PROJECT NUMBER 289706

Start date of the project: 01/12/2011, duration: 54 months

Assessing and Monitoring The Impacts of Genetically Modified Plants on Agro-ecosystems

Project acronym: AMIGA

Funding Scheme: Large Cooperative Project

Deliverable 3.1 A central GIS and associated database

Authors: Guillaume Huby (Geosys), Geoff Squire (James Hutton Institute), Antoine Messéan (INRA)

Organisation names of lead contractors for this deliverable: Geosys

Project leader: Dr. Salvatore ARPAIA
ENEA - Trisaia Research Centre
Tel:+39-0835-974714
E-mail: Salvatore.arpaia@enea.it
Project website address: www.amigaproject.eu

June 2016 **Dissemination Level: Confidential**

Contents

Abstract

A central database and GIS was constructed in AMIGA by iterative development and testing and then populated with a wide range of datasets covering scales from point samples in field trials to national trends in agronomic and environmental variables. A database manager, liaising with partners, ensured consistent procedures were developed for formatting and upload. The AMIGA database and GIS offers a prototype for a future central point of information for planning, executing and monitoring GM crop trials in Europe.

1 Introduction and access to the software

The assessment and monitoring of the impacts of a new technology or a new agricultural practice on agro-ecosystems require the collection of a wide range of data that are very heterogeneous both in content and collection procedure. In the AMIGA project, several work packages have been collecting datasets: WP2 (baseline data), WP3 (evolution of agro-ecosystems over time), WP4/5 (lab and field trials), WP7 (monitoring) and WP8 (IPM field trials). Other datasets are also available from previous projects on GM crops (e.g. EU SIGMEA) or external sources (EFSA fauna database). By gathering all data in a single database shared by all partners and potential users and thematics, this task aims to demonstrate the value of geolocalising observations based on common reference terms as well as the access to a wide range of datasets about receiving environments and impacts of GM (Genetically Modified) crops on the environment.

The objectives of the central GIS (Geographic Information System) and associated database can be described as follows.

- Store observations or events, from the description of the evolution of cropping systems to observations of insects made on individual plants of a field trial.
- Geolocalize all observations and operations so as to be able to retrieve and aggregate information at a given geographical scale.
- Use a common referential (crops, varieties, active ingredients, species, etc) to ensure interoperability across countries and sources of information.
- Be flexible and manage access rights so as to be able to store information from various sources while protecting property rights.
- Develop a user interface that facilitates import of existing datasets and allows exports for specific needs.

In order to fulfill these objectives, the type and resolution of data collected in AMIGA were first considered. Then the most appropriate GIS and database structures were designed to cover protocols for collecting and entering information, analysis required to link historical and regional data and new information from experiments and monitoring.

The work has been shaped through the selection of representative datasets that have been loaded into the database.

To access the database, users can download the software on a FTP (File Transfer Protocol) server: ftp://Amiga2:Am!g@2$!@livraisons.geosys.com, then get a user name and password required to access the database. Together with the software, the user will find an installation guide, tutorial and contextual help with the interface.

This document presents first, a summary of the database and software description, second, illustrative examples of datasets loaded into the database and finally, representative results that a user can get from this application.

2 Database and software description

This part of the report presents a summary of the database model, of the software layout and the way data ownership is managed in the software. Detailed information can be found within the full software help documentation.

2.1 Database model

The first requirement was to be able to store observations or events of very heterogeneous types. For example, there may be observations on NTOs (Non Target Organism like butterflies …) or presence of disease (Potato Blight) or yields. These observations are made in a context that can be can be localized more or less precisely. For example, they can be made from traps in trial plots or full commercial fields or surveys in administrative areas. They can also be identified temporally at some date or during a crop season.

The second requirement was to be able to retrieve information in the database according to search and aggregation criteria. For example, the following criteria were defined with AMIGA partners (Paris meeting 2015, January 8th):

- Get raw data per year, country, crop
- Aggregate data by
  - Genus, Family, Class …
  - Functional group
  - GM / non GM
  - IPM (Integrated Pest Management)

In order to fulfill these requirements, a relational database was selected. This type of database allows entities with a set of properties and connection of entities to each other. In the AMIGA database, there are three main entities:

1. Territorial unit is a geographic entity, which is spatially and temporally defined. It ranges from the highest level (continent) to the lowest level (plant). Territorial units can be linked to each other by specifying a parent (a field trial can be linked to a field, the latter being associated to a farm or a region); the European NUTS (Nomenclature of Units for Territorial Statistics)[1] classification has been implemented to define geographical areas.
2. Territorial unit history corresponds to a grouping of events that occurs on a given territorial unit: it can be a period of observation, a period of crop growth (from seeding to harvest) or a crop rotation in case of multi-year experiments.
3. Events are the unitary concept for all observations. Events share common properties and might have specific properties depending of their type. Three main types of events are considered:

   Crop management: all events related to field management operations (sowing, crop protection, fertilization, etc).

   Receiving environment characteristics describing the climate or regional cropping systems and their evolution.

[1] NUTS are administrative units used for statistics at European level. Since each country has its own convention names and levels of administrative units, statisticians split them in comparable “levels”. Level 0 is the country, Level 1 can be regions or states depending of the country and so on.
See https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics

Observations: all other ground observations for example on plants, microbes and invertebrates.

Territorial units and events can also be linked to documents like import files, original user datasets or protocols.

The relationship between these three types is illustrated on Figure 2-1. A territorial unit can have several histories. For example several crop seasons or several crops. Each history can have many events and each can be of a different type.

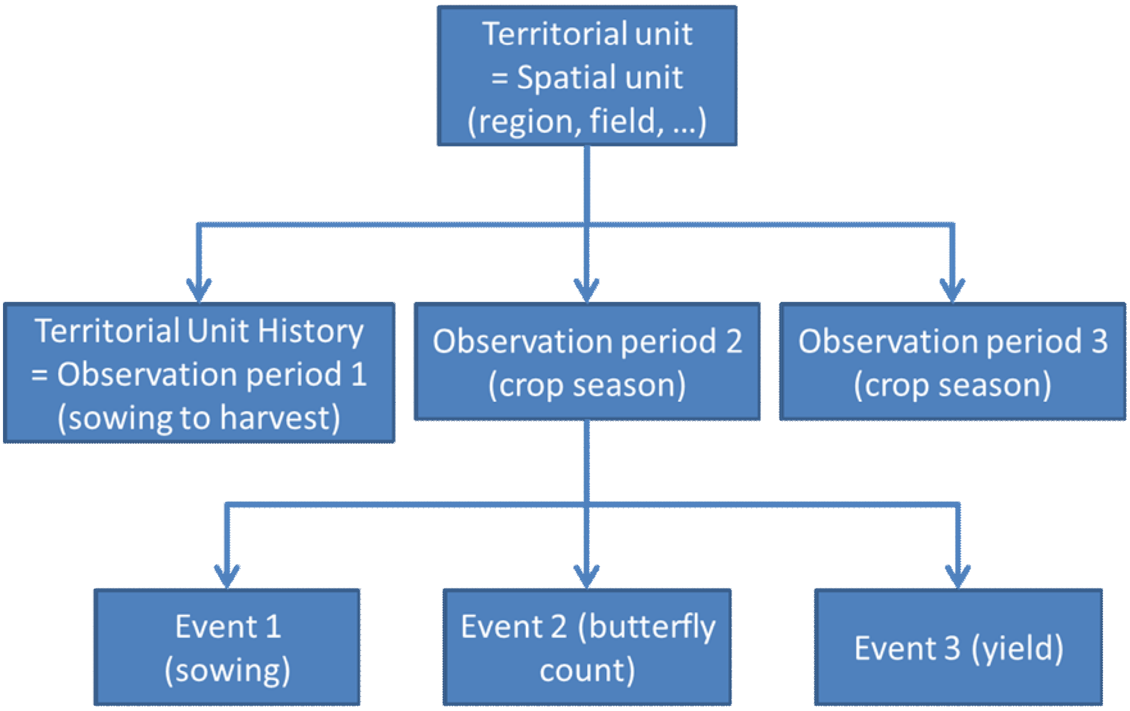


Figure 2-1 Summary of the database model: main architecture

Events can be grouped into entities called datasets. The dataset is a base entity to allow manipulation of data for import, query and export.

On top of these three main entities, other shared entities are created.

- Referential data are all data that are not dependent of time or geography (names of crop and varieties, names of crop protection practices, names of insects, stages of insects). These referential data are the common vocabulary for all entities and all users. This common vocabulary aims to prevent inconsistency, for example, if one user recorded events with the crop “Corn” and another with the crop “Maize”. If both names were allowed, a database query using just one of the names would return incomplete data.
- The second type of shared entity is a ‘document’. Territorial units and events can be linked to documents like import files, an original user dataset, or a measurement protocol that can be shared by all observations made according to this protocol.

The relations between entities are explained on Figure 2-2.

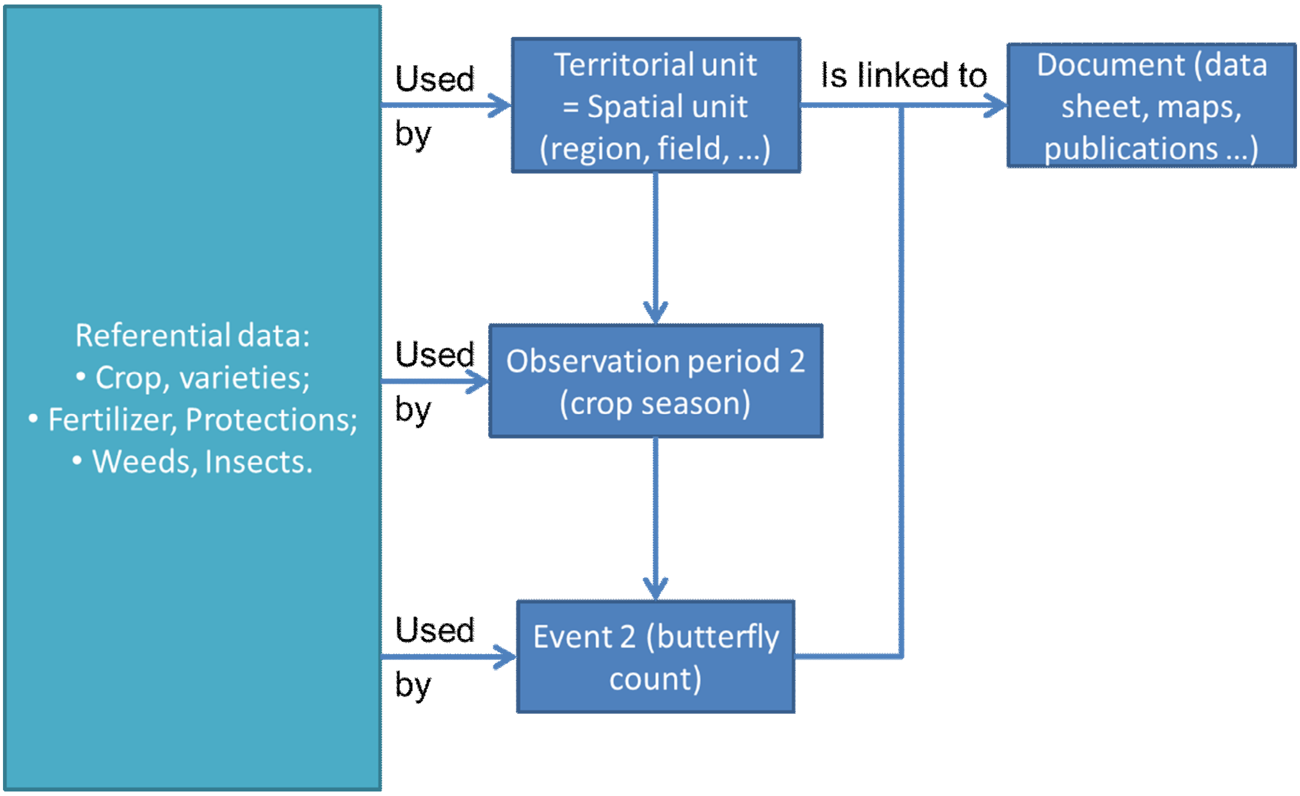


Figure 2-2 Summary of the database model: main objects and relations

The details of objects and properties of the database model are available in the software documentation.

2.2 Software description

To access the database, the user has to download the software to their computer. Since the database is hosted on a central server, the database will be synchronized with the user's computer when the user opens the application. Once synchronized, the user gets a User Interface (UI) that allows them to access the different modules to manage, import or export of data (see Figure 2-3 and Figure 2-4).

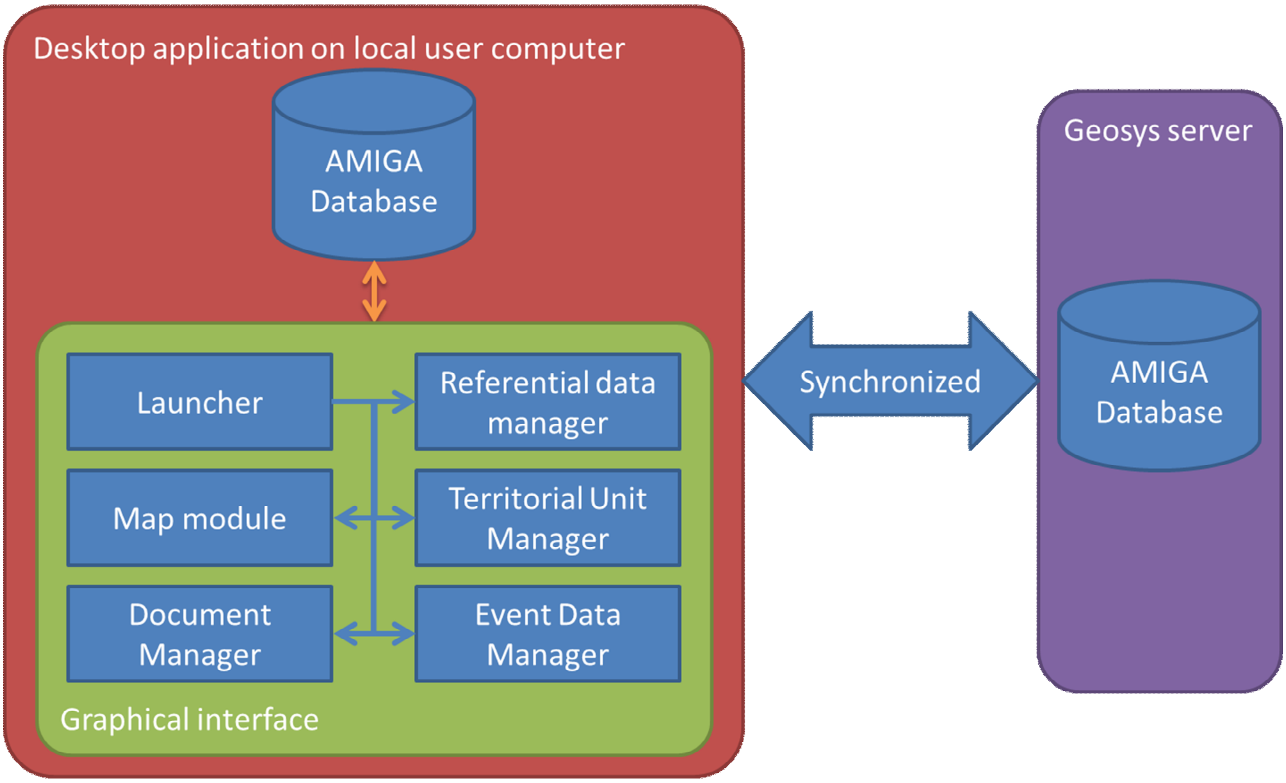


Figure 2-3 Software architecture overview

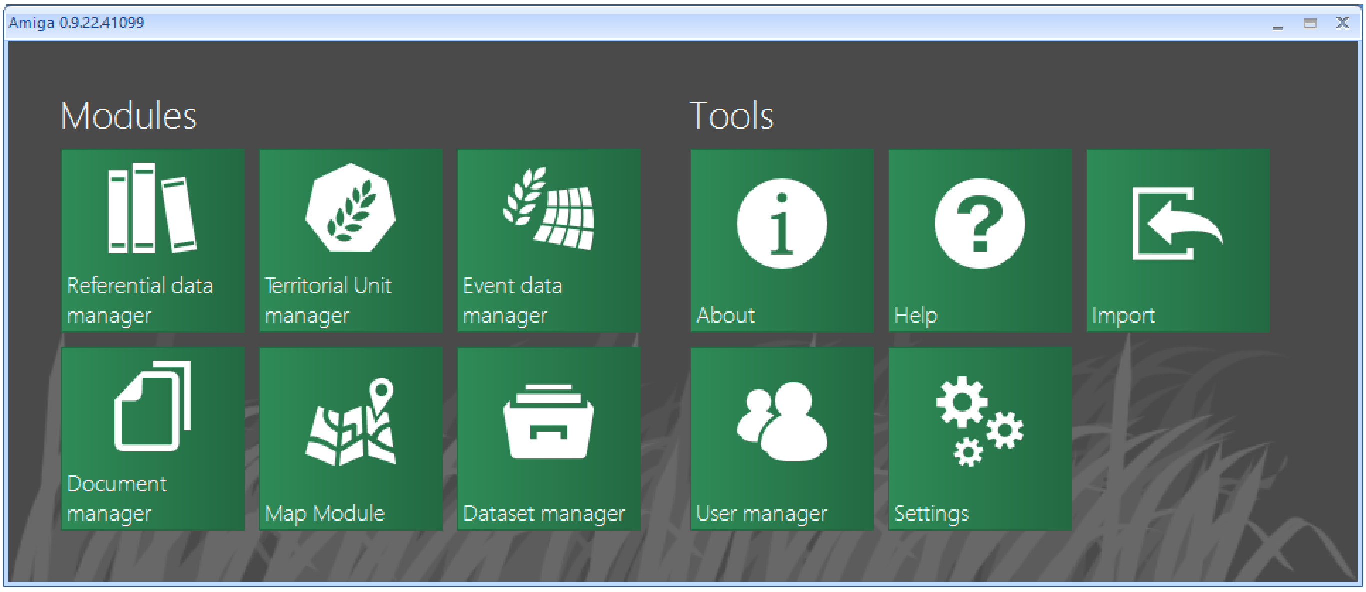


Figure 2-4 AMIGA software main user interface

The detailed functionalities and use are described in a software tutorial that includes documentation that can be easily reached from each module.

Each time the user asks or closes the application, their modifications are synchronized back to the central server.

2.3 Access management

To be able to synchronize and access the data within the software, the user has to use a personal login and password. The login and password can only be created by application administrators (see Figure 2-5).

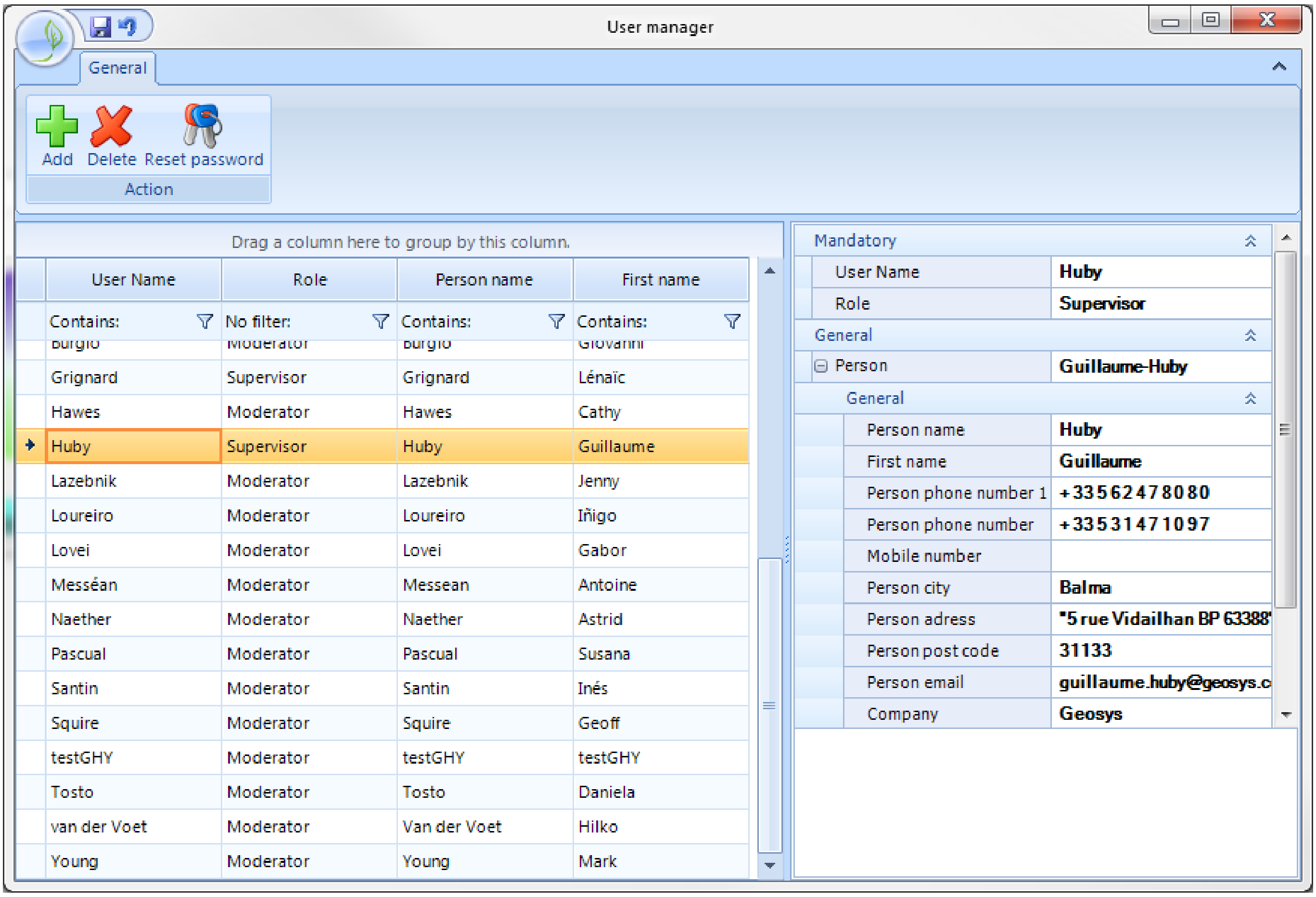


Figure 2-5 User management interface

Furthermore, the user account is used to identify data ownership. If a user creates or imports observations, all observations are associated with that user. In order to share data, the user has to ‘publish’ the datasets. The published data can be modified only by the owner. If another user has questions or finds some error in the data, they can contact the owner within the application so that the error can be fixed (see Figure 2-6).

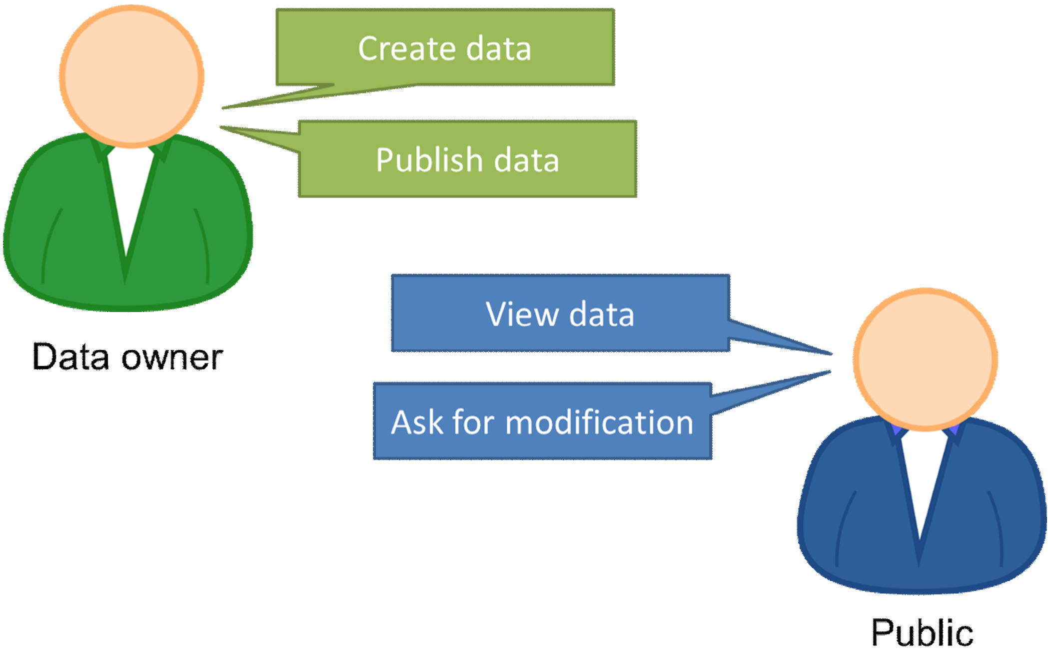


Figure 2-6 Data ownership and possible actions

3 Received and loaded datasets

In order to populate the database and test its functioning, a set of representative datasets from the project has been uploaded (see Figure 3-1).

Dataset name	Type of events	Nb Data	Person	Years	Countries	Land uses/Land covers	Status
Contains:	Contains:	Equals:	No filter:	Contains:	Contains:	Contains:	Contains:
Irlande 2013 Potato trial - Fertilizer	Fertilizer	1	Guillaume-Huby	2013	Ireland	Potato	Published
Irlande 2013 Potato trial - Harvesting	Harvesting	108	Guillaume-Huby	2013	Ireland	Potato	Published
Spain 2012 Corn trial - Harvest events	Harvesting	1	Guillaume-Huby	2012	España	Corn	Pending
Irlande 2013 Potato trial - Irrigation	Irrigation	108	Guillaume-Huby	2013	Ireland	Potato	Published
Bulgaria 2013 Potato Trial - Crop protection events	Protection	11	Mariana-Radkova	2013	Balgarija	Potato	Pending
Irlande 2013 Potato trial - Protection	Protection	24	Guillaume-Huby	2013	Ireland	Potato	Published
Spain 2012 Corn trial - Protection events	Protection	3	Guillaume-Huby	2012	España	Corn	Published
Irlande 2013 Potato trial - Soil preparation	Soil preparation	5	Guillaume-Huby	2013	Ireland	Potato	Published
Spain 2012 Corn trial - Soil preparation events	Soil preparation	5	Guillaume-Huby	2012	España	Corn	Published
Irlande 2013 Potato trial - Sowing	Sowing	108	Guillaume-Huby	2013	Ireland	Potato	Published
Spain 2012 Corn trial - Sowing events	Sowing	20	Guillaume-Huby	2012	España	Corn	Published
Bulgaria 2013 Potato trial - NTO Arthropods observation	Target Observation	228	Mariana-Radkova	2013	Balgarija	Potato	Pending
Butterflies 2013 Vikhög Sweden	Target Observation	172	Andreas-Lang	2013	Sverige	Transect	Pending
Dataset Messéan TARGET OBSERVATION 03/05/2016 12:29:30	Target Observation	1	Antoine-Messean	2013	Danmark	Corn	Pending
Dataset Messéan TARGET OBSERVATION 03/05/2016 13:16:48	Target Observation	1	Antoine-Messean	2013	Sverige	Corn	Pending
Denmark 2013 - Visual samplings observations	Target Observation	1224	Erna-Borbath	2013	Danmark	Corn	Pending
Denmark 2013, 2014 and 2015 - Pitfall traps observations	Target Observation	2836	Erna-Borbath	2015, 2014, 2013	Danmark	Corn	Pending
Denmark 2013, 2014 and 2015 - Pitfall traps observations	Target Observation	1077	Erna-Borbath	2015, 2014, 2013	Danmark	Corn	Pending
Import_Messéan_AMIGA NTO Slovakia 2012 - Coll1_ELxlsx_13/05/2016 15:00:52	Target Observation	880	Antoine-Messean	2012	Slovensko	Corn	Published
Import_Messéan_AMIGA NTO Slovakia 2012 - Coll2_ELxlsx_13/05/2016 15:16:24	Target Observation	880	Antoine-Messean	2012	Slovensko	Corn	Published
Import_Messéan_NTO Slovakia 2012 - BT Trial - Cara_ ELxlsx_04/04/2016 16:16:12	Target Observation	880	Antoine-Messean	2012	Slovensko	Corn	Published
Import_Messéan_NTO Slovakia 2013 - IPM Trial - Cara_ ELxlsx_14/04/2016 14:41:42	Target Observation	432	Antoine-Messean	2013	Slovensko	Corn	Published
Import_Messéan_Survey_Agroecosystems_Slovakia 2013-corn Cara_ELxlsx_22/04/2016 14:40:55	Target Observation	480	Antoine-Messean	2013	Slovensko	Corn	Published
Irlande 2013 Potato trial - Target Observation	Target Observation	324	Guillaume-Huby	2013	Ireland	Potato	Published
Netherland 2013 Potato trial - Pitfall traps counts	Target Observation	2866	Jenny-Lazebnik	2013	Nederland	Potato	Pending
Spain 2012 Corn Trial - Bacteria observations	Target Observation	60	Guillaume-Huby	2012	España	Corn	Published
Spain 2012 Corn trial - Pitfall traps observations in functionnal groups counts	Target Observation	480	Guillaume-Huby	2012	España	Corn	Published
Spain 2012 Corn trial - Weeds observations	Target Observation	120	Guillaume-Huby	2012	España	Corn	Published
Spain 2013 Corn trial - Pitfall traps observations	Target Observation	5565	Iñigo-Loureiro	2013	España	Corn	Pending
Spain 2013 Corn trial - Visual sampling observations	Target Observation	1337	Iñigo-Loureiro	2013	España	Corn	Pending
Spain 2013 Corn trial - Weeds observations	Target Observation	187	Iñigo-Loureiro	2013	España	Corn	Pending

Figure 3-1 Screenshot of datasets loaded into database (26/05/2016)

To facilitate testing, 20 datasets were uploaded containing more than 20 000 observations on agronomy and biodiversity. Emphasis was given to the year 2013 for which there were 12 datasets, but 2012 was represented by 6 datasets and 2014/15 by 2 datasets. 16 datasets are on corn crops, 3 on potatoes and one is based on transect observations. Datasets originated from 7 different countries (see details in Table 3-1).

Table 3-1 Target observation datasets per country

Country	Dataset number
Bulgaria	1
Denmark	4
Spain	6
Ireland	1
Netherland	1
Slovakia	5
Sweden	2
Total	**20**

The datasets also contains 6 different types and topics of observations detailed in Table 3-2.

Table 3-2 Topics and types of observations

Topic and method	Dataset number
Arthropods - Pitfall traps	13
Arthropods - Visual	2
Bacteria - Soil sampling	1
Butterflies - Visual	1
Disease - Visual	1
Weeds - Visual	2
Total	**20**

With these datasets loaded into database, it was possible to demonstrate the possibilities to query the database, to combine data to create new datasets and to apply statistical methods to support experimental design.

4 Results – examples

This part of the report presents an overview of results that can be obtained using the software interface, using GIS, creating new datasets or applying statistics.

4.1 Use of the GIS

Since observations are localized or linked to some spatial entity, the database allow a user to query and display data on a map module. The following screenshots shows how queries can be refined, in this example for observations made on maize plots in Spain during 2013.

The first query asks to display data at a continent scale (Fig. 4.1).

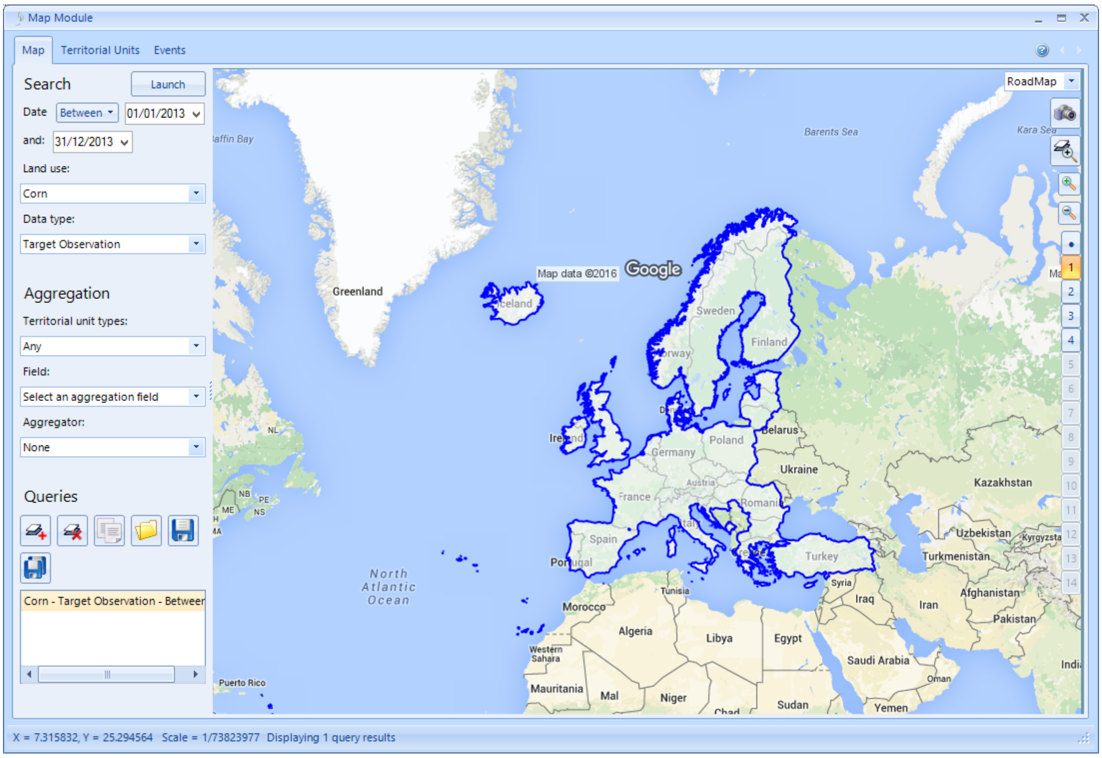


Figure 4-1 After input of search criterias, results are displayed at continent level

The second query narrows down the search to a country scale (Fig. 4.2) and then the third query asks to display the area of trials data or observations within a country (Fig. 4.3). Clicking a country allows a user to view in detail the type of observation available (Fig. 4.4, 4.5).

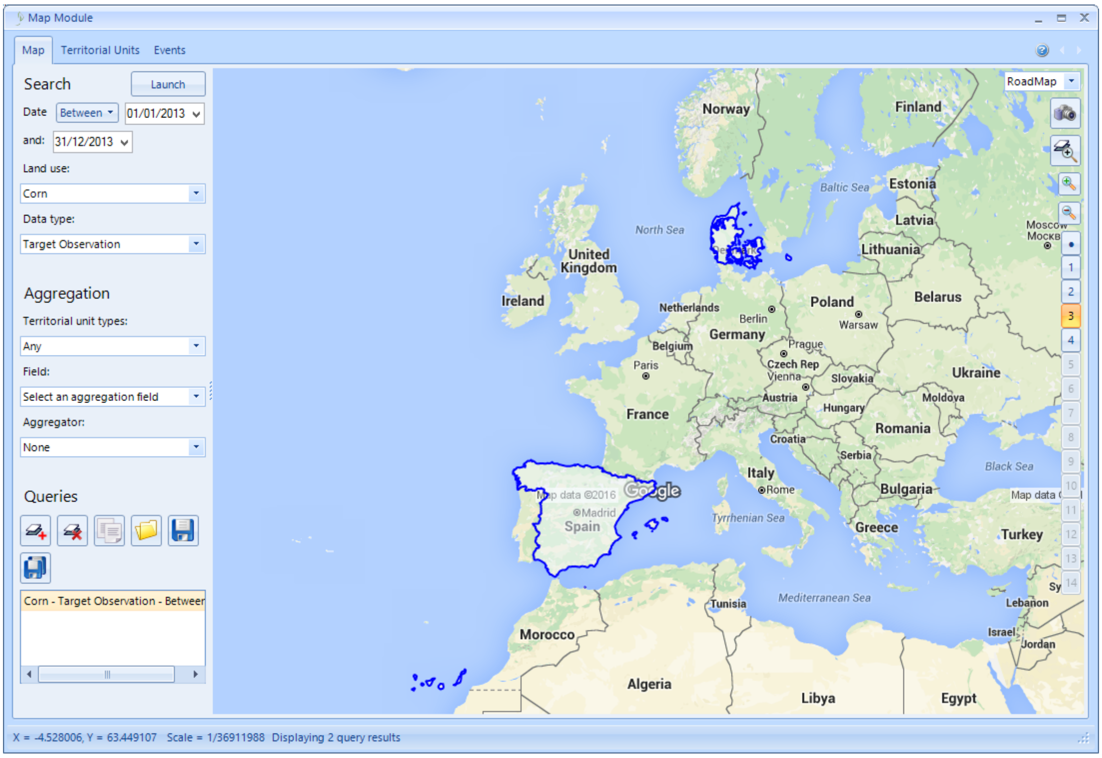


Figure 4-2 User can display results at country level

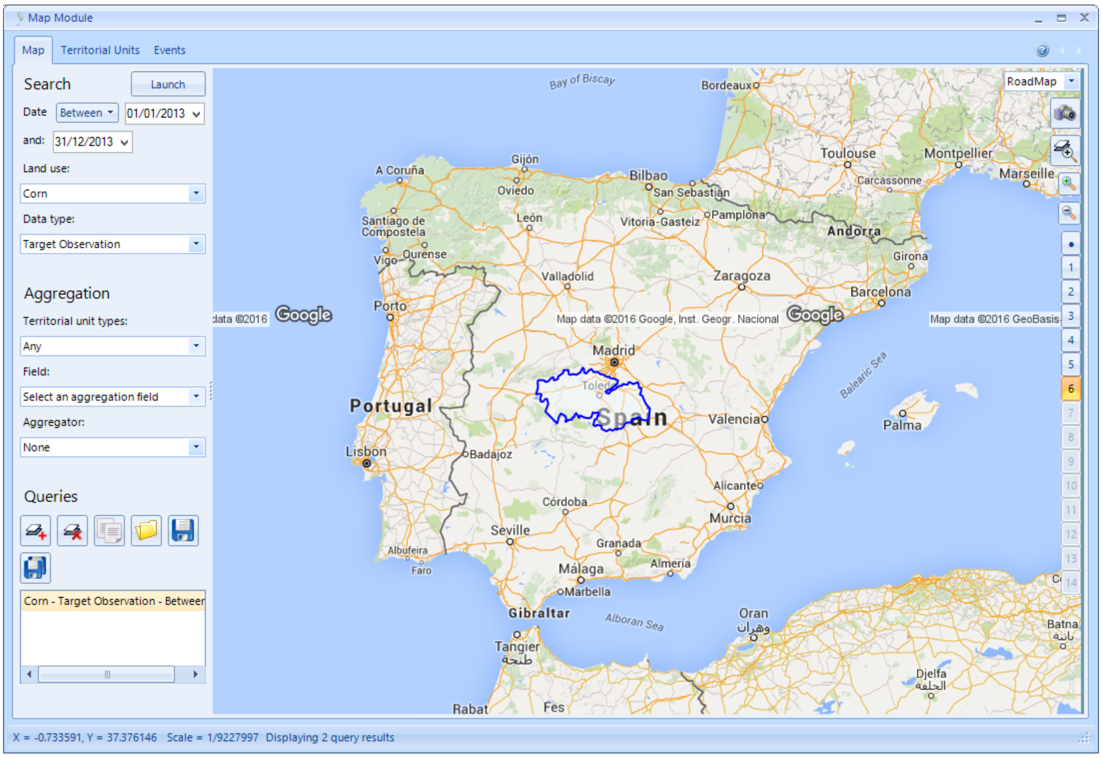


Figure 4-3 User can also zoom to access lower display levels like NUTS level 3 for example

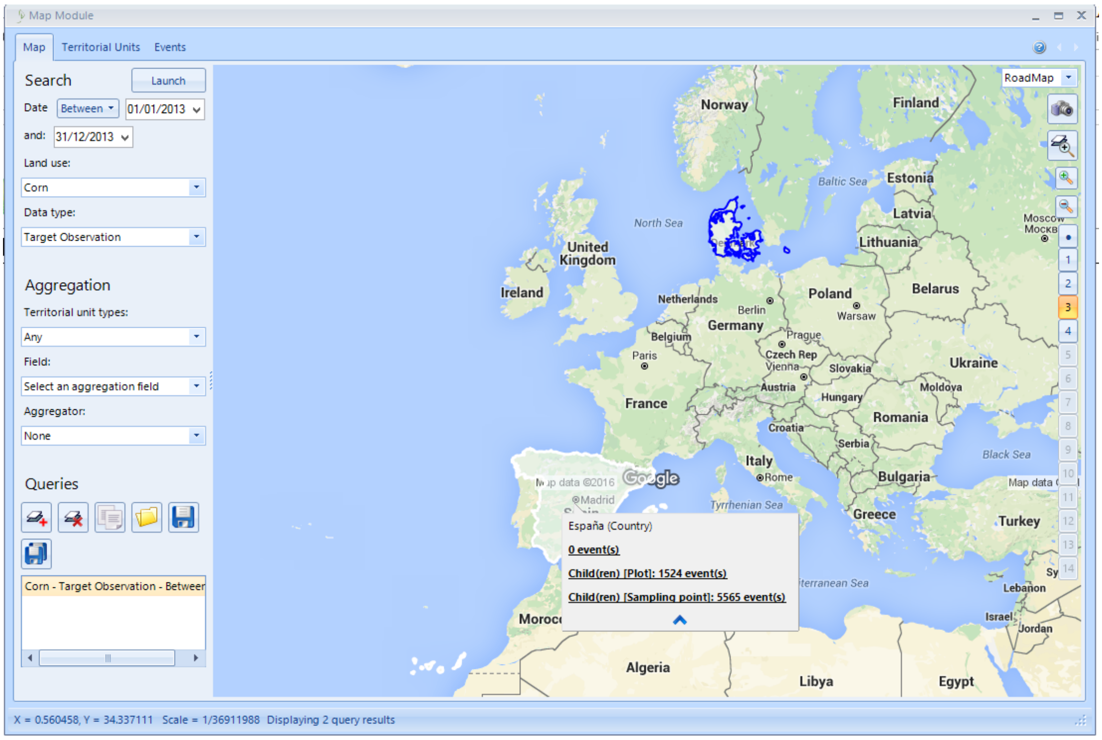


Figure 4-4 Clicking a country allows user to view the number of entries found per sub levels

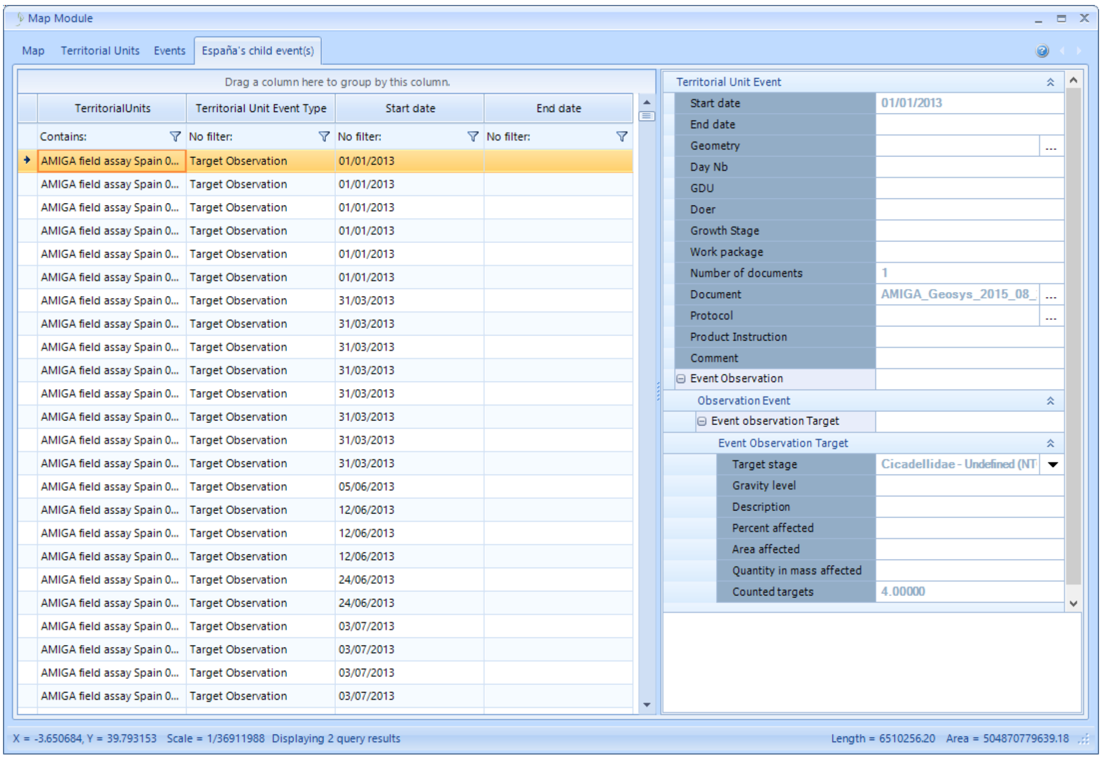


Figure 4-5 When user clicks a level, he can view all entries found in a table view

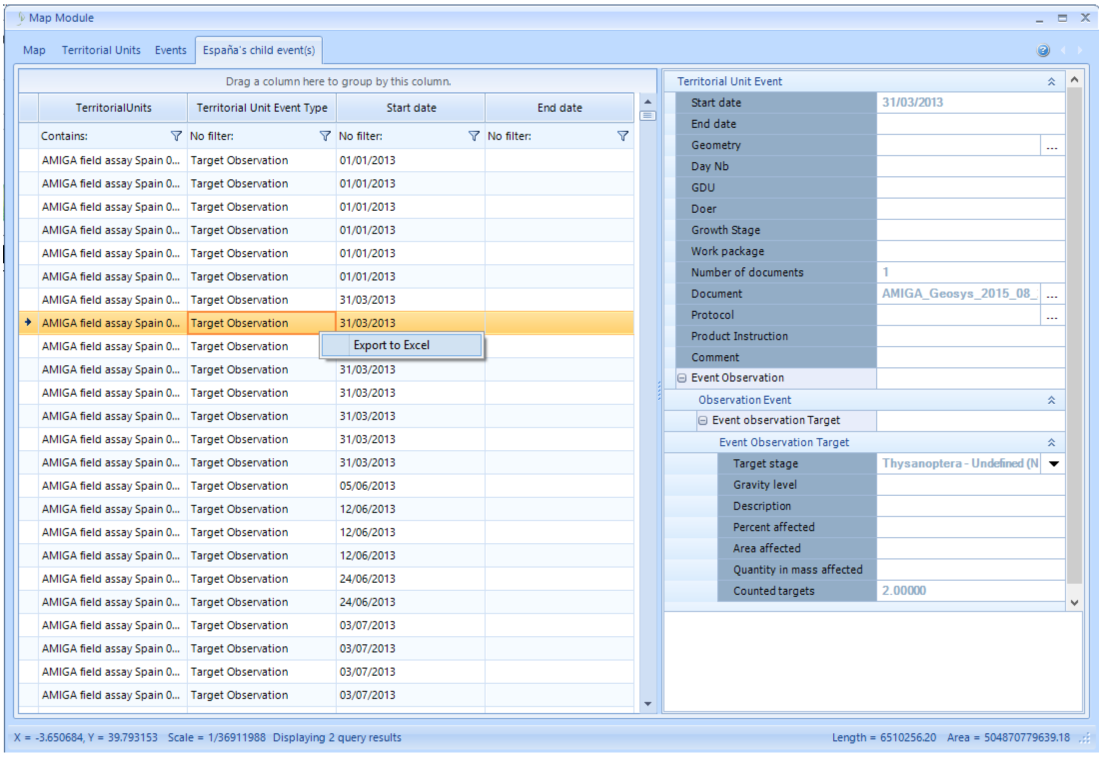


Figure 4-6 Right clicking in the table allows to export the entries to Excel file

4.2 Aggregation and export of data

After selecting one or more datasets (see Figure 4-7), the user clicks the "Export" button and reaches the main screen of the export module. On this screen (see Figure 4-8), the user can reach the screen to select the properties to export, select output file name, preview the data that will be exported and process export.

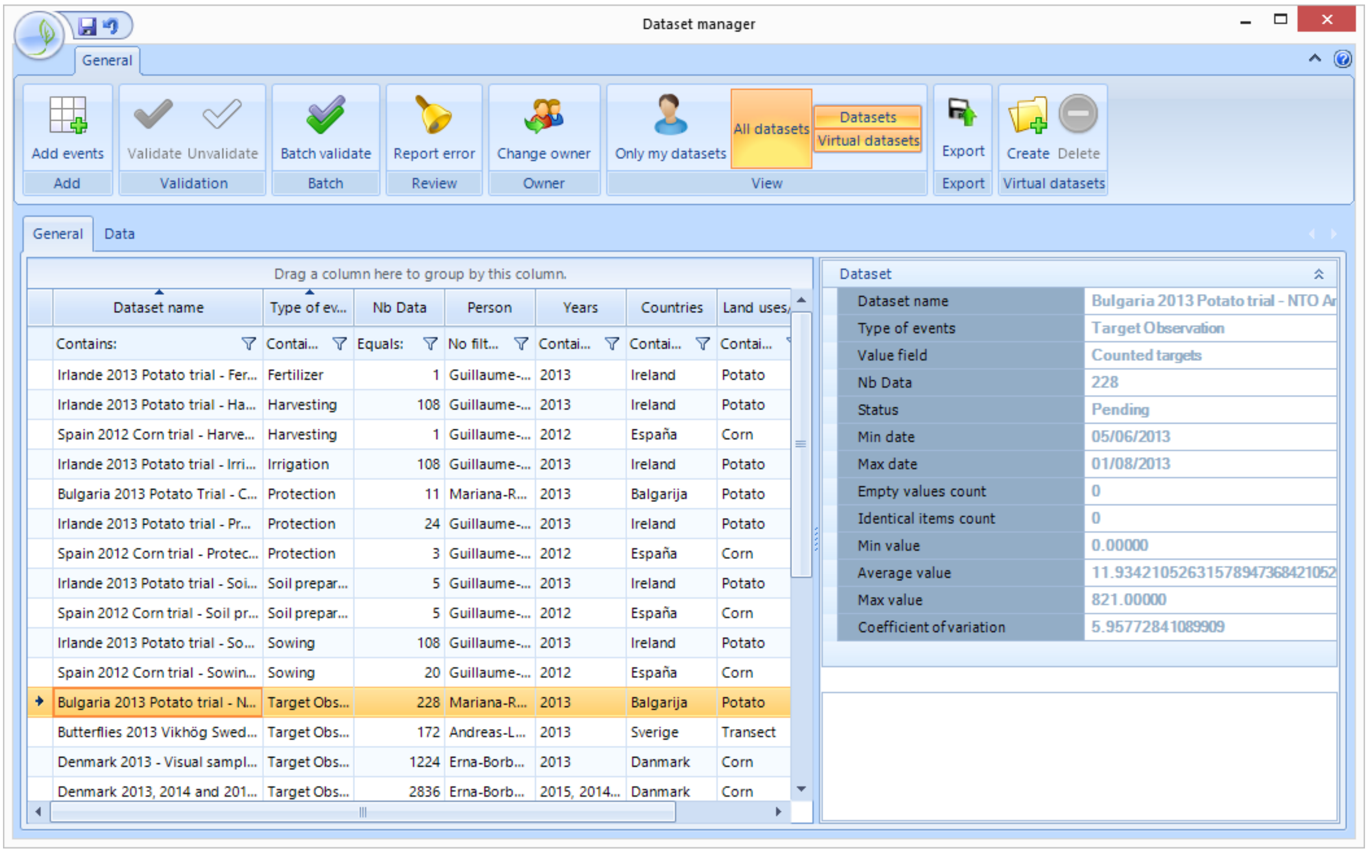


Figure 4-7 Selection of dataset to aggregate and export

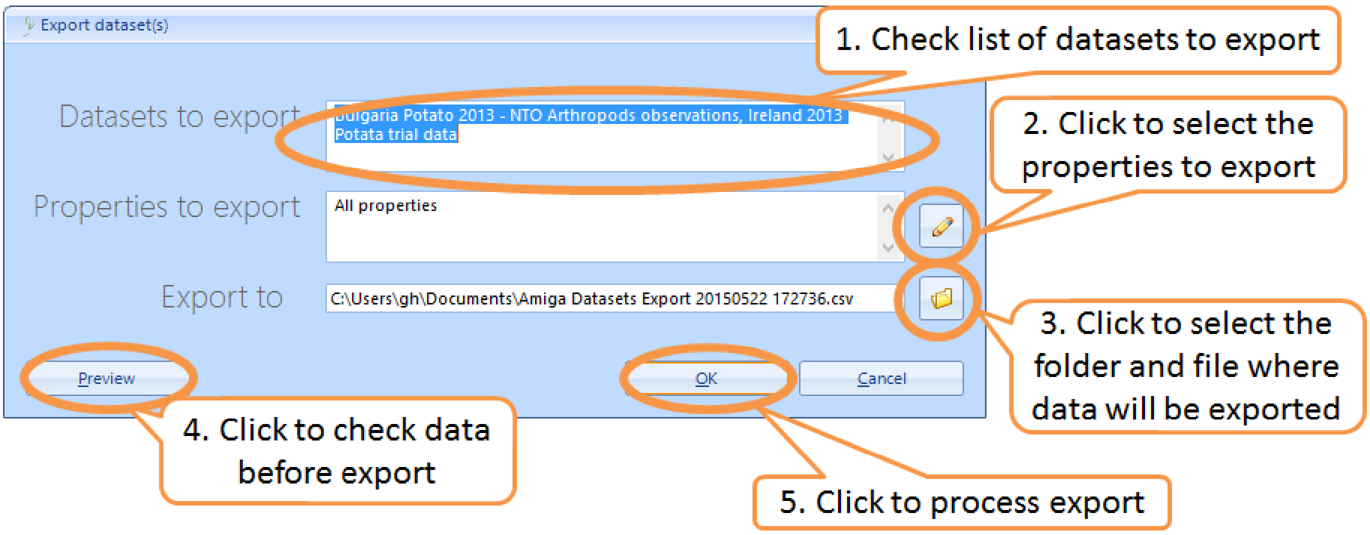


Figure 4-8 Steps to select properties, agregate and export dataset

The user can display the list of properties that can be exported and select them one by one (see Figure 4-9).

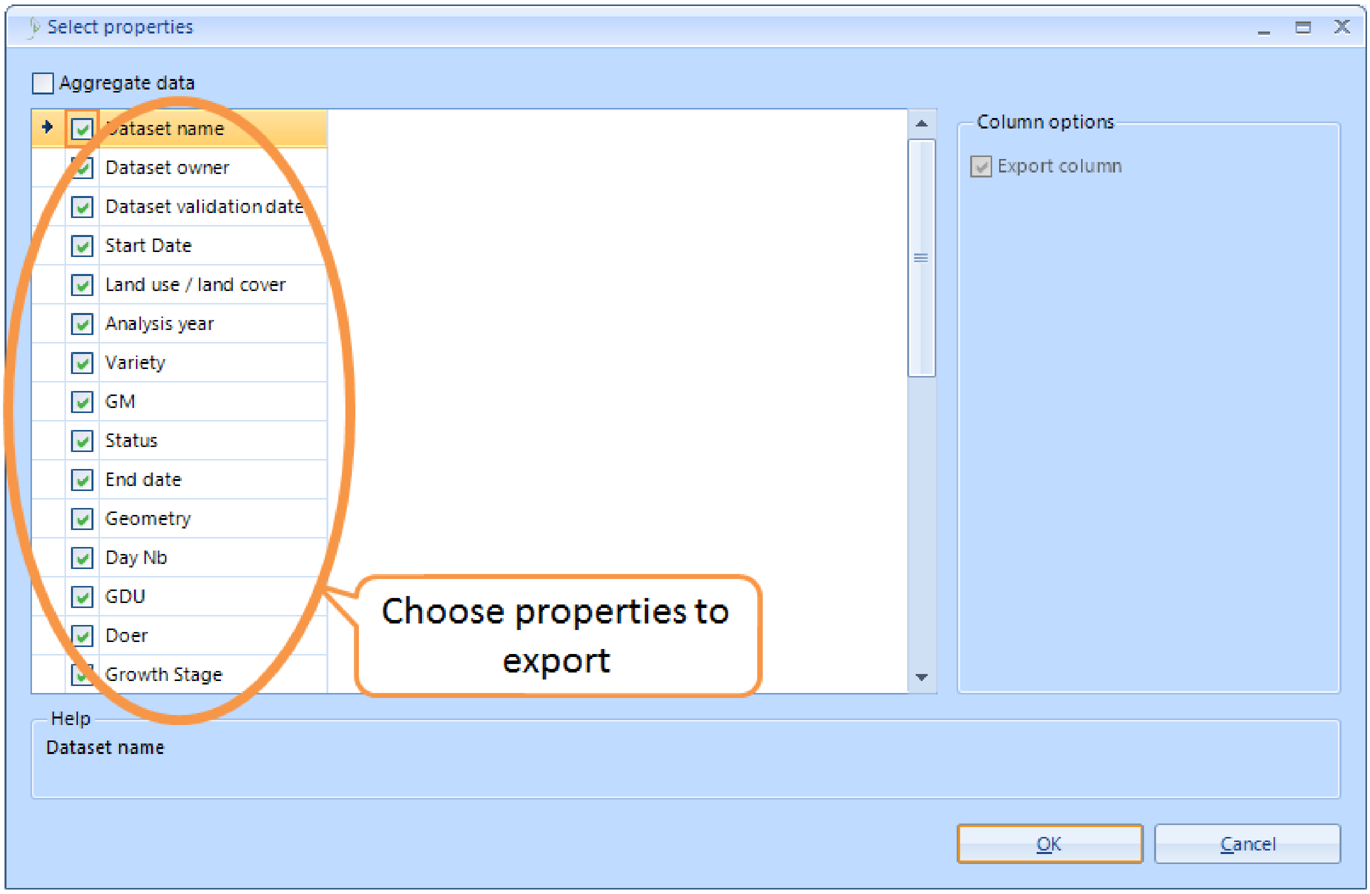


Figure 4-9 Select properties to export

The user can choose to aggregate values and configure aggregation parameters in the window to select properties (see Figure 4-10).

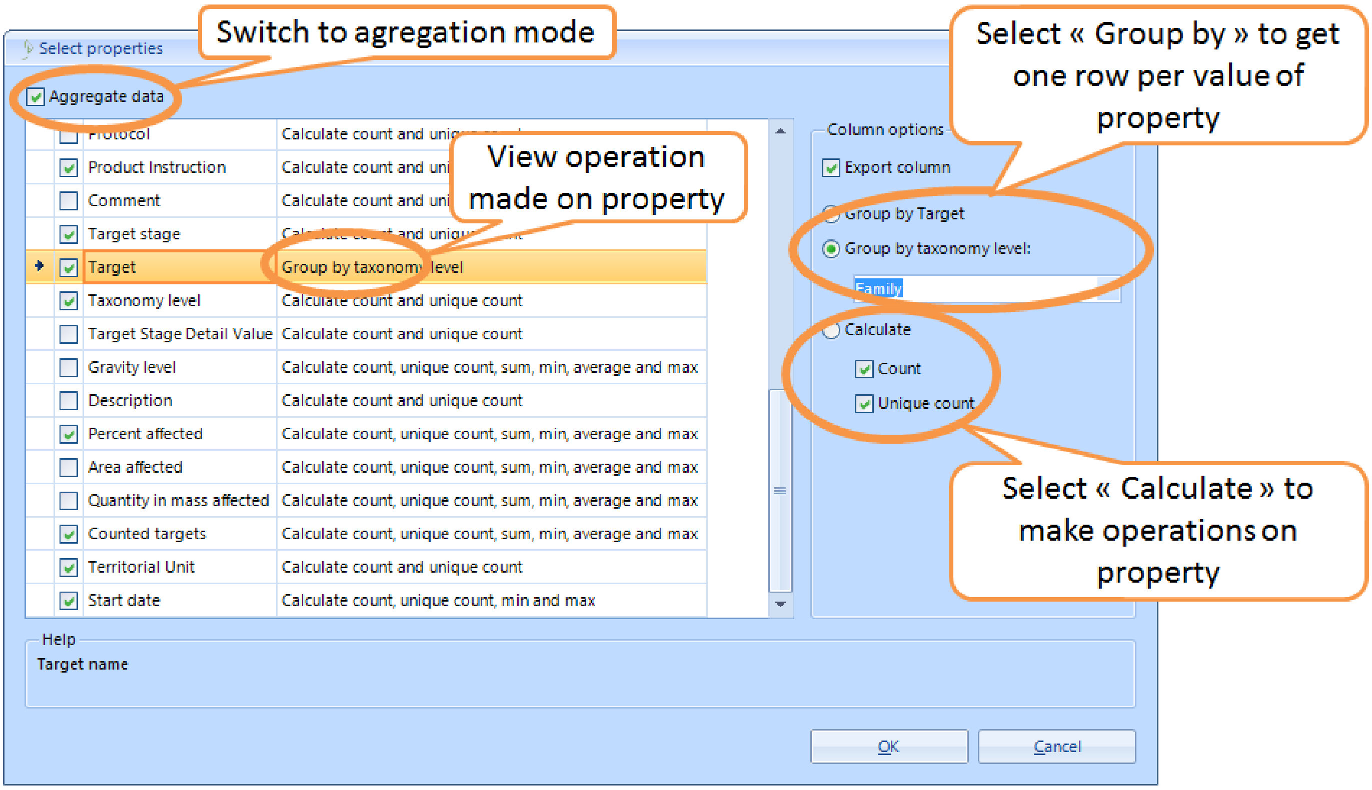


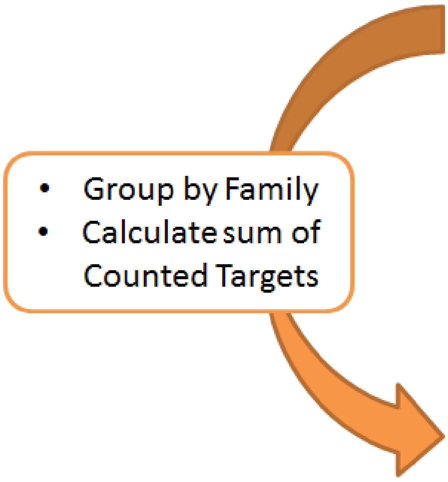
Figure 4-10 Configure agregation rules

Two types of operation can be done on each property while aggregating:

- “Group by”, that will extract rows per values of the selected property;
- “Calculate”, that will allow computing of new values based on the selected properties.

For example, the user might want to get counts of observed targets for each family (see Figure 4-11).

Target	Family	Counted Target
Bembidion quadrimaculatum	Carabidae	94
Formica cinerea	Formicidae	10
Formica rufibarbis	Formicidae	1
Bembidion quadrimaculatum	Carabidae	239
Formica cinerea	Formicidae	3
Formica rufibarbis	Formicidae	2



Family	Counted Target
Carabidae	333
Formicidae	16

Figure 4-11 Sample agregation

4.3 Virtual dataset and statistics for WP9

A user can select several datasets (see Figure 4-7) and combine them in a new and “virtual” dataset. “Virtual” is used here to differentiate original datasets created by the data owner from those created to group data in order to analyze and use them.

The user can view directly some statistics in the software, such as Average value and Coefficient of variation that can be used to build an experimental plan or to set up limits of concerns as described in AMIGA work package 9 “Statistical analysis and design for standardised environmental risk assessment”.

5 Discussion on the challenges faced in loading datasets

This task has successfully built a transversal database that allows AMIGA users to store their datasets on receiving environments, biodiversity and experiments carried out in Europe. Furthermore, the database can in principle host datasets from various sources, including previous EU projects. The collected data can be of many types and the database can be easily extended to store types that might not be already included. The data can be geolocalised or linked to spatial entities so that users of the database can search data at different spatial levels (Country, region, farms …). The data are temporally identified to allow temporal filtering in a search. One other essential point is that data are connected to the owner and can be linked to protocols. The database can be accessed using the Internet and a user can have real-time updated data and interact with dataset producers directly to ask questions or fix errors.

To date, a representative sample of AMIGA datasets has been uploaded to demonstrate the feasibility and potential uses of such a database. The uploading of those datasets has made it possible to identify some issues that should be taken into account in the future development of the database.

The structure and form (template) of the data provided by scientists can be very heterogeneous. Scientists usually collect data in their specific file format and ontology and within their capacities to identify species. Since each dataset may have a specific format, it would be much too expensive and unnecessary to develop import modules for each of them. The consequence is that each dataset needs to comply with the common referential and should be formatted accordingly. Within AMIGA, each researcher sent datasets to “database managers” who performed the task or formatting the data. While this procedure ensured consistency within the project, it relies on the continued presence of an active database manager.

One alternative would be to use the software interface to input observed data directly into the database. It might also be feasible to develop a smartphone or tablet application to allow input directly to the database. Such developments would require new funding.

Querying of the database is another feature that would need refinement. So far, only a limited number of queries have been developed and tested; but depending on the future of the database, additional ones can easily be developed upon requirements of end-users.

Some features can also bring improvements in application developed, such as dataset versioning. Dataset versioning consist in automatically creating a new version of a dataset each time a change is made. This allows users to view older versions of datasets and make comparisons to view changes. This brings greater traceability of work and allows users of datasets to be alerted to changes and to decide to take them or not into account. It would also be interesting to build a data warehouse from database. A data warehouse is a translation of a database that is made to get better performance on specific queries. It could be linked to some Business Intelligence (BI) solution in order to let the user explore data in a more interactive way and create their own reports or exports.

Projects such as AMIGA and previous related research have generated huge amounts of data that have a high potential for reference and re-use. They can be interrogated and analysed to provide new knowledge and can offer a spatial and temporal context for new experiments.

The database built in AMIGA offers the potential to become a much broader repository of data from all experiments and field trials on GM crops in Europe. It would be feasible to connect the spatial capability in AMIGA to other spatial databases, for example on soils, land use and climate. The result would be a unique source of reference for selecting receiving environments, planning experiments, ERA (Environment Risk Assessment) and PMEM (Post-Market Environmental Monitoring).