth University; Arnaud Taton, Center for the Study of Biological Complexity, Virginia Commonwealth University; Andy Whittam, Dept. of Information Technology Leadership, Washington & Jefferson College; JP Massar, CollabRx, Inc.; Jeff Shrager, CollabRx, Inc.

## **6. The ACGT Project: Towards the Pan-European Biomedical Grid Infrastructure**

The presentation summarizes the original results of the ACGT integrated project focusing on the design and development of an European Biomedical Grid infrastructure in support of multicentric, post-genomic clinical trials on cancer. ACGT (the initials stand for "Advancing Clinico-Genomic Trials" or for the four bases in DNA) is a project funded by the European Commission within the 6th Framework Programme. The ultimate objective of the ACGT is the provision of a unified technological infrastructure which will facilitate the seamless and secure access and analysis of multi-level clinical and genomic data enriched with high-performing knowledge discovery operations and services. In achieving this objective ACGT will: deliver a European Biomedical GRID infrastructure offering seamless mediation services for sharing data and data-processing methods and tools; deliver advanced security tools including anonymisation and pseudonymisation of personal data according to European legal and ethical regulations; develop an ACGT Master Ontology and use standard clinical and genomic ontologies and metadata for the semantic integration of heterogeneous data (clinical imaging genomic proteomic metabolomic and other as well as open source data from the Web); develop an Ontology based Trial builder for helping to easily set up new clinico-genomic trials, to collect clinical, research and administrative data, and to put researchers in the position to perform cross trial analysis; deliver data-mining services in order to support and improve complex knowledge discovery processes; and validate the technological platform with of two clinico-genomic trials and an in silico experiment. ACGT presently creates and tests an infrastructure for cancer research that will connect clinical research centers and investigators in a way that allows the common needs of interdisciplinary research to be met. To this end, ACGT intends to collaborate with several biomedical Grid projects and initiatives, such as the caBIG, NCRI cancerbioinformatics, CancerGridUK, myGRID and Intelligrid.

#### **AUTHORS AND AFFILIATIONS:**

Dimitris Kafetzopoulos, PhD, FORTH-IMBB, Greece; Manolis Tsiknakis, PhD, FORTH-ICS, Greece; Norbert Graf, MD, University Hospital of Saarland, Germany; Cristine Desmedt, PhD, Institute Jules Bordet, Belgium; Matthias Brochhausen, PhD, INFOMIS University of Saarland, Germany; Thierry Sengstag, PhD, Swiss Institute of Bioinformatics, Switzerland; George Potamias, PhD, FORTH-ICS, Greece

## **7. Enabling Data Sharing across Borders with NAACCR and caBIG**

The North American Association of Central Cancer Registries (NAACCR) is a professional organization made up of members from the United States and Canada. It is devoted to the promulgation of uniform data standards for cancer registries. These are used to improve cancer surveillance, cancer control and cancer research. NAACCR, along with a growing number of other institutions and programs, are encoding variables according to the ISO 11179 metadata standard. This standard is implemented in the cancer Data Standards Repository (caDSR) that is maintained by the National Cancer Institute (NCI). The caDSR enables researchers to develop systems that pass standard NAACCR data elements via NCI's Cancer Biomedical Informatics Grid (caBIG™). The NAACCR standard elements can be accessed using the Common Data Element (CDE) browser at <http://cdebrowser.nci.nih.gov/CDEBrowser/>. The standard is listed on the left side of the portal under the directories of NCI Population Sciences & Cancer Control, Classifications, and the Division of Cancer Control and Population Sciences. Under the NAACCR directory there are 17 objects that cover demographics, tumors, treatment, hospitals, staging and prognostic factors for cancers. Within these objects are 382 CDEs. These CDEs specify definitions, data types and/or enumerated values for the NAACCR variables. Such detailed, internationally accessible, electronically coded metadata enables the sharing of consistent high quality data between any institutions that use this standard to encode their data. Researchers at the Utah Cancer Registry, Huntsman

Cancer Institute, Intermountain Healthcare, and Utah Cancer Specialists are developing a distributed data repository using caBIG™ to examine details of clinical care for patients with colorectal cancer across Utah. The NAACCR standard will be the core component for data sharing among the participating institutions.

**AUTHORS AND AFFILIATIONS:**

Lewis Frey, PhD, Huntsman Cancer Institute & the Biomedical Informatics Department at the University of Utah; Antoinette Stroup, PhD, Utah Cancer Registry and the Huntsman Cancer Institute; Tao He, Huntsman Cancer Institute; Martin Cryer, Biomedical Informatics Department; Stephane Meystre, MD, PhD, Biomedical Informatics Department; Kerry Rowe, PhD, Intermountain Healthcare; Arthur Hartz, MD, PhD, Huntsman Cancer Institute and Intermountain Healthcare

## **8. IM.Grid, a Grid Computing Approach for Image Mining of High Throughput-High Content Screening**

Image processing and analysis has become essential for both cell biology research and drug discovery since the advent of High Content Screening (HCS) technologies. In this context, the Grid technology is a good opportunity to solve intensive computing problems with large data set. In addition, exploitation of the Grid is not a simple task for many users. Nowadays, an important issue is to provide a simplified use of Grid resources. In this paper, we present IM.Grid, a grid computing extension of our in-house image analysis software called IM (Image Mining) that provides capabilities to simultaneously access visual data of NAS (Network-Attached Storage) and extract knowledge from the raw information by customizable image processing pipeline in a parallel way. In a plug-in development period, each user can design an individual pipeline using specific built-in image processing libraries with proper thresholds. The plug-in becomes an actual processing unit when Grid starts to analyze multiple images retrieving them from the centralized storage at a time. The user can receive the output results as fast as the number of computational grids are occupied. We apply this method to reduce the time of image processing and analysis of cell biological images for drug discovery within the context of High Throughput-High Content Screening (HT-HCS) because the processing time is growing dramatically as the size of the images becoming huge on account of many factors like multi-channel, and high resolution. To address these constraints, we propose a .NET framework based high-performance computing environment that helps to improve productivity not only in developing phases but in HT-HCS platforms.

**AUTHORS AND AFFILIATIONS:**

HongKee Moon, MSc, Institut Pasteur, Korea; Auguste Genovesio, PhD, Institut Pasteur, Korea

## **9. Populomics**

Comprehensively understanding health and disease pathogenesis requires integrating knowledge regarding determinants which exist across several levels of analysis (molecular, cellular, clinico-behavioral, neighborhood/community, cultural and societal). Historically, comprehensive scientific understanding has been hindered in part by what may be called disciplinary perspectivism, nonstandard terminology or nomenclatures, and a focus primarily on clinical, cellular and molecular mechanistic analyses. While this approach has enabled significant advances, the rise in global health concerns generally, the increasing recognition of the importance of socioenvironmental and cultural determinants of health and the inability to reduce or eliminate disparities in health outcomes all highlight the limitations of what has become known as the “medical model”. In the future, clinical, cellular and molecular determinants of disease will need to be understood within the context of broader determinants which operate primarily at the population level. Populomics is emerging as a transdisciplinary approach to scientific investigation and discovery that integrates knowledge across individual and population levels of analysis to yield novel insights to health, disease causation and therapeutics particularly at the population level. This presentation will discuss our work to develop a transdisciplinary research framework that facilitates data integration/harmonization and populomics oriented research within the context tobacco related lung cancer in the US.

**AUTHORS AND AFFILIATIONS:**

Michael Gibbons, MD, Associate Director, Johns Hopkins Urban Health Institute

## 10. Orchestrating caGrid services in Taverna

For the empowerment of users from biological or medical domains in creating and executing their workflows efficiently, the caGrid Workflow team, with the ICR working group, has selected the Taverna workbench and successfully created a prototype to orchestrate caGrid Data and Analytical services for ICR workflows. This prototype is the first step towards achieving our goal of providing an easy-to-use workflow authoring and submission tool that will be capable of orchestrating caGrid data and analytical services in executing workflows. Now, we commit ourselves to provide caGrid Workflow builder and Workflow Service as a tool which will eventually support caBIG users across workspaces in creating and executing their domain based workflows.

### AUTHORS AND AFFILIATIONS:

Wei Tan<sup>1</sup>, Ravi Madduri<sup>1, 2</sup>, Kiran Keshav<sup>3</sup>, Baris E. Suzek<sup>4</sup>, and Scott Oster<sup>5</sup>, Ian Foster<sup>1</sup>

## 11. caArray: Supporting the Federated Exchange of Array Data in caBIG™

caArray is an open-source data management system that supports the exchange of array data using a federated model of local installations across the Cancer Biomedical Informatics Grid (caBIG™). In its second generation, the web-based system provides a graphical user interface through which users can submit and access array data and associated annotations. Users can control public access to experiment- and sample-level data and can create collaboration groups to support data exchange among a defined set of partners. caArray supports bulk data import using MAGE-TAB, a spreadsheet-based format for array data which has proven to be a popular method for data submission to the system. Programmatic access to experiment data is available through a caGrid service as well as a remote Java API. Users of GenePattern and geWorkbench, key genomic analysis platforms in caBIG™, are able to seamlessly acquire data from local or remote instances of caArray. The simplified installation of caArray and its complete reliance on open source software has promoted increased adoption as evidenced by over 20 institutions having installed the software within the first 6 months of the 2.0.0 release. In accordance with caBIG™ principles, caArray is intended to promote open development. Initial collaborations to support community-based development contributions are underway. The NCI Center for Bioinformatics and Information Technology hosts an installation of caArray at <https://array.nci.nih.gov> for inspection and use. Interested parties are encouraged to review the installation package, documentation, and source code available from <http://caarray.nci.nih.gov>.

### AUTHORS AND AFFILIATIONS:

Juli, Klemm 1, Anand, Basu 1, Xiaopeng, Bian 1, Jill, Hadfield 1, Brent, Gendleman 2, Eric, Tavela 2, Rashmi, Srinivasa 2, Todd, Parnell 2, Scott, Miller 2, William, Mason 2, Daniel, Kokotov 2, Makiko, Duncan 2, Paul, Duvall 3, Levent, Gurses 3, Tom, Boal 4, Ron, Keene 4, Leonie, Misquitta 4, Don, Swan 5, Robert, Wysong 5, Alan, Klink 5, Andrea, Johnson 5, Gerald, Fontenay 6, George, Komatsoulis 1.

1 NCI Center for Biomedical Informatics and Information Technology, 2 5AM Solutions, Inc., 3 Stelligent, Inc., 4 NARTec, Inc., 5 TerpSys, Inc., 6 Lawrence Berkeley National Laboratory.

## 12. cancerGrid cgMDR for Decentralized Development and Registration of caBIG™ Compatible Systems

There is an emerging requirement in caBIG™ for a decentralized, peer-to-peer metadata registry solution, allowing distributed registration and management of data elements describing studies that have yet to be undertaken or completed. These data elements can be developed and used within an institution or collaborative group before or alongside the processes of wider review and adoption facilitated by a central registration authority. The NCI's cancer Data Standards Registry (caDSR) has proved extremely effective in support of central registration, but the existing technology base does not lend itself to local installation and distributed operation. In contrast, the CancerGrid metadata registry (cgMDR) has proved effective as a lightweight, desktop solution, interoperable with caDSR, but targeted at the day-to-day needs of cancer researchers, data managers, and software developers. The cgMDR technology which is based, like caDSR, upon the ISO/IEC 11179 metadata standard is being installed and evaluated for use in the development of caBIG solutions, with an emphasis upon qualification for silver-level compatibility. The intention is to simplify systems integration and adoption while remaining interoperable in terms of services and data/metadata representation with the existing caBIG architecture. This poster presents the key features of the cgMDR registry and its surrounding toolset, and outlines the processes surrounding its installation, use and integration within caBIG.

**AUTHORS AND AFFILIATIONS:**

Denise Warzel, BBA, NCI CBIIT, USA; Christophe Ludet, MS, Oracle ,USA; Jim Davies, DPhil, Oxford University Computing Laboratory, UK; Steven Harris, PhD, Oxford University Computing Laboratory UK; Andrew Tsui, MSc, Oxford University Computing Laboratory, UK

### 13. Connecting Skills to Learning Needs: The Educational Perspective of caBIG™

A community's ability to use biomedical data standards and software tools to improve interoperability is enhanced through the use of training. The National Cancer Institute Center for Biomedical Informatics and Information Technology (NCI CBIIT) has developed an ambitious curriculum to serve as the educational foundation for Cancer Biomedical Informatics Grid or caBIG™ activities. Curriculum courses have been reengineered over the past year into a principally self-paced format to meet the needs of the users. Courses are designed to meet the role of the person using or developing caBIG™ tools and resources, and range from novice to expert in complexity. As the cancer grid has become a reality, content to create and activate a grid node has been added to the curriculum. Periodic immersions into educational content in the form of 'Boot Camps' have become a hallmark of caBIG™ community outreach, focused on the software developer or those who want their tools or data sets to be caBIG™ compliant. Appealing educational formats, applicable content, and community availability make training a dynamic and ever-evolving aspect of caBIG™.

**AUTHORS AND AFFILIATIONS:**

Dianne M. Reeves, NCI CBIIT; Denise Warzel, NCI CBIIT; Jennifer Brush, ScenPro, Inc.; Becky Angeles, ScenPro, Inc.; Jamie Parker, ScenPro, Inc.; Leslie Derr, Ph.D., NCI CBIIT

### 14. Enhanced Cancer Models Database Accommodates More Species and Provides Preclinical Trials Information

The Cancer Models Database (caMOD) is a web-based resource that provides information about animal models for human cancers. Key capabilities are Submission, Search, and System Administration. The application can be accessed at <http://cancermodels.nci.nih.gov>; With version 2.3, caMOD customizes the application for zebrafish and rat models in addition to mouse models. Depending on the selected species, the application uses species-specific vocabularies for anatomy and disease throughout the application. Links to the Zebrafish Model Organism Database (<http://www.zfin.org>) and to the Rat Genome Database (<http://rgd.mcw.edu/>) allow the user to retrieve information about the modified alleles. Data about other species can be submitted to caMOD, but no vocabularies for anatomical or diagnostic terms are provided at this time.;caMOD 2.3 further expands the use of vocabularies provided by NCI Thesaurus (<http://nciterns.nci.nih.gov>) by introducing the staining method vocabulary to the imaging portion of the application.;;The most recent release of caMOD connects the cancer models database with caELMIR, the Cancer Electronic Laboratory Management Information Resource (<http://caelmir.compmed.ucdavis.edu/caelmir/>). Users can retrieve data generated from preclinical trials and review information on the study, experiment, and individual animal level.;;The for Fall 2008 scheduled release 2.5 will connect caMOD to caGRID.;;For the future, we envision developing caMOD into a cancer preclinical study information resource that will provide structured, searchable access to information about preclinical study protocols and outcomes, linked to detailed information about the animal models used, to related human clinical trials information, and to other molecular, pathology, and compound resources. The new system will integrate preclinical and clinical data and enable comparison across systems (animal model, cell line, yeast, xenograft, human).

**AUTHORS AND AFFILIATIONS:**

Heiskanen, Mervi, Ph.D. NCI CBIIT,Wagner, Ulli, SAIC Frederick,Pandya, Sima, SAIC,Duncan, Maki, 5AM Solutions,Hadfield, Jill, NCI CBIIT,Thulasiraman, Kavitha, SAIC,Galvez, Jose, MD, UC Davis,Tarnowski, Betty, Ph.D., NCI DCB,Marks, Cheryl, Ph.D., NCI DCB,Klemm, Juli, Ph.D., NCI CBIIT

## 15. Forms Annotation and Registration in caDSR Facilitated by cancerGrid cgMDR

There is a requirement to centrally register, harmonize and share Case Report Forms (CRFs) and associated data elements in NCI's cancer Data Standards Registry (caDSR). Forms can be developed by the community using incompatible forms authoring tools, each using different forms information models and exchange formats. Decomposition of CRFs into data elements is tedious and can be error prone, taking as long as two to three hours per CRF question. Microsoft excel spreadsheets offer an effective means to aid CRF developers mapping questions on CRFs to caDSR data elements. The caDSR Curation Tool and CDE Browser are proven aids in this process, but their use requires 'cut and paste and manual data entry between spreadsheets and caDSR user interfaces, one data element at a time. In contrast, the cancerGrid's cgMDR Microsoft Excel plug-in has proven extremely effective in supporting insertion of existing data elements or parts of data elements directly into spreadsheets, capturing all the necessary information to ensure accuracy of the recorded information and speeding preparation for registration. The completed spreadsheet can then be transformed into a format for use with the existing caDSR SIW/UML Loader to streamline the review and registration process. The UML loader will also organize the content within the caDSR for easier recall and reuse. The poster presents the overall CRF decomposition and registration workflow using features of the cgMDR and SIW/UML Loader to improve the process end-to-end highlighting the time saved with this new process and changes to existing systems to make it work.

### AUTHORS AND AFFILIATIONS:

Denise Warzel, BBA, NCI CBIIT, USA; Christophe Ludet, MS, Oracle, USA; Dianne Reeves, MSN, RN, NCI CBIIT, USA; Jim Davies, DPhil, Oxford University Computing Laboratory, UK; Steve Harris, PhD, Oxford University Computing Laboratory, UK; Andrew Tsui, MSc, Oxford University Computing Laboratory, UK

## 16. National Cancer Imaging Archive (NCIA): Leveraging Imaging in Biomarker Discovery and Clinical Research

The National Cancer Imaging Archive (NCIA) provides access to in vivo cancer images. The goal of the NCIA is to provide the cancer research community, industry, and academia with access to an imaging data archive that will assist in the development and validation of analytical software tools supporting: lesion detection and classification, accelerated diagnostic imaging decision throughput, and quantitative imaging assessment of drug response. The repository aims to provide access to imaging resources that will improve the use of imaging in today's cancer research and practice by: increasing the efficiency and reproducibility of imaging supporting cancer detection, improving the accuracy of imaging supporting cancer diagnosis, leveraging imaging to provide an objective assessment of therapeutic response, and ultimately enabling the development of imaging resources that will lead to confident clinical decisions in patient care. Access to the NCIA system deployed at NCI is available online at <http://imaging.nci.nih.gov/> and also via a caGrid data service, <http://imaging.nci.nih.gov/wsrp/services/cagrid/NCIACoreService>. The NCIA system at NCI houses millions of in vivo images from clinical trials. Currently the archive contains DICOM image data from over 2,430 cases, spanning 2.8 Terabytes, in many cases linked to the associated clinical and annotation data. Recognizing the widespread implementation of the DICOM standard, the NCIA object model is based on the DICOM information entities. Recently, under review for caBIG Silver Compliance, there have been significant challenges in the attempt to closely adhere to DICOM while meeting the caBIG standards for semantic interoperability. Particular examples are highlighted. Those institutions interested in deploying an NCIA system may download the NCIA software bundle; <http://ncicb.nci.nih.gov/download/#NTTools>, and refer to the NCIA Wiki as a starting point for more information; [https://wiki.nci.nih.gov/display/Imaging/NCIA+ System](https://wiki.nci.nih.gov/display/Imaging/NCIA+System).

### AUTHORS AND AFFILIATIONS:

Jennifer Zeng, Science Applications International Corporation (SAIC); Qinyan Pan, SAIC; Thai Le, SAIC; Eric Kascic, SAIC; Surendra Poranki, SAIC; Jim Zhou, SAIC; Carl Blake, SAIC; Larry Holeman, Northern Taiga Ventures Incorporated (NTVI); David Palmer, NTVI; Sharon Gaheen, SAIC; John Freymann, SAIC-Frederick; Carl Jaffe, NCI Cancer Imaging Program (CIP); Laurence Clarke, NCI CIP; John Perry, MIRC Committee, Radiological Society of North America (RSNA); Eliot Siegel, VA MD Healthcare System, MIRC RSNA, UMD Dept of Diagnostic , Radiology; Anand Basu, NCI-CBIIT; George Komatsoulis, NCI-CBIIT; Ken Buetow, NCI-CBIIT

## 17. caLIMS2: A caBIG™ Compliant Laboratory Information Management System

The purpose of the caLIMS2 initiative is to create a next generation open source Laboratory Information Management System that is caBIG(TM) silver compliant. caLIMS2 will complete the caBIG(TM) bench-to-bed model by bridging the gap between biospecimen repositories, data repositories and analysis tools. The application is designed to allow easy customization by users and facilitate integration with data sources, analytical tools and laboratory equipment. caLIMS2 is highly flexible making it suitable for use by research labs, high throughput core facilities, and public health labs. caLIMS2 will be able to support multiple laboratory domains, including genomics, proteomics, environmental studies and nanoparticle characterization. caLIMS2 will support human and non-human biospecimens, environmental samples, and artificial specimens. caLIMS2 will allow a project to draw sample information from caTissue, track the experiments in the lab, and then generate the required files to combine with collected data and metadata to upload into appropriate data repositories such as caArray. There are five core modules in caLIMS2: administration, inventory, workflow, reports, and analysis. Adaptive workflow will allow laboratory staff to organize, track, and annotate the experiments and data in caLIMS2. We have completed the beta release 0.5 object model which features administrative and sample management functionalities. We will present the release 0.5 caLIMS2 object model, describe our progress in developing the release 1.0 object model (which will incorporate adaptive workflow), and provide a demonstration of the release 0.5 caLIMS2 application. caLIMS2 will help further translational cancer research through the organization of laboratory workflow, tracking of specimens, acquisition of laboratory data and metadata, and the appropriate sharing and dissemination of the data to support subsequent GRID-enabled workflows and analyses.

### AUTHORS AND AFFILIATIONS:

Bob Clifford, PhD, NCI/LPG; Jenny Kelley, MA, NCI/LPG; Cu Nguyen, BS, NCI/LPG; Henry Zhang, PhD, NCI/LPG; Sasikumar Thangaraj, MBA, SAIC; Anand Basu, MS/MBA, NCI CBIIT; Ken Buetow, PhD, NCI/LPG, NCI CBIIT LIMS Consortium

## 18. caGrid 1.2 Overview

In this poster we will present a general overview of caGrid 1.2 covering the components available in the 1.2 release and its salient features. The poster will cover the infrastructure components, available services, tools, APIs, and applications.

### AUTHORS AND AFFILIATIONS:

Scott Oster, MS, Ohio State University; Stephen Langella, MS, Ohio State University; Shannon Hastings, MS, Ohio State University; David Ervin, BS, Ohio State University; Tahsin Kurc, PhD, Ohio State University; Joel Saltz, MD, PhD, Ohio State University; Ravi Madduri, MS, University of Chicago/Argonne National Laboratory; Ian Foster, PhD, University of Chicago/Argonne National Laboratory; Joshua Phillips, BS, Semantic Bits, LLC; Manav Kher, BS, Semantic Bits, LLC; Kunal Modi, BS, Ekagra Software Technologies

## 19. caGrid Security Infrastructure (GAARDS)

The GAARDS security infrastructure provides enterprise services and tools for the administration and enforcement of security policy in an enterprise Grid. GAARDS is the official security infrastructure for caGrid and has been adopted and deployed in the caBIG™ production environment.

### AUTHORS AND AFFILIATIONS:

Stephen Langella, MS, Ohio State University; Scott Oster, MS, Ohio State University; Shannon Hastings, MS, Ohio State University; David Ervin, BS, Ohio State University; Justin Permar, BS, Computer Science, Ohio State University; Kunal Modi, BS, Ekagra Software Technologies; Tahsin Kurc, PhD, Ohio State University; Joel Saltz, MD, PhD, Ohio State University

## 20. caGrid Service Authoring Toolkit (Introduce)

Introduce enables users to graphically design analytical and data services that can be used in the caGrid environment. caBIG revolves strongly around shared data models and semantic interoperability through its use of caDSR and EVS. Introduce will enable the user to utilize data types from the caDSR which have been semantically annotated with

concepts from the EVS as input and output data types of a grid service. Introduce will also enable the use of the caGrid security infrastructure to protect the service and its data.

#### **AUTHORS AND AFFILIATIONS:**

Shannon Hastings, MS, Ohio State University; Scott Oster, MS, Ohio State University; David Ervin, BS, Ohio State University; Stephen Langella, MS, Ohio State University; Tahsin Kurc, PhD, Ohio State University; Joel Saltz, MD, PhD, Ohio State University

## **21. The Cancergrid Metadata Registry and Toolset**

CancerGrid has developed a set of caCORE compatible user-oriented tools for the management and consumption of metadata. These tools have been used by the project and its immediate partners to: design and generate case report forms and develop annotated UML models from existing data elements; capture interoperable bioinformatic data in Microsoft Excel 2007; develop metadata standards; and uplift pathology data and tissue metadata into RDF for meta-analysis and the assembly of a tissue banks from multiple sources. This poster introduces key aspects of the work: a personal/workgroup metadata registry application that can be installed in minutes; the (included) Query Service Manager and its associated plug-ins that allow desktop software users to access terminologies, and create and consume metadata elements without leaving their application environment; and the SQIV toolset that uses W3C SAWSDL markup to uplift, transform, validate, cross-tabulate and reason about clinical and bioinformatic data using semantic-web technologies. Examples presented include: importing metadata elements from existing registries such as the NCI caDSR and the NHS Data Dictionary; UML modelling using CDEs in Enterprise Architect; "type a column range from a CDE" in MS Excel; the use of CDE and concept references from caDSR, EVS and workgroup metadata sources for designing and generating case report forms in MS InfoPath; and how run-time transformation, inference and validation between data and metadata can implement privacy policies for trial participants in data warehouses.

#### **AUTHORS AND AFFILIATIONS:**

Andrew Tsui MSc.; Charles Crichton BSc; Steve Harris PhD; Jim Davies, DPhil, Oxford University Computing Laboratory

## **22. Web-Based Application for Digital Pathology and Molecular Analysis**

Virtual Microscopy to Microarray (VM2M) is a web-based application which combines virtual microscopy images and corresponding microarray data for cancer research. The technology to develop VM2M is a service oriented architected (SOA) application. JSP 1.0 was utilized to develop the graphical user interface (GUI) and Java 1.4 was used as the programming language for functionality development. A J2E central data web service (Tomcat 5.5) connects the GUI to the MySQL 4.3 database. The application runs on Apache 2.0 HTTP server. The VM2M application was designed from a use case developed within the Children's Oncology Group (COG). The use case included following a pediatric cancer protocol and scanning all cases in the protocol and the associated slides. The same cases included molecular analysis developed utilizing the Affymetrix platform. VM2M creates a web-based system where Virtual Microscopy data is made available alongside microarray data. The researcher can inspect the corresponding whole slide image using a web-based image viewer allowing the user to zoom and pan to the desired parts of the image. The researcher has the ability to access the microarray data that corresponds to the same tissue sample, along with results from an automated analysis of the digital image. The first iteration of VM2M includes biology and molecular components for the COG IRSG (Rhabdomyosarcoma) protocol. The VM2M application currently contains 192 cel files (expression data) and 146 corresponding images. In conclusion, the VM2M application combines disparate data sources in a web-based, user-friendly foundation. The future iterations will include additional COG studies with corresponding biology and molecular data, integrated Electronic Lab Notebook (ELN) functionality, and additional components to the framework of the application (Radiology, Proteomics, etc.)

#### **AUTHORS AND AFFILIATIONS:**

Dave Billiter, BA, PMP; The Research Institute at Nationwide Children's Hospital; Kathy Nicol, MD, The Research Institute at Nationwide Children's Hospital; Thomas Barr, BS, The Research Institute at Nationwide Children's Hospital; Mark Plaskow, BS, The Research Institute at Nationwide Children's Hospital

## 23. caBIO: An Integrated Resource for Genomic Annotations in caBIG(TM)

caBIO is an integrated repository of genomic information available on caGrid. Information from key genomic data and annotation providers - including UniGene, Entrez Gene, and UniProt - is stored in this caBIG(TM)-compatible database and infrastructure and is updated monthly. Recent additions included annotations from popular microarray platforms and curated literature information from the Cancer Gene Index. caBIO data is accessible through a variety of programming interfaces and the FreestyleLM, a utility that provides a "Google(R)-like" approach to searching the database. In addition, a range query supports searches based on genomic location. Multiple applications rely on caBIO for genomic information, including caMOD, CMAP, and Rembrandt.

### AUTHORS AND AFFILIATIONS:

Konrad Rokicki, SAIC; Lalitha Viswanath, SAIC; Jim Sun, SAIC; Hong Dang, PhD, Alpha Gamma Technologies; Ye Wu, SAIC; Ying Long, SAIC; Sharon Gaheen, SAIC; Krishnakant Shanbhag, NCI; George Komatsoulis, PhD, NCI

## 24. Developing a Collaborative Environment Supporting the Application of Nanotechnology in Biomedicine

The application of nanotechnology in cancer promises advancements in early detection, targeted therapeutics, and cancer prevention and control. The use of nanotechnology in biomedicine involves the engineering of nanoparticles to act as therapeutic carriers, targeting agents, and diagnostic imaging devices. To assist in expediting and validating the use of nanoparticles in biomedicine, the NCI Center for Biomedical Informatics and Information Technology (CBIIT), in collaboration with the NCI Nanotechnology Characterization Laboratory (NCL) and other Cancer Centers of Nanotechnology Excellence (CCNEs), has developed a data sharing portal called caNanoLab. caNanoLab facilitates data sharing via the use of caBIG technologies enabling semantic interoperability. caNanoLab data services are currently available via the caBIG grid (caGrid) and are enabling information exchange between the NCL and CCNEs including Washington University, Georgia Institute of Technology, and Stanford University. caNanoLab is based on a nanotechnology object model (nano-OM) which acts as an initial standard representation of nanoparticles and their physical (e.g. size, molecular weight) and in vitro (e.g., cytotoxicity, immunotoxicity) characterizations. The nano-OM leverages and extends concepts from the NCI's Enterprise Vocabulary Services (EVS) and the nanotechnology ontology designed by Washington University. The nano-OM provides a model for representing the composition of diverse nanoparticle types (e.g., dendrimer, fullerene, quantum dot, carbon nanotube) and associated functionalizing entities (small molecules, antibodies). These functionalizing entities allow particles to achieve the desirable therapeutic or diagnostic functions and enable personalized medicine via the administration of targeted therapies.

### AUTHORS AND AFFILIATIONS:

Anand Basu, NCI CBIIT; Frank Hartel, PhD, NCI CBIIT; Piotr Grodzinski, PhD, NCI OTIR; Anil Patri, PhD, NCL; Marty Fritts, PhD, NCL; Sharon Gaheen, MBA, SAIC; Sue Pan, SAIC; Shuang Cai, SAIC; Qina Tan, SAIC; Elizabeth Hahn-Dantona, PhD, Lockheed

## 25. HealthGrid - An International Initiative for Collaborative eHealth

Our poster describes the international non-profit HealthGrid Initiative, its goals, collaborating organizations, and a very brief history. We include examples of HealthGrids (including, of course, caBIG and European exemplars) and opportunities for advancement. We encourage international participation including in the 2009 HealthGrid Annual Conference in Berlin, Germany.

### AUTHORS AND AFFILIATIONS:

Howard Bilofsky, PhD, University of Pennsylvania; Mary Kratz, MS, University of Michigan; Yannick Legr, HealthGrid Association, Europe; Jonathan Silverstein, MD, University of Chicago

## 26. caAERS - Cancer Adverse Event Reporting System

caAERS is an open source, standards-compliant application designed to collect, assess, and manage adverse events (AEs) in cancer clinical trials. It is web-based, uses a controlled vocabulary, and enables multiple users to access, search for, and report on adverse events. caAERS operates as both a repository for capturing and tracking routine and serious adverse

events, as well as a tool for preparing and submitting expedited adverse event reports in-house and to regulatory agencies. Adverse events can be coded in caAERS using either CTCAE or MedDRA. caAERS participates in CCTS and thus is developed to integrate with other caBIG-compliant CTMS components. In addition, caAERS has the ability to integrate with legacy systems by importing XML files.

**AUTHORS AND AFFILIATIONS:**

Biju Joseph; Srinivasa Akkala; Karthik Iyer; Ram Chilukuri; Vinay Kumar; Edmond Mulaire, SemanticBits

## **27. caGrid and GAARDS Infrastructure in the caBIG Clinical Trials Suite (CCTS) v1.0**

CCTS is an enterprise clinical trials system that is comprised of a number of interoperable modules covering a broad range of key clinical workflows. In this poster we focus on the supporting infrastructure including the deployment and security architectures. CCTS is supported by a number of caGrid components that must act in concert in order for messages to be routed and security to be maintained. GTS ensures trust, Dorian manages grid identities and provides authentication, CDS and CAS handle delegation and single sign-on, and caGrid tooling provides message serialization. Furthermore, third-party components can be leveraged to hook into the infrastructure, such as Acegi. We describe caAERS as a case study for this infrastructure.

**AUTHORS AND AFFILIATIONS:**

Manav Kher, Patrick McConnell, Ram Chilukuri, Vinay Kumar, Edmond Mulaire, SemanticBits

## **28. The Cancer Central Participant Registry (C3PR) v2.0**

C3PR is a web-based application used for end-to-end registration of patients to clinical trials. This includes capturing the consent signed date, eligibility criteria, stratification, randomization, and screening. Clinical workflows are enabled by both subject- and study-centric views into the registration process. C3PR can be run in a standalone mode where study definitions, investigators, study personnel, and sites are entered into the system, or C3PR can be run in an integrated mode with the caBIG Clinical Trials Suite (CCTS). C3PR also enables multi-site clinical trials where registration information is entered locally at affiliate sites and the registration is completed by call-out to the coordinating site.

**AUTHORS AND AFFILIATIONS:**

Patrick McConnell, Duke University/SemanticBits; Robert Annechiarico, Duke University; Vijaya Chadaram, Duke University; A. Jamie Cuticchia, Duke University; Manav Kher, SemanticBits; Ram Chilukuri, SemanticBits; Kimberly Livengood, Wake Forest University; Robert Morrell, Wake Forest University; Sharon Elcombe, Mayo; Steve Riorden, Westat; Kimberly Johnson, CALGB

## **29. Utilizing caGrid Infrastructure for Pathological Image Analysis**

The caGrid toolkit provides a middleware infrastructure for rapidly creating analytical and data services that can be shared across multiple institutions located at geographically diverse locations. These caGrid services allow researchers and clinicians to access data remotely through a common interface, and use public and authorized computational resources including analysis algorithms written by researchers around the world. The caGrid toolkit is being used by the National Cancer Research Institute (NCRI) in the UK and the National Cancer Institute (NCI) Cancer Bioinformatics Grid (caBIG™) for the development of interoperable cancer research grids. This interoperability is illustrated by our demonstration project between the University of Leeds, the UK National Grid Service (NGS), and the Department of Biomedical Informatics at the Ohio State University (OSU). The Leeds-NGS-OSU demonstration project focuses on the sharing of pathology image analysis algorithms from both organizations in a distributed grid environment. The project involves the creation of caGrid analytical services that internally executes algorithms developed at the University of Leeds and OSU. The services access remote Aperio ImageServer (Aperio, Vista, CA) instances to retrieve and process the images. The results of the analysis algorithms are stored and made accessible to the researcher. The initial implementation leverages existing Matlab (Mathworks, Natick, MA) based algorithms for analyzing digitized, whole-slide neuroblastoma histological sections. The services are installed at Leeds, NGS, and OSU, and allow remote invocations while leveraging the caGrid authentication and authorization infrastructure. The demonstration project represents a first step towards full interoperability between the UK and US cancer research grid infrastructures. The immediate

benefits to the pathology researchers include access to larger collections of pathology images and existing analysis routines for validation and exploratory purposes.

#### **AUTHORS AND AFFILIATIONS:**

Tony Pan, MS, Department of Biomedical Informatics, The Ohio State University; Jason Lander, National Grid Service, University of Leeds; Martin Waterhouse, Section of Pathology and Tumour Biology, Leeds Institute of Molecular Medicine, University of Leeds; Alexander Wright, Section of Pathology and Tumour Biology, Leeds Institute of Molecular Medicine, University of Leeds; Shiv Kaushal, National Grid Service, University of Leeds; Terry Harmer, PhD, Belfast eScience Center, Queens University of Belfast; Ashish Sharma, PhD, Department of Biomedical Informatics, The Ohio State University; Joel Saltz, MD, PhD, Department of Biomedical Informatics, The Ohio State University; Phil Quirke, MD, Section of Pathology and Tumour Biology, Leeds Institute of Molecular Medicine, University of Leeds; Metin Gurcan, PhD, Department of Biomedical Informatics, The Ohio State University; Darren Treanor, MD, Section of Pathology and Tumour Biology, Leeds Institute of Molecular Medicine, University Of Leeds

### **30. Guidelines on Minimum Information Collection for Antibody Therapy Experiments**

Research groups developing antibody therapies are generating diverse data sets, however the value of these individual sets could be greatly compounded by amalgamating all data sets. Analysis to detect interactions between complex sets of parameters may yield further reaching interactions when performed on larger scaled data sets. Valid comparisons between experiments would facilitate the detection of research areas not yet fully explored. In order to achieve functional amalgamation, standards for collecting data from experiments must be defined. According to the NCRI Planning, these standards need to be defined particularly in the areas of pre-clinical and clinical development of antibody-based drugs. This poster describes the creation of common data elements (CDEs) as guidelines for collecting minimum information from experiments on antibody therapies. The CDEs conform to the ISO11179 standards, concepts used to build the definitions and properties of the CDEs are sourced from the controlled vocabulary provided by NCI Thesaurus. CDEs are created for fields identified on the Guidelines for Information About Antibody Therapy Experiments (GIAATE) tree, developed by the Antibody Society. The CDE value domains link to existing databases identifiers where possible. For example, the protein structure of the antibody target is a field which links into Protein Data Bank through the protein structure identifier. The CDEs are presented through forms, each form containing a field of data to be collected, the field's definitions and the context in which the field is to be used. The forms also present acceptable data values and formats. A server will be hosted to allow researchers to download forms, to propose changes to current CDEs in terms of data values, formats or contextual use and to allow researchers to develop new CDEs, specific to their areas of antibody therapy research.

#### **AUTHORS AND AFFILIATIONS:**

Richard Begent, Professor, Cancer Institute University College, London; May Yong, PhD, Cancer Institute University College, London; Igor Toujilov, PhD, Cancer Institute University College, London; Sylvia Nagl, PhD, Cancer Institute University College London

### **31. ONIX Semantic Federated Query Infrastructure**

Data search and semantic integration are central problems in biomedical grids. NCRI ONcology Information eXchange (ONIX) is designed to facilitate users in searching multiple UK and international data sources based on the resources structure, syntax and semantics. ONIX aims to interoperate with caBIG resources, whose semantic representation is maintained in a metadata registry (caDSR) and the semantic interpretation is based on mappings of caDSR data elements to the NCI Thesaurus (NCIt). This study presents the design of the ONIX semantic federated query infrastructure. Semantic queries, which can be expressed using concepts, rely on ontological representations of the related data resources. The benefit of this infrastructure is its flexible approach to semantic data integration. It supports maintainability while allowing evolution of existing data sources (in content, structure or semantics) and simple addition of new data sources. In the case of caBIG data sources, an OWL ontological representation is automatically generated from their UML models and their semantic representation, which is stored in caDSR. By using this knowledge representation methodology and exploiting the existing caGrid federated query infrastructure, ONIX provides semantic queries where join conditions are given by reasoning over the ontologies. An example is presented where two ontological representations of related resources are built and semantic queries run against them. This allows cancer researches to perform exploratory scientific investigation and hypothesis testing. Apart from supporting concept-based

queries over multiple related data sources, additional advantages of the ontological representation are examined including: detection of query inconsistencies, checking query containment, and verification of UML model consistency. To conclude, this study demonstrates how to exploit cutting-edge semantic web technologies and current caGrid metadata to provide ONIX semantic querying functionality.

#### **AUTHORS AND AFFILIATIONS:**

Alejandra Gonzalez Beltran, PhD, University College London; Anthony Finkelstein, Beng, MSc, PhD, Ceng, FIET CTP FBCS, University College, London; Jeff Kramer, FREng, FCGI, Imperial College, London; J. Max Wilkinson, PhD, NCRI Informatics Initiative

### **32. eCRF Designer: Intuitive Dynamic Semantically Interoperable Case Report Form Designer**

One of the most challenging tasks for running a randomized controlled trial (RCT) is to design data-aware and action-enabled case reports forms (CRF) to follow up and collect participants' data. Traditionally, CRFs are paper-based although recently more electronic-based CRF (eCRF) are being used in clinical trials. Currently simple eCRFs are either created using specific templates based on proprietary software, such as Microsoft Excel, or bespoke systems. However, more complex eCRFs, that can validate data and include metadata annotations, require intelligent programming support, and are usually created by a professional programmer. Various clinical trial management systems started adding tools to create eCRFs but with limited capabilities. The caBIG Form Builder was amongst the first CRF designers to enable the creation of CRF-structures using common data elements (CDE) concepts. Using CDEs enables creating CRF-structures that are more semantically interoperable and re-usable across studies. However, these CRF-structures provide only structural content and lack data awareness, dynamic action invocation and validation, and graphical interface features and capabilities. The ePCRN eCRF Designer has been created to overcome these limitations. It enables creating eCRFs from ISO-11179 compatible CDEs as its basic elements, which can be newly created or imported from an ISO-11179 store. These CRFs can be enriched by meta-data and graphical user interface attributes. The created forms are then automatically generated and dynamically deployed within the ePCRN clinical management system. These forms can be exported for re-use or imported directly from the caBIG caDSR repository.

#### **AUTHORS AND AFFILIATIONS:**

A. Taweel ; Delaney BC; Peterson K; Zhao L; Arvanitis T; Speedie S; Janowicz M.; Sim I; Hobbs FDR, University of Birmingham Edgbaston

### **33. Intuitive Graphical User interface for Capturing Clinical Trials Eligibility Criteria Using NCI Enterprise Vocabulary Service**

One of the principal tasks for completing a randomized controlled trial (RCT) successfully is to recruit the required number of participants. For RCTs that will recruit from a pool of prevalent cases, it is possible to conduct searches of individual electronic health records (eHRs) held in clinics, visiting the clinic and running a search on the clinical record system, but this is a laborious process, and results may not be comparable between systems. The electronic Primary Care Research Network (ePCRN) is an NIH Roadmap funded project designed to construct an electronic platform for conducting clinical trials in primary care. In order to link searches with the trial eligibility criteria it is necessary to use a controlled vocabulary, allowing a choice of clinical concepts and target codes. This paper presents a prototype system that was implemented to capture various elements of eligibility criteria in primary care. The system provides an intuitive interface that allows capturing different elements of the eligibility criteria. The interface is driven by an underlying generic research object model for primary care clinical trials. It dynamically links to the Electronic Vocabulary Service (EVS) from CaBIG through a custom developed GUI interface that enables retrieving terminology and the respective available codes and coding information. Once all elements of an eligibility criterion are captured, the system allows generating an XML and/or SQL query based on a CCR structure. This query can then be submitted through the ePCRN research workbench and grid-based infrastructure to retrieve counts of eligible patients that meet the define eligibility criteria.

#### **AUTHORS AND AFFILIATIONS:**

A. Taweel ; Delaney BC; Peterson K; Zhao L; Arvanitis T; Speedie S; Janowicz M.; Sim I; Hobbs FDR, University of Birmingham Edgbaston

## 34. Applying AstroGrid Techniques to the Analysis of Tissue Microarrays

Antibody hybridization and image scanning of tissue microarrays (TMAs) is a highly automated process, but the subsequent manual scoring of TMAs by a trained pathologist is a major bottleneck in their analysis. We can overcome this bottleneck by applying techniques developed for astronomical imaging data, as part of the global Virtual Observatory initiative, to the analysis of TMA images, and build pipelines which automate the analysis, acquisition and querying of TMA data. We describe an initial test case application for the automated 'scoring' of Estrogen Receptor (ER). ER is an important regulator of mammary growth but is also known to play a role in breast cancer. Assessing ER in the clinical setting enables decisions to be made as to which care programmes should be followed by patients. Pipeline automation, the basis of our 'PathGrid' system, has been implemented by adapting components from the UK's AstroGrid virtual observatory project. The processing pipeline, built from these components, can be executed as a workflow using a plug-in for Taverna ("Astro-Taverna"). The processing algorithms themselves are adapted from those developed to handle deep sky astronomical imaging programs. In our initial pilot study we assess ER in approximately 500 tumors from a large population-based clinical trial (SEARCH; part of the Anglia Breast Cancer Study). The poster describes the workflow covering the login process, file transfer, image conversion, image analysis, generation of results, and storage of resulting images and files on either a local or virtual file system. We also validate our ER scoring algorithm by comparing the results with scores manually assigned by a pathologist. The Pathgrid system should have a significant impact in improving the quantitative analysis of a range of TMA markers, providing increased throughput and objective assessment of expression.

### AUTHORS AND AFFILIATIONS:

Nicholas Walton, PhD, University of Cambridge; James Brenton, MD, PhD, University of Cambridge; Carlos Caldas, MD, University of Cambridge; Mike Irwin, PhD, University of Cambridge; Asif Akram, PhD, University of Cambridge; Peter Macallum, PhD, University of Cambridge; Nikita Makretsov, MD, University of Cambridge; Lorna Morris, PhD, University of Cambridge

## 35. Clinical Data Integration for METABRIC

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) Project collects clinical and genomic data on breast cancer tumors from five different hospitals/research centers in the UK and Canada. The aim is to discover molecular signatures for breast cancer and better classify tumors by analysing high throughput molecular profiling experiments on the tumor samples. The clinical information associated with tumors differs between hospitals and between tumors from the same hospital. The lack of common standards for cancer data representation and of any means to relate one set of data values to another makes data sharing and integration a challenging task. The caGrid project has created a set of tools for the standardization of datasets, using inference rules to describe relationships between different datasets, and enabling querying across diverse datasets. Here, we create Common Data Elements (CDEs) for each field of the original clinical datasets. The poster shows how the standardization tool converts the original clinical dataset (in XML format) into RDF using the SAWSDL markup on XML schemas to attach metadata identifiers to the resulting RDF file. In order to enable queries across differently defined fields in the datasets, we formulate inference rules that describe the relationships between the different CDE values. The poster demonstrates an example of one of these inference rules and the result of applying the inference rule on a small sample of clinical data. Finally we show how the Query tool can be used to extract data from RDF back into XML format for further processing. This approach ensures no information loss, since the datasets are stored according to their original definitions and classification schemes. It should have significant impact on the re-use of data from a variety of sources, enabling their integration and comparison, thereby facilitating their use for meta-analysis.

### AUTHORS AND AFFILIATIONS:

Irene Papatheodorou, PhD, University of Cambridge; Charles Crichton, BSc, University of Oxford; Lorna, Morris, PhD, University of Cambridge; Peter Maccallum, PhD, University of Cambridge; Jim, Davies, PhD, University of Oxford; James Brenton, MD, PhD, University of Cambridge; Carlos, Caldas, MD, University of Cambridge

## 36. Linking Tissue Microarray Core Image and Clinical Data

High-throughput genomic methods, such as expression microarrays studies on frozen tissue samples, have resulted in the discovery of many novel gene signatures that are correlated with clinical outcomes in cancer treatment. It is essential that these potential biomarkers are validated on large numbers of independent samples prior to clinical use. Tissue microarrays (TMAs) from paraffin-embedded tumor samples are an ideal reagent for validation experiments. In the CancerGrid project, we have been working on techniques for the automatic integration of data from TMAs with the clinical data obtained in large phase III studies. This poster describes the tracking and exchange of TMA and clinical data between different institutions involved in running a phase III study in breast cancer. The annotation of data with Common Data Elements (CDEs) allows automatic comparison and integration of data from a variety of sources. The poster also outlines a minimum set of CDEs for TMA layout and immunohistochemistry for a TMA slide, analogous to the MIAME (Minimum Information About a Microarray Experiment) standard for expression microarrays.

### AUTHORS AND AFFILIATIONS:

Lorna Morris, PhD, University of Cambridge; Steve Harris, PhD, University of Oxford; Charles Crichton, BSc, University of Oxford; Irene Papatheodorou, PhD, University of Cambridge; James Brenton, MD, PhD, University of Cambridge; Carlos Caldas, MD, University of Cambridge; Jim Davies, PhD, University of Oxford

## 37. Lowering the Barriers to Cancer Imaging

Cancer imaging researchers deal with a variety of issues including mutual understanding and the sharing of methods, data and information during the development of solutions for medical image analysis (MIA). Our goal is to lower the barriers to cancer imaging by providing mechanisms to enhance collaboration among clinicians and MIA researchers, alleviate the frustration of non-IT members at not being able to interact with medical imaging applications with reasonable effort, and maximize the efficiency of MIA researchers during the development and deployment of applications for clinical trial. Generally, MIA researchers spend approximately 30% of their research time testing the performance of methods previously developed by other researchers. The lack of available information about such methods combined with the significant implementation time could contribute to unsatisfactory results. Moreover, for deploying a method for clinical trial, additional time is spent to generate a suitable user interface. We are designing a framework built on cloud computing concepts, well known image processing toolkits, existing middleware and, where possible, on Microsoft tools. Furthermore, we believe that the use of multi-touch and interactive technology could be a subtle and intuitive way of engaging clinicians with the use of imaging tools and for enhancing collaboration among clinicians and MIA researchers. We have already proved that the use of graphics tablet technology enhances manual segmentation tasks performed by MIA researchers and clinicians in Oxford.

### AUTHORS AND AFFILIATIONS:

Maria Susana, Avila Garcia, PhD, University of Oxford; Anne E. Trefethen, Prof., University of Oxford; Michael Brady, Prof., University of Oxford; Fergus Gleeson, Dr., Churchill Hospital, University of Oxford

## 38. Identifying Collaboration Opportunities for Scientists from Coded Research Descriptions

Introduction: Initiatives such as caBIG™ increase the potential for fruitful collaboration between diverse communities of cancer clinicians and researchers by reducing data- and knowledge-sharing barriers. However, increased availability of data without adequate analysis tools may yield "information overload" rather than benefit. We hypothesize that collaboration opportunities can be identified by an automated clustering strategy that separates researchers into expertise domains and then matches them using descriptors derived from electronic sources. As an initial step, we studied whether concept sharing in the work of basic and clinical scientists precedes future collaboration. Methods: We analyzed six years of MEDLINE research articles from major basic and clinical journals. Research descriptors were derived from articles by basic and clinical scientists in the first five years who co-published for the first time in the sixth year (pre-collaboration) or who did not co-publish (non-collaboration), using MeSH headings that describe concepts in the NCI Thesaurus. Similarities of the basic versus clinical descriptors in pre-collaboration versus non-collaboration work were computed by the cosine measure. Results: Conceptual overlap in basic and clinical publications was substantial ( $r=0.74$ ). Concept similarities in publications by pre-collaboration scientists [ $0.048\pm 0.064$  (mean $\pm$ SD)] and in scientists' own work ( $0.076\pm 0.14$ ) were both greater ( $p < 0.0001$ ) than in publications by non-collaborating scientists ( $0.010\pm 0.038$ ). Pre-collaboration and non-collaboration publications were conceptually distinguishable by ROC analysis (AUC=0.74).

Conclusion: Scientists from two broad domains whose publications share research descriptors have an increased probability of successful collaboration. Pruning and/or weighting classes of concepts by relevance to collaboration may permit better separation. With additional work in deriving research descriptor lists for scientists from their publications, and in incorporating other coded information sources containing scientists research descriptions, this method could provide a foundation for collaboration discovery tools useful to scientists and research programs.

**AUTHORS AND AFFILIATIONS:**

Andrew R. Post, MD, PhD, Department of Public Health Sciences (Clinical Informatics Division), University of Virginia, Charlottesville, VA; James H. Harrison, Jr., MD, PhD, Departments of Public Health Sciences (Clinical Informatics Division) and Pathology, University of Virginia, Charlottesville, VA

### **39. cancer Bench to Bedside (caB2B): A caGrid Client to Facilitate Translational Research**

caB2B is a caGrid client that permits scientists to leverage caBIG™ compatible data and analytical services through a user friendly GUI. Metadata-based query interface enables end user to search virtually any caGrid data service. The goal of the tool is to aid translational scientists in combining data from tools like a caBIG™\_ biospecimen repository service, caTissue Core/Suite, with data from a microarray data repository like caArray and use the analytical services to analyze and visualize the results. caB2B has two modules, the administrative module and the end user client. The administrative module of caB2B is a GUI where the administrator can customize a particular instance for end users. For example, an administrator can select models and service instances that may be queried, curate paths between classes in models, and create inter-model joins. The end user client is a Java application that enables end users to query for and persist data available on the caGrid. It also allows for the analysis and visualization of this information. The query component consists of a diagrammatic view that allows the user to create a directed acyclic graph of the query that is to be executed and also helps the user to connect two or more classes to be searched. The data obtained from the query may be saved as a virtual experiment™ and customized. The end user may also visualize data in the experiment by using various graphical components, thus allowing comparison, variation and co-relational analysis of the data. Thus, caB2B is a caGrid client capable of querying and analyzing available data on the caGrid. The complex combinatorial, visual and statistical analyses that it currently enables and will facilitate in the future will result in the better understanding of the complex pathophysiology of polyfactorial diseases and will accelerate the identification of potential diagnostic markers and therapeutic targets for these illnesses.

**AUTHORS AND AFFILIATIONS:**

Mukesh K. Sharma, Washington University School of Medicine; Chandrakant Talele, Persistent Systems; Srikanth Adiga, Persistent Systems; Rakesh Nagarajan, Washington University School of Medicine