The EUROHARP data management system

M. Van Liedekerke¹, F. Bouraoui¹, P. Panagos¹ & A. Nogueira²
¹Joint Research Centre of the European Commission, Ispra, Italy.
²Terma Srl, Italy.

Abstract

In scientific support to the implementation of the Water Framework Directive, the EC-funded EUROHARP project studies the harmonisation of tools to quantify nutrient losses from diffuse sources. Within such a project, effective data management is crucial to its success since all other activities draw from it. On the basis of data requirements for all quantification tools, a data protocol has been defined for the compilation, formatting and interchange of data. Detailed map and attribute data for more than 500 parameters, organised in 22 main themes, are collected from 17 catchments. Attribute data are quality checked and, if compliant with the protocol, inserted into a single database.

The database maintains the link between attribute data and corresponding map features, for instance data on land management practices are linked to land management units (polygons) on the land management map. The map data and the database are the sources to a dedicated GIS-like application, which allows the user to inspect visually and in a harmonised manner across catchments all data available in the database. The tool allows typical GIS operations and data drilldown through a cascaded sets of data tables. For time series and table data, graphing functionality has been provided, giving the user a synthesised view on the data. This tool which, developed using royalty free software, is also bundled with the data and is to be shipped to the EUROHARP community to serve as a common data platform. This paper describes in detail the different elements of the Euroharp Data Management System and its potential to serve as a common platform for similar endeavours.

1 Introduction

The EC-funded Euroharp project studies the harmonisation of modelling tools to quantify nutrient losses from diffuse sources. To do so, the performance of nine quantification tools (models) is tested on a number of catchments. A total of 17
catchments are available to the Euroharp modelling community but the projects has been organised such that each tool will be tested on 6 catchments, 3 of which are common to all tools. The final outcome of the project will be a 'toolbox' which should provide guidance to decision makers for the selection of suitable tools for nutrient loss quantification, an exercise which will be required by Member State (MS) authorities in the context of the WFD implementation. Euroharp is one project within an EC-organized cluster of projects, which touch upon similar issues (CATCHMOD cluster).

In order to run the models effectively a common base of data is needed in order to allow comparison among models and catchments. The project is organised such that a dedicated Data Management workpackage ensures proper catchment data collection before any model is run. The collection of the data is not only useful for running the models but will also indicate whether catchment data owners can easily comply with a set of data requirements for specific tools and will thus indicate the feasibility, in a WFD context, for MS authorities to collect such data. When successful, the database and system, as developed under the project, could also serve as a model for WFD implementation.

Most of the models require data in a spatial context. Euroharp data management is about the collection of maps and attribute data (related to the map features) from catchment data owners, about inserting these data into a system which acts as a single data reference point for the project, and about means of distributing these data in a usable form to the modellers (as separate map and attribute files, as a database and also as data embedded in a standalone tool).

This paper is structured as following: section 2 describes the data requirements (as dictated by the models itself) and the system requirements (as dictated by all users). Section 3 describes how these requirements have been turned into a number of system components: the data model, the database, data entry templates, the standalone tool and Web access. In section 4, the data collection and insertion process is explained and section 5 draws preliminary conclusions and outlines areas for future work.

2 Data and system requirements

2.1 Data requirements

The 9 model owners were asked to specify which precise data (parameters) were required in order to run their model. Some of the required data were common to all models; some were specific to a particular model. A list of more than 500 parameters was compiled which could be split in 22 subsets according to themes (Soil, Weather, Surface Water Quality and Quantity observations, Land Cover, Land Management, etc.), each theme carrying a map component and a set of attributes related directly or indirectly to the map features. E.g. parameters for the Soil theme include soil-type and soil-layers, each soil-type consisting of one or more layers, each soil-type having a number of characteristics (such as ParentMaterial, etc.) and each layer corresponding with a set of water characteristic curves and textural properties.
2.2 System requirements

From the start of the project, it was decided that the data should be embedded in a data management system, the components of which are illustrated in figure 1.

![Diagram of Euroharp data management system](image)

Figure 1: Euroharp data management system.

At the core of the system is the central catchment database (CDB) hosting both map and attribute data as provided by all catchment data owners. This database is seen as the single reference point to extract data from for further use in the standalone tool and for Web access. The standalone tool (SAT) was envisioned as a GIS like application, holding all or subsets of the data and giving project partners the opportunity to view both map and attribute data within a single application and across different catchments. A Web component is foreseen to distribute to users selective portions of all available data on a password-protected basis.

3 Development

Given the complexity of the required data, given the fact that project partners have at their disposal different systems for map and attribute data storage and
production, and finally given the requirement that all data needed to be stored in a single system and in harmonised way, it was thought that the most effective way to proceed to data collection was to specify to data providers data delivery formats which are a fair trade-off between own formats and formats required for the database.

Maps should be delivered in one of a number of possible formats. Attribute data should be delivered through a number of prepared spreadsheet documents. The link between maps and attribute data is made through a "linking attribute": individual features are labelled with a linking attribute value in the map and these values are reported on the spreadsheets to make the correspondence with the attribute data. The values for linking are up to the data provider. This freedom is reflected in the design and implementation of the database.

3.1 Object Model

On the basis of the data requirements, a suitable Object Model was constructed (using Rational Rose) along the themes identified in the requirements phase; this model served conceptualisation and visual representation of the future database, suitable for discussion with the 'domain experts'; it is a platform- and language-neutral interface which identifies the logical entities of the project, organises them by groups and defines their relationships and properties.

3.2 Data Model

On the basis of the Object Model, a Data Model was produced, first through automatic generation from the Object Model, then manually for further refinement and optimisation; the Data Model was then implemented in a RDBMS (MS SQL Server). During the course of the project, the Data Model was adapted a few times to take into account small differences in the interpretation of the data requirements.

Both in Object and Data Model, it was assumed that map data would reside as independent files in a directory structure while attribute data would entirely be stored in the database tables; this is reflected in the way the objects and data tables are constructed.

On the basis of the resulting database tables, spreadsheet (MS Excel) templates were constructed and a Data Protocol was specified; the protocol included specification of map formats, the spreadsheet files and a description of how data providers should make the link between maps and attribute data. There is one Excel document per theme, each document consisting of many sheets, each sheet corresponding more or less with a table in the database.

For instance for the Soil theme, the Soil.xls consists of six sheets; one sheet describes all possible SoilTypes for the catchment, in terms of a number of attributes (columns in the sheet). Another sheet describes per SoilType all possible SoilLayers, each layer in terms of a number of attributes. Three more sheets allow the provider to specify for each SoilLayer hydrology related curves;
one sheet is a lookup table for ParentMaterial, the values of which are to be used in the ParentMaterial column of the SoilTypes sheet.

3.3 Database structure

It is beyond the scope of this paper to go into each aspect of the database, however through an example for the Soil theme, it will be shown how data are structured into the database.

Assume a Soil map composed of polygons, e.g. an ESRI shapefile; each polygon carries an internal ID. A shape file has also a corresponding Feature data file (FDF) that contains data for the individual features on the map. It was
required by the data protocol that one data-column of the FDF would serve as the *Linking Attribute* (LA), containing values which maintain the correspondence between the map features and the attribute data in the Excel attribute files.

Each feature on the map has an internal ID (which can not be directly seen in the FDF but which can be extracted through programming); in terms of the database this ID is generally called the *GeoLinkFeature* (GLF); also in terms of the database, the LA is called *GeoLinkAttribute* (GLA). The GLF and GLA values of each map feature for each catchment and each theme are stored in the database in one table called FEATURE. So the table FEATURE contains the following columns:
- a record Feature_ID identifying uniquely the entry of the catchment-theme-GLF-GLA values;
- the GeoLinkFeature;
- the GeoLinkAttribute;
- an ID referring to a (theme, catchment) combination in the CatchmentTheme table.

Most attribute data in the Excel files correspond to features on the map, to which they refer with a LA value. For instance: in a polygon soil map polygons with internal ID 2000, 2001 and 2004 could carry the soiltype id = 'A' (linking attribute value). In the SoilTypes sheet of the Soil Excel file, a SoilType with ID = 'A' would be defined, and store for it many characteristics (HOSTCLASS, etc.)

In the database, the link (relationship) between the table FEATURE and the theme Attribute data stored in the database is maintained through a Theme related Feature Table (TFT table). For each Theme there is a TFT table to make the link between the table FEATURE and the main attribute data table for that theme. For theme <T>, the TFT is called <T>Feature. Examples: for the Soil theme there is the SoilFeature TFT, for the Administrative theme there is AdministrativeFeature TFT.

Coming back to the example for the Soil theme:
Assume a Soil polygon soil map with polygons with internal ID 2000, 2001 and 2004 which carry the A as the linking attribute value and polygons with internal ID 2002, 2003 and 2005 which carry the B as the linking attribute value. In the SoilTypes sheet of the Soil Excel file, there are 2 rows, respectively carrying A and B in the SoilType ID column and other data in the other columns.

Table 1. Part of database table Feature related to Soils theme.

<table>
<thead>
<tr>
<th>Feature_ID</th>
<th>CatchmentTheme_ID</th>
<th>GeoLinkFeature</th>
<th>GeoLinkAttribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345</td>
<td>12</td>
<td>2000</td>
<td>A</td>
</tr>
<tr>
<td>12346</td>
<td>12</td>
<td>2001</td>
<td>A</td>
</tr>
<tr>
<td>12347</td>
<td>12</td>
<td>2002</td>
<td>B</td>
</tr>
<tr>
<td>12348</td>
<td>12</td>
<td>2003</td>
<td>B</td>
</tr>
<tr>
<td>12349</td>
<td>12</td>
<td>2004</td>
<td>A</td>
</tr>
</tbody>
</table>

The values in the Feature_ID column are internal database identifiers and have no relation with the map or attribute data.
Table 2. Part of the Theme related Feature Table (TFT).

<table>
<thead>
<tr>
<th>Feature_ID</th>
<th>SoilType_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>12345</td>
<td>764</td>
</tr>
<tr>
<td>12346</td>
<td>764</td>
</tr>
<tr>
<td>12347</td>
<td>765</td>
</tr>
<tr>
<td>12348</td>
<td>765</td>
</tr>
<tr>
<td>12349</td>
<td>764</td>
</tr>
</tbody>
</table>

Table 3. Part of the SoilTypes Basic Table.

<table>
<thead>
<tr>
<th>SoilType_ID</th>
<th>Name</th>
<th>Description</th>
<th>GleyedLayer_depth</th>
<th>HostClass</th>
</tr>
</thead>
<tbody>
<tr>
<td>764</td>
<td>Soil type A</td>
<td>Sandy loam</td>
<td>30</td>
<td>...</td>
</tr>
<tr>
<td>765</td>
<td>Soil type B</td>
<td>Sandy clay</td>
<td>43</td>
<td>...</td>
</tr>
</tbody>
</table>

The ID (record identifier, SoilType_ID) in the basic attribute table is generated by the system (not provided by data Owners). The ID (record identifier) stored in the table Feature is also generated automatically by the system. The ID provided by the data owner is stored in the Feature table as the GeoLinkAttribute value (read from the Excel files). The Link (association) between the Attribute Basic Table and the Feature Table is done through the TFT table.

3.4 Data reporting/statistics

In support of the database administration and in order to give the Euroharp community the possibility to track progress in the updates of the database with new catchment data, a number of data reporting/statistics routines where added to the system which report per catchment on data availability at various levels: are there any data for a theme, are there any data for a certain database table (corresponding to a spreadsheet) within a specific theme, are there any data for a certain parameter (column in database) within a specific theme.

3.5 Stand Alone Tool (SAT) - GISViewer

The purpose of the Euroharp SAT application, dubbed GISViewer, is to give to the user a map-based tool to inspect map and attribute data that are part of the database. It offers a map-based interface, which allows the user to:
- Select a catchment of interest;
- Select map layers for display within the selected catchment;
- Navigate the maps by panning and zooming;
- Select among the visible layers/themes an 'Active Theme' for which attribute data can be retrieved through the selection of map features on the map;
- View the attribute data for the selected map features of the selected Active Theme, in a grid containing values in the cells of single-valued data columns.
and icons in the cells of multiple-valued columns (e.g. in case of tables and time series);
- View some of the data as charts (e.g. time-series)
- View attribute data related to the catchment but not to particular features on the map (e.g. a table of SoilTypes defined for the catchment)
- View look-up data which are valid across catchments (e.g. all CropTypes)

The tool has been developed in Visual Basic using ESRI's MapObjects library for including the GIS functionality. The development started from scratch and did not build on an existing GIS package because it could not be taken for granted that EuroHarp partners had a specific GIS package at their disposal and because it is not straightforward to program additional functionality (e.g. viewing data in tables or as charts) on existing GIS applications. The GISViewer is under further development to add additional functionality such as import and export facilities and scenario comparison.

Figure 3: Snapshots from the GISViewer.

The GISViewer can be configured such that it reads data directly from the central SQL Server database or from a stand-alone MS Access database obtained through export from the central database. Other GISViewer features related to the 'look' of maps are configurable as well.
So far, the GISViewer has been distributed by Web, requiring download of 20 MB for the application and an additional 20 to 30 MB for each catchment (data in MS Access plus separate shapefiles).

3.6 Web-based data access

A password protected Web portal is under construction to give access to the data in the central database and to the delivered maps and attribute files. Concerning file based access, for each theme it is possible to download the original map and attribute files as received from the catchment data provider, a quality-controlled version (used directly for data import in the database) of the same files and set of files which comment on the data. Additionally, the attribute data all catchments are exported to a Microsoft Access database, for download and use by the modellers. This Access database comes with a large number of example queries.

4 Data collection and insertion

The Data Protocol document details the procedure and format for exchanging data. Received data not complying with the protocol has to be returned to the sender for negotiation with the database administrators and for correction. As each catchment data owner has his own data conventions and local specifications, it sometimes has proven to be difficult to adopt the rules set by the data protocol.

The map and attribute data finally ready to be imported into the database may be the result of 2 to 10 rounds of negotiation. Once compliant with the data protocol, the data import process into the database starts following a strict procedure, partly automatic, partly manual.

Technically, the SQL Server allows constructing Data Transformation Packages (DTPs), a software mechanism to import data in an SQL Server database from another format. Those packages have been developed for each table and operate as a bridge, which transfers the data from the Excel Templates to the EuroHarp database. Moreover, the packages act as a first quality control mechanism, which doesn't allow data input if the data are not compliant according to format definitions. Apart from the pure attribute data, also parts of the map data have to be imported in the database, so part of the input process is to export the GeoLinkFeature (GLF) and the GeoLinkAttribute (GLA) from the map file to intermediate files and to import these data in FT and TFT; again specific DTPs have been created for that.

4.1 Status of data input

It should be mentioned that catchment data owners were not obliged to send data for all themes; some catchments provided more, other less data. In the current phase of the project, data have been imported successfully for 9 out of the 17 catchments (status: complete). For 6 catchments, data have been inspected and negotiations are still under way to reach final conclusion (status: pending). For 2 catchments no data have been provided yet. The total status is also reflected in
the database reports/statistics. For the 10 complete catchments, more than 750,000 rows and 4,000,000 data values have been entered into the database.

5 Conclusions

In order to be successful, the Euroharp project needs rigorous and timely data management because consecutive work packages in the project require the data for modelling. This paper has described how data management has been tackled and which products have been developed: the central database, the standalone tool to view map and attribute data, the Web access to data files and database, the tools to import to and export from the central database and the reporting facilities. It is particularly useful to have one reference point for the data since models need to be compared between each other and across catchments. Data distribution (as files, as database and as standalone tool) has only recently started with data for two Core Catchments, so it is too early for reporting user comments. Whether the approach of data management, the data protocol and the implemented products could be of future use to the WFD is still to be evaluated; to this end, a questionnaire, distributed among the catchment data owners is under way and will allow evaluation of how easy or difficult it is to collect the data as specified in the project. Also, at some stage later in the project, data issues should be discussed with peers in other projects in or outside the CATCHMOD cluster. Immediate future work will complete the data collection and insertion into the database, the extension of the GISViewer with additional functionality and the provision of advanced Web access to the database. In a second stage, all data will be rigorously quality checked (e.g. removal or correction of outliers or impossible values) in negotiation with the catchment data owners, leading to an updated and definite dataset for the Euroharp project.

6 Acknowledgements

This work was carried out in the frame of EUROHARP (Towards European Harmonised Procedures for Quantification of Nutrient Losses from Diffuse Sources) financed by the European Commission under contract EVK-CT-2001-00096.