Life08 ENV/IT/ 000399

Report on Data Stored in DB

Peterseil Johannes

Environment Agency Austria

Life08 ENV/IT/0003999

< Blank page >

Report on data stored in the database

Status report 2011

Deliverable number	***PD_A1.4.4***
Delivery date	***11/2011***
Status	*Rev 0.5*
Authors	*Johannes Peterseil, Tomas Kliment, Alessandro Oggioni, Herbert Schentz, Paola Carrara*

With the contribution of the
LIFE financial instrument of the
European Community

Title	Report on data stored in the database
Creator	*Johannes Peterseil (EAA), Tomas Kliment (CNR-ISMAR), Alessandro Oggioni (CNR-IREA), Herbert Schentz (EAA), Paola Carrara (CNR-IREA)*
Creation date	30/07/2011
Date of last revision	30/11/2011
Subject	
Status	☐ Draft ☒ Final
Publisher	EnvEurope project
Type	Text
Description	This document describes the formats and flows for the collection of real data from monitoring and observation used in the analysis actions of the project. The report reflects the current status and an outlook on the planned improvements of the data management system.
Contributor	
Revised by	Johannes Peterseil (EAA)
Format	Doc
Source	
Rights	☐ Restricted ☒ Public
Identifier	PD_A1_D1_4_4_Peterseil_et_al - DataManagementStatus - Rev0.5.docx
Language	En
Relation	
Coverage	Project duration

These are Dublin Core metadata elements. See for more details and examples http://www.dublincore.org/

Contact information:

Johannes Peterseil – Editor
Umweltbundesamt GmbH (EAA)
Spittelauer Lände 5
1090 Vienna
Austria
E-mail: johannes.peterseil@umweltbundesamt.at
Tel +43 1 31304 3443
Fax +43 1 31304 3533

Herbert Schentz
Umweltbundesamt GmbH (EAA)
Spittelauer Lände 5
1090 Vienna
Austria
E-mail: michael.mirtl@umweltbundesamt.at
Tel +43 1 31304 5308
Fax +43 1 31304 3533

Paola Carrara
National Research Council (CNR-IREA, Uos Milano)
Via Bassini, 15
I-20133 Milano
Italy
E-mail: carrara.p@irea.cnr.it
Tel +39 02 23699 295
Fax +39 02 23699 300

Alessandro Oggioni
National Research Council (CNR-IREA, Uos Milano)
Via Bassini, 15
I-20133 Milano
Italy
E-mail: oggioni.a@irea.cnr.it
Tel +39 02 23699 299
Fax +39 02 23699 300

Tomas Kliment
National Research Council (CNR-ISMAR)
Arsenale-Tesa 104
Castello 2737/F
I-30122 Venezia
Italy
E-mail: tomas.kliment@gmail.com

Abbreviations

CB	Coordinating Beneficiary	**SC**	Steering Committee
AB	Associated beneficiary	**AC**	Advisory Committee
AR	Action responsible	**QA/QC C**	Quality Assurance and Control Committee

Version	Date	Author	Task
0.1	2011-11-08	Peterseil	Compilation of the existing documents into a consistent report
0.2	2011-11-16	Core Team A1	Definition of the structure
0.3	2011-11-23	Peterseil	Inclusion of comments
0.4	2011-11-28	Carrara	Revision from CNR
0.5	2011-11-30	Peterseil	Summarising the comments from the authors and final version

Content

< Blank page >

1 Introduction

In order to enable data discovery, interpretation and, if applicable, data analysis, information about the "how, where, when, what, who …" needs to be captured in an accessible and understandable manner. If adequate metadata is available, this data can be reused after years or decades, either on its own or in combination with data from other sources (Karasti & Baker 2008, Karasti et al. 2007). Obtaining sufficient and well described data is therefore a challenging task.

The EnvEurope project joins beneficiaries with a highly diverse background in data management ranging from simple file based data storage to highly developed web based data management solutions. This is also true for the LTER Europe as well as for the international ILTER network. Providing widely accepted solutions in data collection and data exchange, as well as data documentation are therefore important tasks also on this level. Despite the wide range of solutions for data integration and exchange not many examples are in place to be adopted of in the domain of long term ecological research. On the European as well as on the global level the focus is mainly on the provision of metadata with EML (Ecological Markup Language, described by Michener et al. 1997) being one of the well-known examples in the domain of long term ecological research on the dataset level. On the site level in Europe the LTER InfoBase provides metadata about sites which describe the LTER Sites and LTSER Platforms as a whole (Adamescu et al. 2010, Vadineanu et al. 2006) and which form the site network of ALTER-Net and LTER Europe (see Haberl et al. 2006, Mirtl & Krauze 2007).

Action 1 Data Management of EnvEurope (from now on A1) tries to setup a working use case for data storage, management and exchange for the long term ecological monitoring in Europe. The results of this work can be used for the further work and implementation for SEIS, the Shared Environmental Information System, on the European level.

Using a step wise procedure in its implementation the solution tries to meet the short term requirements and data needs on the one side and the long term vision towards a service-based architecture on the other.

For the design of the proposed solution the following criteria were adopted:

a) Capability - the solution must meet the existing requirements, but should also pave the path to future needs;
b) Costs - software must be freeware. But nevertheless there are not only the direct costs, which have to be taken in account, but also all indirect costs, within the lifecycle of application, where the installation and maintenance are big parts;
c) Availability - easiness for beneficiaries of EnvEurope to download, install and apply the solution, as well as to obtain support. Therefore also the existing skills to install and run the software (see below) determine the availability of the solution.
d) Skills within the community - Even small institutes which have to provide access to their data must be able to install the software.
e) Experiences - experience within the community or at least within close communities are essential, in order not to start from the scratch

Therefore the current report focuses on the description of the solution adopted and the implemented work flow limited to a file-based data collection using a Data Reporting Format defined for EnvEurope. This approach will be further developed in the forthcoming period of the project towards more advanced solutions such as a service-based data exchange.

The report comprises the work of several activities within Action 1:

- A1.2.2 Collection and review of existing data management tools
- A1.3.2 Collection and review of data models from the beneficiaries
- A1.3.3 Analysis of existing data models

- A1.3.4 Establishment of a generic data model and application schema for data exchange
- A1.4.1 Setup data flows and data management structures
- A1.4.3 Identification of relevant datasets and data sources

The tasks were carried out either in small working groups defining the core models as well as in discussions with all beneficiaries. The results were discussed within the technical meetings, which were attended by most of the beneficiaries. Testing the conversion of local data structures to the requested Data Reporting Format was done by the beneficiaries.

The current report includes an overview on the data management solutions used by the beneficiaries as the basis for further considerations for data exchange, the data reporting format used for the data collection and exchange, the current status of data uploaded as well as an outlook to further steps in the development of the data management framework in EnvEurope.

2 Terms and Definitions

The following terms are used in the context of the report and are therefore explained.

Community or LTER-Europe Community
reflects the community composed by all Long Term Ecological Research sites. It focuses on different types of ecosystems, i.e., marine, lacustrine, riverine and terrestrial. The mission of Long Term community is: to track and understand the effects of global, regional and local changes on socio-ecological systems and their feedbacks to environment and society; to provide recommendations and support for solving current and future environmental problems (http://www.lter-europe.net/).

Data management
the term data management is referring to all methods of storing, managing and archiving data being digital or analogue.

Dataset
is a collection of data. In the LTER compound the dataset is a collection of single parameters stored in a specific site. The dataset is not time dependent; each dataset can cover different time period with different frequency. The term dataset is describing a concrete dataset of an observation or a sum of observations (e.g. vegetation releveés from permanent plots, soil temperature measurements from a plot, etc.).

Metadata
are data about the dataset; data providing information about one or more aspects of the data. Metadata are used to search, locate, evaluate and discovery a dataset.

Site
The term site is referring to an observation place, which is defined and listed in the LTER InfoBase (see Site Identifier).

SOS (Sensor Observation Service)
is a web services proposed by the Open Geospatial Consortium (OCG), a global standardisation initiative for geographic applications and data to provide observation data (time and space).

Station
a station is the location where an ecological phenomenon (e.g. soil temperature) is observed or monitored within a site. A spatial group of observations can be repeated in time at the same station. Examples of stations are sampling plots, observation plots, and plots with sensors installed, etc.

WFS (Web Feature Service)
is a web services proposed by the Open Geospatial Consortium (OCG), a global standardisation initiative for geographic applications and data to provide spatial data.

WMS (Web Mapping Service)
is a web services proposed by the Open Geospatial Consortium (OCG), a global standardisation initiative for geographic applications and data to provide spatial data as map.

3 Current status

As a first step toward an integrated data management and data collection within the EnvEurope project a collection of the existing data management solution within the consortium was done. This contributes to provide the basis to select the best data management options for the project. The results allow for a targeted solution for the short term as well as for a long term implementation plan and outlook beyond the runtime of the project.

To perform this task a questionnaire was defined to collect this information from the beneficiaries. The questionnaire is divided into six parts and tries to collect information and expertise of EnvEurope beneficiaries about the data management done at the sites which are part of the EnvEurope site network. The questionnaire is attached as Annex to the internal report.

- The first part deals with questions about the *Background information* (Q1-Q8). This collects information about the contributor and the list of sites for which the questionnaire was answered. The list of sites is based on the verified site list of EnvEurope. Because of the organisation of the national LTER networks in some cases information on a sub site level was provided.
- The second part deals with questions about the current *Data management* (Q9-Q10). This includes questions about the format as well as the tools and software used to manage the data.
- The third part deals with questions about the *Data Access and Distribution* (Q11-Q13). The questions aimed to get an overview how data are technically shared and which were the main target user groups for the data.
- The fourth part deals with questions about the *Data Sharing Policy* (Q14-Q18). The questions aimed to get an overview how data sharing policy is implemented and which share and cost models are applied at the different sites.
- The fifth part deals with questions about *Requirements* (Q19-21). These questions aim to get an overview about the expectations and requirements from the associated beneficiaries to the action A1 on Data Management.
- The sixth part allowed *General Comments* (Q22) about the questionnaire in plain text format.

The questions were numbered and short explanations for each question were given. In addition a prefilled example of the questionnaire about a concrete site in the EnvEurope site network was provided. If no other option was indicated, multiple answers were possible. A comment field was provided to give further explanations about the answer if needed.

One questionnaire could be provided for a number of sites which showed a similar data management or the data management was done by the same institution. In this case each answer was only counted one.

The questionnaires were collected centrally and the results entered in a Microsoft Access Database to do the further analysis. During the import routine a consistency check of the answers was made and missing answers were inserted with “no” or “N/A”. The resulting database is provided in the members’ area of the EnvEurope web page.

In total n=45 questionnaires were sent back for the analysis. The overall response rate to the questionnaire was 100% of the partner countries in the EnvEurope project. For some of the sites the information about the data management was given on a sub site level as the sites listed in the EnvEurope site network were collections of different sub sites. This was the case for some sites of

Italy. The total number of sites for which information on data management was provided is 58. These are 93% of the EnvEurope sites.

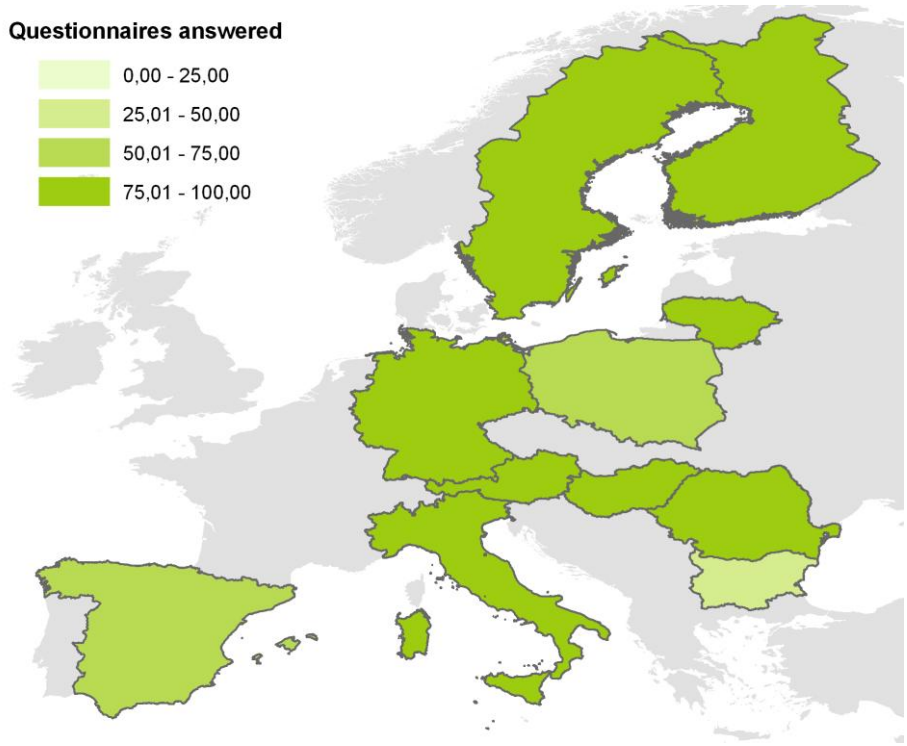

Figure 1 Overview on the replies to the questionnaire. The figure shows the proportion of questionnaires for the sites in relation to the total number of sites listed in the EnvEurope Site Network.

3.1 Data management

The questionnaire collected information on the data formats and available staff for data management. The following section provides the result on that.

Data format

All data managed at the site level are provided in digital format. Only some of the oldest data are in printed versions or in the original recording sheets. This was also the result of a quick overview at the A1 technical meeting in Vienna.

Nearly all sites provide data in Excel format. These files are structured according to the needs of the project and are structured to extract the information needed. About 60% provide data in structured databases, which allow a systematic inquiry of the data. Potentially these data sources are also able to be connected online via services. The database systems used are either Access (8 comments) but also Oracle (4 comments), PostgreSQL (3 comments), MySQL (1 comment) and Microsoft SQL Server (1 comment).

About 40% provide data as unstructured text files which are difficult to analyse. About 80% of questionnaires provide spatial information which are either organised as simple shape files (about 60%) or in a spatial database (about 20%). The main software listed are ArcGIS Geodatabase, ArcSDE and PostGIS. The GIS software used ranged from ESRI ArcGIS, ESRI ArcView to the open source products as GRASS and QGIS.

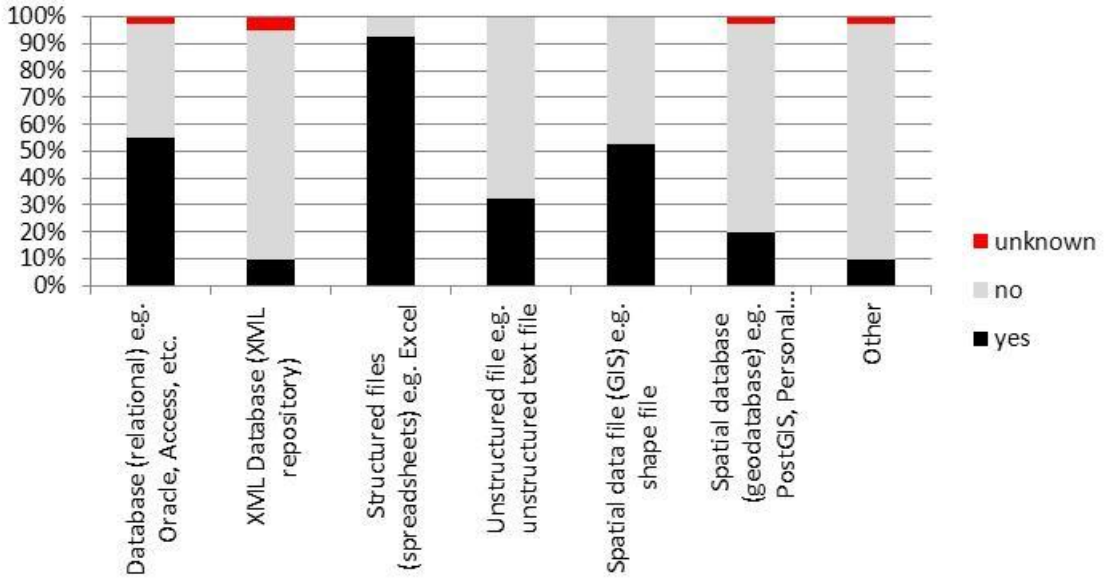

Figure 2 Question 9.1.1 Data provided in digital format?

Despite the digital data management which is present at nearly all of the sites, quite a number of sites also provide data in non-digital format or in proprietary formats which can be exchanged only with difficulties. About 25% of the sites provide data in paper or printed format which are mainly for historic data.

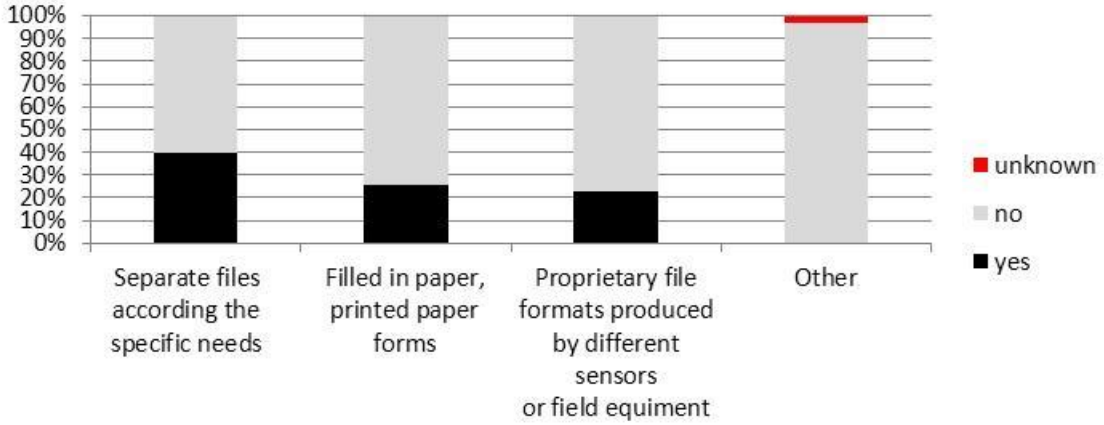

Figure 3 Question 9.2.1 Data provided in non-digital or proprietary formats?

Regarding digitising attempts for older data which are present in non-digital format, about 35% of the replies answered for yes. This is planned either in on-going projects within the institutions or partly in EnvEurope for requested data.

Regarding the architecture of the data management, about 65% have either a central data management (~20%) or the data sources are distributed within the same institution (~46%). This situation at least allows an easier access to the data regarding the contacts and access rights. About 30% of the questionnaires answered replied that the data are distributed over multiple institutions.

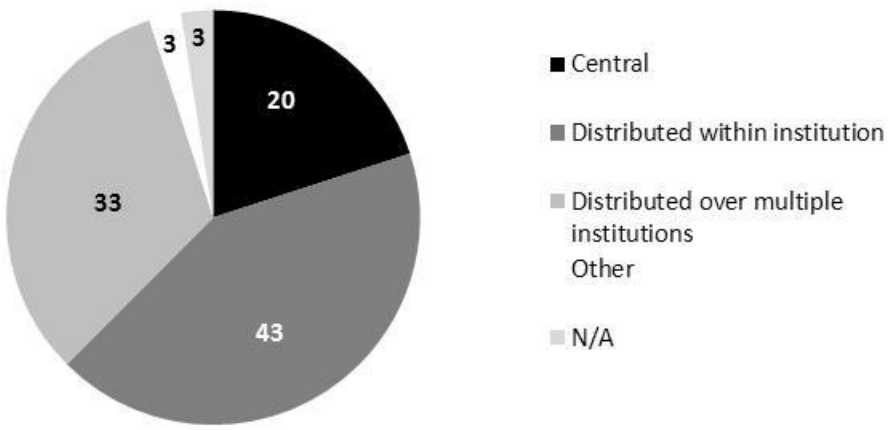

Figure 4 Question 9.1.2 Places of Data Storage in % of the total answers. Total answer n=37

If the data are distributed over multiple institutions the number of places ranges from a view up to more than 30. This also reflects quite well the heterogeneous situation of the data management within the LTER institutions. This fragmentation of the data sources can either reflect the data ownership or the thematic orientation, as e.g. different databases can occur for different topics (e.g. air quality measurements, biodiversity, etc.).

Further on, especially in the situation of ICP Forest and ICP Integrated Monitoring, a separation of raw data versus aggregation data on a monthly basis is given for some of the databases. The raw data mostly remain in the realm of the institution doing the monitoring whereas the aggregated data are in the realm of the central data management. Therefore the data are also distributed over multiple institutions.

Data management staff

Looking on the availability of staff for the data management, most of the institutions show a "scientific oriented approach" in the organisation of the data management. This means that scientists are managing their data themselves. Most of the institutions managing long term data have mainly scientific staff for doing this job (~80%). About 49% also have technical staff at the scientific departments in dealing with data management topics. Only 23% have a separate data management department dealing with the management of long term monitoring and research data.

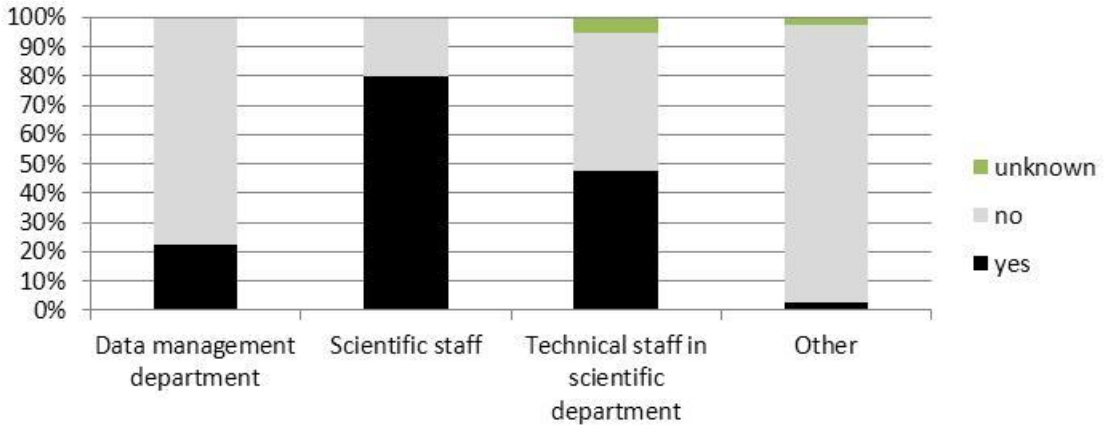

Figure 5 Question 10 – Staff available for data management

This pretty much reflects the situation in the long term monitoring community in Europe where most of the responsibility and expertise in data management is part of the scientific oriented departments.

3.2 Data access

The questions on the data access aim to get information on how the data can be accessed and which services are provided by the institutions. These questions divided into two sections – one on the services provided and a second on the user groups which were served.

Data services

About 27% of the questionnaires answered that a data portal is provided. This means that a single point of access is provided to view and download the data (this question was not directly addressed). Examples for these data portals are

- http://apps.iecolab.es/linaria/
- http://gamta.lt/cms/index?rubricId=48d34fea-0b25-4738-bc21-f7c7cecbc78a
- http://giida-biodiv.ise.cnr.it/
- http://icts.ebd.csic.es/
- http://www.slu.se/en/faculties-and-departments/faculty-of-natural-resources-and-agricultural-sciences/about-the-faculty/departments/department-of-aquatic-sciences-and-assessment/environment/
- http://www.tereno.net/
- https://secure.umweltbundesamt.at/eMORIS/

21% answered that also data services were provided to view and/or download the data. The data services provided are on the one side metadata services (like OGC CSW – Catalogue Service Web) and on the other hand the OGC access services like WFS (Web Feature Service) and WMS (Web Map Service) as well as Sensor Web Enablement (SWE).

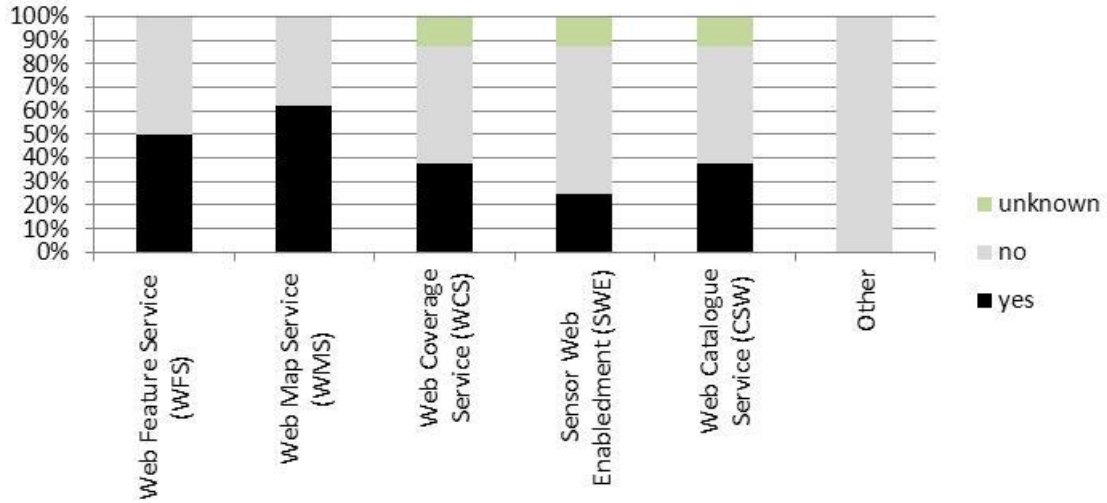

Figure 6 gives an overview on the data services provided within the EnvEurope consortium.

About 38% of the answers showed that metadata services for the site are provided. The software used for the metadata services are either GeoNetwork (ISO19115 compliant) or MetaCat (EML).

For the OGC access services 62,5% of the questionnaires answered that a Web Map Services (WMS) is provided allowing to access and view the data via a GIS-like client. 50% provide a Web Feature Service (WFS) which is suggested as download service in the INSPIRE/SEIS architecture. Here not only vector maps but also feature information is provided. A smaller share, about 38% is also providing a data service for raster data (Web Coverage Service, WCS). The above OGC services are implemented with either open source software such as GeoServer, MapServer, or MiraMon or with commercial software as ArcGIS Server.

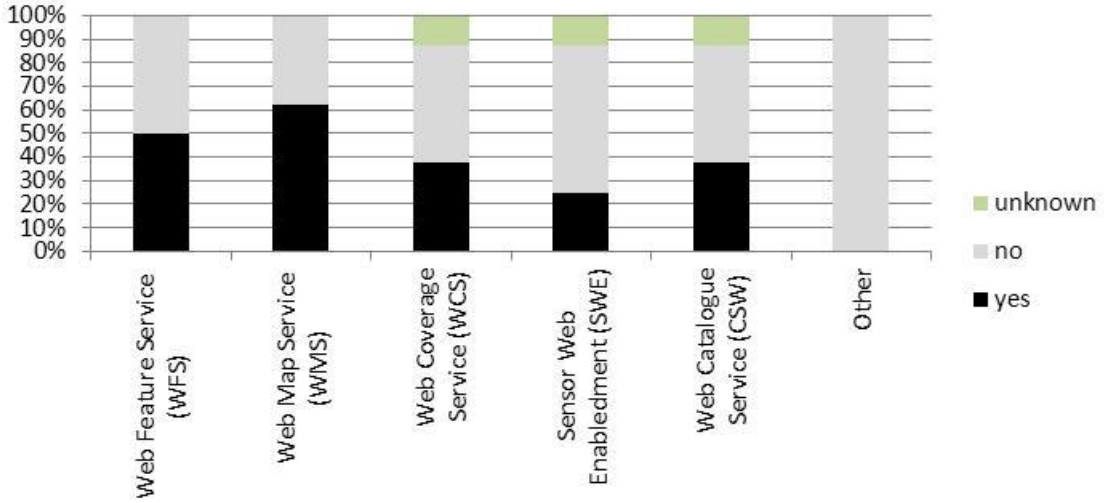

Figure 6 Question Q12.1 Data services provided (8 questionnaires). WFS (Web feature service), WMS (Web map service), WCS (Web coverage service), SWE (Sensor Web Enablement) and CSW (Catalogue Service Web)

In addition to these services some of the institutions are working on the implementation of SWE (Sensor Web Enablement) services (25% of positive answers). The service tested is mainly Sensor Observation Service (OGC SOS) implemented by the facilities of 52°North.

Despite the existence of data portals and services the majority of the data requests are handled in a very traditional way. The question 17 aimed to collect information how the data can be accessed. About 72% of the answers showed that the data only could be requested by a direct request either by mail or telephone. These correspond to either specific persons to be addressed which act as scientific managers or data managers of the project or formal requests by filling a form and sending it to the central coordination.

About 22% are providing an online access at least to the metadata to send the data request. Only 3% of the answers showed really an inline access to the data which allows the discovery and download of the data by a single point of access.

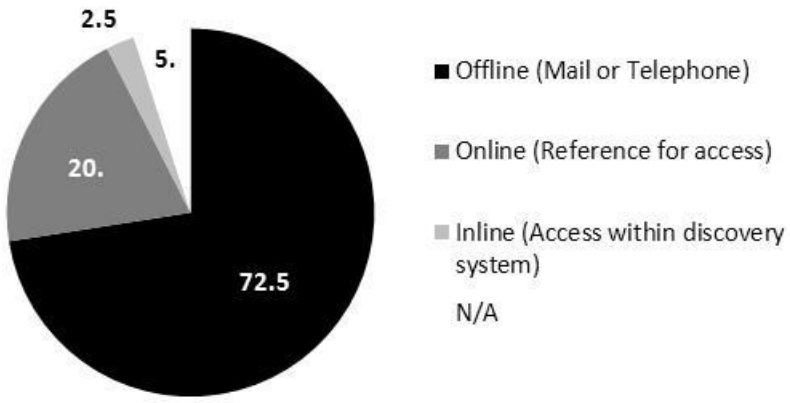

Figure 7 Question 17 – How can the data be requested? Total answers n=37

User groups

To get an overview on the audience for the data access, the user groups were asked to be identified in the two sections of the questionnaire. First in the "Data access" section to get an overview on the intended audience (Q13) and second in the section on "Data Sharing Policy" to get an overview on the actual audience (Q16).

The main target groups were asked in the question about "To whom might your data be useful? (Q13)". The result is shown in Figure 8. It shows that as the main target group for the data scientific and research institutions are seen as well as for educational purposes. About 92% of the questionnaires showed the answer "yes" for the first and around 81% for the second group. But also the two groups "Administration" (68% positive answers) and "Public" (65% positive answers) were seen as quite important.

The relatively low answer for the group "EnvEurope Community" of 51% is surprising. But this might be an artefact in the questionnaire.

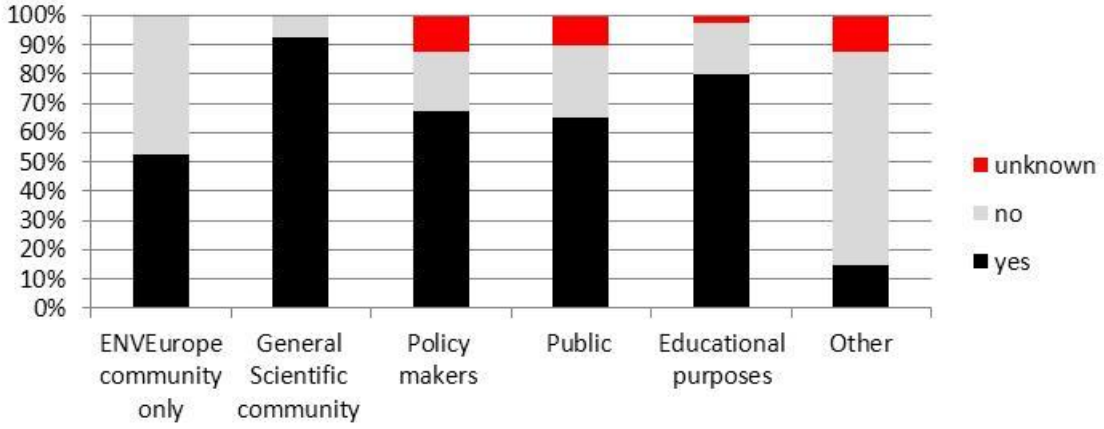

Figure 8 Question 13 – To whom it might your data be useful?

The question Q13 was targeted to get an insight to whom the data might be of interest. It did not reflect the current use of the data. This was done in the next section on the data sharing policy with the question "What are the main user groups? Please also indicate the users more specifically (e.g. universities, municipalities, etc.)? (Q16)".

The results are shown in Figure 9. The research community is the main user group of the data (95% positive answers). This is similar to the results of Q13 about the target groups. In the comments especially universities and research institutes are named as the main users. The Education group shows a much smaller use of the data (46% of positive answers) than the intended importance.

Administration and Public also have a lower actual use of the data than the intended importance as data users.

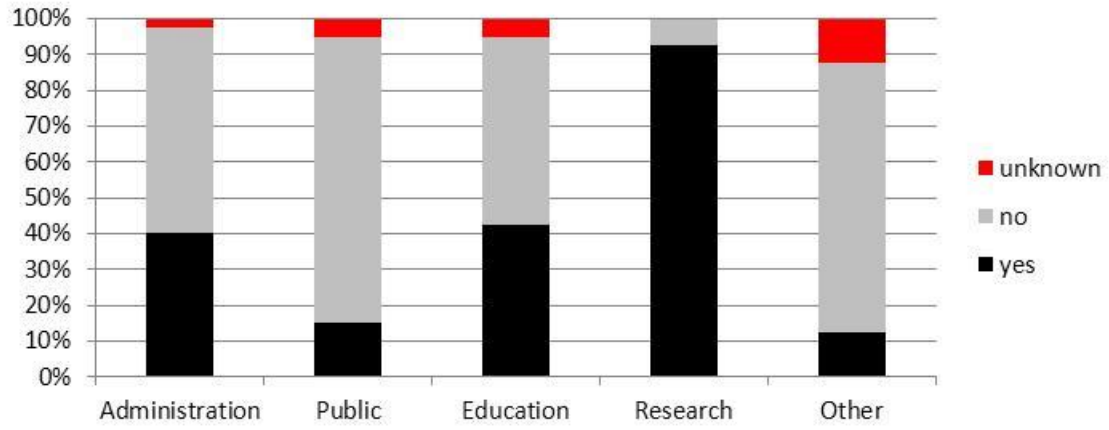

Figure 9 Question 16 - What are the main user groups? Overview on the real users of the data

The comparison of the two questions Q13 and Q16 shows that the research community is still the main data user.

3.3 Data sharing policy

The section on data sharing policy tried to collect information on the data license and the cost model for the data sharing. The results are shown in the following sub-chapter.

The questionnaire provided the following data license model were “free”, “free upon request”, “restricted”, and “no access”. Free in this sense meant that the data can be used by everybody under specified terms of use (e.g. notification) but no major restrictions were applied for the use. Free upon request means that the terms of use are negotiated case by case, but the data are in principal free to use. Restricted means that the use of the data is only for either a restricted group or a restricted set of purposes; the terms of use are specified. No access means that the data are not free to use. If other terms of use are used it was possible also to specify this.

The main data license model for data sharing is “free upon request”. This means the data are free under certain conditions which are negotiated during the request phase. This allows a certain control over data by the data owner. It also allows an overview about who is using the data. The questionnaire showed no difference in the user groups. For administration and public as well as for research and education the same data sharing model is applied. The results are shown in Figure 10.

The data access model “free” is mainly used for the user groups’ administration and public. But the differences to the user groups’ research and education are quite small.

The listed users for the user groups ranged on all levels of the European administration, e.g. from European environmental administration to the local environmental administration. The user group of “public users” seems not to be defined very clear. For the user group research and education mainly universities and research institutes were listed. For the user group other e.g. commercial scientific research companies were named as examples.

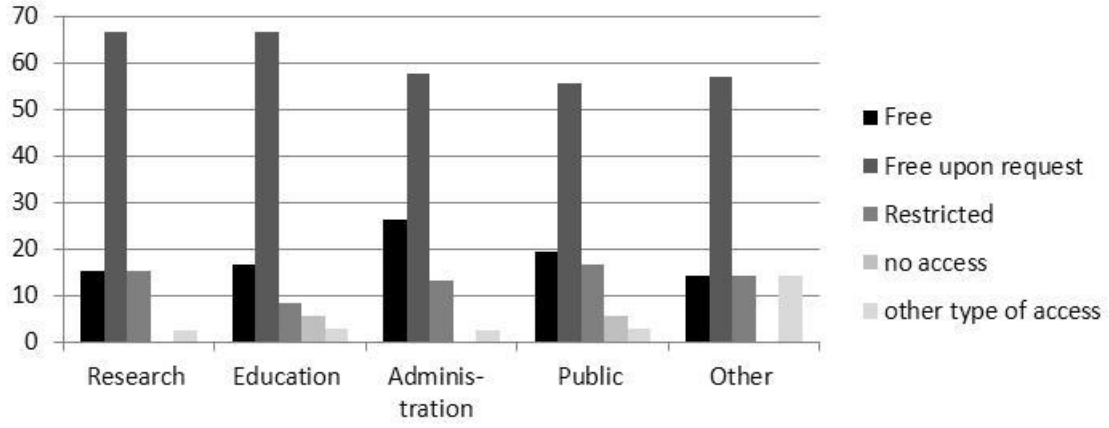

Figure 10 Question 14 – Are your data freely available in % for the different user groups: a) Research (n=38), b) Education (n=36), c) Administration (n=36), d) Public (n=39), e) other user groups (n=7).

Regarding the cost model for data sharing the options "no costs", "data manipulation costs", and "data creation costs" were questioned. The "no cost" model means that no additional costs are charged for the data sharing. The "data manipulation" cost model means that only costs for the data manipulation, e.g. query time, transfer time, extract time, etc., are charged in the data sharing process, but no costs for the data creation. The "data creation" cost model means that costs for the data creation are charged. In addition it was possible to list other cost models if necessary.

The cost model "data manipulation cost" was the most frequent for all different user groups (50-58% positive answers). Only for the user group "Others" the cost model "data creation costs" was dominant with 57% positive answers.

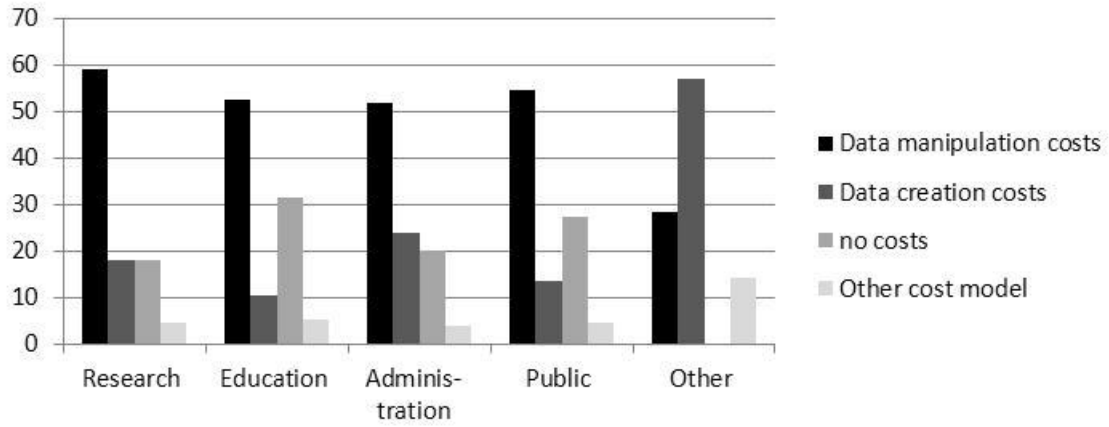

Figure 11 Question 15 – Cost Model for Data Sharing in % for the different user groups: a) Research (n=22), b) Education (n=19), c) Administration (n=25), d) Public (n=22), e) Other user groups (n=7).

About 58% of the questionnaires answered that they have sensitive data (Q18). 14% of the answers could not answer on that. Sensitive data would need additional data security levels if the data are distributed.

In total the replies showed that the main data use model for all user groups was the model "free upon request" which means a case to case negotiation. This hinders an automatic seamless access to data on a "data market square" as negotiated terms of access for at least bigger user groups are

needed. And these agreed terms need to be accepted in the course of the data download. The main cost model, except for commercial use of the data, is the “data manipulation cost model”. This would be compliant to the INSPIRE directive which allows to charge costs for the manipulation of the data.

3.4 Expectations

The questionnaire also tried to collect expectations from the EnvEurope community to the data management of the project. About two third of the questionnaires (66,7%) answered the question about a data portal (Q19) with “yes”. The same is true for data services. 61% answered the question (Q20) with “yes” to expect data services within the project runtime. Therefore the expectations are quite high.

The general requirements listed in the comment section of this question ranged from the use of open source standards and tools, to the provision of aggregated information using data portals and services.

Regarding the services requested a similar picture to the existing services could be shown. Here the same distinction between metadata services (CSW) and data services can be made.

About 36% of the answers requested a metadata service. This is part of the EnvEurope work plan and the tasks of action 1 on data management. About 50% requested OGC standard GIS services like Web Feature Service (WFS) and Web Map Services (WMS). The Web Coverage Service (WCS) for raster data seems to play and smaller role.

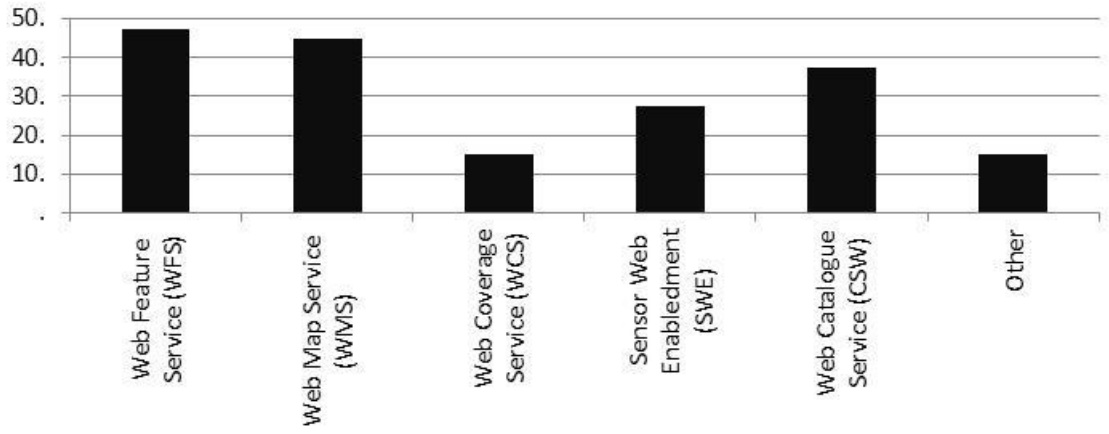

Figure 12 Question 20.1 – Which kind of data service do you expect within EnvEurope?

With about 28% of positive answers also SWE (Sensor Web Enablement) seems to play a smaller role. But this solution is showing maybe a solution to exchange time series data in quite good way.

One important comment on the requirements was that a system on data quality and data access rights needs to be implemented and assured. Also that, as the simple version of data exchange, as simple file based data download at least for the runtime of the project should be possible.

Regarding the intended user groups for the EnvEurope “data management solution” a similar picture could be shown as in the earlier questions. Administration, research and education seem to be the target user groups who should be addressed by the EnvEurope data management. These can be interpreted as the user and the creator of information about the state and trends of the environment of Europe.

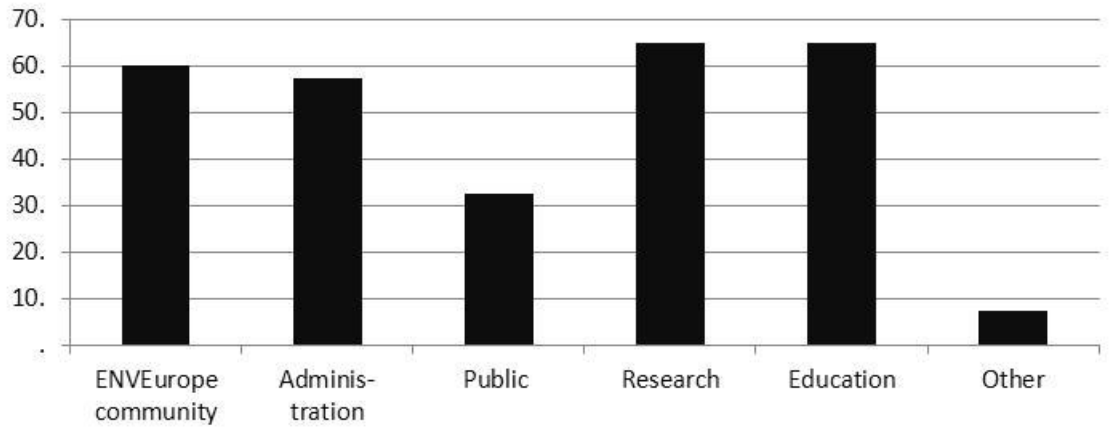

Figure 13 Question 20.2 - Who should be able to access and use the data pool of EnvEurope?

It also means that not only scientific data can be downloaded but also aggregated information, as e.g. small explanations and simple figures, need to be provided. Here the action 1 interrelates with the action 6 on dissemination.

4 Data reporting

Based mainly on the results of the questionnaire, a first, short term solution for the collection and further management of the datasets in EnvEurope has been proposed. It consists in collecting data in the form of Excel files to be uploaded in a central ftp repository maintained at EAA premises. In a further step, data will be entered in a relational database implemented by MySQL. An online user interface developed using DRUPAL will be implemented as front end to upload and query the data.

To the above purposes a simple Data Reporting Format was developed which could be used by all beneficiaries. The data reporting format tries to include all data elements proposed as necessary for describing commonly used datasets identified in the EnvEurope[1] project.

For the EnvEurope project in general, depending on the identified indicator, monthly or annual data should be reported for selected parameters identified by Action 3. This includes physical or chemical analysis (e.g. meteorology or air quality) as well as vegetation observation data.

To define the Data Reporting Format, relevant monitoring programmes were analysed on the way how data upload and collection is done. This included UNECE ICP Integrated Monitoring, UNECE ICP Forest, and UNECE ICP Waters, generally using a file based data collection and upload to a central database.

As many of the beneficiaries participate in these monitoring programmes, EnvEurope adopted the standard used in the UNECE ICP Integrated Monitoring Programme[2]. The format for the data collection seems to be promising as data are reported in a sequential format where every line contains one observed parameter and time. This allows for a flexible data reporting which can easily adapted to the needs of the data required without changing the data format and structure.

As far as possible existing reference lists (enumerations) were taken from these standards as they are normally maintained by a central organisation.

4.1 Generic data model

Based on the existing data reporting formats a generic data model for the upload of the data was developed and described. This data model was then translated into a reporting format using Microsoft Excel to collect the data from the beneficiaries.

The different data elements are described in the following section.

The main object classes are:

Metadata
: contains information about the dataset uploaded.

Data
: contains the observation or monitoring data and providing the relevant Meta information about the observation (e.g. time, method, unit, etc.).

Station
: contains information about the observation plot or monitoring plot to identify sub structures for the observation and monitoring within the site.

Method
: contains information about the methods applied in the field to collect the data and method in the office to aggregate the information to the requested aggregation level.

[1] See http://www.enveurope.eu/
[2] See http://www.ymparisto.fi/default.asp?node=6329&lan=en

Enumerations

provide the reference lists linked to the relevant fields in the object class **Data**. In Figure 14 only exemplary entries are given for the different enumeration classes.

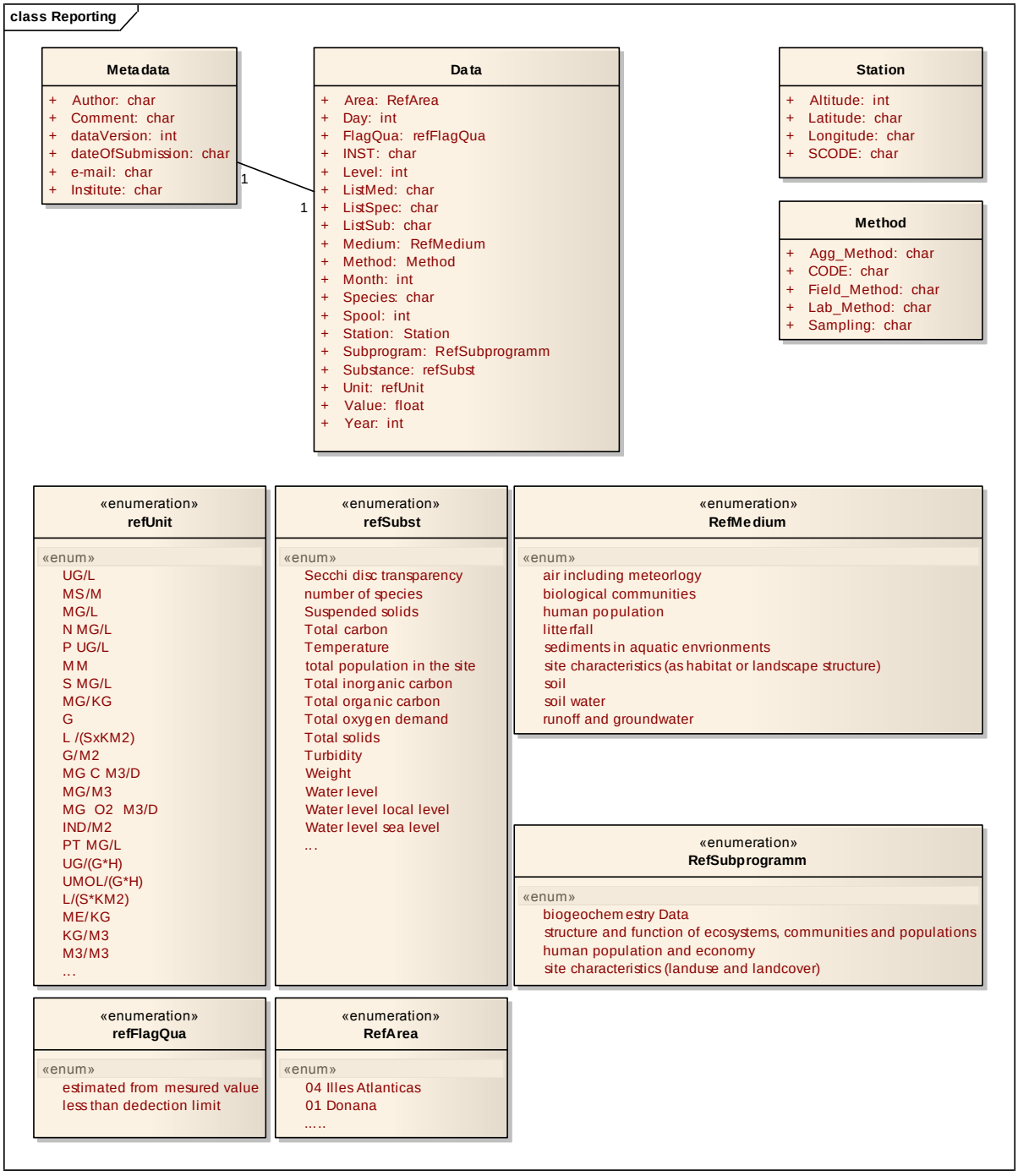

Figure 14 Generic data model for data reporting

4.2 Data specification

All proposed elements are defined in form of table with following information:

- The **name of data element**
- The **column name** used in the reporting file
- A short description and d**efinition** of the data element
- **Obligation/condition** for the data element
- **Multiplicity** of data entries, meaning if more than one information could be given for the entry (e.g. 1 or 1-n)
- A description of the **format** and **reference lists** used for the data reporting
- An **example** from EnvEurope domain dataset

In EnvEurope two different formats for the datasets are proposed:

a) Data about the chemical and physical conditions of the observed part of the ecosystem, e.g. meteorology, soil temperature, soil water analysis, litter analysis, etc.
b) Data about the biological composition of the observed part of the ecosystem, e.g. vegetation plots

In the following the two reporting formats are described. Commonly used fields are only described for the chemical and physical conditions and there is only a reference to that.

4.2.1 Reference lists

The reference lists (except for the species) are provided directly in the reporting file. They provide the general codes. If codes are missing the user can add additional ones at the end of the list. There a grey shaded area can be found where additional codes can be added. These additional codes can then also be used in the drop down list for the reporting.

4.2.2 Metadata

The section on metadata contains the main information about the source of the data. A separate metadata description of the dataset has to be given with the EnvEurope Metadata Editor.

4.2.2.1 Individual name

Metadata element name	**Individual name**
Column name	IND_NAME
Definition	Name of person who submits the data to EnvEurope. This is also the point of contact for the dataset in case of any questions, e.g. usage rights or questions about the methodology
Obligation/condition	Mandatory
Multiplicity	1
Example	Dirnböck, Thomas
Format	Text → <Last name>, <First name>
Reference list	None

4.2.2.2 Organisation name

Metadata element name	**Organisation name**
Column name	ORG_NAME

Definition	Name of the institute who submits the data to EnvEurope. This is also the point of contact for the dataset in case of any questions, e.g. usage rights or questions about the methodology
Obligation/condition	Mandatory
Multiplicity	1
Example	Environment Agency Austria (EAA), Austria
Format	Text → <Institute>, <Country>
Reference list	None

4.2.2.3 Electronic mail address

Metadata element name	**Electronic Mail address**
Column name	EMAIL
Definition	Email address of the contact person who submitted the data for EnvEurope
Obligation/condition	Mandatory
Multiplicity	1
Example	Thomas.Dirnböck@Umweltbundesamt.at
Format	Text → <name>@<domain>
Reference list	None

4.2.2.4 Dataset publication date

Metadata element name	**Dataset publication date**
Column name	DATE OF PUBLICATION
Definition	Date of the data submission
Obligation/condition	Mandatory
Multiplicity	1
Example	12.10.2011
Format	Date → DD.MM.YYYY
Reference list	None

4.2.2.5 Data version

Metadata element name	**Data version**
Column name	DATA VERSION
Definition	Version of the data in case the data are updated
Obligation/condition	Mandatory
Multiplicity	1
Example	V1.0

Format	Text → V<main version>.<sub version>
Reference list	None

4.2.2.6 Comments

Metadata element name	Comments
Column name	COMMENTS
Definition	Any comments regarding the data, e.g. usage restrictions, methods, etc.
Obligation/condition	Optional
Multiplicity	1
Example	Only internal use allowed
Format	Text
Reference list	None

4.2.3 Stations

The **station** is an observation plot or measurement plot within the site. Basic metadata about the station, if different from the site as such is given.

4.2.3.1 Station code

Metadata element name	Station code
Column name	SCODE
Definition	Code for the station within the site. A station is any measuring unit such as a sampling plot or a meteorological station. If the station equals to the site, meaning that only one station is used within the site, only the site identifier is provided in the data recording sheet.
Obligation/condition	Mandatory
Multiplicity	1
Example	IP1
Format	Text
Reference list	None

4.2.3.2 Longitude

Metadata element name	Longitude
Column name	LONGITUDE
Definition	Longitude of the sampling plot. The data is provided if necessary for the data reporting and the plot is a sub-unit of the site; e.g. in the case of very big sites
Obligation/condition	Conditional if different from the site
Multiplicity	1

Example	
Format	Text → DD MM SS, Datum WGS84
Reference list	None

4.2.3.3 Latitude

Metadata element name	Latitude
Column name	LATITUDE
Definition	Latitude of the sampling plot. The data is provided if necessary for the data reporting and the plot is a sub-unit of the site; e.g. in the case of very big sites
Obligation/condition	Conditional if different from the site
Multiplicity	1
Example	
Format	Text → DD MM SS, Datum WGS84
Reference list	None

4.2.3.4 Altitude

Metadata element name	Altitude
Column name	ALTITUDE
Definition	Altitude of the sampling plot. The data is provided if necessary for the data reporting and the plot is a sub-unit of the site; e.g. in the case of steep altitudinal gradients within a site
Obligation/condition	Conditional if different from the site
Multiplicity	1
Example	935
Format	Text → [m a.s.l.]
Reference list	None

4.2.4 Methods

The section contains information on the methods used in the observation. The method is referenced in both – the chemical and physical observations as well as for the species observations. The method section should give an overview on the sampling, the field method and the method used in the lab to create the data value. This part will be specified in future work and undergo a standardisation.

4.2.4.1 Method code

Metadata element name	Method code
Column name	METHOD_CODE
Definition	Code for the method. This code is used in the data reporting sheet for the data to reference to the method.

Obligation/condition	Mandatory
Multiplicity	1
Example	METH_BB
Format	Text, max 10 characters
Reference list	None

4.2.4.2 Dataset sampling description

Metadata element name	Dataset sampling description
Column name	SAMPLING
Definition	Short description on the sampling procedure (selection of plots, observation points, etc.)
Obligation/condition	Conditional • If a sampling procedure was applied this should be stated here
Multiplicity	1
Example	Random sampling of spruce stands in the entire area of the site; 5 regularly spaced (10 m) positions on a transect; etc.
Format	Text
Reference list	None

4.2.4.3 Dataset methods description - Field method

Metadata element name	Dataset methods description - Field method
Column name	FIELD_METHOD
Definition	Short description of the method used in the field either to collect the samples or to do the observation
Obligation/condition	Mandatory
Multiplicity	1
Example	Volume weighted mixing from 5 bulk sampler, 2 weeks interval of sampling, cooled transportation of the samples
Format	Text
Reference list	None

4.2.4.4 Dataset methods description - Lab method

Metadata element name	Dataset methods description - Lab method
Column name	LAB_METHOD
Definition	Short description on the procedures and methods applied in the lab, e.g. filtering, analysis, etc.
Obligation/condition	Mandatory
Multiplicity	1

Example	45µm filtered; ICP-OES
Format	Text
Reference list	None

4.2.4.5 Dataset methods description - Aggregation procedure

Metadata element name	Dataset methods description - Aggregation procedure
Column name	AGG_METHOD
Definition	Description of the procedure how the values has been aggregated from primary values; for primary data the aggregation procedure is "NONE".
Obligation/condition	Mandatory
Multiplicity	1
Example	X
Format	Text
Reference list	None

4.2.5 Observations

This section contains data on any observation or measurement in the different compartments of the ecosystem. It includes bio-geochemical measurements as well as biotic observations

4.2.5.1 Sub programme

Metadata element name	Sub programme	
Column name	SUBPROG	
Definition	Code for the sub programme for which the data are reported, e.g. BIOCHEM for "biogeochemical data" within the site. This refers to the parameter groups used in the EnvEurope context.	
Obligation/condition	Mandatory	
Multiplicity	1	
Example	BIOCHEM	
Format	Text → LOV Ref_SUBPROG	
Reference list	BIOCHEM	biogeochemistry data
	STRUCTU	Structure and function of ecosystems, communities and populations
	HUMANEC	human population and economy
	SITECHA	site characteristics (land use and land cover)

4.2.5.2 LTER Europe Site Code

Metadata element name	LTER Europe Site Code
Column name	SITE_CODE

Definition	code of the site according to LTER InfoBase
Obligation/condition	Mandatory
Multiplicity	1
Example	SI001496
Format	Text → LOV Ref_AREA
Reference list	Identifier according to the LTER InfoBase for the EnvEurope site which is provided on the ftp-repository
Exception	if the LTER InfoBase Code is not known please use the site name instead of the site identifier

4.2.5.3 Organisation name

Metadata element name	**Organisation name**
Column name	ORG_NAME
Definition	Name of the institute providing the data. This could be different from the institute doing the data submission
Obligation/condition	Mandatory
Multiplicity	1
Example	EAA
Format	Text
Reference list	None

4.2.5.4 Station code

Metadata element name	**Station code**
Column name	SCODE
Definition	Code for the station within the site. A station is any measuring unit such as a sampling plot or a meteorological station. If the station equals to the site, meaning that only one station is used within the site, only the site identifier is provided in the data recording sheet.
Obligation/condition	Mandatory
Multiplicity	1
Example	IP1
Format	Text
Reference list	None

4.2.5.5 Medium

Metadata element name	**Medium**
Column name	MEDIUM
Definition	code for the sampled medium in the observation

Obligation/condition	Mandatory
Multiplicity	1
Example	AIR
Format	Text → LOV Ref_MEDIUM
Reference list	AIR air including meteorology SOIL soil SOILWAT soil water WATER runoff and groundwater SEDIMENT sediments in aquatic environments LITTER litter fall BIOCOM biological communities HUMPOP human population SITECHAR site characteristics (as habitat or landscape structure)

4.2.5.6 Reference list for medium

Metadata element name	**Reference list for medium**
Column name	LISTMED
Definition	Medium code list
Obligation/condition	Mandatory
Multiplicity	1
Example	EnvEurope
Format	Text → LOV
Reference list	EnvEurope IM ICP Integrated Monitoring DB

4.2.5.7 Altitude or depth - maximum

Metadata element name	**Altitude or depth - maximum**
Column name	MAX_LEVEL
Definition	measurement level in [cm]; the soil/rock surface is the zero level; in case of aquatic systems it also could be given as from to level (e.g. 0 - -20)
Obligation/condition	Mandatory
Multiplicity	1
Example	-20
Format	Number in [cm]
Reference list	None

4.2.5.8 Altitude or depth minimum

Metadata element name	Altitude or depth minimum
Column name	MIN_LEVEL
Definition	measurement level in [cm]; the soil/rock surface is the zero level; in case of aquatic systems minimum depth to sampling
Obligation/condition	Mandatory
Multiplicity	1
Example	0
Format	Number in [cm]
Reference list	None

4.2.5.9 Size

Metadata element name	Size
Column name	SIZE
Definition	Size of the sampling plot where the observation takes place or the size of the area for which the aggregated values are representative (e.g. the site or part of the site such as the forested area)
Obligation/condition	Mandatory
Multiplicity	1
Example	100
Format	Number in [m^2]
Reference list	None

4.2.5.10 Year of observation

Metadata element name	Year of observation
Column name	YEAR
Definition	Year of the measurement or the year for which the measurements were aggregated the year of an observation (e.g. plants)
Obligation/condition	Mandatory
Multiplicity	1
Example	2004
Format	Number → YYYY
Reference list	None

4.2.5.11 Month of observation

Metadata element name	Month of observation
Column name	MONTH
Definition	Month of the measurement or the month for which the

	measurements were aggregated; the month of the observation (e.g. plants)
Obligation/condition	Conditional if monthly observations are provided, in case of yearly reporting the MONTH is left blank
Multiplicity	1
Example	12
Format	Number → MM
Reference list	None

4.2.5.12 Day of the observation

Metadata element name	**Day of the observation**
Column name	DAY
Definition	Day of the measurement or observation; usually not provided as monthly sums or means are reported
Obligation/condition	Conditional if the daily observations are provided, in case of monthly reporting the DAY is left blank
Multiplicity	1
Example	12
Format	Number → DD
Reference list	None

4.2.5.13 Hour of the observation

Metadata element name	**Hour of the observation**
Column name	HOUR
Definition	Hour of the measurement or observation
Obligation/condition	Optional Mandatory in the case of sensor (e.g. meteorological station, probe in water, etc.)
Multiplicity	1
Example	14
Format	Number → HH
Reference list	None

4.2.5.14 Minute of the observation

Metadata element name	**Minute of the observation**
Column name	MINUTE
Definition	MINUTE of the measurement or observation
Obligation/condition	Optional

	Mandatory in the case of sensor (e.g. meteorological station, probe in water, etc.)
Multiplicity	1
Example	53
Format	Number → MM
Reference list	None

4.2.5.15 Second of the observation

Metadata element name	**Second of the observation**
Column name	Second
Definition	Second of the measurement or observation
Obligation/condition	Optional Mandatory in the case of sensor (e.g. meteorological station, probe in water, etc.)
Multiplicity	1
Example	43
Format	Number → SS
Reference list	None

4.2.5.16 Spool of the observation

Metadata element name	**Spool of the observation**
Column name	SPOOL
Definition	spatial pool as the number of devices (e.g. sensors, sampling units, etc.) or sampling plots (e.g. subplots of a bigger plot area) used to measure a parameter
Obligation/condition	Mandatory
Multiplicity	1
Example	5
Format	Number
Reference list	None

4.2.5.17 Dataset taxonomic rank value

Species names are defined according to the species lists provided by ICP Integrated Monitoring, which are based on international standards (e.g. the Flora Europea). Not all species of all sites will be found in these lists. If so, add your own list, which includes the species name and nomenclature. The specific species lists have to be reported with the species data. Please be careful with synonyms and check if your species really doesn't exist in the provided list.

The species code lists are not directly worked into the reporting file. Please refer to the directory _ref_list_species to select the appropriate species for the reporting. The species lists are provided as textfiles (*.EXP) – please rename them to *.txt to open them or import them directly to Excel.

Metadata element name	Dataset taxonomic rank value
Column name	TAXA
Definition	Name of the taxa. For the specie use a 3 (genus) + 4 (species) letter code, and for another taxa level use the first two letters of taxa rank name.
Obligation/condition	Mandatory for species observation data (e.g. vegetation plot, biotic samples) – in case of bio-geochemical data this field remains empty
Multiplicity	1
Example	FAG SYLV
Format	Text → LOV
Reference list	see on the ftp-repository the directory __ref_list_species

4.2.5.18 Reference list for the taxa

Metadata element name	Reference list of the taxa
Column name	LISTTAXA
Definition	Code for the reference list of the species used
Obligation/condition	Mandatory for species observation data (e.g. vegetation relevees) – in case of bio-geochemical data this field remains empty
Multiplicity	1
Example	DB
Format	Text → LOV
Reference list	EnvEurope

4.2.5.19 Parameter observed

Metadata element name	Parameter observed
Column name	SUBST
Definition	substance code (chemical elements) or parameter (physical measurement) observed in the measurement
Obligation/condition	Mandatory
Multiplicity	1
Example	PH
Format	Text → LOV Ref_SUBST
Reference list	LISTSUB SUBST Name DB ALK Alkalinity DB BOD Biochemical oxygen demand DB TC Total carbon DB CODCR Chemical oxygen demand COD-Cr DB CODMN Chemical oxygen demand COD-Mn DB DC Dissolved carbon DB DIC Dissolved inorganic carbon

	DB	DOC	Dissolved organic carbon
	DB	DOD	Direct oxygen demand
	DB	NH3	Ammonia
	DB	NH4	Ammonium
	DB	NH4N	Ammonium as nitrogen
	DB	NKJ	Kjeldahl nitrogen
	DB	NO2	Nitrite
	DB	NO23	Nitrite nitrate
	DB	NO23N	Nitrite nitrate as nitrogen
	DB	NO2N	Nitrite as nitrogen
	DB	NO3	Nitrate
	DB	NO3N	Nitrate as nitrogen
	DB	NOXNDO	Nitrogen oxides as NO2
	DB	NTOT	Total nitrogen
	DB	O2	Oxygen
	DB	O2D	Dissolved oxygen
	DB	O2S	Oxygen saturation
	DB	PO4	Phosphate
	DB	PO4P	Phosphate as phosphorous
	DB	PTOT	Total phosphorous
	DB	TIC	Total inorganic carbon
	DB	TOC	Total organic carbon
	DB	TOD	Total oxygen demand
	DB	COND	Conductivity
	DB	DEPTHB	Depth of sampling from bottom
	DB	DEPTHS	Depth of sampling from surface
	DB	DEPTHT	Depth to bottom
	DB	EH	Redox potential
	DB	FLOW	Flow
	DB	HH	Humidity
	DB	LENGTH	Length
	DB	PH	pH
	DB	SDT	Secchi disc transparency
	DB	TEMP	Temperature
	DB	TS	Total solids
	DB	SS	Suspended solids
	DB	TURB	Turbidity
	DB	WL	Water level
	DB	WLL	Water level local level
	DB	WLS	Water level sea level
	DB	BPP	Biological primary production net
	DB	BPY	Biological primary productivity net
	DB	CP	Chlorophyll a
	DB	PREC	Precipitation
	DB	DISCH	Discharge
	DB	P	Phosphorus
	DB	LDEP	litter deposition (weight)
	DB	WEIGHT	Weight
	*	BIOMASS	biomass of
	*	SPNB	number of species
	*	AB	abundance of species
	*	THP	total human population in the site
	*	HDENSITY	density

	*	HAGESTR	age structure
	*	HECONACT	main economic activity
	*	HINCOME	average income
	*	LANDUSE	land use
	*	LANDCOV	land cover
	IM	COVE_T	species cover tree layer
	IM	COVE_S	species cover shrub layer
	IM	COVE_F	species cover field layer
	IM	COVE_B	species cover bottom layer

4.2.5.20 Reference for parameters

Metadata element name	**Reference for parameters**
Column name	LISTSUB
Definition	code list for the substances or parameter
Obligation/condition	Mandatory
Multiplicity	1
Example	EnvEurope
Format	Text → LOV
Reference list	EnvEurope IM ICP Integrated Monitoring DB definition needed

4.2.5.21 Method code

Metadata element name	**Method code**
Column name	METHOD_CODE
Definition	code for the methods defined in the table METHOD
Obligation/condition	Mandatory
Multiplicity	1
Example	METH_1
Format	Text → LOV
Reference list	See table METHOD

4.2.5.22 Value

Metadata element name	**Value**
Column name	VALUE
Definition	value of the measurement or observation
Obligation/condition	Mandatory
Multiplicity	1
Example	2,23

Format	Number
Reference list	None

4.2.5.23 Unit

Metadata element name	**Unit**
Column name	UNIT
Definition	unit of the observation or measurement
Obligation/condition	Mandatory
Multiplicity	1
Example	µg/l
Format	Text –LOV Ref_UNIT
Reference list	see Data Reporting Format sheet

4.2.5.24 Quality flag

Metadata element name	**Quality flag**
Column name	FLAGQUA
Definition	data quality flag
Obligation/condition	Mandatory
Multiplicity	1
Example	E
Format	Text → LOV
Reference list	L less than detection limit E estimated from measured value

5 Data Reporting - Data upload

The data reported can be directly uploaded by the beneficiaries and then accessed and downloaded using the ftp-repository at the Environment Agency Austria/Umweltbundesamt. The link to the ftp-repository is:

ftp://ftp.umweltbundesamt.at/KnownUsers/2253_2/EnvEurope_DataCollection/.

5.1 Access to the ftp-repository

The ftp-repository is password secured.

- Username: **KU2253_2**
- password: **lter_member**

There are several ways to connect to the ftp-repository

Windows Explorer

The direct access to the ftp-repository in the Microsoft Windows Explorer is possible (preferably NOT the Internet Explorer). For this copy the link
ftp://KU2253_2:lter_member@ftp.umweltbundesamt.at/KnownUsers/2253_2/EnvEurope_DataCollection/
in the address space of Explorer. This link already includes username and password.

FTP-Client

Alternatively any ftp-Client can be used to access the ftp-repository. We recommend Core-ftp-Lite which is a freeware ftp-client. This client can be downloaded from http://www.coreftp.com/download.html.

When using Core ftp Lite the ftp-connection can be specified using the menu item “FILE” → “CONNECT”. There the details about the connection can be specified, e.g. ftp, user name and password. Further specify under “ADVANCED” in this window the start directory. This needs to be set to /KnownUsers/2253_2/EnvEurope_DataCollection/.

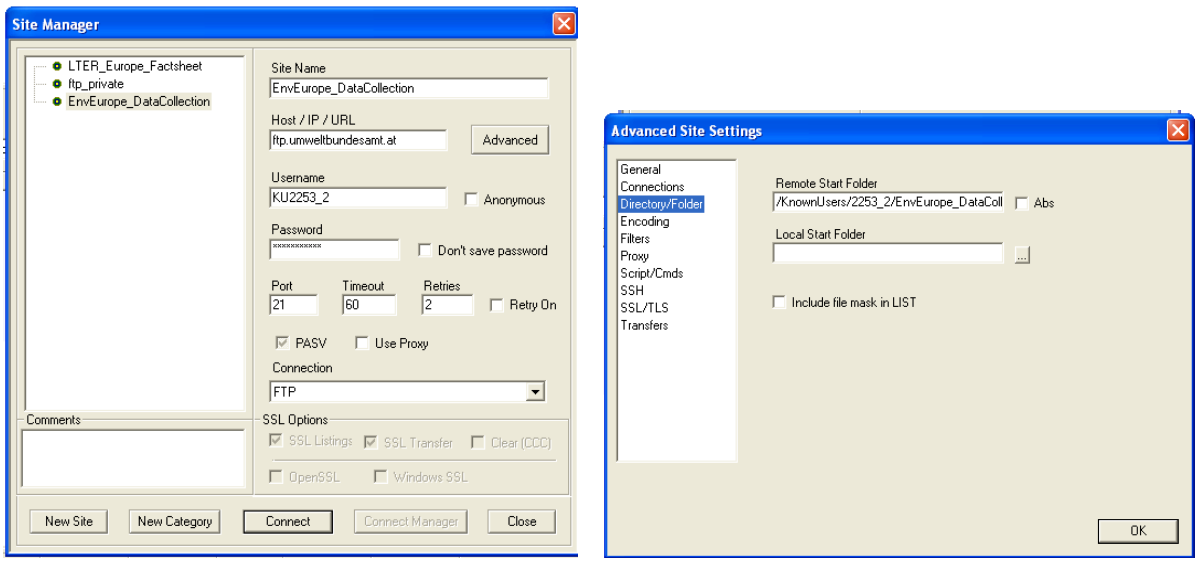

Figure 15 Specifying the ftp-link, user name and password with Core-ftp Lite as well as the remote start folder

5.2 Directory structure

The repository contains sub directories for the different countries. Each beneficiary can upload her/his file in her/his country folder, naming it by using the specified naming convention specified here following.

In addition to the country directories, a directory containing the reference lists for taxa, substances and methods used by the UNECE ICP Integrated Monitoring are provided for the data reporting.

Naming convention

The names of the data file needs to be set to

<Country code, 2 letters>_<Site code>_<Date of data submission>.xls

e.g. AT_SI001496_20111021.xls

In addition if files are split according to different sub-programmes, e.g. Vegetation, the sub-programme should be used in the file name

e.g. AT_SI001496_Vegetation_20111021.xls

6 Selection of relevant datasets

The selection of relevant datasets is one of the results of the Bucharest meeting. Prior to the meeting a data collection format was sent to beneficiaries to test the future collection of datasets (EnvEurope Data Reporting Format). The data reporting format is in fact containing a list of 65 parameters (see attached list).

During the meeting in Bucharest no other parameter was added to the list, but a number of project proposals have been submitted (14 project proposals). The list of parameters and possible datasets needed from different beneficiaries is found on the attached project list. The list proposed can be amended by project needs in future.

Table 1 Data collection – List of parameters (defined by Action 3)

Parameter	Name	Description
ALK	Alkalinity	
BOD	Biochemical oxygen demand	Essential to state incubation time - see pre-treatment list: Incubation.
TC	Total carbon	
CODCR	Chemical oxygen demand COD-Cr	
CODMN	Chemical oxygen demand COD-Mn	
DC	Dissolved carbon	
DIC	Dissolved inorganic carbon	
DOC	Dissolved organic carbon	
DOD	Direct oxygen demand	
NH3	Ammonia	
NH4	Ammonium	
NH4N	Ammonium as nitrogen	
NKJ	Kjeldahl nitrogen	
NO2	Nitrite	
NO23	Nitrite nitrate	Nitrite and nitrate
NO23N	Nitrite nitrate as nitrogen	Nitrite and nitrate as nitrogen
NO2N	Nitrite as nitrogen	
NO3	Nitrate	
NO3N	Nitrate as nitrogen	
NOXNDO	Nitrogen oxides as NO2	
NTOT	Total nitrogen	
O2	Oxygen	
O2D	Dissolved oxygen	
O2S	Oxygen saturation	The amount of oxygen dissolved in the water compared to what theoretically can be dissolved at the same temperature expressed as percentage value.
PO4	Phosphate	
PO4P	Phosphate as phosphorous	
PTOT	Total phosphorous	

TIC	Total inorganic carbon	
TOC	Total organic carbon	
TOD	Total oxygen demand	
COND	Conductivity	
DEPTHB	Depth of sampling from bottom	
DEPTHS	Depth of sampling from surface	
DEPTHT	Depth to bottom	
EH	Redox potential	
FLOW	Flow	
HH	Humidity	
LENGTH	Length	
PH	pH	
SDT	Secchi disc transparency	
TEMP	Temperature	
TS	Total solids	
SS	Suspended solids	
TURB	Turbidity	
WL	Water level	
WLL	Water level local level	Water level compared to a local point.
WLS	Water level sea level	Water level compared to the sea.
BPP	Biological primary production net	
BPY	Biological primary productivity net	
CP	Chlorophyll a	
PREC	Precipitation	
DISCH	Discharge	
P	Phosphorus	
LDEP		litter fall amount (oven dry weight)
WEIGHT	Weight	
BIOMASS	biomass of	groups; to be specified by user; to be described in the metadata
SPNB	number of species	groups; to be specified by user; to be described in the metadata
AB	abundance of	groups; to be specified by user; to be described in the metadata
THP	total population in the site	human population
HDENSITY	density	human population
HAGESTR	age structure	human population
HECONACT	main economic activity	human population
HINCOME	average income	human population
LANDUSE	land use	
LANDCOV	land cover	

7 Status on data reporting

The data upload is currently tested and the data reporting formats are going to be adapted to the needs of the projects requirements. A sub-set of the beneficiaries already tested the work flows and the conversion of the locally stored data to the data reporting format.

The uploaded data can be accessed at the ftp-repository. User name and password is specified in the chapter5.

The data upload at the moment had the following task:

- Testing of the Data Reporting Format
- Testing of the central data repository
- Evaluation of the time needed for the data reporting

Currently exemplary datasets from eight countries were uploaded to check the data reporting format and the data flow. The uploaded files and the status of the data files are described in Table 2.

Table 2 List of data files uploaded to the central data repository

Country	Site	Dataset status
Austria	SI000049	Sample dataset of selected parameter
Bulgaria	SI001483	Sample dataset of selected parameter
Finland	SI001186	Sample dataset of selected parameter
Germany	SI000315	Sample dataset of selected parameter
Germany	SI001513	Sample dataset of selected parameter
Italy	SI001223	Sample dataset for selected parameter
Lithuania	SI000457	Sample dataset for selected parameter
Romania	SI000706	Sample dataset for selected parameter
Spain	SI001345	Sample dataset for selected parameter

During the EnvEurope Plenary Meeting held in Bucharest (11/2011) problems and difficulties with the current version of the data upload were discussed and the data reporting format was adapted to the needs defined by the project. This updated version of the Data Reporting Format will be used to collect datasets used by Action 3.

The parameters for the datasets needed are specified until mid of December 2011. The main part of the data upload is therefore expected until end February 2012.

The currently collected data will only be used internally for cross domain and cross site analysis made within EnvEurope by Action 3. Any other use of the data has to be negotiated with the data providers. A common data policy needs to be defined as a next step in the project.

8 Future outlook

As stated in the previous chapters the long term implementation plan for data exchange in EnvEurope will focus on web based technologies. To this aim, reference models and implementation strategies from biodiversity relevant ICT projects on the European level (e.g. LifeWatch, SANY, etc.) and other biodiversity related projects (e.g. EBONE, TERENO, etc.) needs to be analysed for their relevance to the project. The most promising being XML based OGC services as well as the emerging RDF based Linked Data technologies. On this basis, the final strategy for the future development of the data management in EnvEurope, will be defined in early 2012 in order to allow for prototypal implementation and testing.

This chapter shortly outlines different options that can be considered for the further development of data management in the project. It does not count for completeness.

8.1 WFS / WMS / SOS

Geo-referenced data (i.e. data that can be associated to a location on Earth) are traditionally managed by means of Geographic Information Systems (GIS). Internet has been a powerful innovation engine also in this field of spatial data, and a new type of online applications has been spread, namely Web GIS. As a matter of fact they are usually Web mapping facilities, where users can visually inspect thematic maps, managed as overlaid layers, and perform simple operations like pan and zoom. Moreover in usual Web GIS applications, each repository that serves the layers is strictly associated with a client interface, so that different repositories must be accessed by different user interfaces.

A further advance has been introduced by the development of geo-services, i.e., Web services that serve geographic data and tools. They are the building blocks of the so called Spatial Data Infrastructures (SDI), the ICT infrastructures to share and consume spatial data in an interoperable way. In fact, geo-services are based on standard interfaces and allow decoupling the functions of data serving and data accessing. As usual for service oriented architecture (SOA), on one side a geo-data service can serve its content to multiple clients, provided they cope with the same standards; on the other one, a client can access at the same time the data of different and distributed standard geo-services. This approach has been adopted and recommended in the INSPIRE Directive of the European Community, which aims at developing the European SDI, to share environmental data of the whole continent.

WFS, WMS and SOS are popular Web services proposed by the Open Geospatial Consortium (OCG), a global standardisation initiative for geographic applications and data. The first two are included in the INSPIRE recommendation, while SOS is recommended by GEOSS (Global Earth Observation System of System).

Short description of terms:

WMS (Web Map Service)
is an OGC standard service for serving geo-referenced images over the Internet that is provided by a map server. Both vector and raster images can be visualised by this service.

WFS (Web Feature Service)
This OGC spatial service provides an interface to query, as well as to perform transactions of spatial features of vector maps. Responses are encoded using the Geography Markup Language (GML). With respect to WMS, WFS allows not only to visualise maps but also to access data associated to map objects.

SOS (Sensor Observation Service)
is an OGC standard and one of the specifications of the Sensor Web Initiative; SOS defines web service interfaces for requesting, filtering, and retrieving observations and sensor information.

Within EnvEurope data to be served by the beneficiaries are mainly observations/parameter values with a spatial (x,y,z) and a temporal (t) assignment. Only in some cases this information can be provided as maps, by example if they are thematic maps from remote sensing observations or they are the result of a spatial processing over data from one or more punctual stations. Therefore WMS and WFS will probably have limited usage in the realm of EnvEurope. SOS is instead a more promising standard, the more as in this approach data (observations) are not treated as files (like the layers included in map servers) but as records of databases, thus allowing more flexible analysis and an easier management of data with high temporal granularity. Very useful client applications are available for visualising data served by SOS[3].

Though SOS applications have been already implemented and tested, also in European Projects like SANY, OSIRIS, and Mobesens, a great effort is still required to offer them as easy to use, plug-and-play facilities to end users. Some very important issues are still under debate, such as the treatment of the z dimension which is of great importance in EnvEurope.

In EnvEurope, serving data by Web services represents a first step towards some valuable objectives, i.e. data interoperability and sharing by multiple and independent data repositories, where observations collected by the different beneficiary institutions can be stored and maintained without the need of huge centralized IT storage facilities. A service-based solution can be a first step towards more advanced solutions such as LinkedData.

8.2 LinkedData

Based on XML, on the one hand XML-Schema and on the other hand, RDF / RDFS and OWL / SKOS have been developed and were adopted by W3C as recommendations. The basic idea of RDF (resource description framework) is, to meaningfully interlink any resource across the World Wide Web, where resource can be a single value, word, graphic element or a whole document, picture, or any resource which can be identified by an URL and be presented as http – document.

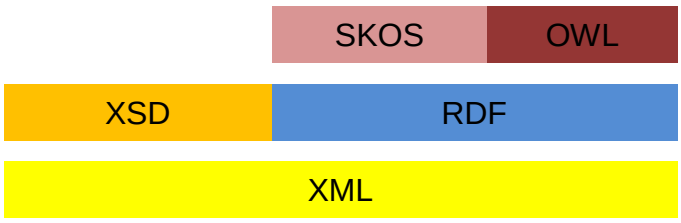

Figure 16 Connection between different W3C Standards

It can be easily understood as the idea, to extend the very successful hyperlinks by meanings. The meaning of hyperlink can be understood as "see also", but that meaning is never declared, whereas links with the meaning "is author of", "has dimension", "is taxon XXXX", can be declared within linked data architecture and thus promote links which can be used for exact, machine readable definitions. The definition languages for RDF are RDFS, OWL and SKOS, which again are expressed through RDF.

A remarkable general feature within linkedData architecture is that definitions and their instantiations need not be separated, but can be contained within one document. Linked data services are based on REST Services. There are three technological bases for linked data: 1) any

[3] E.g. see http://52north.org/communities/sensorweb/clients/Thin_SWE_Client/Version_2.0/

resource is identified by a URI, mostly a URL; 2) the resources are interlinked by RDF; and 3) The request for a resource is answered by an http document.

A growing number of sources, exposing data with linked data services and interlinking their resources to other resources, are presented in the so called linked data cloud.

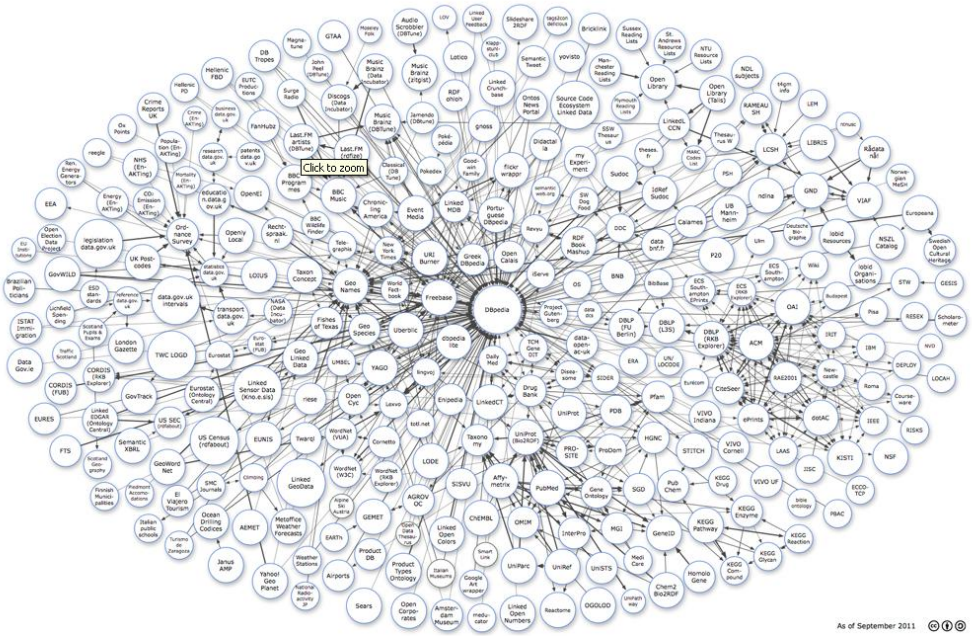

Figure 17 Linked Data Cloud

Currently the European Environment Agency (EEA) is publishing several code lists, taxonomic lists within the linked data technologies.

In the domain of long term ecological research (LTER) the LTER Taiwan network currently published the use of linked data for integrating observation data (see Mai et al. 2011, in press; see Figure 18). For that example implementation 4 different databases, an ontology, and the metadata derived from EML documents were interconnected.

Any of the databases have a data access service, so called RDF wrapper, where the mapping to existing RDF structures (mostly ontologies) takes place. The RDF resources could be accessed by RDF clients directly, but mainly for performance reasons, they stored in a cache, realized by a triple store.

Once published in the RDF data cloud, the data can be accessed by any RDF search-, display- and download client.

The details of the Taiwanese architecture will be published in December and the included data will be, as announced, further on extended.

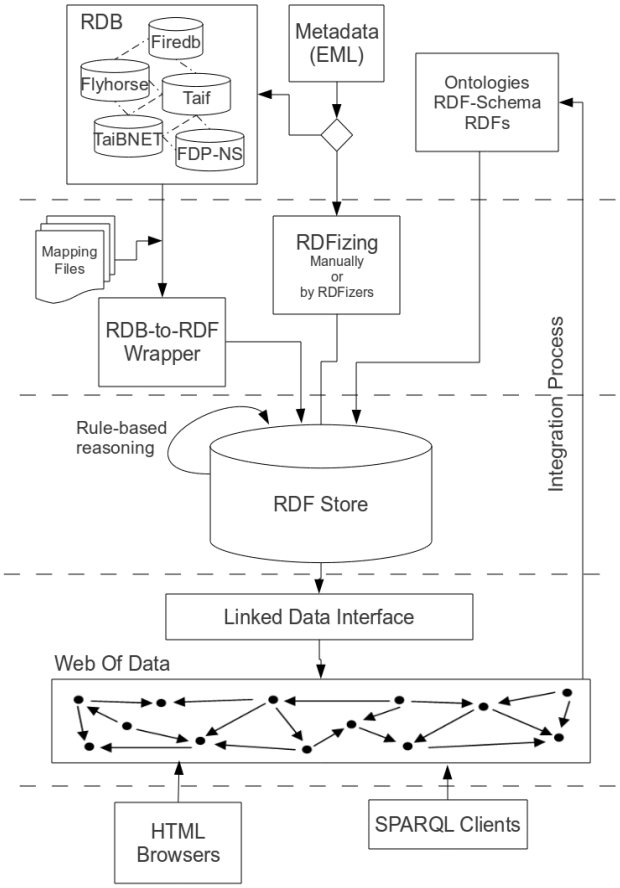

Figure 18 Workflow of Linked Open Data of Ecology (LODE) as example from LTER Taiwan (see Mai et al. 2011)

9 Next steps

A. Data Reporting Format

The Data Reporting Format provides the general model for the definition of the Excel files and the feeding of the collected data. The Data Reporting Format proofed to be sufficient to the selected test datasets. Nevertheless adaptations to future needs could be necessary in the following period.

B. Data management tools

The currently used ftp repository is only the first step to collect the datasets and to allow for changes in the Data Reporting Format without causing further technical implications. As a next step for the data management, a web based data upload and storage using an interface implemented in DRUPAL as web front end will be developed. The general design is already laid out and first implementation steps are done. In this way, since some metadata management facilities have been already developed as DRUPAL interfaces, a unique Web environment will be available to EnvEurope for the description of the dataset with metadata and the upload of the data. This DRUPAL application, with its client facilities, will constitute the EnvEurope Data Portal (see Figure 19).

In this second step, central data storage will be provided, where datasets will be stored in a database. It is offered as central data repository for any data which cannot be directly accessed via a service or need to be cached because of connection problems. In Figure 19 this is indicated by the data cache storing the centrally collected data. Nevertheless this data will be published using web services, like Sensor Observation Service (SOS) or Web Feature Service (WFS) / Web Mapping Service (WMS).

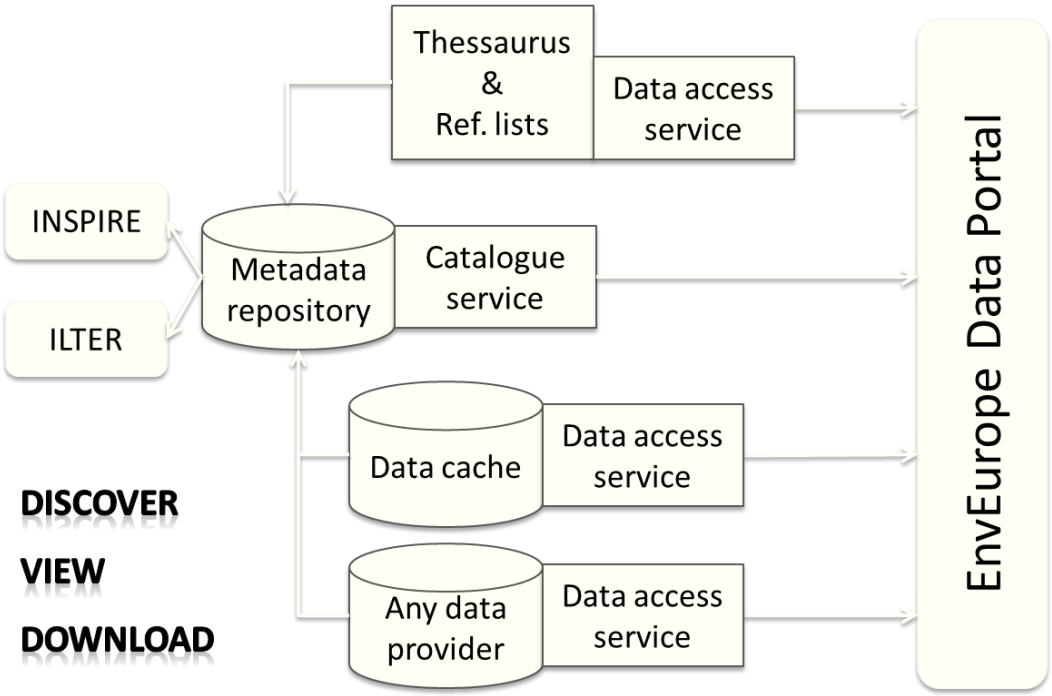

Figure 19 General EnvEurope System Architecture

This general architecture allows for the integration of the different services provided by the beneficiaries and providing an integrated data pool for EnvEurope. Nevertheless, the service based data exchange is the long term vision in the EnvEurope project.

To decide which architecture should be adopted by EnvEurope, we made the following **decision for short and medium term solution in the EnvEurope data management architecture**

In the short term a solution based on OCG services cannot be proposed since there are neither skills in the community nor experiences in related communities available. To opt for this solution would have meant a rather basic IT experiment.

In the medium term the decision between the classic OGC services and linkedData architecture is not that easy. However, some on-going proposals of OGC services supporting linkedData are in the phase of testing, and can pave the path towards a hybrid solution.

SOS allow for the inclusion of:

- acquiring, storing, accessing and visualising of in-situ observations;
- sharing the observations in a standard and interoperable way (Sensor Observation Service - SOS);
- defining and share the metadata of sensors and observations;
- obtaining the features of sensors (geo position, properties, time extent, quality of observations, etc.).

Due to interoperability Sensor Web (SW) in EnvEurope let to:

- delivery and share observations from their own repositories without duplication on centralized data centres;
- discovery the features of observations collected and delivered;
- having common clients that let explore, access and visualize observations distributed by all beneficiaries.

However, to the best of our knowledge, linkedData is a more flexible and future oriented architecture: It allows for the establishment of stable and yet extendable controlled vocabulary, the inclusion of other communities, and interlink to heterogeneous data sources.

Therefore, in the medium term a couple of use cases of application of both architectures and technologies in the realm of EnvEurope will be tested in order to better evaluate their advantages and drawbacks within the community. CNR could take care of the SWE use case, while EEA of the linkedData case. The use cases and the evaluation criteria will be designed in conjunction between A1 core partners and the final solution to be proposed to the project beneficiaries will be discussed on the basis of this evaluation.

C. Data validation

Based on the web based data upload validation routines checking the data (e.g. span of values for a defined parameter) will be needed to be implemented.

Acknowledgements

This deliverable could not have been developed without the valuable and fruitful help and contribution from many members from the working groups of individual actions within the EnvEurope project, and other LTER-Europe project involvements.

The work in Action 1 Data Management in EnvEurope is closely linked to Information Management initiatives especially from the US LTER and ILTER as well as projects on the European level. Therefore we want to thank David Blankman for his input to the metadata profile and application and providing the link to US LTER, Inigo San Gil for providing the prototype of DRUAPL MD Editor for EML and the help in getting along the way, John Porter for the US LTER Controlled Vocabulary, Chau-Chin Lin on providing an insight on Linked Data in LTER Taiwan and the colleagues from EXPEER in having fruitful discussion on the way how to proceed in a European perspective on linking data management approaches from long term monitoring and long term experimental sites.

We also want to thank all beneficiaries for their contribution to the “boring” work of data and information management.

10 References

Adamescu, M., Cazacu, C., Peterseil, J., Datcu, Sabina., Schleidt, K. (2007). Report on LTER InfoBase. [Download 2009-02-02 from http://www5.umweltbundesamt.at/ALTERNet/index.php?title=Image:Report_LTER_InfoBase_version3_UNIBUC.zip]

Adamescu, M., Peterseil, J., Dactu, S., Cazacu, C., Vadineanu, A. (2010). Elements for the design of a General Ecological Database. In: Maurer, I. and Tochtermann, K. (eds.) Information and Communication Technologies for Biodiversity and Agriculture. Shaker Verlag, Aachen. pp. 49-66.

Haberl, H., Winiwarter, V., Andersson, K., Ayres, R.U., Boone, C., Castillo, A., Cunfer, G., Fischer-Kowalski, M., Freudenburg, W.R., Furman, E., Kaufmann, R., Krausmann, F., Langthaler, E., Lotze-Campen, H., Mirtl, M., Redman, C.L., Reenberg, A., Wardell, A., Warr, B., Zechmeister, H. (2006). From LTER to LTSER: Conceptualizing the Socioeconomic Dimension of Long-term Socioecological Research. Ecology and Society 11(2):13. [online URL: http://www.ecologyandsociety.org/vol11/iss2/art13/]

Karasti, H., Baker, K.S. (2008). Digital Data Practices and the Long Term Ecological Research Program Growing Global. The International Journal of Digital Curation 3(2):42-58.

Karasti, H., Baker, K.S., Schleidt, K. (2007). Digital Data Practices and the Long Term Ecological Research Program. Third International Digital Curation Conference, 11-13 Dec 2007, Washington, DC, USA (http://interoperability.ucsd.edu/docs/07Karasti-Baker-Schleidt-DCC07.pdf).

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G. (1997). Non geospatial Metadata for the Ecological Sciences. Ecological Applications 7(1):330-342.

Mirtl, M., Krauze, K. (2007). Developing a new Strategy for Environmental Research and Monitoring: The European Long-Term Ecological Research Network's (LTER Europe) role and perspective. In: Chmielewski, T.J. (Ed.) Nature Conservation Management: From Idea to practical Results. Lublin – Lodz – Helsinki – Aarhus. pp. 36-52.

Vadineanu, A., Datcu, S., Adamescu, M., Cazacu, C. (2006). The state of the art for LTER activities in Europe. (ALTER-Net) Project n. GOCE-CT-2003-505298. [online 2009-02-02 from http://www.alter-net.info/POOLED/DOCUMENTS/a208973/l3023v02_LTER_facilities_report_UNIBUC.pdf]

Michener, William K., James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. 1997. Nongeospatial metadata for the Ecological Sciences. Ecological Applications 7:330–342.

Nogueras-Iso, J. et al. (2005): Geographic Information Metadata for Spatial Data Infrastructures - Resources, Interoperability and Information Retrieval. Springer Verlag

Mai, Guan-Shuo, Wang, Yu-Hwang, Hsia, Yue-Joe, Lu, Shang-Shan, Lin, Chau-Chin (2011). Linked Open Data of Ecology (LODE): A New Approach for Ecological Data. (in preparation)

11 Annex: Data Reporting Format

The data reporting format is attached to the report as Microsoft Excel File.

EnvEurope
Life Environment Project LIFE08 ENV/IT/000399

EnvEurope Data submission

For the EnvEurope project monthly or annual data should be reported for selected parameters identified by Action 3. This includes physical or chemical analysis (e.g. meteorology or air quality) as well as vegetation observation data.

The data will only be used internally for cross domain and cross site analysis. Any other use of the data is prohibited and have to be negotiated with the data holder.

AUTHOR	
INSTITUTE	
EMAIL	
DATE OF SUBMISSION	
DATA VERSION	
COMMENTS	