The GENOMA-KB platform: Queries over integrated of linguistic resources

Margarita Hospedales, Manel Rodríguez

Serveis i Plataformes Orientades al Coneixement (SPOC)
Sant Antoni, 36-38, 3er 3ª, 08001 – Barcelona. Spain
mhospedales@spoc.com, mrodriguez@spoc.com

Abstract

The human genome knowledge base (GENOMA-KB) platform integrates a variety of linguistic resources in a unified environment, independent from the origin and nature of the data it retrieves, thereby offering the visitor one sole access point for different heterogeneous sources of information.

GENOMA-KB platform has been developed by Serveis i Plataformes Orientades al Coneixement (SPOC) in order to carry out an initiative presented by the Institute for Applied Linguistics (IULA).

One of the requirements of the solution was to minimize its impact on the normal working procedures of the linguist team. For this reason the data input processes have been left intact.

GENOMA-KB's user interface, developed under the premise of ease of navigation and aesthetic quality, offers the visitor a unified search interface and transparent navigation between different information sources, achieved in an agile and simple fashion.

1. Introduction

The GENOMA-KB platform integrates a variety of linguistic resources in a unified environment independent of the origin and nature of the data thereby offering the visitor one sole access point for different heterogeneous sources of information.

1.1. To whom it is addressed

Human Genome Knowledge Base (GENOMA-KB) is a helpful instrument for different kinds of users such as: translators, terminologists and lexicographers; information science experts; specialized writers and journalists; researchers and scholars; linguists.

GENOMA-KB's platform (denoted GENOMA), developed under the premise of ease of navigation, intuitive use, modernity and aesthetic quality, offers the visitor: A unified query interface, transparent navigation between different information sources, a scalable framework for future development and a high level of precision in the information retrieved, achieved in an agile and simple fashion.

2. Goal of the paper

The objective of this paper is to describe the architecture that supports GENOMA.

First follows a brief description of the structure and nature of the data sources that can be accessed by the portal, followed by a description of the architecture used in the applied solution, and finally a brief summary of the user interface.

3. A working environment for linguistic teams. The data sources.

The principal objective of the GENOMA-KB project, that has been conceived by the IULATerm group, was the construction of a biomedical knowledge base for the human genome.

Within GENOMA-KB linguistic teams work within different applications allowing them to consult and maintain linguistic data of various types.

These applications form a modular structure that, although they store related information, does not establish relations or communication protocols between modules.

A brief description of each module follows.

3.1.1. Corpus textual

A Corpus constructed from documents of various origins and related to different thematic areas, one of them being the human genome, which is used as a data source for GENOMA.

Both the maintenance and the consultation of the Corpus is carried out from the Corpus WorkBench®[1] which offers various tools that provide support to lexicographic and terminological work.

It also includes a public environment for making queries, bwanaNet. Developed by the IULA and accessible via Internet bwanaNet uses the Corpus WorkBench tool CQP (Corpus Query Processor) to query the Corpus itself

3.1.2. Ontological and Terminological module

The ontological and terminological module collects (lexicographic) terms and concepts related to the human genome. This module is formed by two databases with intimately related data: The Terminological database and the Ontological database.

Linguistic teams use the Ontoterm[2] to maintain these databases.

OntoTerm does not provide a interface for consulting data but allows the IULA users to export data to an HTML format.

[1] Corpus WorkBench, developed by the IMS of Stuttgart University allows the retreival of complete texts from an extensive resource of texts.

[2] OntoTerm is a Terminology Management System (TMS) based in ontology.

OntoTerm is installed under Microsoft Windows in workstations within the IULA's local network and accesses databases via LAN connections.

Terminological database (Genoterm)

A database formed by specialized knowledge units related to the human genome.

The principal characteristics of this database are its interrelation with the ontological database and the existence of predefined concepts associated to the terms.

Terms are interrelated by the concepts that are associated to them.

Ontological database (Ontology)

Contains the concepts that are used to relate terms and the relations between these concepts.

3.1.3. Documental and Factographic module (Gendofac)

Consisting in bibliographical data for documents that form part of the textual corpus, works referenced by the terminological database and other documents selected by specialists in the field.

In addition it contains data related to people institutions, companies, products and methodologies related to the human genome.

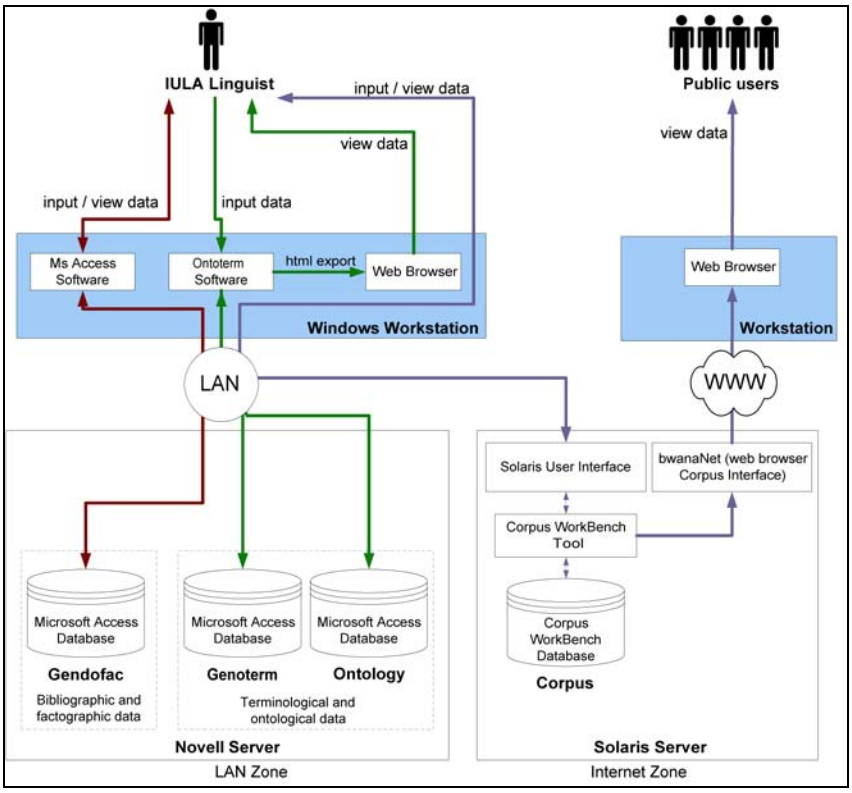

Figure 1: Before GENOMA, the only way to query the Corpus for public users was via BwanaNet. The maintenance and querying of other data sources (Bibliographic, Terminological, Ontological) was carried out from within the IULA's local network.

Module	Maintenance	Local network querying	Web querying	Data bases	Server
Corpus textual	Corpus WorkBench	Corpus WorkBench	BwanaNet	Corpus WorkBench (proprietary format)	Solaris
Ontological and Terminological module	Ontoterm	Exportation to HTML	No	Genoterm and Ontology (Ms.Access)	Novell
Documental and factographic module	Gendofac	Gendofac	No	Gendofac (Ms. Access)	Novell

Table 1: Summary of the initial situation

4. Architecture supported by the GENOMA web interface

The applied solution offers an integrated query environment, unifying the different data sources and is accessible via Internet.

The solution was developed in Java™ using J2EE™ architecture. This provides the additional advantage of being platform independent should the machinery of the IULA change in the future, for example from Sun/Solaris to Intel/Linux.

Due to the fact that the solution is being deployed in an Internet environment a Sun/Solaris server that the IULA already use was chosen to host GENOMA.

4.1. Application structure

The J2EE™ application, developed by SPOC, follows an n-tiered architecture. This architecture, adopted by the majority of fabricants is based in the separation of the application in tiers.

The presentation tier and interaction with the users is based in web browser technology. The components that integrate this tier are: a web server**,** a web container and the web browser used by the user.

The logical tier formed by a diversity of components based in Java, is responsible for the implementation of integration services for the different data sources that are accessed by the GENOMA. This tier contains the search procedures available to the users. Separating the presentation tier from the subjacent technical solutions.

The data or resource tier formed by a relational database and Java components that encapsulate access to non-relational systems (for example the Corpus Workbench), is responsible for obtaining data from the different data sources and providing a homogeneous interface to the logical tier. This homogeneity "hides" the complexity inherent in working with databases of differing nature. Taking into account the fact that the new database is only used at runtime for querying data (loading of new data is achieved through batch migration processes) the data structure has been redesigned in order to improve response time.

This structuring of the application into tiers is totally transparent for the user to whom it offers a unified interface, independent of the origin of data source being queried and the nature of the data requested

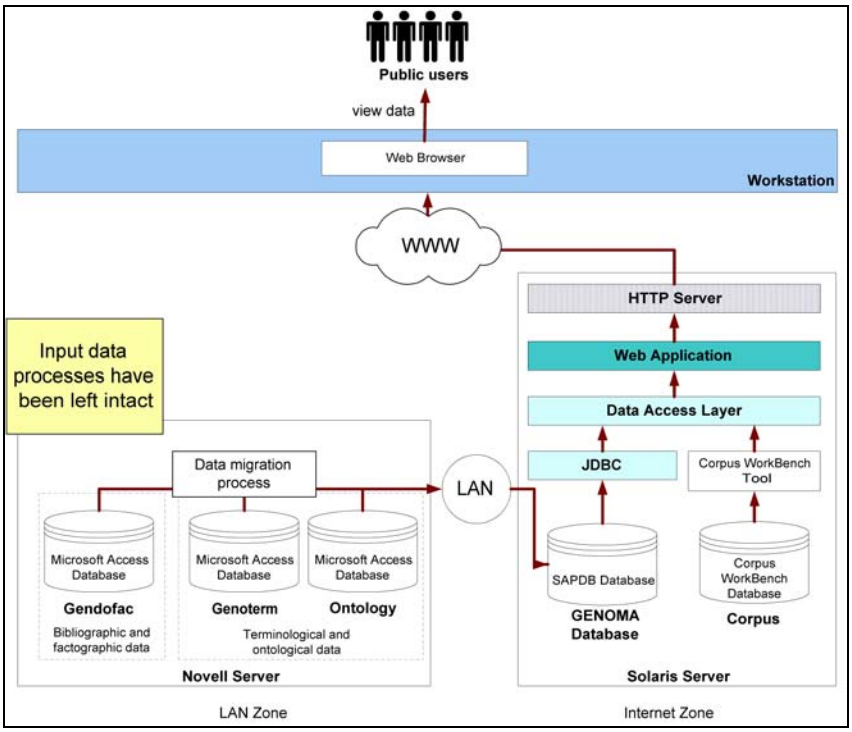

Figure 2: GENOMA integrates the consultation of different data sources in one interface, regardless of the origin of the data and without interfering in the normal working procedures of the linguists.

4.2. Software

The software used to implement the applied solution is platform independent and could easily be installed in both Intel – Windows and Intel – GNU/Linux platforms.

The majority of the tools used are freeware (with the exception of the Solaris operating system which already belonged to the IULA).

Functionality	Product
Web server	Apache 1.3.27
Servlet, JSP container	Tomcat 4.18
Java runtime environment	JDK 1.4
Database server	SAPDB 7.4

Table 1: Software used for implement the applied solution

5. GENOMA's interface

The user interface was developed under the premises of ease of navigation and intuitive use, offering the visitor:

- Transparent querying and navigation between different data sources
- Agile and simple interfaces for information enlargement and precision.

5.1. The basic search

The basic search is GENOMA's principal utility and allows users to find information within a data source for a given search condition[3].

Access to search in all four of data sources is provided in the home page.

5.1.1. Terminological Banc (TB)

Search for terms. Starting from a list of terms that are retrieved as result of a text search it provides the user with the option to:

- Visualize a term's terminological data (Definition, language, grammatical category, etc.)
- Consult a term's ontological relations
- Consult a term's variants and equivalents
- Find contexts for the term in the Corpus Textual.

Figure 3: Appearance of GENOMA's home page where the user has access to the different data sources.

Figure 4: Consultation of the terminological data of a term.

5.1.2. Corpus

The corpus offers two basic options: Context searches and frequency calculations.

The option "Contexts" allows a user to search for contexts for a word, sequence of words or truncated words.

Via the option "Frequency" the user is able to view the frequency with which a word, sequence of words or truncated words appear in the Corpus.

If the context found corresponds to a registered document in the Documental Banc the application offers the user a link to the bibliographical reference.

5.1.3. Documental Bank (DB)

Allows the user to find bibliographical registers that fulfil certain search conditions. The search options offered are: Search by Author, Search by Title, Search by Subject, Search by ISBN/ISSN and Search in All Fields.

5.1.4. Factographic Bank (FB)

Allows the user to find factographic registers (entities, and people related to the human genome) that fulfil certain search conditions. The search options offered are: Search by Institution, Search by Responsible, Search by Services and Search in All Fields.

References

M. Teresa Cabré, Carme Bach, Judit Feliu, Rosa Estopà, Gemma Martínez, Jorge Vivaldi. (2004). The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities.

[3] In future versions the option of undertaking complicated searches introducing different conditions and combining different data sources has been anticipated.