



NCRI  
INFORMATICS  
INITIATIVE



NCRI  
National  
Cancer  
Research  
Institute

# Newsletter

AUTUMN 2008 ISSUE 11

## Welcome to Issue 11 of the NCRI Informatics Initiative Newsletter

The Informatics Initiative is hosting the 'Data Sharing Debate' on Monday 6<sup>th</sup> October, during the forthcoming NCRI Cancer Conference in Birmingham. The debate, which will be chaired by Informatics Task Force Chairman, Prof. Richard Begent, will bring together representatives from industry and academia to discuss the perceived pros and cons of data-sharing from different perspectives.

Knowing that there is a growing obligation for researchers to make their data more widely available, but appreciating that there is a continued uncertainty regarding the practicalities / risks of doing so amongst the general research community, the intention of the debate is to attract a wider audience than those with an established interest in biomedical informatics.

Through this debate we hope to gauge the perception of medical researchers and physicians towards data sharing as a principle and in practice. We hope that the output generated from this debate will allow us to have a better understanding of the research community and their concerns through which to develop tools and programmes to assist the community with their data-sharing obligations.

We very much hope that you will join in the debate to help us build a constructive dialogue to help us further assist the research community in this area.

For more information on the NCRI Cancer Conference visit: <http://www.ncri.org.uk/ncri-conference/>.

We hope you enjoy this issue!

### INSIDE THIS ISSUE:



SPECIAL FEATURE  
**Life Sciences Data  
Analysis Tools**



UNIT ACTIVITIES  
**Exploiting  
the informatics  
revolution**



MEETING UPDATE  
**2<sup>nd</sup> annual  
caBIG™/NCRI  
Informatics joint  
conference**

UNIT NEWS  
**New Starters**

If you have any queries or comments on any article in our newsletter, or would like to contribute to the next issue, please contact us at:

NCRI Informatics Initiative  
PO Box 123  
61 Lincoln's Inn Fields  
London, WC2A 3PX

Email us at:  
[info@cancerinformatics.org.uk](mailto:info@cancerinformatics.org.uk)

## caBIG™ Life Science Data Analysis Tools

There is a tremendous need to integrate life sciences datasets from disparate sources to enable translation of critical information from bench to bedside and back. Hence, a critical factor in the advancement of biomedical research and delivery of care is the ease with which data can be integrated, redistributed and analyzed, both within and across functional domains.

### The need

Let's consider a typical use case in cancer research. High-grade gliomas, which include glioblastoma and anaplastic astrocytoma, are the most common intrinsic brain tumours in adults and are nearly uniformly fatal. Can we correlate gene expression profiles with clinical outcome in these tumours? Are there specific pathways that are impacted in these subtypes?

Many efforts are underway to understand the molecular genetics of these tumours and answer questions such as those raised above. The US National Cancer Institute and the National Human Genome Research Institute are currently creating a comprehensive catalogue of the gene changes that underlie multiple forms of cancer called [the Cancer Genome Atlas \(TCGA\)](#). In the pilot phase of the programme, data on large scale sequencing, gene expression, DNA fragment copy number changes, loss of heterozygosity, and the epigenetic state of the genome are being generated on a common collection of biospecimens

that will be annotated with rich clinical information. The TCGA data pipeline has thus far generated approximately 1 terabyte of heterogeneous glioblastoma datasets that need to be stored, shared and semantically connected so that researchers can extract scientific findings.

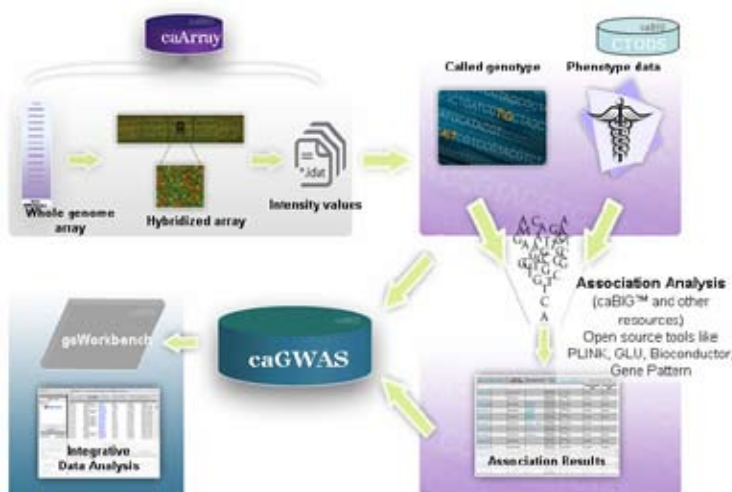
### caBIG™'s Life Sciences Distribution (LDS)

To address such critical needs in the translational research space, the US National Cancer Institute has recently released the Life Sciences software bundle as part of the cancer biomedical informatics initiative, caBIG™ initiative. The LDS includes five caGrid-enabled software tools: [caArray](#), a microarray data management system; [caGWAS](#), a system to manage data from genome-wide association studies; [caTissue](#), a grid-enabled version of the caTissue biobanking management tool; the [National Cancer Imaging Archive](#), a searchable repository for DICOM-based images that are integrated with clinical and genomic data; and the [Clinical Trials Object Database System](#), a repository for clinical trial information. The authentication/authorization security framework implemented by the LSD products helps ensure the protection of research participants in a manner consistent with consent forms, and data are shared appropriately via tiered open and protected access levels. The LSD tools support a variety of capabilities from tracking and managing biospecimens, to analyzing and integrating microarray data that are critical to large-scale cancer genomics projects like TCGA. Together, these tools enable cancer researchers to more easily integrate, analyze and share data from many different sources.

### Integrated portals

Biology is an information-driven science. Large-scale data sets from genomics, proteomics, population genetics and imaging are driving research at a dizzying rate. Simultaneously, interdisciplinary collaborations among experimental biologists, clinicians, statisticians and computer scientists have become the key to making effective use of these data sets. However, too many researchers have trouble accessing and utilizing these electronic data sets and tools effectively. It is not sufficient to provide infrastructure to store the life sciences datasets; it is equally important to serve the data through integrated user-friendly portals. caBIG™'s [Cancer Molecular Analysis \(CMA\)](#) portal addresses this need by integrating heterogeneous datasets using a user-friendly interface. The CMA portal utilizes the LSD's infrastructure and standard interfaces to bring these datasets together in an easily mine-able fashion. The current version of the CMA portal provides web access to tools for the analysis and visualization of gene expression, gene copy number, SNP data, methylation data and DNA sequence. The molecular data can be analysed within the context of clinical information, including treatment history, pathology status, tumour site and surgical history. Using CMA, a researcher can browse the mutations found in a particular cancer and then use the pathway maps to look for genes that are impacted by those mutations in the context of a signaling pathway. By simply selecting a biomarker and an anomaly type (expression, mutation etc), researchers can view Kaplan-Meier survival charts for patient cohorts, thus integrating genomic datasets seamlessly with clinical annotations with a click of a button. Examples of analysis and views in CMA portal are shown in Figure 2: Viewers for integrative cancer genomic analysis – Genes to Clinical outcome

Cancer statistics reveal that more than half a million cancer deaths occur annually. We need to put the tools in the hands of physician scientists who are generating new hypothesis for their next clinical trial or study. However, busy clinician scientists will not adopt tools that are difficult to use and have a steep learning curve. This is the challenge that the LSD infrastructure and integrated portals such as CMA are beginning to address with both summary level findings and analytic capabilities via user-friendly tools. ■



**Genome-wide association study (GWAS) workflow.** The usage of the LDS helps store and search for the most significant SNPs associated with cancer occurrence.

UNIT ACTIVITIES

## Exploiting the Informatics Revolution

### ONcology Information eXchange - 3rd User Group session

During the first half of 2008 the NCRI Informatics Initiative in partnership with University College London's Platform Architecture and Requirements Testing (PART2) project conducted a third round of ONIX user sessions. Users were asked to provide examples of their day-to-day work and were asked specific questions regarding the reasoning, meaning and results of their activity together with the purpose of such activity. This particular round was aimed at eliciting specific requirements on searches for resources and data as well as searches over data, semantics and data integration. To this end a subset of the ONIX user group was selected specifically because they are engaged in informatics activity in their day-to-day work. The users were bioinformaticians (both core services and experimentalists), clinical data managers and research scientists that use informatics resources to fulfil aspects of their objectives or research goals. This work is currently being analysed for specific requirements. This article describes some general themes that were identified.

#### What the users think...

Within the group, users often query the same resources for different reasons. Users require different functionality based on their research purposes or specific tasks and that these purposes were often complex. For example, certain 'ag-

gregator' websites were accessed regularly though these were generally quite specific in the data they aggregate (e.g. PubMed, geneCards, Ensembl to name a few). In addition a large amount of time was required for repeating simple and generic tasks such as locating and accessing data, scripting analysis code and algorithms and cleaning data. Many of these tasks were achieved through a variety of software but popular packages were identified.

Data sharing was generally seen as a good thing amongst the users and many of them described barriers to effective data sharing that regularly took time to overcome. There was often concern that data sharing did not compromise the custodianship of sensitive clinical data or the intellectual property that researchers attain in generating empirical data. Almost all the users identified authenticity and authority of data as a serious concern to data sharing. The preference was for raw data with associated metadata that could be validated either on the fly or in-house. Where access to raw data was not possible clear evidence supporting processed data was required.

All users wished to be able to efficiently construct queries with the confidence that the system could 'understand' what they were looking for and could be found with ease. Several users mentioned they were open to the concept that such queries would search sources that they were not familiar with and locate data they had not previously considered.

So with the preliminary analysis and elicitation of this user round the PART2 project set about designing an architectural component that centred around semantic querying. This functionality is

being designed to fulfil the requirement to construct semantically rich queries, to distribute these queries over multiple resources and that in constructing and executing such queries users would be confident that ONIX could find what they meant. The following article describes what involves to be able to do semantically-rich queries.

#### Have you found what you meant?

In the biomedical domain, scientists constantly generate different types of data to perform their research activities and access a variety of information sources to put their data into a broader context. More often than not, a single source of information is not enough and multiple sources need to be generated, found and/or accessed.

Given this setting of distributed data sources, scientists not only face the problem of access but also the issues of validation, interpretation and usage of the information they find. Different data sources are likely to use disparate data formats, divergent terms to refer to the same entity (e.g. carcinoma and malignant epithelial tumour are synonyms) or even use the same term to refer to different concepts (such as insulin, which may refer to a gene, a protein or a drug depending on the context). Additionally, distinct biological data sources might store information at different levels of detail depending on the particular studies (e.g. the gene for BRAC1 can be associated with DNA, RNA or protein sequence in one resource while another resource may hold predilection and association data for disease occurrence). To further complicate this landscape, subjective interpretation is common and consensus can be difficult to resolve.

For all the reasons stated above, the information about the interpretation of the data is as paramount as the data itself. Without knowing about the meaning of the raw data, usually referred as metadata because it is data about the data, it becomes impossible to use the raw data. Thus, computer science topics such as knowledge representation and semantics are extremely relevant in the environment of distributed data sources.

This exact picture of data heterogeneity and distribution was painted by the users, while performing the user group sessions as described in the previous article. Although scientists are currently using different Web-accessible tools that



The NCI Terminology browser. Search results for 'carcinoma'.

provide capabilities to access biological databases, these are still lacking in functionality. In general, there is no uniform query interface to access several resources and researchers need to deal with different interfaces during much of their work. There was an agreement on the usefulness of sites that aggregate information from different sources. However, these tools generally deal with sub-domains rather than exposing cross domain data. Semantic relationships are usually not explicit and most tools still do not provide support for retrieving semantically related data. The scientist is largely in charge of manually combining the results in order to achieve their goals. A general concern was the lack of metadata annotations in some of the available resources.

Through the collaboration with University College London (UCL), the NCRI Informatics Initiative is exploring solutions for the challenges presented here to be incorporated to the Oncology Information eXchange (ONIX). The aim is to provide a consistent way of constructing and performing searches for data and data sources. Such queries will take into account semantic relationships and retrieve the data relevant to the meaning of the search. The approach exploits the existing computing infrastructure of the cancer Biomedical Informatics Grid (caBIG™) project run by the U.S. National Cancer Institute. In particular is the programme of work ca-GRID, a technology that permits data and resource sharing and collaboration across organisations. While this system is rich in metadata, the infrastructure needs to be extended to exploit it further, as it currently does not support semantically rich queries. The PART2 project is working towards the design of a semantic query engine for ONIX. An additional challenge for PART2 is in adapting this semantic query engine to consider UK data sources, and their own distinctive uses of metadata. ■

## MEETING UPDATE

### The 2<sup>nd</sup> annual caBIG™/ NCRI Informatics joint conference September 2<sup>nd</sup> - 3<sup>rd</sup> 2008, Bethesda MD, USA.

The conference's aim was to initiate discussion centred on the theme "Biomedical Informatics without Borders: Enabling Collaboration to Strengthen Research and Care". This was explored through a range of talks, a panel discussion, breakout sessions, poster presentations and technology demonstrations giving delegates the opportunity to examine the technical, scientific, social and legal issues associated with achieving the seamless sharing of information that is required for effective collaboration.

The conference sessions highlighted:

- The work of Grid initiatives in supporting scientific collaboration;
- The need for a pragmatic approach to the legalities of data sharing;
- The importance of standards;
- The necessity for technologists to develop user-friendly research tools;
- The need to integrate information from research and clinical care;
- The importance of scientists to share data, adopt new technology and utilise standards;
- How shared data can be utilised by publishers, libraries and software developers to diversify its use and extend its potential; and
- The need for some flexibility to prevent research being stifled by the technical and legal requirements of data sharing.

For more information and to view the presentations and poster abstracts please visit: <https://cabig.nci.nih.gov/nci-ncri2008conference>. ■

## UNIT NEWS

### New Starters

#### ALEJANDRA GONZÁLEZ BELTRÁN

Dr Alejandra González Beltrán joined University College London (UCL) to work with Prof. Anthony Finkelstein (UCL) and Prof. Jeff Kramer (Imperial College London), in collaboration with the NCRI Informatics Initiative.

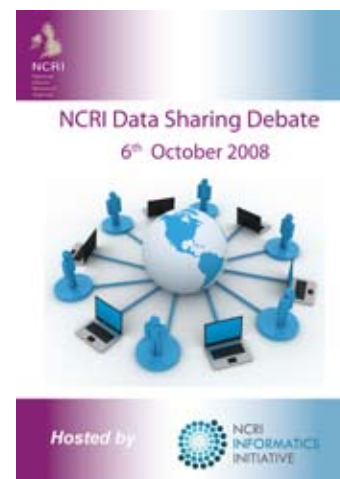
Alejandra received the degrees of Licentiate in Computer Science (Honours) from the National University of Rosario, Argentina, and Ph.D. from Queen's University Belfast, UK. Currently, she is working in the PART2 (Platform Architecture and Requirements Testing) project on semantic data integration of heterogeneous data resources. In particular, this involves supporting semantic federated queries over shared data resources in grid environments.

#### JONATHAN BULL


Dr Jonathan Bull joined the NCRI Informatics Initiative as a Research Manager in July 2008. Jonathan graduated with a PhD in Cutaneous Biology from Queen Mary College, University of London. Since then he has worked in academia as a Postdoctoral Researcher at Bart's and The London School of Medicine and Dentistry and at the Institute of Cancer Research. He has a strong background in cell and molecular biology, and a broad knowledge of cancer informatics through firsthand experience of high throughput technologies.

### In the next issue...

Update on the  
Informatics session during  
the NCRI Cancer Conference  
in Birmingham, UK.



NCRI  
Data Sharing Debate  
6<sup>th</sup> October 2008

Hosted by 



Presentation of best poster award to Steve Harris from Cancergrid during the joint caBIG™/NCRI Informatics conference.