

The universal genetics database: information sharing in genetics and beyond

by Dr D. Widdows and Prof M. Barmada

In many empirically intensive fields, researchers spend much of their time marshalling, formatting, and moving impenetrable blocks of data from one place to another. To solve this problem and pave the way for future-proof development, the University of Pittsburgh Graduate School of Public Health teamed up with MAYA Design, a research and technology lab spun out of Carnegie Mellon University. The result is the first prototype of a Universal Genetics Database, which automatically combines background information from genetic databases with experimental results in particular studies.

THE NEED FOR INTEGRATION

Biomedical informatics has become a field with many huge databases that contain valuable common information and countless small studies with information kept in local databases or text files that are spread across different file systems and exchanged by e-mail. Data integration has become a huge challenge in itself, with a variety of relational databases, markup languages, and heterogeneous ontologies vying for attention.

Researchers in the Department of Human Genetics at the University of Pittsburgh's Graduate School of Public Health (GSPH) are acutely aware of the information integration problems in the field of genetic epidemiology. With the advent of the genomics age ushered in by the Human Genome Project has come a multitude of databases, each with their own unique syntax. New graduate students with a background in biological sciences have to become familiar with scripting and query languages just to do the data plumbing that has become prerequisite to almost any scientific inference.

To pave the way towards solving these problems once and for all, the GSPH started an ongoing collaboration with MAYA Design, a Pittsburgh technology lab that has been developing and helping to deploy collaborative information architectures for the past 15 years.

THE UNIVERSAL DATABASE ARCHITECTURE

While the problems currently facing biomedical informatics are unique in scale and com-

plexity, they are similar to the increasing problems faced in many other growing fields. MAYA Design has spent much of the past 15 years researching, developing, and deploying systems for information collaboration, and during this period found endemic infrastructure patterns that were hampering integration.

Electronic information is dispersed across a huge number of locations, and to find that information, users need to request copies of files directly from those locations. Every time a file is moved from one machine to another, or even to a different location on the same machine, it changes its identity and becomes a different file. Imagine if every time a book moved to a different bookshelf, it became a different book! Library systems would never work, and the reliable transmission of knowledge that enabled modern science to develop would probably never have happened.

MAYA's core innovation is an information system in which information itself, and not its physical location or transmission medium, is the primary currency. Because information retains its identity wherever it is found, the information system is described as the Universal Database. The Universal Database is a peer-to-peer system in which all information is broken down into data objects called u-forms, which can be moved independently and replicated to many places at once [Figure 1].

Each u-form has a universally unique identi-

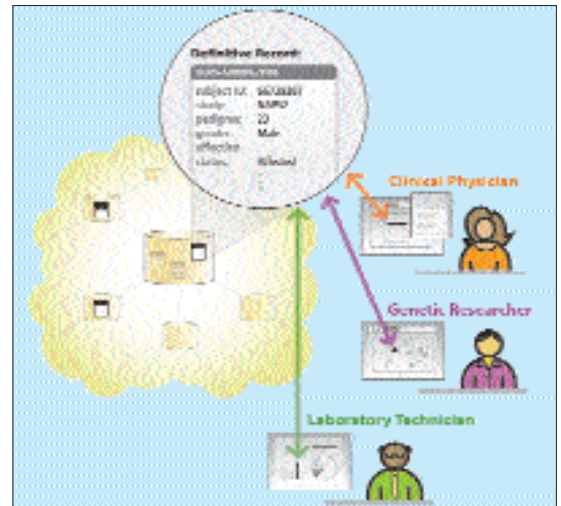


Figure 1. The same information objects can be replicated to many locations and seen in different ways by different users.

fier, so that any data object can refer to any other data object, even if they are from completely different datasets on different peers in the network. New variables or attributes can be added to any u-form, so the format of each data item can evolve as the need arises. These two requirements (data identity and data extensibility) have been recognized as vitally important by other developing frameworks such as the Semantic Web, and since (UUID, attribute, value) triples can be mapped directly to (URI, predicate, object) triples, the Universal Database can automatically be used to express the Resource Definition Format of the Semantic Web. The key difference

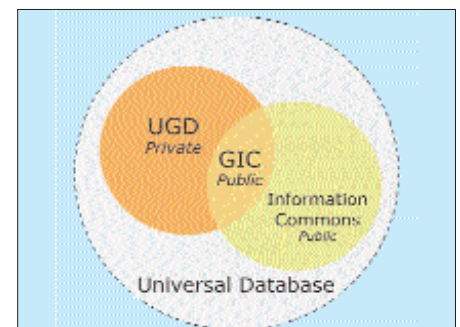


Figure 2. The Genetics Information Commons is the part of the Universal Genetics Database that is in the public domain.

between these two frameworks is that the Universal Database encourages data liquidity, i.e. the flow of information to wherever it is needed. Because the identifiers in the Universal Database are not physical locations, a u-form with a particular UUID can often be found in many locations, and can be replicated to a user's own venue. This is vital for supporting offline activity, and for optimizing analytical work with small portions of large datasets.

Because objects from authoritative datasets can be replicated to many different locations, MAYA uses the Universal Database architecture as a platform for disseminating the Information Commons, through which definitive and authoritative public data can be obtained and massively replicated.

THE UNIVERSAL GENETICS DATABASE AND THE GENETICS INFORMATION COMMONS

Researchers at the University of Pittsburgh's Graduate School of Public Health teamed up with MAYA Design to test the viability of a Universal Genetics Database (UGD), which uses the Universal Database as a platform for collection and fusing together genetic information and making this information and accompanying tools automatically available to researchers.

By studying the workflow and data needs of a typical genetic study, researchers from the GSPH and MAYA created an information architecture and data fusion tools to fulfill all of the information representation needs of several end-to-end studies, using the Universal Database architecture. (See attached case study). Part of the novelty of this approach is that information from publicly available genetic databases is represented using the same infrastructure as study-specific data about particular genetic samples, markers, and pedigrees. Where necessary, personally identifiable information is protected by encryption and by preventing its flow outside of the GSPH network. In this way, the public parts of the UGD become a Genetics Information Commons (GIC), contributing to the growing wealth of Information Commons data [Figure 2].

THE GREATER GIC VISION

Together with MAYA Design, GSPH

UGD CASE STUDIES AT THE GRADUATE SCHOOL OF PUBLIC HEALTH

Founded in 1948, GSPH is world-renowned for contributions that have influenced public health practices and medical care for millions of people. The Human Genetics Department within the GSPH is concerned with identifying genetic susceptibility loci for common complex disorders, and with understanding the impact of those susceptibility loci on disease prevention and public health.

Carrying out a genetic study to identify which genes are responsible for different effects involves collecting and collating information about family trees, which family members are affected, what biologic samples are taken from each individual, which "versions" (alleles/haplotypes) of each gene are present in each sample, and what information is available in standard genetic databases concerning genes of interest. This process could take up to 6 person weeks for a single study, and the integration of information from a plethora of local and remote databases, spreadsheets, and lab results in different file formats. To complicate matters more, typical studies generate several potentially overlapping partial data sets which must be compiled together to form a whole.

To streamline this process, MAYA Design and the GSPH worked together to create import and data fusion tools that do the heavy-lifting involved in creating a genetic study. As well as creating data import and expert tools, this process involved importing datasets into the GIC, including the National Center for Biotechnology Information's dbSNP and Iceland's Decode Genetics databases that keep track of known genetic loci and their positions on the chromosome. These datasets are indexed along useful (and easily extended) dimensions such as name and genetic distance in base pairs. As well as different portals for viewing the information space, the main functionality created was data import and export tools. Using simple commands, whole collections of input data can be brought into the system. As well as uniting data from several different file formats into a single model, this action caused data to be linked to important background information, such as public databases of information about genetic markers.

The GIC tools were used to import and analyse previous studies concerning Ulcerative Colitis and Crohn's Disease (the two common subtypes of Inflammatory Bowel Disease), as well as data from several population- and family-based studies of acute and chronic pancreatitis. The automatic import, indexing and fusion tools within a single information infrastructure enabled results to be obtained much more quickly, enabling researchers to spend more time on data analysis and less time on data integration.

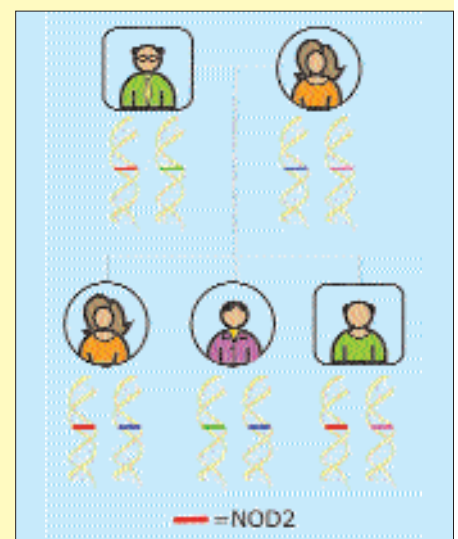


Figure 3. An example of the type of data used in a genetic epidemiology study. In addition to the information about who is related to whom, and what condition each person has, genetic studies need to deal with information on the markers typed in each individual (like the NOD2 marker, shown here, which is a known risk factor for inflammatory bowel disease). As the figure demonstrates, individuals carry different "versions" (or alleles) of markers within genes of interest, and the patterns of inheritance can give information on the likelihood that the gene is involved in the disorder (when combined with information about the pedigree and the disease). In addition, public databases such as those housed at the U.S. National Center for Biotechnology Information (NCBI) can give additional information, both about the disease of interest and about the gene or markers that are used in a study. All of this information must be integrated in a meaningful fashion for a genetic study to have a hope of identifying genes for complex human genetic diseases.

researchers are extending the GIC tools to address a richer variety of human conditions, and to fuse data across a wider collection of datasets from different domains that are already part of the Information Commons. The boundaries of the Genetics Information Commons are not fixed. Instead there is an integrated heterogeneous information space through which researchers can access demographic, textual, environmental and economic datasets, all fused around common points of reference such as shared spatial or temporal concepts.

The long-term goal of this system, as it develops, is to create an information infrastructure in which researchers spend less time tracking down, parsing and organizing different datafiles ("data plumbing"), and more time analysing and publishing scientific results. Universal identity, extensibility, and data liquidity gradually encourage an environment in which the importing and formatting of data by one researcher naturally makes it available to others if the initial research team so desires. Data reuse and availability across multiple dis-

ciplines will eventually enable scientists from many fields to collaborate and explore information at a much more granular level than is possible using the traditional journal article as the main means for information publication. Analysts using the Universal Database do not spend time researching the artificial data mismatches that result from data evolving separately in separate machines. This leaves the much more important problem of researching the relationships between phenomena in the real world.

TAKING PART IN GIC DEVELOPMENT

Researchers interested in joining the Genetics Information Commons can do so in a number of ways. Currently, MAYA Design is offering pilot versions of the Universal Database with command-line tools and APIs to select researchers who are interested in joining the GIC. In addition, MAYA Design is seeking partners to collaborate in pioneering the development of the GIC through joint projects that will add useful data and tools to the GIC for distribution to the research commu-

nity. For more information, contact Josh Knauer, Director of Advanced Development (knauer@maya.com).

THE AUTHORS

Dominic Widdows, D.Phil.

Senior Research Engineer

MAYA Design, Inc.

Building 2, Suite 300,

2730 Sidney Street

Pittsburgh, PA 15203

USA

Tel: +1 412 488-2900

e-mail widdows@maya.com

Michael Barmada, Ph.D.

Associate Professor of Human Genetics

Director, Center for Computational Genetics

Co-Director, Bioinformatics Analysis Core Services,

Graduate School of Public Health, University of Pittsburgh,

130 Desoto Street, Pittsburgh, PA 15261

USA