# BIMS: An information management system for biobanking in the 21st century

G. Ölund
P. Lindqvist
J-E. Litton

Although the mapping of the human genome has allowed researchers to integrate genomic data with demographic and life-style data for use in epidemiologic studies, the problem of merging and accessing data that originate in heterogeneous sources is yet to be overcome. In this paper we describe Biobank Information Management System (BIMS), a system designed to integrate data from various research studies in which data formats and data collection methods vary widely. In addition, BIMS handles data that are continually updated, provides a user interface that is easy to use and does not require programming skills, and controls access to data according to well-defined policy rules. We outline the current information management challenges in biobanking, describe the BIMS architecture and its main components, and discuss the extent to which it addresses the stated challenges.

## INTRODUCTION

Collecting biological specimens and samples and storing them into biorepositories have played an important role in the advancement of medical science. At the close of the 20th century, medical research was increasingly concerned with the genetic components of disease and their applicability to personalized medicine, which led to a heightened interest in the biobanking of human DNA (deoxyribonucleic acid).[1,2]

Providing a large, well-annotated repository of biological samples for scientific research is the main idea of biobanking. Although the storing of the actual human tissue is currently a prerequisite for preserving the information for future scientific studies, it is the information itself, contained within the tissue, that is key to understanding the genetic background of human traits. Of equal or higher importance is the information about the tissue donors themselves, that is, their phenotypes.

The possibility of combining molecular characterization of biological samples, obtained with high-throughput genotyping technologies, with demographic and life-style data collected through modern communications technologies is now on the verge of being realized on a large scale. Recording and long-

term storing of this information for future scientific research are our primary goals for biobanking in the 21st century.

Whereas most biobanks are designed exclusively for managing biological samples, for which purpose they make use of laboratory information management systems (LIMS), we aim to connect the sample information with genotypic, phenotypic, and environmental information. Toward this goal we implemented the Biobank Information Management System (BIMS), which is designed to integrate research data originating from many sources. The data may originate with various research studies, small and large, where data formats and data collection methods vary significantly. In contrast, in most biobanks, data collection has limitations imposed by the biobank system, an approach that allows for much greater uniformity and control over the data. In addition to integrating data from various sources, BIMS is designed with these additional goals in mind: handle data that are continually updated, query data with ease, and control access to data.

BIMS implements important advancements in the definition, structure, and standardization of information that has been gathered from a multitude of sources, such as population-based registries, biobanks, patient records, and large-scale molecular measurements. Even though no similar system for epidemiologic research has previously been described in the scientific literature, there are a number of related efforts, most notably UK Biobank[3] and work at the Mayo Clinic.[4] Other work, which attempts to address part of the problem, includes warehouse solutions such as MONICA (MONItoring of trends and CArdiovascular disease),[5] EPIC (European Prospective Investigation into Cancer and Nutrition),[6] and GenomEUtwin.[7]

The rest of the paper is organized as follows. In the next section we present a brief overview of the challenges involved in integrating biological data from heterogeneous sources. Then, in the main section of this paper, we describe BIMS, an information management system for biological data. We describe the system architecture, its main components, the user interface, and the policy issues that affected its design. In the concluding section we evaluate BIMS with respect to the previously described challenges for integrating biological data and outline directions for future work.

## DATA INTEGRATION CHALLENGES

The amount of information generated by biomedical research has increased almost exponentially due to the use of modern information technology (IT) tools. Particularly in the field of epidemiology, a discipline tightly associated with the use of biobanks, there is a growing need to manage the huge quantities of data collected from the study subjects themselves, their medical records, and their biological samples.

In describing the challenges of dealing with large amounts of data, we focus on the following aspects of data integration because these play a vital role in the management of medical data:

1. *Quality and comparability of data*—When dealing with data from several sources, it is important to first determine if it is possible to compare them. Data may have been collected by methods that are not consistent. It is possible that data variables which at first seem identical may not convey the same information at all. In addition, the quality of data varies as very few scientific data sources are regulated by a quality control system. Mistakes are often made in the data collection process. Data quality and comparability are not primarily technical issues, but IT can provide the mechanisms to detect errors and enforce quality control.

2. *Differing data models*—In different data sources, data may be structured according to different data models. These data models usually range from completely normalized relational data models to very simple spreadsheet-based models. The same type of information can be stored under different models, and when data is integrated, it is important to be able to handle these differences.

3. *Differing ontologies*—Data sources often make use of differing ontologies and taxonomies to represent the same information. This lack of standardization is a general problem that undermines the widespread use of biospecimens. Mapping ontology A to ontology B is a challenging research problem. However, once the mapping exists, the ability to integrate data based on different ontologies has important practical uses.[8]

4. *Deidentification*—For a variety of reasons data integration often requires the deidentification of data. The process of deidentification (sometimes

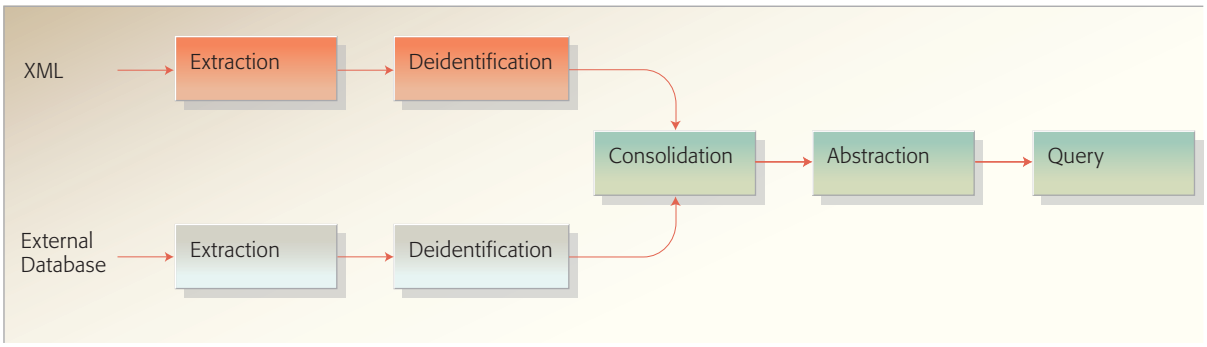


**Figure 1**
BIMS data flow

known as anonymization) generally involves the removal of information that can identify the person associated with a medical record. The deidentification process is sometimes reversible.

5. *Differing data formats*—Medical data is stored in a variety of formats, including spreadsheets, text files, data sets produced by statistical applications in proprietary formats, relational databases, images, various binary streams (such as instrument data), and Internet-based formats. All these data formats have to be accommodated by the information management system.

6. *Ownership of data*—When integrating data from sources under different ownership, it is important to be able to control access to data and track ownership. This is especially important for gaining the trust of data source owners and for promoting a spirit of collaboration.

   When integrating data from two data sources, the combined data set often contains more information than the two sources viewed separately from one another (a fact that is the primary driving force behind the integration of information). Thus, it is important to be able to control authorization (i.e., access to data and services) at a very detailed level. As an example, user U is granted access to data set A and to data set B, but U is not allowed to access A and B at the same time.

7. *Legal and ethical aspects*—Integration of data from multiple sources requires that we comply with the legal and ethical requirements, such as complying with the wishes of the study participants as expressed in the consent agreements. The main challenge is to comply with these requirements and, at the same time, make the data as useful as possible to the research community.

## BIOBANK INFORMATION MANAGEMENT SYSTEM

The BIMS functional requirements are supported by a number of components framed as sequential processes: *extraction, deidentification, consolidation, abstraction*, and *query*. ***Figure 1*** shows the flow of data through BIMS. Starting on the left, data is processed through the extraction, deidentification, consolidation, abstraction and query components. We now describe these components in more detail.

### Extraction

The extraction component receives data from a number of external data sources, whose data is not under the control or ownership of the BIMS administration. The data may be in different formats, such as relational databases, spreadsheets, and text files, and may originate in physically separate locations. The data may be organized according to differing data models. There may be vast differences in skills among the people responsible for managing the data, and the frequency of data updates may vary widely.

New data are added to BIMS by using one of two methods: (1) federated database connections and (2) XML documents. Federated database connections[9,10] are in effect direct database-to-database connections between the BIMS database and external data sources. Data from a federated data source is copied directly and stored inside BIMS in a more or less unaltered form, akin to what occurs in a data

warehouse. If the external data source is a relational database, then the federated approach is the preferred integration mechanism in BIMS. These data are continually updated and maintained, and these updates are uploaded in BIMS, provided that the data owners allow this kind of access to their system.

The more common scenario, however, is one in which the source of data is not a state-of-the-art database but research data in a variety of formats, such as text files and spreadsheets. In order to import this type of data into the system, a set of XML schemas, supporting different types of data, is provided. Researchers can generate XML documents conforming to these schemas in order to import their data into the system. The major advantage of this data input interface is that it moves the responsibility of cleaning up the data from the BIMS administrator to the persons who know the data best, the researchers themselves. An added benefit is that the schemas allow for certain types of validation to be performed on the data before it has even entered the system, thereby increasing their quality. As in the federated approach, XML documents can be used either for one-time import of data from a data source or for updates to that data.

XML documents can enter the system in one of two ways: either they are manually uploaded by the user, or they are automatically imported from other systems through the BIMS Web interface. Because not all researchers are skilled in XML, BIMS also provides a set of simple tools that help convert text data into XML.

## Deidentification

Every data item in BIMS is associated with a person. Whether a blood pressure measurement, a food-related habit, a birth date, or a genotype, the data item must contain some sort of identifier that allows it to be connected to the person it describes. We refer to such an identifier as a *person identifier* (PID). All data sources use their own PID when data is collected—we refer to it as a *data-collection-specific PID* (*collection PID*, for short). Other PIDs originate with an analysis of data collected elsewhere. We refer to the latter as a *data-analysis-specific PID* (*analysis PID*, for short). Researchers can use these PIDs to track the identity of a person within their data. In BIMS however, these identifiers by themselves are not sufficient to establish a connection between persons' identities across data from different external data sources. Fortunately a mechanism for establishing such a connection can be implemented based on the fact that all Swedish residents possess a *civic registration number* (***Figure 2***).

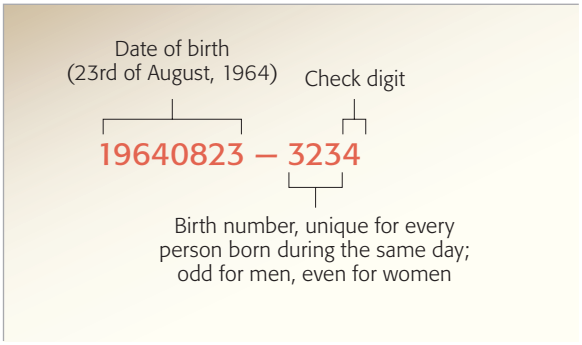


**Figure 2**
Swedish civic registration number

The Swedish civic registration number ("personnummer" in Swedish) is a uniform identifier, similar to the Social Security number in the United States, that has been in use since 1947. Once assigned, it is used throughout a person's life for identification purposes in health care, insurance, banking, driver's licenses, military service, and credit card transactions. By virtue of its widespread use it can also be used as a tool for epidemiological research.

Because the civic registration number must be highly protected due to its being tightly coupled to personal integrity, it is not stored within BIMS. Yet, in order to fulfill the requirements of the Swedish law known as the Biobanks in Medical Care Act,[11] BIMS must support the retrieval of all data items associated with a given person. This function is required, for example, when persons withdraw their consent to have their medical information stored in BIMS.

For these reasons all data stored in BIMS have to be deidentified; that is, the civic registration number and any other information that can possibly lead to the identification of the person must be removed. Deidentification can be accomplished in one of two ways: exclusion or alteration. In the first approach, we simply omit the identifying data during the extraction process. In the second approach, we alter the data by means of a one-way hashing algorithm,

SHA-512.[12] The civic registration number is not stored in BIMS—just the result of the hashing function. We refer to the PID thus obtained as the *global PID*. The civic registration number can still be used to locate BIMS data associated with the person (as a specific civic registration number always produces the same hashed product), enabling a one-way mapping of an incoming civic registration number to its altered self. If a person requests that his or her data be tracked or removed from BIMS, that person's civic registration number is used by the BIMS administration as input to the same algorithm to locate the data in question. This solution provides a dynamic way of deidentifying sensitive data (currently only the civic registration number, something that can be expanded if necessary) until a European counterpart to HIPAA (Health Insurance Portability and Accountability Act), a United States law, is widely accepted.[13]

Using a third party for deidentification, in which an external organization is responsible for the deidentification process, is not implemented in the current BIMS solution. However, the BIMS architecture can accommodate a solution, should the need for it arise from a legal or ethical viewpoint, in which the entire deidentification component is hosted by a third party.

## Consolidation

Consolidation is the act of mapping the imported data sets to a unified form. The process involves two data manipulation steps: model consolidation and identity consolidation.

Model consolidation aims to fit the data into either a *shared data model* or a *source-specific model*. The shared model is so named because it stores data from different data sources according to a common (shared) normalized data model. Such a model provides a framework that facilitates the combining of research data from different studies. It also makes the data administration task easier and allows the use of a single set of tools for quality control (data integrity checks, data consistency checks, etc.). Even though the data is stored according to a shared data model, each data item has a link that points to the data source from which it originated. Examples of data currently based on the shared model include genotypes and consent information.

The source-specific model holds data from a particular data source that is deemed too specific to warrant inclusion in the shared data model. Such data usually is complex, comes from a very specific research area, or consists of biological sample information from a biobank LIMS. Because the data are stored according to their own data model, the need for mapping the data onto another is limited.

The process of identity consolidation links the PIDs in BIMS data to the person they correspond to. Some of these PIDs are collection PIDs and others are analysis PIDs. The collection PID and the analysis PID are linked by the global PID as shown in ***Figure 3***.

Note that the same person can be associated with several collection PIDs and to several analysis PIDs, depending on how many collection data sources and analysis studies the person is a part of. Figure 3 illustrates the linking mechanism and the associated database tables used in BIMS to associate data items from different data sources with a person. All the IDs in the shaded area form part of the data access control mechanism. The data which can be accessed from within a particular analysis study is determined by the collection data sources to which the study is linked and the analysis PIDs that belong to the study (see the section "Security" later).

If several data sources use the same system of collection PIDs, that is, if they belong to the same *data source group*, it is sufficient to have the deidentified civic registration number (global PID) for one of the data sources in the group; the identity consolidation process then provides the link to the collection PIDs in the remaining data sources. In a corresponding fashion, the analysis PIDs are also linked to the global PID.

## Abstraction

Following consolidation, the data are in effect ready to be queried by researchers. It is beneficial, however, to insert an additional abstraction layer that supports the data presentation and the control of data access. Because these functions are intertwined, one mechanism is needed that supports both.

The abstraction layer is specified by an XML document that defines the mappings to be applied to database variables. This information is parsed at runtime and used for displaying information to the BIMS user in the graphical user interface. In effect, it is this step that actually links collection PIDs with

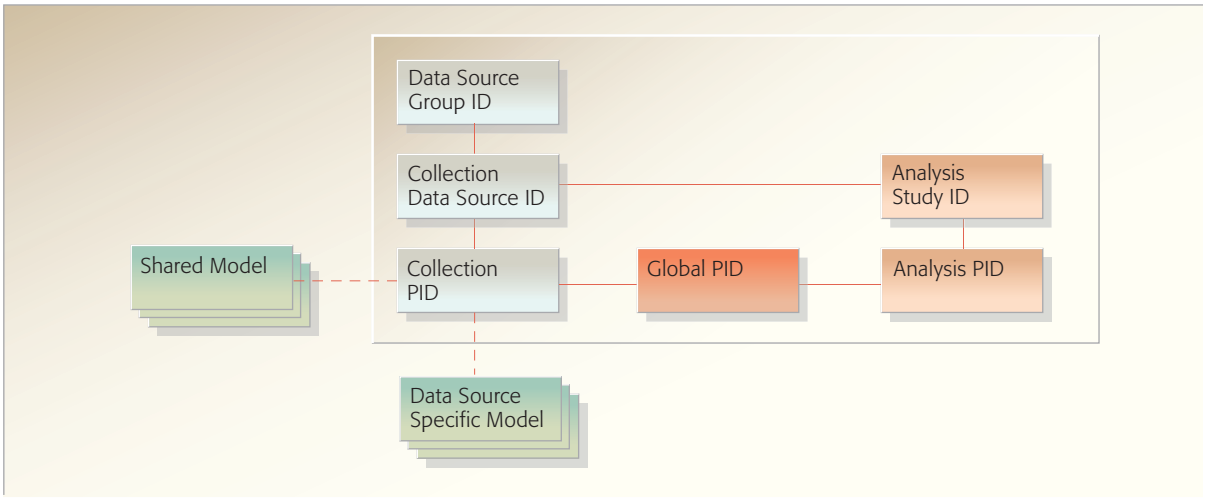


**Figure 3**
Linking of PIDs in the BIMS consolidation process

analysis PIDs, as this connection is not usually enforced by database foreign key constraints.

## Query

For querying the database, BIMS provides the easy-to-use query interface shown in ***Figure 4***. The query interface is Web-based and allows the user to construct queries and submit them to the database engine. It does not require SQL (Structured Query Language) programming skills. Variables can be assigned more descriptive names, which makes for a friendlier user interface. For example, *numpacks-year* is shown as *number of cigarette packages smoked per year*. The system allows for defining variables as a function of other variables by performing functions on variable values and by grouping similar variables into logical categories (e.g., new variable *age* is defined based on variable *birth date*). The process can be repeated, new variables can be defined, and thus various customized views of the data can be created.

Figure 4 shows a screen capture of the BIMS query interface. It shows a query that is built by navigating the category tree ("Select Condition Field" on the left) and the emerging plain text question that results ("Condition Summary" on the right).

## Security

Our decision to allow access to BIMS from anywhere on the Internet had a profound impact on the system design. It required an emphasis on security throughout the development process. A number of measures have been taken to ensure the integrity of the data, and the foremost of these requires that BIMS be protected by its own network zone.

***Figure 5*** shows the BIMS network architecture. The BIMS Controlled Zone (BCZ) is a network zone with its own Internet Protocol (IP) addresses and hardened security, in which most BIMS servers are located. In particular the database server, which contains all the consolidated data, is located in the BCZ. Users log in from the Internet (Unrestricted Zone) by way of an HTTP (Hypertext Transfer Protocol) server located in the External Restricted Zone. The action is marked (1) in Figure 5. This server acts as login broker to the Web portal that authenticates incoming requests and allows access to the portal server in BCZ (2). At the same time, access to the query interface application is granted (3). Federated data can be pulled into the central database by opening certain ports in the BCZ firewalls combined with VPN (virtual private network) tunneling (4).

Data import using federated database connections is only allowed when initiated from inside the zone, over specific ports, between specific IP addresses, and by using secure communication. If necessary the federated connections can be passed through VPN tunnels in order to further secure the direct

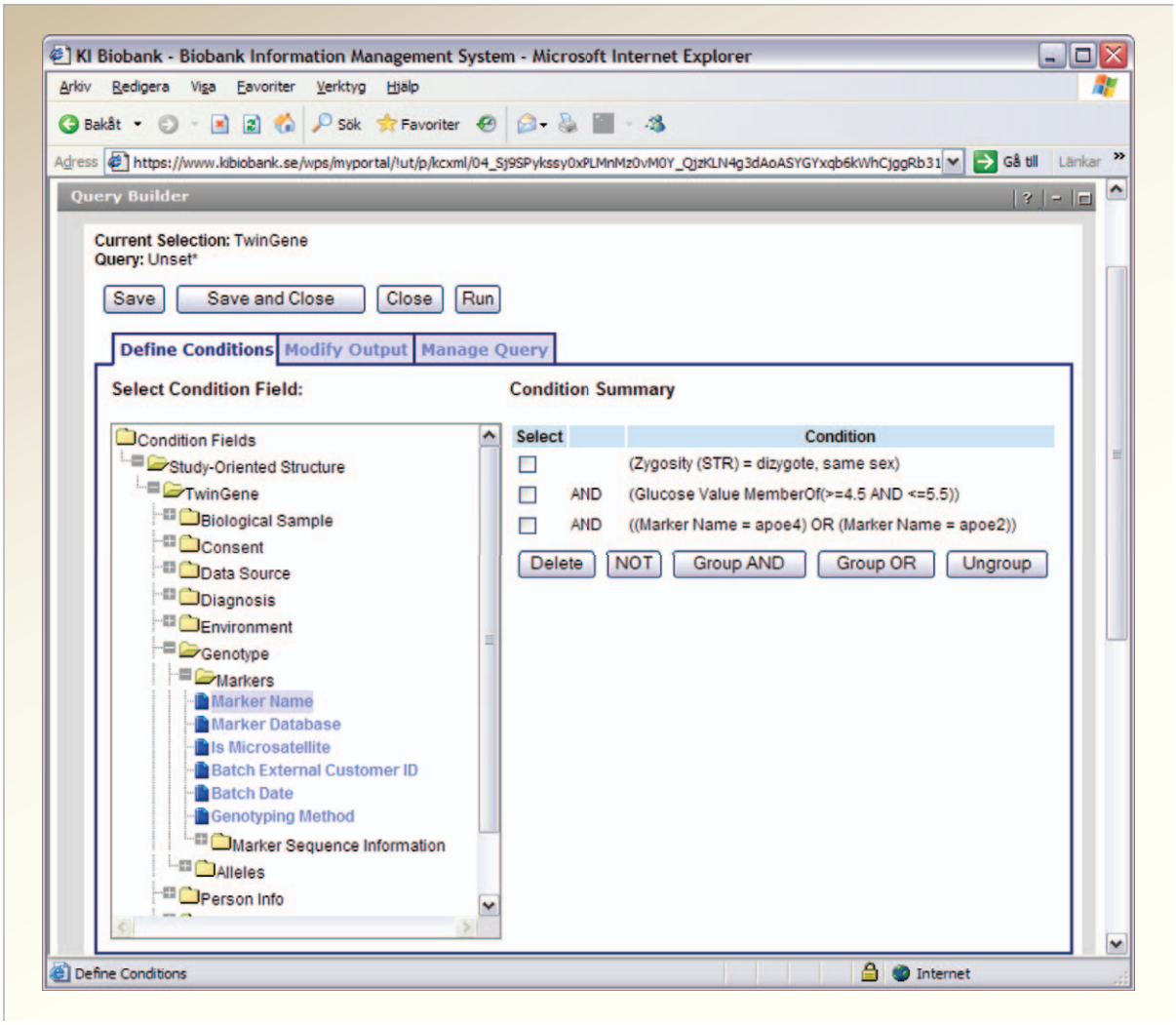


Figure 4
BIMS query interface (screen capture)

database access.[14] XML documents must be uploaded with secure connections and are protected by using public key 1024-bit encryption.

User access to BCZ can only be granted by having the user present a valid digital certificate to the HTTP server. Once logged in, the user is restricted to what services can be accessed in the portal.

The kinds of query that a user is allowed to perform (using the Web interface) are regulated on several levels. A user can have access to one or more analysis studies, and a user who has access to more than one study is allowed access to only one study at a time. Within each study, access to data is further constrained by the mechanism illustrated in *Figure 6*. Assume there are three data sources, A, B, and C, and a population that consists of three persons, X, Y, and Z. Figure 6 shows a scenario in which the user has access to a study with sources A and B and to a subset of the population that consists of the single person X. Thus, the access to data is limited to two data items within the rectangle highlighted in Figure 6. In addition to this access control mechanism, all data stored are deidentified, further contributing to improved security.

Whereas most data are only available to qualified users, some limited demographic data from the data source is publicly available (e.g., the number of samples in a biobank).

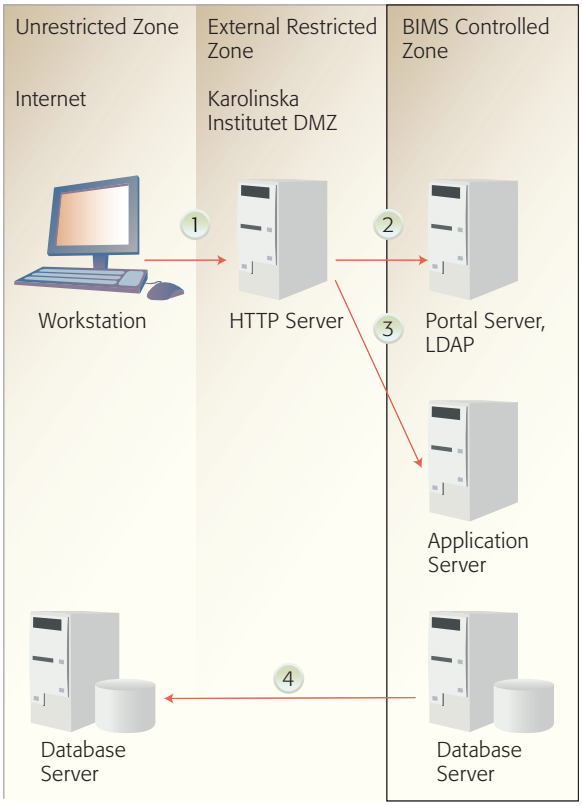


**Figure 5**
BIMS network architecture

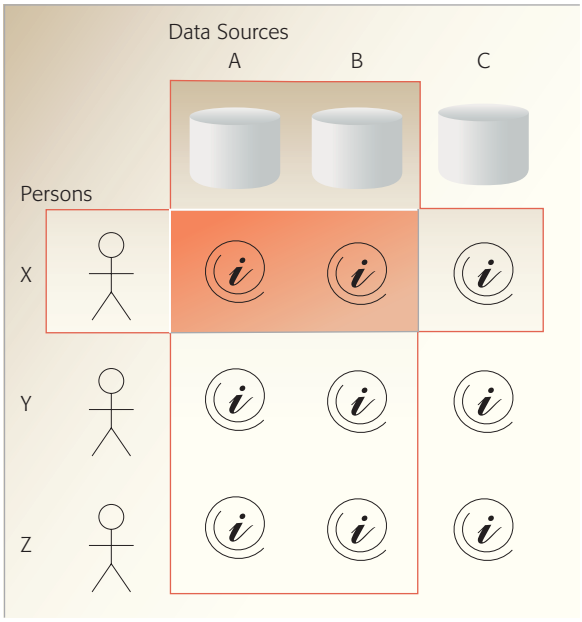


**Figure 6**
BIMS control of data access (conceptual view)

## Implementation

The realization of BIMS, a joint collaboration between Karolinska Institute Biobank and IBM, represents approximately a 5-man-year effort to date.

The guiding principle in the development of BIMS has been to build a solution by making use of existing off-the-shelf components wherever possible. The central database warehouse and the federated database connections are built using IBM WebSphere* Information Integrator.[15,16] The abstraction layer and the query interface use IBM Data Discovery and Query Builder,[17,18] and the entire Web presentation layer is contained within IBM WebSphere Portal, with authorization information stored in a Microsoft Active Directory** solution.

The extraction processing of XML documents, deidentification, and consolidation are all performed by custom-developed Java** components. All XML documents are converted to Java objects using JAXB (Java Architecture for XML Binding).[19] Dynamic class loading is used in order to allow for the inclusion of new deidentifier and consolidator modules at runtime.

At the time of writing, BIMS integrates data from seven collection data sources, most notably the Swedish Twin Registry (phenotype data on 170,000 twins) and our biobank's LIMS (data on over 100,000 samples), but also genotype and phenotype information originating in spreadsheet formats. Data access is provided by five different analysis studies dealing with epidemiological twin research and with genetics in multiple sclerosis, and our users consist of researchers in these fields. A first release of BIMS has been running for over a year, and the current version detailed in this text has been in production for a few months.

## DISCUSSION

Although information management for the medical research area is very active and expanding in terms of function and available data sources, it is clear that some of its challenges will not be met for quite some time. Nevertheless, it is possible to evaluate the extent to which BIMS addresses the challenges listed in the next subsection "Addressing data integration challenges."

### Addressing data integration challenges

BIMS represents not only a technical solution but also of a set of internal policies that deal with

aspects of data integration and control of data access. Thus, most of the design issues need to be dealt with from a technical as well as a policy perspective. The typical case involves the issue of quality and comparability of data. Policies specify that before any new data source is added to BIMS, there must be a thorough inventory of the data, which includes the need to document the meaning of each variable.

The ability to compare different variables lies within the extraction, consolidation, and abstraction components. For example, genotypes need to be adapted to fit our XML schema, but the actual adaptation is facilitated by providing Microsoft Excel** scripts (extraction). Further, presenting each research group with a familiar notation of genotypes enables them to make meaningful comparisons (abstraction). As for the anticipated difficulty of dealing with the quality of data, we found that much of the problem was solved by the process of incorporating the data into BIMS. Just having to go through the data (first by taking an inventory and documenting the data; then by validation, model consolidation, and querying the data) helped enormously in finding erroneous or missing data.

Coping with differing data models is similar to the way we handle comparability of data. A careful documentation process—responsibilities lying firmly with the data source owner, but aided by the BIMS administration—paves the way for easy mapping of data in the consolidation phase. Handling differing data formats was one of the easiest challenges to accommodate. It is a purely technical issue, which we addressed by using IBM WebSphere Information Integrator and various XML generation tools. In contrast, dealing with the challenge of differing ontologies is hard (e.g., when comparing medical diagnoses), and the issue is not addressed to any significant extent in BIMS except that the generic framework for storing ontologies enables multiple ontology standards to coexist. Using abstraction, it is then possible to map different terms to the same entity, but this does not solve the larger problem of interpreting the terms and deciding if they are alike or not. Rather, work must be done in the realm of meta-taxonomies like UMLS[20] to improve the chances of meaningful comparisons.

The deidentification issue is effectively dealt with by the BIMS deidentification process, which minimizes the level of sensitive information stored within the system. This results in increased trust on the part of our partners and collaborators and establishes a means for the safe exchange of information with other information systems.

The amount of effort invested in the issue of security and access control in BIMS has paid off. A stable implementation of the authorization module has resulted, which is based on the concept of linking collection PIDs and analysis PIDs through global PIDs.

One of the concerns at the start of the project was the issue of data ownership and the extent to which data owners would be willing to entrust their data to us. If many researchers did not agree to store their data in BIMS, the system would be of limited use. As it turned out, this was not a serious problem. In fact, most researchers, once they understood the benefits of having their data better organized and safely stored, had no qualms about participating in the enterprise. In large measure this was due to the investment we made in tight security mechanisms, which provided the needed level of trust. As a result of the high level of security, the deidentification process, and the firm internal policies on consent procedures (consent has to be recorded within BIMS as a condition for importing the data), the legal and ethical concerns have been manageable so far.

## Scope

BIMS is not easily compared to other information management solutions for epidemiological data because it tries to tackle most of the data integration challenges mentioned and because it relies on many (sometimes competing) research groups to provide content. It goes further than the data-warehouse-type solutions, such as MONICA and EPIC, by providing support for a wider range of services.

Closer in scope is UK Biobank, a community-based, prospective cohort study involving some 500,000 volunteers. Having a much more streamlined data model, its information management system had less difficult challenges to overcome. UK Biobank also benefits from governmental funding and support for the computerization of United Kingdom health records.[21] In terms of integrating vast amounts of information and providing clinical researchers the possibility to query that information, Mayo Clinic

has employed a solution similar to BIMS (in fact, one that goes further in some areas, most notably built-in result analysis and text mining of medical records). BIMS strengths come from having a more general approach to integrating data, using both federation and pushed XML content as information carriers, focusing on precise access control, and catering to the needs of individual research groups as they fulfill the roles of both information producers and consumers. Because of its modularized architecture, BIMS can scale up to a system able to handle research studies that involve the entire population of Sweden.

We should point out that even though the BIMS project attempts to provide many functions for biobanking research, it is not meant to replace the LIMS. On the contrary, a robust quality-controlled LIMS complements BIMS and serves as an excellent data source.

What kinds of new research questions does BIMS support? Whereas integrating epidemiological data from multiple sources is not new, BIMS allows a researcher without special computer skills to make use of data originating in heterogeneous sources by simply selecting the data for study. A clear opportunity to conduct more data-driven research thus emerges. Examples of the types of queries for which researchers can use BIMS include:

- Select all single nucleotide polymorphism markers between chromosome position X and chromosome position Y for all participants that have been diagnosed with A but not with B.

- Identify all cell samples with at least X amount of synovial fluid cells from patients undergoing anti-tumor necrosis factor therapy. What are their cyclic citrullinated peptide and rheumatoid factor status?

- Select all twins wherein one has disease X and the other is unaffected and both have consented to participate in the clinical trial. Produce a list of their eating habits and correlate with their glucose levels.

The types of questions that can be posed and the usefulness of the results should increase as more data are fed into the system.

In summary, the current BIMS solution acts as a powerful facilitator in modern biobank research. Population-based studies and high-throughput molecular technology constitute only two, albeit important, components of the research infrastructure that will need to be addressed as relevant future research questions concerning the health of populations. BIMS addresses the data management needs of these components. Other components include rapid and accurate collection over time of exposure information and intermediate morbidity endpoints from large population segments. Although many challenges remain, BIMS provides a good start as a solution for connecting large population-based studies in the future.

## Future steps

Beyond addressing the challenges of managing epidemiological information and integrating multiple data sources, future work includes the ability to handle data in a standardized format, a prerequisite for exchanging information with other biobanks (which is needed, for example, when there are not enough samples in a single biobank to study a certain disease). Other areas of high priority include high-performance data structures to handle large-scale genotyping efforts and mechanisms that can navigate and query family structures as a data source.

Additionally, we plan to investigate the need for post-query built-in analysis, to offer researchers as a service a set of simple LIMS functions that might avoid the expense of purchasing a LIMS, to provide processing of Web questionnaire input, and to provide a new general data model for phenotypic data.

It is our belief that by using systems such as BIMS, researchers will not only be more productive, but through collaboration with other researchers, the quality of their work will improve. Connecting BIMS with other biobank centers around the world is the first step toward worldwide biobanking.

BIMS is a challenging and ambitious undertaking that tackles difficult issues in biobanking research, focusing on information management for medical research. To date, no system similar in scope to BIMS has been described in the scientific literature. By combining genotype and phenotype data with the actual samples in a standardized fashion, BIMS

enables optimal use of collected samples. The project aims for breakthroughs in ensuring the value, quality, and usefulness of large quantities of medical data. A significant portion of this effort relates to defining, structuring, and standardizing the information that has been gathered in clinical and epidemiological studies. If successful, the project might significantly improve the use of collected data in a cross-disciplinary fashion and in contexts other than those anticipated.

## ACKNOWLEDGMENTS

Many people have helped in building the BIMS solution. In particular, we would like to express our gratitude to Richard Dettinger, lead developer of Data Discovery and Query Builder, and Douglas Del Prete, DB2* database expert. In addition, Benedikt Furrer, Kjell Larsson, Maria Andér, and Michael Hehenberger made many valuable contributions to BIMS.



## CITED REFERENCES

1. Z. Zimmerman, M. Swensson, B. Reeve, F. Betsou, M. Ferguson, B. Jallal, and J-E. Litton, "Biobanks: Accelerating Molecular Medicine—Challenges Facing the Global Biobanking Community," Biobank Summit II, Tarrytown, NY, November 2004, IDC Special Study, IDC#4296, International Data Corporation, pp. 1–31 (2004).
2. L. J. Palmer, "The New Epidemiology: Putting the Pieces Together in Complex Disease Aetiology," *International Journal of Epidemiology* **33**, No. 5, 925–928 (2004).
3. UK Biobank, http://www.ukbiobank.ac.uk.
4. R. Rhodes, "A Healthy Approach to Data," *IBM Systems Magazine*, August 2002, http://ibmsystemsmag.com/.
5. *MONICA, Monograph and Multimedia Sourcebook*, H. Tunstall-Pedoe, Editor, World Health Organization, Geneva (2003).
6. E. Riboli, "Nutrition and Cancer: Background and Rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC)," *Annals of Oncology* **3**, 783–791 (1992).
7. J-E. Litton, J. Muilu, A. Björklund, A. Leinonen, and N. L. Pedersen, "Data Modeling and Communication in GenomEUtwin," *Twin Research* **6**, No. 5, 383–390 (2003).
8. B. M. Knoppers and M. Saginur, "The Babel of Genetic Data Terminology," *Nature Biotechnology* **23**, 925–927 (2005).
9. L. M. Haas, P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice, and W. C. Swope, "DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources," *IBM Systems Journal* **40**, No. 2, 489–511 (2001).
10. L. M. Haas, E. T. Lin, and M. A. Roth, "Data Integration through Database Federation," *IBM Systems Journal* **41**, No. 4, 578–596 (2002).
11. Biobanks in Medical Care Act, Ministry of Health and Social Affairs, Sweden, http://www.sweden.gov.se/content/1/c6/02/31/26/f69e36fd.pdf.
12. *Secure Hash Standard (SHS)*, FIPS Publication 180-2, National Institute of Standards and Technology (August 2002), http://csrc.nist.gov/publications/fips/fips180-2/fips180-2.pdf.
13. Health Insurance Portability and Accountability Act Of 1996, Public Law 104-191, United States Department of Health and Human Services, http://www.hhs.gov/ocr/hipaa/.
14. J. Johansen and J-E. Litton, "Security Policies for TwinNET. Draft 2," 2005, a GenomEUtwin document, http://www.meb.ki.se/twinreg/genomeutwin/documents/Policy%20document_TwinNET.pdf.
15. C. M. Saracco, M. A. Roth, and D. C. Wolfson, "Enabling Distributed Enterprise Integration with WebSphere and DB2 Information Integrator," *IBM Systems Journal* **43**, No. 2, 255–269 (2004).
16. WebSphere Information Integrator Content Edition, IBM Corporation, http://www-306.ibm.com/software/data/integration/db2ii/editions_content.html.
17. J. W. Tenner, "Hungry for Data? Serve Yourself," *IBM Systems Magazine* (February 2004), http://ibmsystemsmag.com.
18. Data Discovery and Query Builder, DB2 for i5/OS, IBM Corporation, http://www-03.ibm.com/servers/eserver/iseries/db2/ddqb.html.
19. JAXB 2.0 Project, java.net, https://jaxb.dev.java.net/.
20. B. L. Humphreys and D. A. Lindberg, "The UMLS Project: Making the Conceptual Connection between Users and the Information They Need," *Bulletin of the Medical Library Association* **81**, No. 2, 170–177 (1993).
21. W. Ollier, T. Sprosen, and T. Peakman, "UK Biobank: From Concept to Reality," *Pharmacogenomics* **6**, No. 6, 639–646 (2005).

*Accepted for publication July 6, 2006.*
*Published online December 27, 2006.*

*George Ölund*
*KI Biobank, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, SE-171 77 Stockholm, Sweden (george.olund@ki.se).* Mr. Ölund is a senior software engineer and lead developer of the BIMS solution. He received an M.S. degree in biomedicine from Karolinska Institutet in 1999 and an M.S. degree in computer science from the Royal Institute of Technology in 2005. He has many years of software development experience as well as background in medical research.

*Pontus Lindqvist*
*KI Biobank, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, SE-171 77 Stockholm, Sweden (pontus.lindqvist@ki.se).* Mr. Lindqvist is IT Manager at Karolinska Institutet Biobank and the project manager for the BIMS project. He has a Master's degree in medical science from Karolinska Institutet and a Master's degree in computer science from the Royal Institute of Technology in Stockholm. With approximately 10 years of

experience in biomedical research and software engineering, he currently leads the IT team at Karolinska Institutet Biobank. He is also involved in the nationwide integration of biobanks in Sweden and in many other international collaborative projects.

*Jan-Eric Litton*
*KI Biobank, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, SE-171 77 Stockholm, Sweden (jan-eric.litton@ki.se).* Dr. Litton, a Professor of Biomedical Computing Technology since 2002, is Director of Informatics, Karolinska Institutet Biobank, and head of IT and Computing, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. He received a Ph.D. degree in medical science in 1983 from the Karolinska Institutet. He leads an infrastructure group in the EU Coordination Action, whose purpose is to make best use of population-based biobanks. Dr. Litton is also a member of the steering group in the $P^3G$ project and co-directory and responsible for the Swedish LifeGene initiative, a prospective cohort-based biobank. Dr. Litton also heads the development of e-epidemiology by using the Internet, cellular telephones, digital paper, and digital TV for collecting epidemiology data. ■