A global information portal to facilitate and promote accessibility and rational utilization of *ex situ* plant genetic resources for food and agriculture

Fawzy NAWAR [1] and Michael MACKAY [1]

1. Bioversity International, Understanding and Managing Biodiversity Programme, Rome, Italy.

Abstract

The plant genetic resources for food and agriculture conserved in 1,500 *ex situ* genebanks are critical sources of biodiversity. They are not only important for conservation and agricultural production, but also for identifying genetic variation that can help to achieve food security and adapt to climate change. Ensuring access to and use of these genebanks depends upon making information about them available to users, including plant scientists and breeders. There are numerous systems with information on genebank accessions; however not all are electronic or available online. Some systems are specific to a single genebank while others cover networks of genebanks based on regions, organizations or crops. Virtually all existing online systems provide only primary identification information – known as passport data – and do not include information for all the data fields of the international standard, the Multi-Crop Passport Descriptors. A global portal is now being developed to address the limitations in existing systems, and to offer integrated and structured solutions. This portal will provide opportunities to: enhance and link existing systems; facilitate the connection of new systems; augment the limited information available in current systems; expand data types to include phenotypic (characterization and evaluation), environmental and spatial information (with the possibility of also including genetic information); and provide realistic access with query functions. Most importantly, the global portal will provide free utilities for genebanks that are not presently available to upload and exercise full control of their data within the portal. This will contribute to the Global Information System mandated by the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) to ensure that unique diversity is effectively and efficiently conserved and made available for use.

Introduction

World genebanks have been collecting plant genetic resources for *ex situ* conservation since early in the 20th Century. Documentation of this collected material is as critical as the plant material itself. The need for a global information system to facilitate access to genebank collections was initially highlighted by the International Undertaking on Plant Genetic Resources in 1983. This was subsequently reinforced in Article 17 of the International Treaty on plant Genetic Resources for Food and Agriculture, which came into effect in 2004.

The initial stages of this effort lacked a global framework as each collecting mission recorded different types of data and most genebanks created their own documentation and data management systems, often in isolation. In 1993, a joint effort between FAO and Bioversity International (Bioversity) aimed to facilitate the exchange of information and produced the Multi-Crop Passport Descriptors (MCPD) (Alercia et al, 2001) and, together with IUCN and UNEP, a technical guide on germplasm collecting (Guarino et al, 1995). To date the major germplasm information networks and national programmes have accumulated around 2 million accessions in information systems, but their records remain fragmented. Some have been published, but many are not yet online or cannot be easily accessed. Although this information has helped decision makers to plan new collecting missions and provides useful information about material conserved in genebanks, it does not enable plant breeders to easily find the specific material needed to develop new plant varieties in their regions.

The need to improve access to information and increase use of genebank holdings necessitated a collaborative project to bring available information together with advanced querying functionality on a global level. This project is supported by the Global Crop Biodiversity Trust (GCDT), the Secretariat of the International Treaty for Plant Genetic Resources for Food and Agriculture (ITPGRFA) and Bioversity International. The project aims to provide plant breeders and other users with a one-stop shop for searching and requesting germplasm using any combination of trait, environmental and passport information.

A first prototype of the global portal (now called Genesys) went online March 2009 for evaluation and feedback. It collated the passport data of 1.2 million genebank accessions of the projects 22 mandated crops from the SINGER, EURISCO and GRIN[1] online information systems within a single portal. Additionally, Genesys includes environmental data for any accession with valid geo-coordinates and 3 million records on 835 phenotypic traits from GRIN.

The Problem

Handling plant characterization and evaluation data at the global level is complex, with an unlimited number of traits measured by genebanks, researchers, plant breeders and others. These traits may differ between crops and the same trait can be measured using many different methods. For example, plant height can be measured as average height of plants at maturity (measured in centimeters from the ground to top of the inflorescence) or with a simpler ordinal scale such as tall, medium or short. The situation is more complex when it comes to traits like disease resistance or abiotic stresses such as drought and salinity tolerance. In addition, some traits are sensitive to environment, which means that different locations or experimental treatments will produce different results. However, there is still a need to provide fast, user-friendly, web-based queries with meaningful and harmonized results as well as metadata, to help users interpret, assess and understand the information.

A second problem is related to the size of the database: 2 million records, each containing 22 fields (MCPD fields) and producing a table size of approximately one gigabyte (of disk space) will need to be joined with 70 fields of environmental data, more than 1,000 characterization and evaluation fields with tens of millions of records, associated metadata and related trait names. A reasonably simple query for specific information within such a database or web portal would result in an unacceptable delay waiting for a response.

Structural and non structural, hyper database model

The database structure was the key to solving the problem: four different types of information were identified, each requiring special handling.

Passport information

The objective was to distribute the data into smaller chunks within the database to facilitate rapid performance. This was achieved by creating a crop template consisting of: one kernel table holding accession passport fields; 12 fields of the MCPD structure; and four other tables, each holding several other MCPD fields (Fig 1). Only one of these tables may have a record related to an accession record, depending on how the accession was acquired by the genebank. In addition, there is one table holding taxonomy data, one table holding environment parameters – this table is created and populated with data 'on the fly' based on the coordinates provided in the MCPD longitude and latitude fields (with data extracted from WorldClim at http://www.worldclim.org/) (Hijmans et al, 2005). Other fields, such as soil parameters, are to be added to this table. The last table holds any other identifiers, or

[1] The System-wide Information Network for Genetic Resources (SINGER) is the germplasm information exchange network of the Consultative Group on International Agricultural Research (CGIAR) and its partners. EURISCO is a European network of *ex situ* National Inventories. GRIN is the Germplasm Resources Information Network of the USDA Agricultural Research Service.

names, the accession might have, such as local names, cultivar names, synonyms, accession numbers in other genebanks and quarantine identifiers.

By apportioning the data in this way, realistic data file sizes were achieved which enhanced the overall performance of both the database and the portal scripts. This produced faster responses to users' queries without violating any system requirements and maintained the option of reconstructing all the data from data templates if necessary.

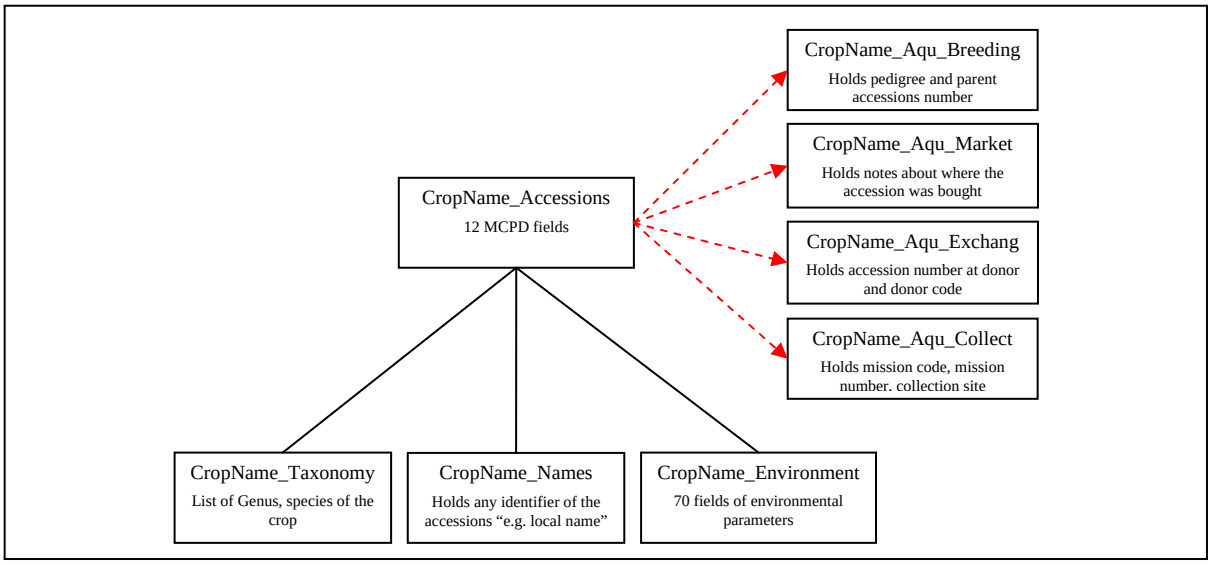

Fig. 1. Crop template structure

General Lookup tables

Replacing countries' full names with ISO codes, and institute names with codes, saves a lot of space and enhances performance. But these codes need to be replaced with the full text when displayed on the portal. For example, SYR002 will be searched in the (fao_institutes) table and displayed as "International Center for Agriculture Research in the Dry Areas" (FAO, 2010).

Add-on tables

Genesys includes data from partner networks in extra tables that are not necessary for the functionality of Genesys. For example, SINGER displays statistics about the material distributed through CGIAR centers and details about collecting missions. Additional information required by networks publishing their information via Genesys can generally be accommodated through the use of add-on tables.

Characterization and evaluation data

Characterization and evaluation data do not have a specific structure. They differ between crops, and the same observation can be 'measured' in many ways. Furthermore, information on experimental conditions is necessary to assist users in interpreting an observation, such as soil type, experimental treatment or if the experiment was conducted in a controlled environment.

The following set of relational tables (Fig 2) was examined for handling an unlimited number of tables holding characterization and evaluation data. It was found to be efficient for the performance and scalability of the system. The following illustration shows that for each crop there are a number of traits, and each trait can have more than one method of measurement. Each method recorded has: (i) a unique identification number; (ii) a method description; (iii) field type; (iv) field size; (v) options; and (vi) range. Genesys uses this information to create a new table that holds the actual measurement. The table name will be the same as the method identification number, and it will contain two fixed

mandatory fields. The first field is “accession_id” to link the score to an accession, and the second is “metadata_id” to link the score to a metadata record on the experiment that contains information about where and how the experiment was conducted (such as use of fertilizer or type of soil). Finally, the score field uses the type and size provided in the method record.

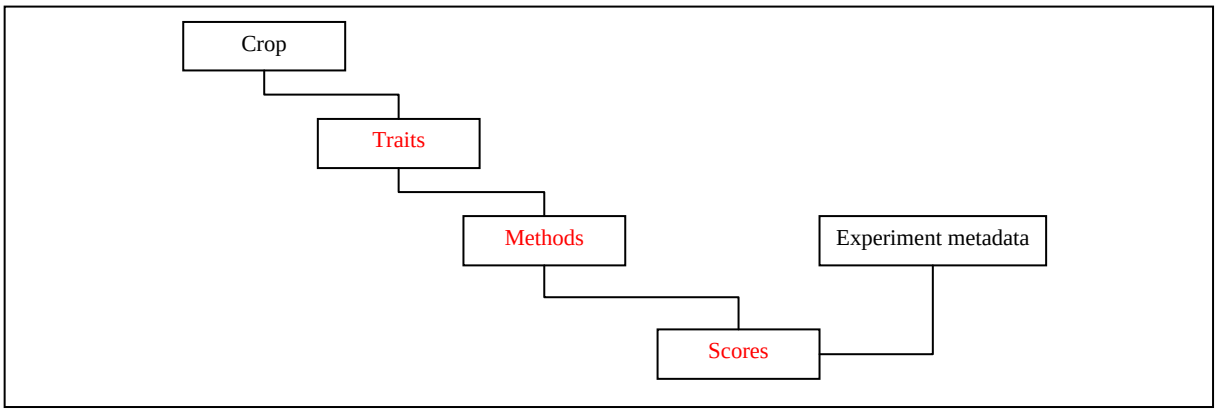

Fig. 2. Characterization and evaluation data in Genesys

Exchange format, upload mechanism, and data quality

Because the MCPD is widely accepted by genebanks and networks as a standard data-exchange protocol on plant genetic resource information, it was decided to keep it as the data-exchange protocol for Genesys. This required more development to extract, slice, and derive data from MCPD, and insert it into the appropriate place within the Genesys structure.

However this protocol does not provide a solution for how to exchange information on characterization and evaluation data – currently, there is no structure or data-exchange protocol for this purpose. Developing such a protocol across crops (like the MCPD) is not feasible. Even developing a protocol for each crop is not feasible because there are multiple methods to measure the same trait. In addition, data-processing procedures to harmonize deferent coding and metadata on the experiment – and the expected volumes of data – make this task even more challenging.

The problems of data quality and completeness arose upon visualizing the currently available data. Table 1 provides an example of the data quality/completeness for rice accessions across the SINGER, EURISCO and GRIN systems in 2009. For example, a significant number of geo-coordinates were found to be in the ocean or outside the indicated country of origin. To address these issues, an automated script was developed to measure data quality/completeness on variables important to the functionality of Genesys.

Table 1. Data quality/completeness of the rice collection on four essential data fields

Network/ System	Total No. of accessions	Missing/Invalid geo-coordinates	Missing/Invalid country of origin	Missing acquisition source	Missing sample status
EURISCO	6306	6305 (99.98%)	565 (8.96%)	5836 (92.55%)	1259 (19.97%)
GRIN	41551	32697 (78.69%)	12 (0.03%)	35771 (86.09%)	0 (0.00%)
SINGER	130026	120883 (92.97%)	11407 (8.77%)	112838 (86.78%)	106687 (82.05%)
All	177883	159885 (89.88%)	11984 (6.74%)	154445 (86.82%)	107946 (60.68%)

Clearly, a novel upload mechanism was needed to serve the following requirements:

1. Accept the MCPD exchange format, perform data processing and upload it to the correct parts of the Genesys data structure;
2. Validate data quality, correctness and completeness;
3. Derive related environmental data;
4. Manage the dynamic characterization and evaluation structure, and facilitate mapping between data providers' coding systems to the Genesys code for any given trait/method;
5. Reports any error while uploading on the server side; and
6. Offer data providers control in maintaining and updating their data in Genesys.

Direct Data Control (DDC), a Java application, was created to implement all of these functions. It gives data providers the ability to upload or update data using an easy drag-and-drop method to map their files to MCPD fields; data providers can also validate their data before submitting. The application provides one-time code mapping between data-provider values and Genesys values. For example, if a genebank codes flower color as 1 for white, 2 pink, 3 yellow and Genesys uses the codes W for white, P for pink, Y for yellow, the utility will validate data against the codes supplied by the data provider (1, 2 or 3), then submit the information to Genesys database using the system counterpart codes (W, P or Y). Figure 3 illustrates the functionality of the DDC utility.

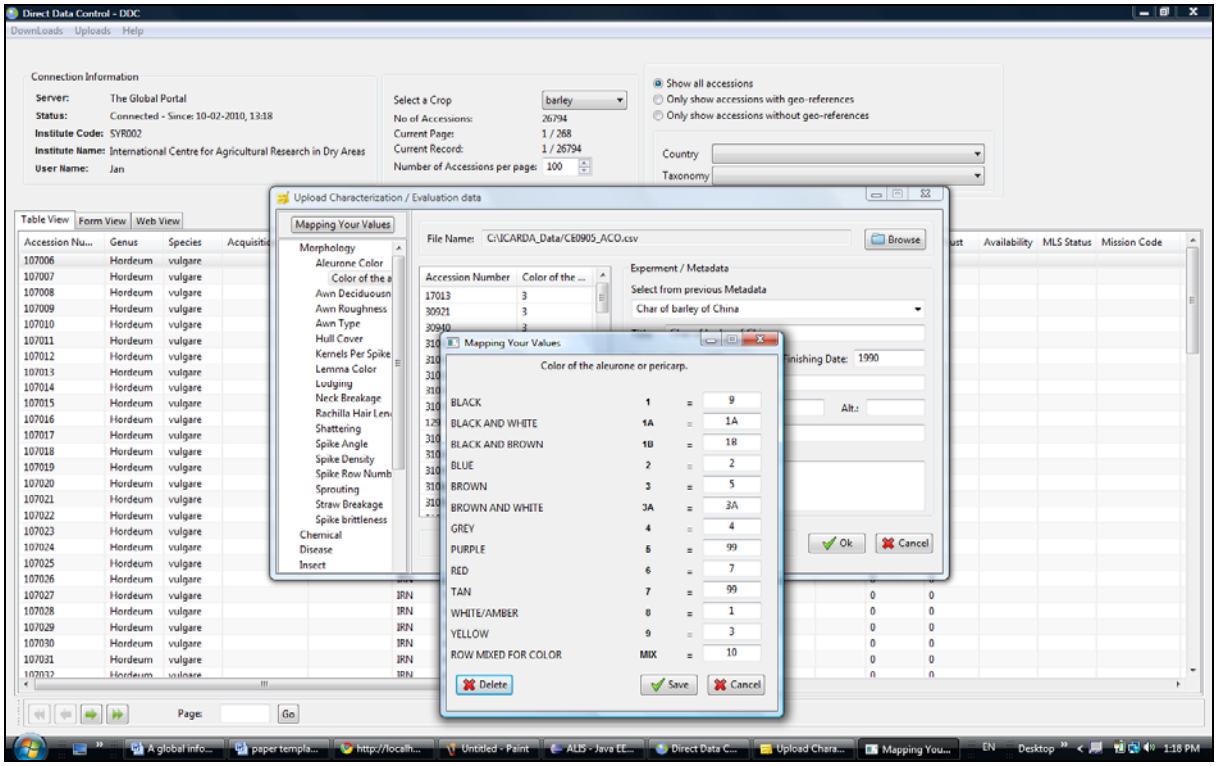

Fig. 3. Snapshot of the DDC characterization and evaluation-upload mechanism.

The system architecture

The Genesys architecture has three major components and is illustrated in Figure 4.

Data providers

Data providers for Genesys could be networks like SINGER or EURISCO, national programmes like the USDA-ARS/GRIN or independent genebanks. Of these, all have established a way to gather passport data, but only GRIN has developed methods to harvest and manage characterization and evaluation data. All data providers will need a method to communicate with the Genesys server to upload and manage their data.

Centralized data warehouse

All the available data from the SINGER, EURISCO and GRIN systems was analyzed and has been compiled into the Genesys data warehouse and its extended tables (including data unique to individual networks such as SINGER collecting mission data); data will continue to be accumulated and updated into the Genesys data warehouse.

Multiple representations (portals)

Each of the participating networks will have access to a centralized data warehouse in order to publish data of interest to their communities on their own customized web portals.

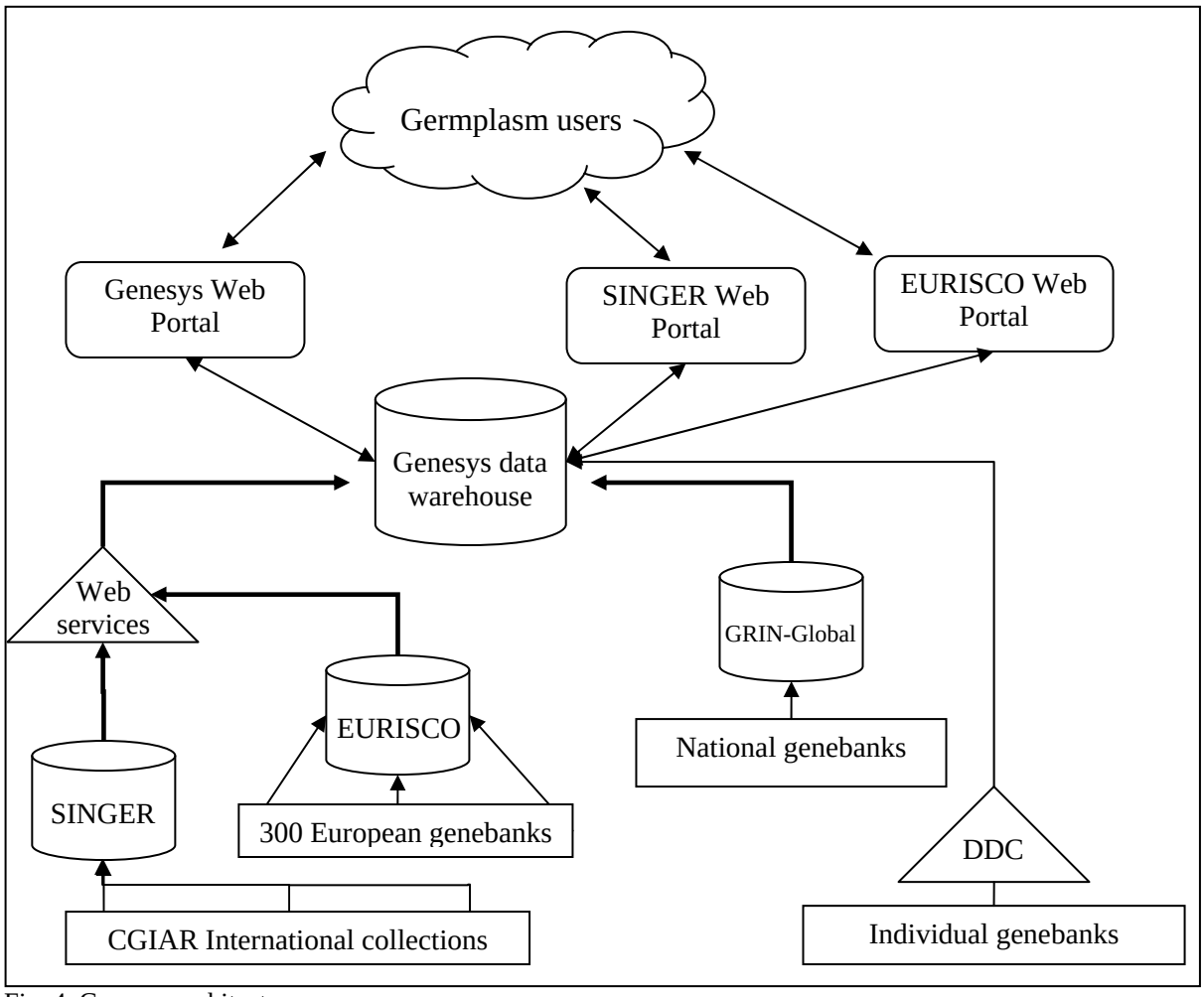

Fig. 4. Genesys architecture

The web portal

Genesys provides a set of functions that enable users to search global genebank holdings using any combination of passport, environmental, and traits scores. For example, a user could select all accessions of *Triticum aestivum* found in areas that receive annual precipitation between 250-400 mm and have white grains with 1,000 kernel weight greater than 45 g. The Genesys prototype query page is currently very large, but it will be reorganized into a tabbed format based on data categories. It provides users with minimum and maximum limits for numerical fields and includes pull-down menus for fields containing ordinal scales or multiple option values. Users can replace their current query or add new criteria to the query. For example, if the current query is (sample status = wild) and the user

selected (sample status= cultivar) and added to the current query, the final result would be (sample status = wild or cultivar).

A list of query filters is available to enable users to visualize their result set after applying each filter in their criteria. This also allows them to remove any of the filters without repeating the querying process. Genesys also provides summery statistics on the query result from different perspectives, such as the number of accessions and the countries in which it was collected, or which institute holds those accessions and the number held.

A data browser provides lists of accessions with different views and check boxes in front of each to select accessions, add them to a 'shopping cart' to request them from genebanks, or to download data. This feature, along with numerous others, is not yet active because of its dependency on the ongoing development of associated modules. Clicking any row on the list will display accession details with all available information about the selected accession.

Genesys also provides a 'tracing' function to find related accessions, sometimes called duplicates, across genebanks. This feature could provide users with an alternative genebank for requesting. A function will also be added to enable users to search for accessions from the same environment.

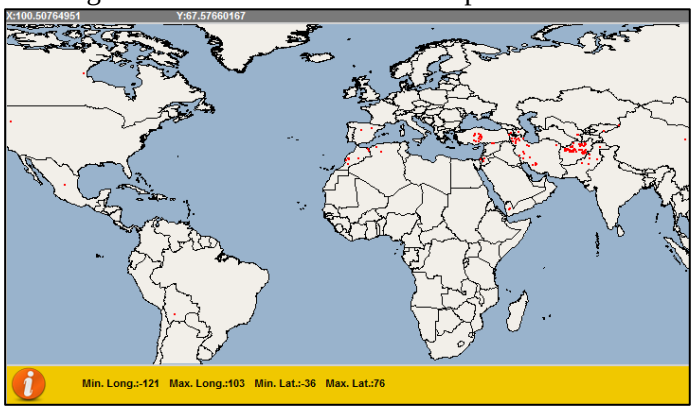

Fig. 5. Collection sites mapped within Genesys.

The system also provides fast mapping of accessions (Figure 5), with the ability to download Google Earth KMZ files for visualizing the topology of collecting sites (Figure 6). Accession details can be sent back to Genesys through a link in Google Earth.

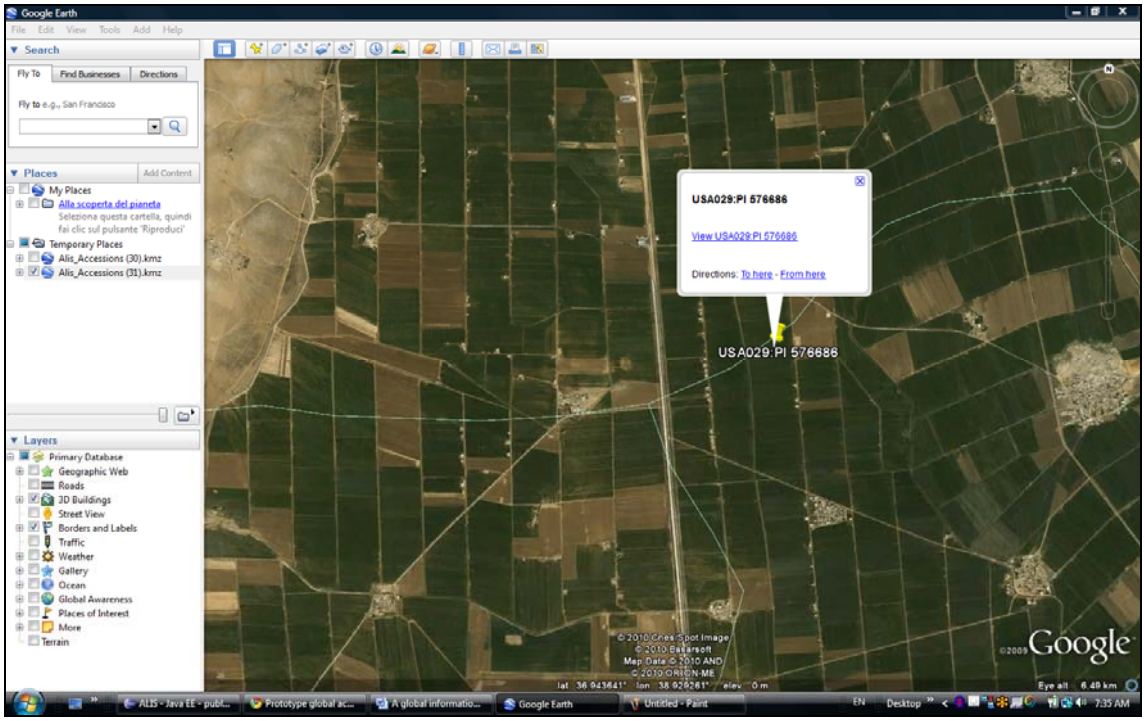

Fig. 6: Viewing topographic and other information associated with Genesys data in Google Earth

Conclusion

Genesys demonstrates the feasibility of incorporating characterization, evaluation and environmental, together with passport, data from genebanks worldwide into a single online system providing fast responses to complex queries. The amount and diversity of the data already online is proof of concept. In the future, it is also envisaged that Genesys will handle genetic information. The long road ahead to capture all of the characterization and evaluation data on genebank accessions worldwide requires the efforts of all genebank curators, network coordinators as well as plant breeders and pre-breeders who generate so much of the phenotypic information.

It is anticipated that Genesys will integrate an enormous amount of data with high dimensionality. As the existing data types are further populated and new types of information are added it will become increasingly difficult to query and analyze it using conventional methods. More novel approaches will be required, such as data mining and knowledge discovery in databases (KDD). Genesys is presented as a significant step towards the global information system envisaged by the ITPGRFA.

References

ALERCIA A., DIULGHEROFF S. and METZ T. 2001. Multicrop Passport Descriptors. FAO/IPGRI Rome.

FAO (2010) World information and early warning system on PGRFA (WIEWS). Internet page: http://apps3.fao.org/wiews/wiews.jsp, accessed 15 February 2010.

GUARINO, L., RAMANATHA RAO, V., REID, R. 1995. Collecting Plant Genetic Diversity: Technical Guidelines. IPGRI, FAO, IUCN and UNEP, Rome.

HIJMANS, R.J., S.E. CAMERON, J.L. PARRA, P.G. JONES and A. JARVIS, 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.