Services for a Genomics Open Distributed Environment

(*Position paper submitted at XEWA'2000*)

N. Boudjlida, M-D. Devignes * and M. Smaïl-Tabbone
LORIA, UHP Nancy 1 (F) - *E-mail: {nacer or devignes or malika}@loria.fr*

1. Introduction

Considering the variety and the heterogeneity of genomics databases or data sources (DS), we figure an "almost-ideal" situation where the following facilities are provided: (i) one can easily locate existing DSs and know about their content and the services they offer, (ii) one can easily identify and select among existing DSs those that can contribute to satisfy one's needs; (iii) one can easily navigate, extract and combine information coming from selected DSs.

Steps toward this situation have been experienced in our previous work on a user-oriented system for collecting and structuring genome information. Xmap, a prototype of this system, has been developed and it is currently in use by biologists. The system is based on a model for information retrieval from heterogeneous databases that includes assistance to navigation, estimation of data confidence degree and structuring of the retrieved data in order to ease their subsequent exploitation. This work is applied to the retrieval of mapping data on human genome, with the ultimate goal to correlate the position of given novel genes with that of orphan diseases. Indeed, since the development of large scale sequencing of human gene transcripts, a gene can be represented by a cluster of partial nucleotide sequences. Markers have been derived from these sequences in order to carry-on mapping experiments whose results are stored in various databases accessible through the Web. The scenario considered in this study consists in querying the appropriate databases to retrieve (i) the list of sequences associated to a given gene, (ii) the list of markers designed from these sequences, (iii) the mapping data available for these markers, (iv) mapping data for genetic markers in the vicinity, (v) data on genetic diseases mapped in the same region of the genome.

Every step has been modelled according to a common generic information retrieval process that provides various facilities such as selection of the database to be queried, query formulation and re-formulation, extraction of useful items from the query results and structuring of the retrieved information. This scheme includes a tight interaction between the system and the user at the step of database selection (e.g.: assistance for unexperienced users), automatic query formulation whenever possible, query results sorting (e.g.: according to a reliability function that uses an estimated quality factor of the database), and various data integration into a common structure that is built all along the retrieval session. The integrated data is called a *session document*.

The session document is structured using the XML (eXtensible Markup Language) standard according to a defined Document Type Definition (DTD) that takes into account existing initiatives in the domain. The various elements of the document are therefore accessible for establishing a session synthesis, for visualizing mapping information, and for analyzing the sessions history.

We feel that the generalization of this approach to reach the "ideal situation" may consist in providing an open infrastructure where DSs can be described, compared and accessed. The infrastructure is viewed as a trading architecture model and it consists in at least three sets of services: (i) a set dedicated to *DS providers*, i.e. institutions and scientists that make their DS publicly available, (ii) a set concerned with *DS availability management*, and (iii) a set dedicated to *DS clients*, i.e. mainly scientists that need to query (a subset of) the DSs.

Let us first give a rapid (and obviously non exhaustive) view of every set of services and then discuss

* LORIA and Genexpress, Villejuif (F)

possible technical orientations and candidate enabling technology.

2. Infrastructure Services

(i) Services for DS Providers : This set of services should enable a DS provider to *publish* the DS availability and capabilities, i.e. to make the DS known to the scientific community. DS publication encompasses information about the DS location, its structure (schema), its scope (domain of the contained data), and the possible additional services that the DS offers. Let us call *DS Identity Card (DS-IdC)* the information that describes a published DS, i.e. DS location, schemata, scope, system support, communication protocols, etc. Services to update and to delete DS-IdCs should be also provided.

(ii) Services for DS availability management : First, one should notice that DS availability management is not DS management: DS management remains under the DS providers responsibility (*autonomy principle*). DS availability management encompasses the set of services to store, query and maintain the information that concern the publicly available DSs, i.e. mainly DS identity cards management. These services are available for DS providers as well as for DS clients. Besides the DS-IdC management services, DS availability management (*transparently* or not) provides facilities to classify data sources. The classification depends on the registered DS contents and may rely on some defined and commonly agreed classification criteria or on some approved ontologies.

(iii) Services for DS Clients : DS clients mainly use the DS availability manager query and navigation facilities. Queries are applied at two levels. The first level concerns the DS classification and its main purpose is to help a client in locating the DSs that can potentially contribute to satisfy its needs, i.e. starting from a client's query this set of services helps in identifying a kind of *scope of the query*. The second level helps a client in navigating through the identified DSs, i.e. inside the identified scope. Navigation is based on information retrieval concepts and technology where answers are more than *yes/no* answers. Furthermore, facilities should be provided to store and re-use query scopes and navigation paths.

3. Technical Considerations and Enabling Technology

Our claim in these considerations is to rely, as far as possible, on existing standards to settle the environment. It is clear that the services for DS availability management are central in our proposal. They first assume the availability of a *common DS description (meta-)model* for DS-IdC registry: OMG and ODMG object models (they are compatible) are good candidates for DS schema description while XML is a good candidate for DS content description, assuming a common agreement on a minimal set of descriptors (XML tags) grouped into a DTD or an XML-Schema. DS providers can obviously extend this minimal set with their own descriptors. DS that offer services, in addition to data, can describe the interfaces to their services using CORBA IDL (syntax) and XML (semantics).

DS-IdC physical storage can be done using classical database technology (relational or OO data server). *DS classification* can be stored as a graph database to ease navigation and search into the classification. Moreover, *Ds-IdC and DS classification databases* can be managed as remotely accessed centralized databases but preferably, as *replicated databases with a primary copy*, i.e. updates are first performed on a distinguished database copy (the primary one) before being propagated to the other copies.

Since DS query language transparency is difficult to achieve, we suggest to base the *query facilities* on existing query languages standards (like SQL or OQL) or on emerging ones (XML-QL). *Query scopes and navigation paths* can be stored locally on the clients sites as they can be stored on the DS availability manager site.

We feel that the production of the various facilities to publish DS identity cards is a short-term perspective, the development of the DS availability services can be a mid-term objective while the development of sophisticated query and navigation services is a longer-term perspective.