



# datamatique

## Rapid, Customized Database Setup

- **Rapid, Customized Database Design and Setup**
- **Optimized for Biological Data but Essentially Data Neutral**
- **Custom GUI for Individual Requirements**
- **Automated Data Capture from Scientific Equipment**
- **Optional Manual Data Entry**
- **Suitable for databases of all sizes (few MB to hundreds of GB)**
- **Robust XML or Relational schema**
- **Integration of Text, Image and Signal (graphs, etc) Data**

# Datamatique:

## Custom Databases Made Easy

Genvea Biosciences offers Datamatique, a platform for rapid design and setup of customized laboratory databases. Datamatique is essentially data neutral and has the intrinsic ability to quickly enable the end user to set up a database of his own with minimal effort. The database set up is optimized for capturing various formats of biological datasets including images, signal information and text.

Datamatique can be based on XML or relational schema.

Relational databases are suitable for stand-alone reliable database setups, where data structure is expected to remain unchanged.

XML designs are favored for compatibility with publicly available high-throughput research data. Moreover, the XML backbone makes the locally set up database ready for integration with public domain databases using our proprietary data integration platform, XMELD.

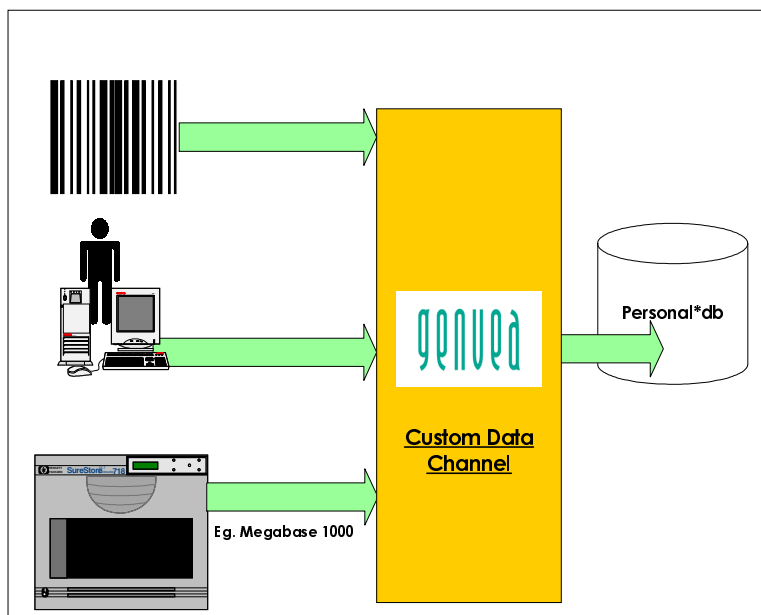
Datamatique is highly scalable from a few Megabytes (MB) to several Gigabytes (GB).

### Introduction

The amount of data generated and analyzed worldwide, especially in computationally-intensive research fields is growing rapidly, as is the need for analyzing and managing it. This data, which lies at the foundation of hypothesis development and experiment validation, can characterize from being small and simple to being voluminous, highly evolving, highly heterogeneous and semistructured. Increasingly holistic and cross-disciplinary research also ensures increasing heterogeneity in data and formats. Moreover, upcoming computational fields suffer from a lack of common standards.

In a typical laboratory set up a lot of detailed planning goes into the experimental set up, protocol design and methodical execution of a set of experiments that generate heterogeneous raw data sets which are vital. These raw data sets are stored in different folders of different hard disks of computers. More often than not different research groups in the same research project may not be even aware of the types of data generated by their own colleagues in a different research group!! The data is scattered all over the laboratory or institute. Once there is a demand for publication or report submission, there is a scramble for data access and data analysis. These tasks are done manually which is highly error prone and subject to individual capacities.

A more systematic and scientific approach to capture, archive, retrieve and manage precious raw data is the need of the hour today. High throughput instrumentation and the dawn of genomics era in life sciences have transformed the way in which the raw data is captured and analyzed. The raw data generation has become a highly cost intensive process. The raw data generated has to be securely managed in the current IPR era. Scientists find it very difficult and cumbersome to engage in data management related issues that interfere with their main scientific research.



In the field of Life Sciences, for instance, it has been claimed<sup>1</sup> that a biologist spends more than half of his time on tasks related to the database set up, integration of data from incompatible databases and software programs.

Moreover, a common requirement in a lot of research laboratories is that the raw data generated from different high throughput research instruments need to be systematically captured

and managed. There arises the need for a standardized but flexible framework for creating a database from different types of data sets that is ready to be integrated with public domain databases.

<sup>1</sup> A. C. Siepel, A. N. Tolopko, A. D. Farmer et al. An Integration Platform for Heterogeneous Bioinformatics Software Components. IBM Systems Journal, 40(2), pp. 570-591, 2001

## Datamatique: A Functional Overview

### User friendly and simple:

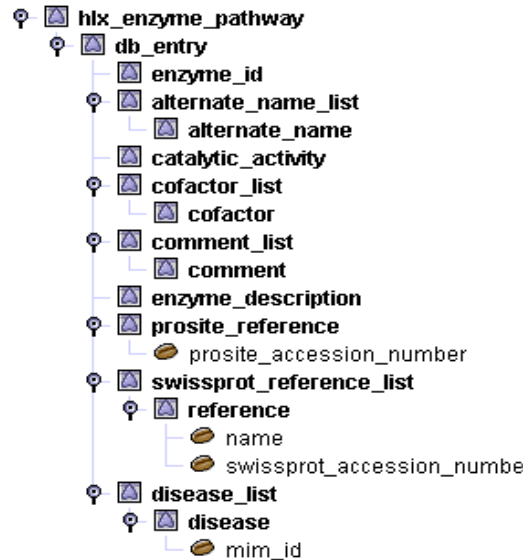
Datamatique is based on the well established client-server architecture. It is a highly user friendly tool with a client side web based GUI (Graphical User Interface) that is customized to individual customer requirements. The client software has the intrinsic ability to take in data in disparate formats such as images (gels, chromatograms, GIS images, microarray images, JPEG images etc.) and built in querying capabilities. Hence, the end user gets an integrated display whenever he queries his database without the frantic scrambling to locate different files in different locations.

### Security and Authentication:

The client can be connected to the server, which hosts an XML based database either through the local LAN or through a dedicated LAN with dedicated network switch between the server and designated clients. For the security sensitive installations that require remote access to the server, SSL (Security Socket Layer) based connection with one or more levels of authentication (password and additional authentication techniques such as encrypted digital certificates) will be provided and customized as per the requirements of the customer.

### Data Integration Compatibility:

The unique selling point of Datamatique is the backbone of the database, which is a highly modified and optimized XML schema that is inherently scalable from a few GB to several TB. It is also highly flexible with a tree like architecture where new attributes and elements can be added and/or deleted without sacrificing the database integrity and performance. Because of its tree like architecture and the added optimizations, querying the database is much faster than any conventional relational databases. Moreover, the database has all the built in enterprise database features such as automatic backup (optional tape backup), disaster recovery, fail over etc.



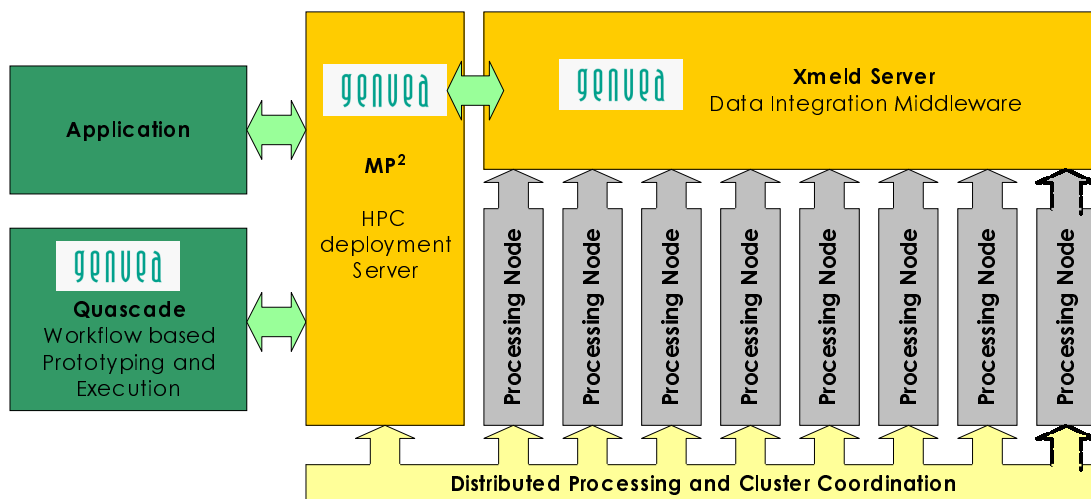
### Key Advantages

Some of the key advantages of using Datamatique are as follows:

- (1) The graphical interface on the client makes constructing an integrated database a breeze. Different members of the same research group or members from different groups can simultaneously put in the data and the database gets automatically populated in real time and is ready for querying by the PI (Principal Investigator)
- (2) The data input feature of the client has the automated data capture from a designated site within the network also, besides the manual entry by individual researchers. Such a feature is of immense importance in laboratories where critical raw data need to be captured and databased for future use.
- (3) Once the database is automatically populated it can become a part of the local warehouse containing several public domain databases of customer's choice. Since all the databases are in a transparent XML format, they can be integrated with the local database for highly efficient data mining and knowledge base creation.

## A Brief Overview of Genvea Data/ Application Middleware

Genvea provides the most compelling data-integration technology devised for data-driven drug discovery, clinical trials analyses and knowledge management for research in the life sciences. Working in conjunction with the MP<sup>2</sup> facility for high-performance computing and application-prototyping tools, it provides an effective foundation for rapid development of scientific applications in the life sciences.



The typical application written with this middleware therefore executes as a multiple tier application, with parts of it executing on the client computer, and parts executing on several server-side, distributed computers.

Our technology allows customers to integrate disparate and heterogeneous forms of data to provide the basis for a holistic approach to life sciences research and the systemic foundations for business intelligence. The technology relies on its ability to put together physically disparate sources of data and then allowing end-users to conduct single-point queries across this vast pool of data, as though it were a single resource.

This solution provides the basis for integrating any form of complex, unstructured research data. Moreover, the solution assumes that such data is largely heterogeneous, physically disparate and available off different platforms (flat files, ftp sites, legacy database systems etc.).

Besides integrating existing enterprise data, this solution also allows the end users to create and maintain their own databases on an ongoing basis.

### A Bird's Eye View to the MP<sup>2</sup> Environment

We believe that the simplicity of the development and deployment environment makes for an important consideration in the successful adoption of new programmable systems. The MP<sup>2</sup> facility operates on a clustered computing environment.

Applications operate from the client's computer by communicating with the cluster through a single computer that hosts a J2EE server. This coordinating server then identifies one or more "processing nodes" which are computers running a small footprint daemon, to perform the task of executing the server-side functionality of the application.

The use of a single coordinating server is beneficial in several ways. The cluster can, at the point of initiation of the application, decide the number of computers that need to be assigned for that application, depending on availability and necessity.

Moreover, the coordinating server also maintains directories of available resources.

The MP<sup>2</sup> Facility has been primarily written using Java, which is increasingly popular within the scientific community. Where performance is a consideration, applications can be written in C or C++ and interfaced with the overall system using the Java Native Interface (JNI).

**Genvea Biosciences**  
53, Craig Road  
#04-01  
Singapore 089691

**Chemoinformatics**  
20, Jalan SS2/66  
47300 Petaling Jaya  
Selangor Darul Ehsan

Phone : +65.62224569  
Email: [info@genvea.com](mailto:info@genvea.com)

Phone : +60.3.78760022  
Email: [info@chemoinformatics.com](mailto:info@chemoinformatics.com)

©2003 Genvea Biosciences. All rights reserved. Specifications subject to change without notice. Genvea Biosciences and the Genvea Biosciences logo, are registered trademarks of Genvea Biosciences worldwide. All other trademarks mentioned herein are the property of their respective owners.