EVOLTREE ELAB - AN INFORMATION SYSTEM FOR FOREST GENETICS

F. EHRENMANN[1], S. GLAUBITZER[2], D. KOPECKY[2], J. SCHMIDT[2], S. FLUCH[2], E. MARIA SEHR[2], A. KREMER[1].

1 INRA, UMR1202 BIOGECO, Cestas F-33610, France; 2 AIT Austrian Institute of Technology GmbH, Health & Environment Dept., Konrad-Lorenz-Straße 24, 3430 Tulln, Austria.
* These authors contributed equally to the construction of the eLab and its components.

The EVOLTREE web portal acts as a platform for information and data storage, retrieval, exchange and communication. It was set up in the early years of the network (between 2007 and 2008) in order to fulfill one of EVOLTREE's objectives of maintaining and reinforcing electronic and physical resources, repositories and infrastructures.
It comprises what is known as the "electronic Lab" (eLab) which was designed as a centralised search engine for databases that are stored in different servers located in different institutions in Europe. The web portal can be accessed via the EVOLTREE website[1].

Background and objectives of the eLab

While some of the portal databases existed beforehand, most of the web applications corresponding to the databases were constructed in the first four years of the EVOLTREE network. These have been continuously updated and populated ever since by the member laboratories which host the databases and carry out this activity as part of their 'in-kind' contribution to the network.

The portal databases are connected through a standardised, HTTP transmittable interface (TAPIR - www.tdwg.org/activities/tapir/), so that queries can be made within the whole set of databases.

Given the number of tree species studied throughout Europe, it was decided to "virtually" subdivide the eLab into three major portals corresponding to the three major botanical forest tree families that are studied: the Quercus Portal (for species belonging to the Fagaceae family), the Pinus Portal (for the Pinaceae family), and the Populus Portal (for species belonging to Salicaceae). Depending on their field of interest, users can therefore enter the system and make queries via three channels:

- The individual database for queries targeting well-focused information.
- The eLab for an overall search across all the databases. Access via the eLab research engine is recommended if users do not know where - e.g., in which database - the information of interest is located.
- One of the family portals, for data corresponding to a particular species, or genera, of the Fagaceae, the Salicaceae, or the Pinaceae family. Databases concerning species not belonging to these families can be directly accessed via the eLab.

Photo: BioGeCo, INRA

The individual databases

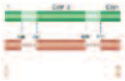
Passport, phenotypic, genetic and genomic data corresponding to different research units (genes, individuals, populations, species) were stored in separate databases, some of which existed before the launch of the network. At the beginning of EVOLTREE it was decided to keep the decentralized structure of the databases and to connect them via an interoperable interface in order to benefit from the already existing resources and the contributions of different partners.

Table 1 provides a summary of the content of the largest and most completed individual databases. All databases can be accessed individually and queries can be made internally without using the overall research engine of the eLab. A few databases offer some additional features and in some cases provide internal data analysis ; for example, genetic or QTL maps can be compared using Cmap. GD2 is dynamically linked with EUFGIS (http://portal.eufgis.org/), the georeferenced database of forest conservation units coordinated by EUFORGEN. The database connection is

1• The web portal can be accessed via the EVOLTREE website: www.evoltree.eu/index.php/e-recources/elab

TABLE 1

The main eLab databases and their content

Acronyms	Main content	Additionnal features	Access via eLab	Access via family portal
Map 	Genetic and phenotypic records of trees belonging to mapping pedigrees.	Contains only data. Direct access to Cmap possible.	Yes	Yes, under the names of QuercusMap, PinusMap andPopulusMap.
Cmap 	Position of markers and QTLs on genetic and QTL maps.	Comparison of different maps of different pedigrees.	Yes	Yes, under the the same name (Cmap) or all three families (Fagaceae, Pinaceae, Salicaceae).
Treepop TreePop	Genetic and phenotypic records of trees belonging to natural or unstructured populations.		Yes	Yes, under the the same name (Treepop) for all three families (Fagaceae, Pinaceae, Salicaceae).
Provenances 	Passport, Genetic and phenotypic records of trees belonging to provenances established in provenance tests.	Also contains climatic data related to the provenance sources.	Yes	OakProvenance (exists only for oak).
GD2 	Georeferenced data of allelic frequencies and diversity statistics in natural populations.	Is connected to the EUFGIS database.	Yes	Yes, under the same name (GD^2) for all three families (Fagaceae, Pinaceae, Salicaceae).
SSR 	Sequences of microsatellites motifs and their flanking regions.		Yes	Only in the Quercus Portal.
SNP 	Sequences of the contigcontaining the SNP and the two flanking regions.		Yes	No
Candidate genes 	Sequences of candidate genes.		Yes	No
ESTs 	Expressed sequence tags of gene transcripts.		Yes	Yes, under the same name (EST) for all three families (Fagaceaea, Pinaceae, Salicaceae).

carried out using the TAPIR interface. The landscape of genetic diversity near conservation units can be drawn as a result of the connection between both databases, thus potentially helping to refine the setup of the conservation units.

Data access is controlled via user accounts and a hierarchy of roles is granted depending on user access rights. A high level of confidentiality is maintained using table fields with assigned values depending on the related user groups. Thus, data can be kept confidential and restricted to a particular group of users before publication.

Data input

In order to deal with the large amount of data created during the EVOLTREE project, it was decided to offer individual database-specific solutions for transferring data, using Excel or comma-separated text files (.csv). Most of the database-driven web applications enable EVOLTREE users and/or database administrators to manage data and use templates to import data of the same type (for example markers, populations or sequences).

Data output

EVOLTREE users can export data files (.csv) from the individual databases. Search menus are available in most of the database-driven web applications, with several criteria to avoid downloading issues for oversized datasets. The "download" buttons or sections are only available for registered users.

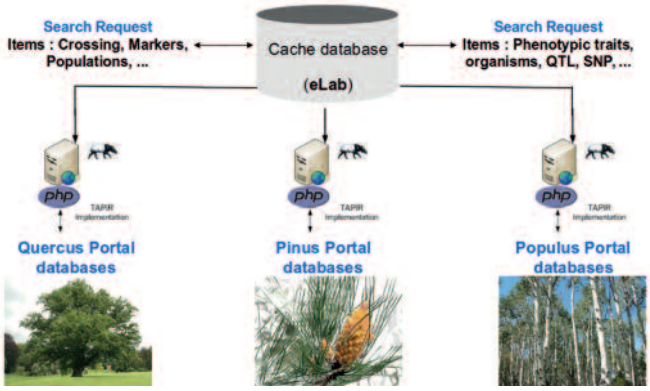
The eLab (electronic Lab)

The eLab gathers the data from all the individual EVOLTREE databases, and provides an interoperable web-interface for

FIGURE 1

General view of the eLab and its components

The eLab search interface consists of three parts:(1) a full-text search interface, (2) a guided search interface, and (3) a tree-view of the data. Within the full-text search interface, users can search for textual terms that occur somewhere in the cache database; whether the term defines a species name, a genus name, an annotation, a marker name or something else (e.g., a comment). The synonyms found during the standardisation process (see above) are also integrated into the full-text search, so that users can retrieve their own terms with their original names.



Data transfer is done over HTML using the TAPIR interface (http://www.tdwg.org/activities/tapir/). TAPIR is a XML-based protocol that can be used for information retrieval in distributed architectures. It is used to collect information from heterogeneous data sources in a pre-defined standardised format. TAPIR operates by harvesting all relevant information of the individual databases.

users to make queries against the data. Thus, the eLab functions as a clearinghouse mechanism enhancing the exchange of information and data throughout the different member labs of EVOLTREE. The relevant data parts of every database were defined when a new database was being integrated into the eLab.

The collected data is first stored locally and then transferred into a virtual cache database. This data collection is carried out at frequent intervals so that the latest information is always available in the cache. While transferring the data into the cache, the data is also merged into a standardised data format by using unique taxonomies; for example, different names for species (e.g., in English, German, French) are standardised so that only one name will be used in the cache database.

The implementation of a proper standardisation system played a major and important part in developing the eLab. The centralised search engine of the eLab only queries the cache database. Therefore, some specific information - only available in the individual databases and which was not considered to be relevant during the integration process - is not visible when using the eLab search engine. If users wish to access such additional information, their search will result in being redirected to the user interface of the corresponding individual database. During the redirection, the user information is encrypted when sent to the database (Figure 1).

Within the guided search interface, users can define more specific search queries. The existing data is presented in a web form and users can select their terms of interest (e.g., species, genus, institution, etc). It is also possible to refine queries further by selecting different pre-defined datatypes (e.g., genetic markers or population). The web-forms are updated dynamically when the selection changes. This way, users can, for example, search for all entries in the cache database that belong to a certain species, to a certain genus, or to a certain institution. The tree-view of the data represents a categorisation of the data available in the cache database. When transferring data into the cache, every data item is categorised according to pre-defined taxonomies. In the tree-view, users can browse through the hierarchical taxonomy and quickly find out how many, for example, data items for a certain sequence feature region exist. It is also possible to detect how many data items belong to a certain EVOLTREE partner.

In addition to the search tool, the eLab offers a reporting service. As they usually contain a large number of entries, the results are grouped together to get a better overview of the data. Each result entry is attributed a description to characterise it. If the user clicks on an entry, he will see all the information that is available for this entry in the cache database, which at this point may not be the "complete" data he is looking for. In order to view the "complete" data, the user can click on a second link and he will then be redirected to the external database the current result entry belongs to.

Photo: EFIATLANTIC

FIGURE 2

Main page of the Quercus Portal

The Quercus Portal comprises two sections:

A static section (left part of Figure 2) that provides general information regarding the biology, biogeography, phylogeny, botany and genetics of the botanical family and the different genera. The static page also comprises information about ongoing research and projects and links to their dedicated web pages.

A dynamic section that corresponds to the different databases related to the species or genera belonging to the Quercus family. They appear as different entry tabs in the headings of the webpage of the portal (upper part of Figure 2).

The family portals

To ease the queries of the user, the different databases were virtually subdivided into three families (Fagaceae, Pinaceae and Salicaceae); thus, users may directly enter one of the three portals (Quercus Portal, Pinus Portal or Populus Portal) and get direct access to the data they are looking for. As mentioned earlier, queries through the eLab retrieve the information stored in the cache database first and not the "complete" data stored in the individual database to which the user can be redirected; access via the portals is therefore much more rapid. As the different portals are designed in the same way, only the Quercus Portal is shown here, being the most complete at this stage (Figure 2).

The Quercus Portal has its own research engine (Global Search) which can be used to make queries across the databases hosted by the portal. An update of the current content of the different databases of the Quercus Portal is available in Table 2.

Current and future use of the eLab and the portals

A web analytic service has tracked and reported the EVOLTREE website traffic ever since the beginning of the network. From 2007 to 2015, 68,838 sessions were recorded by 37,576 users. On average, every time a person visited the EVOLTREE site (a single session), they looked at 4.45 pages for a total pages viewed of 307,000, and an average session duration of 2 minutes and 57 seconds.

While the main databases were constructed in the early years of EVOLTREE and the current portal structure was designed more recently, the main focus is now on the maintenance and regular updating of the databases. We anticipate, however, that very large data sets are still to come as a result of the development and applications of next generation sequencing (NGS) in population genomics of trees. Not all data collections corresponding to forest

FIGURE 3

Interoperability and data flow between information systems in the field of genomics and forestry

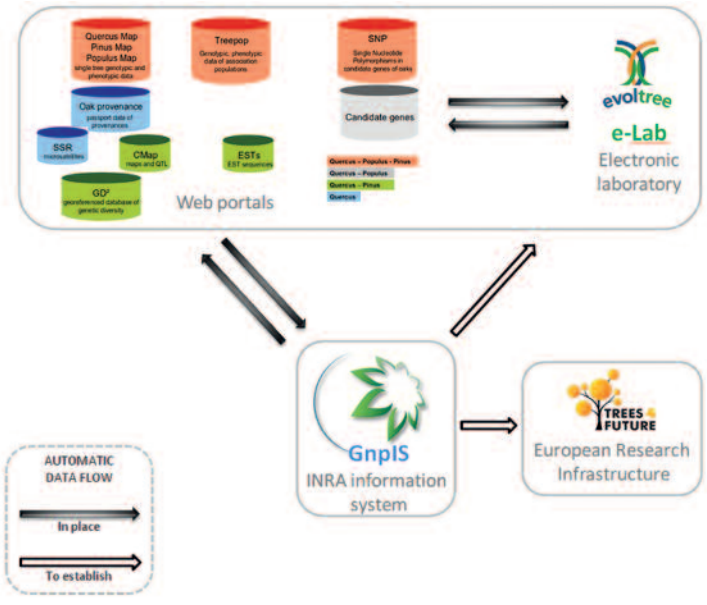


TABLE 2

Update of the content of Quercus Portal (March 31st 2016)

Databases	Taxons	Data types	Features
QuercusMap	*Q. robur, Q. petraea, Q. robur x Q. petraea*	Pedigrees	18
		Genotypes	11,000
		Traits	214
		Genotypic data	515,000
		Phenotypic data	335,000
Cmap	*Q. robur, Q. petraea, Q. robur x Q. petraea, C. sativa*	Geneticmap sets	24
		QTL map sets	13
		Maps	683
EST	*Q. robur, Q. petraea, Q. robur x Q. petraea*	Unigene sets	3
		Contigs OCV1	69,514
		Contigs OCV2	65,712
		Contigs OCV3	91,000
TreePop	*Q. robur, Q. petraea, Q. robur x Q. petraea*	ISS	4
		Association populations	7
		Genotypes	4,729
		Genotypic data	323,784
		Phenotypic data	83,813
GD²	*106 distinct species for Quercus genus*	Populations	4,017
		Trees	24,160
		Frequencymeasures	61,823
		Diversitymeasures	6,902
Oak provenance	*Q. robur, Q. petraea*	Provenances	419
		Provenance tests	60
		Seed lots	464
		Traits	1,874
		Phenotypic data	1,883,677
SSR	*Q. robur, Q. petraea*	Genetic markers	669
Candidate genes	*Q. robur, Q. petraea, Q. robur x Q. petraea*	Genes	648
		Traits	17
SNP	*Q. robur, Q. petraea*	SNP	7,576

trees can be hosted by international databases, such as GenBank, or dbSNP, and thus it is highly likely that in the future new databases will need to be constructed within the eLab.

In recent years, the eLab has also been connected to external data repositories related to either forestry or genomics. This is made possible by the use of a common set of exchange formats and of compatible protocols with the external repositories, similar to the TAPIR interface. Such interoperable protocols have now been installed with GnpIS (a multispecies integrative information system dedicated to genomic data of plants and fungi pests hosted and curated by INRA) and Trees4Future (an Integrative European Research Infrastructure in the field of Forestry) (Figure 3). Meetings, software demos and video conferences are organised to maintain the communication between collaborators and ensure a useful evolution of the information systems.

ACKNOWLEDGEMENTS

We are grateful to Catherine Bastien, Zamira Betancourt, Noémie Emeriau, Audrey Jacques-Gustave, Véronique Jorge, Thierry Labbé, Mélina Millox, Christophe Plomion, Frédéric Raspail, Richard Séverin, and Jean-Paul Soularue for their contribution during the construction of the databases and the eLab.

Photo: EFIATLANTIC