Combination of CP & CSA

D7.1 – ScanDB Software Specification Report

Project No.262693– INTERACT

FP7-INFRASTRUCTURES-2010-1

Start date of project: 01/01/2011 Duration: 48 months
Due date of deliverable: 30/06/2011 (M6) Actual Submission date: 30/06/2011

Lead partner for deliverable: ITU
Authors: Philippe Bonnet, Yannis Ioannidis

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the Consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

Table of Contents

Publishable Executive Summary

One of the goals of the INTERACT project is to "*make archived and new observations more accessible to a wide range of users* *including developing partnerships with the research community, particularly those using experimental and modelling approaches".* While some groups of researchers have made significant progress towards this goal, e.g. the EU-funded Integrated Carbon Observation System (ICOS; http://www.icos-infrastructure.eu/), it has historically been very difficult to achieve in the polar research community at large, e.g. the data management component of the Internal Polar Year did not meet expectations.

Our premise in Task 7.2 is that **Arctic researchers managing long-term monitoring programs or observation projects do not have appropriate tools to manage data and metadata throughout their lifecycle**. More specifically, we have identified two main problems: (1) While there has been a lot of emphasis on data representation, appropriate tools throughout the data lifecycle have been largely ignored, and (2) the issue of provenance (or data lineage, i.e., what data and derivation processes were involved in obtaining a given data set), which is emerging as a key feature to enable data modelling or simply comparison across sites, is not yet well-integrated into current data management software.

In this report (Deliverable D7.1 below), we focus on two case studies (Biobasis observations in Zackenberg and CO2 flux data from Abisko). For each case study, we describe the activities that take place throughout the data lifecycle, we identify the users that are involved in those activities and the tools that are used to support them. Based on these case studies and on a review of the data management tools available today, we define the baseline functionalities of **ScanDB, a** **repository for ecological data products, that should complement existing tools to more efficiently support data products management throughout their lifecycle**.

1. Introduction

1.1.Purpose and Scope

Task 7.2 focuses on the data management software tools that efficiently support researchers throughout the lifecycle of the data they manage. This goal is aligned with the overall goal of the INTERACT project to “make archived and new observations more accessible to a wide range of users including developing partnerships with the research community, particularly those using experimental and modelling approaches.“

Our main hypothesis is that the accessibility of the data collected at the field stations will significantly improve when researchers rely on appropriate tools throughout the data lifecycle:

1. Obviously, the very first step is to make every step of the data lifecycle explicit. In this document, we rely on two case studies to describe the data life cycle. We document data products and activities throughout the data life cycle for two case studies that are representative of INTERACT observations: Biobasis at Zackenberg, and Carbon Observations at Abisko, and we describe how these activities are supported by existing data management software.
2. The second step is to identify the tools that – if available – would significantly improve the way INTERACT researchers manage, access and utilize their data. In other words, we fix the requirements for the initial functionalities of ScanDB (a repository for ecological data products, that should complement existing tools to more efficiently support data products management throughout their lifecycle) and we describe the iterative process that will allow us to define subsequent versions of ScanDB.

This report (based on task 7.2) is deliverable D7.1 which marks the beginning of a process in two areas:

- Based on the two case studies that we describe, we aim at engaging the INTERACT community to reflect upon their data life cycle, the potential shortfalls that might exist and whether the functions we propose for ScanDB address them.
- We aim at fixing a baseline set of functionalities for ScanDB that will be refined and expanded based on the feedback we obtain from the INTERACT community.

As a consequence, this document is not a complete specification of the ScanDB software. In the reminder of this Section, we review the context of our work and introduce a few definitions that will be used throughout the document.

1.2. Context

Back in 2005, Jim Gray et al. [1] articulated the following vision for the field of scientific data management:” In an ideal world there would be powerful tools that make it easy to capture, organize, analyze, visualize, and publish data. The tools would do data mining and machine learning on the data, and would make it easy to script workflows that analyze the data. Good metadata for the inputs is essential to make these tools automatic. Preserving and augmenting this metadata as part of the processing (data lineage) will be a key benefit of the next-generation tools." In the area of ecological data management, there is a growing awareness that "high quality data management systems are critical for addressing future environmental challenges, requiring a new approach to how we conduct ecological research, one that views data as a resource and promotes stewardship, recycling and sharing of data" [3].

In Task 7.2, we focus on the tools that will improve data stewardship, i.e. what tools are needed to help researchers manage, access and utilize the data they collect at the field stations. The issue of data recycling and sharing are largely orthogonal to the issue we consider in this task; they will be partly addressed in Task 7.1 (see [6, 8, 9] for a discussion of the main issues with the sharing of ecological data).

Many ecological data management systems exist today. Most of them, coupled with GIS systems, are targeted at environmental consultants and site managers (e.g., EquIS from Earthsoft or EIM from Locus Technologies). However, these systems are either expensive or require a steep learning curve and end up being too expensive for many researchers[1], that end up relying on more or less elaborate ad-hoc solutions to manage their data.

Even if the issues related to ecological data management have been identified for some years [3,7], there has been much more focus on metadata standardization (e.g., AON standardization efforts) and data sharing infrastructures [11] (e.g., the EU-funded SEIS platform[2]) than on the tools that researchers can use to manage their data efficiently to complement generic tools such as spreadsheets or database systems. A couple of projects are focusing on such tools:

- *Entangled bank from Microsoft Research[3] focuses on integrating ecological data sources. This is orthogonal to our approach as we focus on the management of the data collected by INTERACT scientists.*
- *EcoData Retriever from ecologicaldata.org[4]. This is a prototype still in beta stage whose goal is to* improve the ecologists' ability to quickly access and analyze data by designing database structures for ecological datasets and then downloading the data, pre-processing it, and installing it into major database management systems. This project focuses on providing a repository of dataset and making it easy for researchers to download and install those datasets on their databases. In contrast, we focus on the data that INTERACT researchers are collecting and deriving (e.g., annotation and lineage management are relevant in our context, but not for EcoData Retriever that focuses on installing existing data sets).

[1] We will be looking forward to the feedback from the INTERACT community to quantify this statement.
[2] http://52north.org/envip/content/introduction-environmental-information-systems-and-services
[3] http://research.microsoft.com/en-us/projects/entangledbank/default.aspx
[4] http://ecologicaldata.org/ecodata-retriever

1.3. Definitions

Data Products: Again, citing from J.Gray et al. [1], *"The raw instrument and simulation data is processed by pipelines that produce standard data products. In the NASA terminology5, the raw Level 0 data is calibrated and rectified to Level 1 datasets that are combined with other data to make derived Level 2 datasets. Most analysis happens on these Level 2 datasets with drill down to Level 1 data when anomalies are investigated."*

Data Lifecycle: The data lifecycle is composed of the following phases – data acquisition (resulting in primary level 0 data), data derivation (that leads to level 1 and level 2 data), data modelling, data curation.

Activities: Activities represent the processes or actions that take place throughout the data life cycle (e.g., manual data collection in the field, or loading data into a database).

Data Management Software: The data management software and the tools that are used to describe, store, query, manipulate, visualize data throughout their life cycle.

2. Case Study: BioBasis at Zackenberg

Biobasis[6], the biological monitoring programme within the Zackenberg Ecological Research Operations (ZERO), is run by the National Environmental Research Institute (NERI). The BioBasis programme includes 35 elements of terrestrial plant, arthropod, bird and mammal dynamics in Zackenberg Valley and adjacent valleys besides monitoring of phyto- and zooplankton in two lakes. Together they intend to cover a wide variety of flora and fauna typical for the local High Arctic ecosystem, and the relations between them. Emphasis is on *populations, phenology, reproduction* and *predation*. The variability of these parameters are virtually unknown for High Arctic Greenland (and partly for the Arctic in general), and they are expected to show pronounced reactions on year to year as well as on long-termed abiotic variations and trends.

2.1.Data Life Cycle

From the Biobasis point of view, the data lifecycle consists of three main activities. They are represented on Figure 1:

- Data acquisition: Most data are collected manually, in the field at Zackenberg, based on a sampling protocol[7]. Today, PDAs are used to collect this data in digital form (e.g., for phenology counts); other data are recorded in notebooks and later typed in into spreadsheets. The notebooks and USB keys that contain the data are transported back to NERI (DMU in danish) by the researchers and technicians (A and B on Figure 1) upon

[5] Committee on Data Management, Archiving, and Computing (CODMAC) Data Level Definitions http://science.hq.nasa.gov/research/earth_ science_formats.html

[6] http://www2.dmu.dk/1_viden/2_miljoe-tilstand/3_natur/biobasis/index.asp

[7] http://www2.dmu.dk/1_viden/2_miljoe-tilstand/3_natur/biobasis/biobasismanual.asp

return from Zackenberg (the term sneakernet on the figure refers to the transfer of electronic information, especially computer files by physically couriering removable media[8]).

- Data cleaning: the Biobasis program manager (M on Figure 1) is in charge of cleaning the data that are collected in the field. Again, the procedure for data cleaning is described in a protocol. In terms of tools, this process is largely ad-hoc. The program manager is relying on a database residing on his laptop (Access) to store the data and on a spreadsheet to visualize the data sets and find possible anomalies visually. Once an anomaly is detected, the data set is annotated, and if there is some missing data then extrapolations are used to generate appropriate data points. In summary, the data cleaning process is manual and based on generic data management tools. Once cleaned up, data sets are exported in text format and forwarded to the data technicians.
- Data publication: The data manager (N on Figure 1) loads the data sets obtained from the program manager into a public repository (Oracle). See 2.2 for a discussion of the structure of this repository. The repository is available via a GIS Web interface[9] that allows the user to download complete data sets based on some metadata parameters (e.g., biotic parameter, location and year).

Figure 2 presents a flowchart that decomposes the three activities presented above and maps their relationships. Figure 3 presents the tools used throughout the lifecycle.

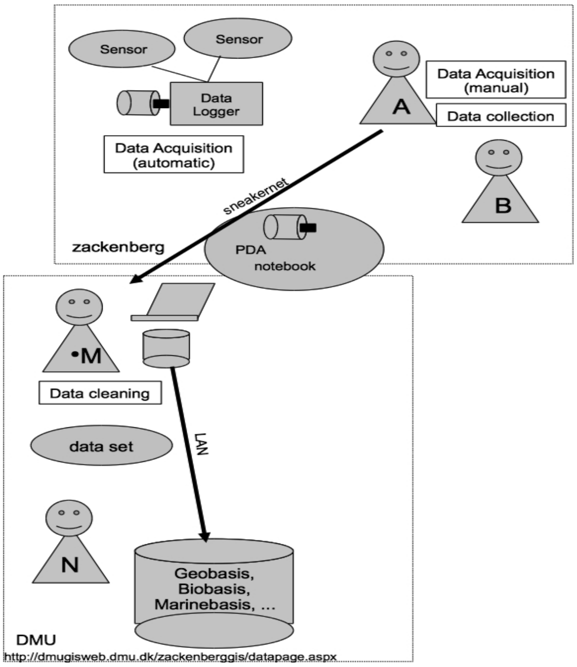

Sensor
Sensor
Data Logger
Data Acquisition (automatic)
Data Acquisition (manual)
Data collection
A
B
sneakernet
zackenberg
PDA
notebook
•M
Data cleaning
data set
LAN
N
Geobasis, Biobasis, Marinebasis, ...
DMU
http://dmugisweb.dmu.dk/zackenberggis/datapage.aspx

Figure 1: Data Lifecycle for Biobasis

[8] http://en.wikipedia.org/wiki/Sneakernet
[9] http://dmugisweb.dmu.dk/ZackenbergGIS

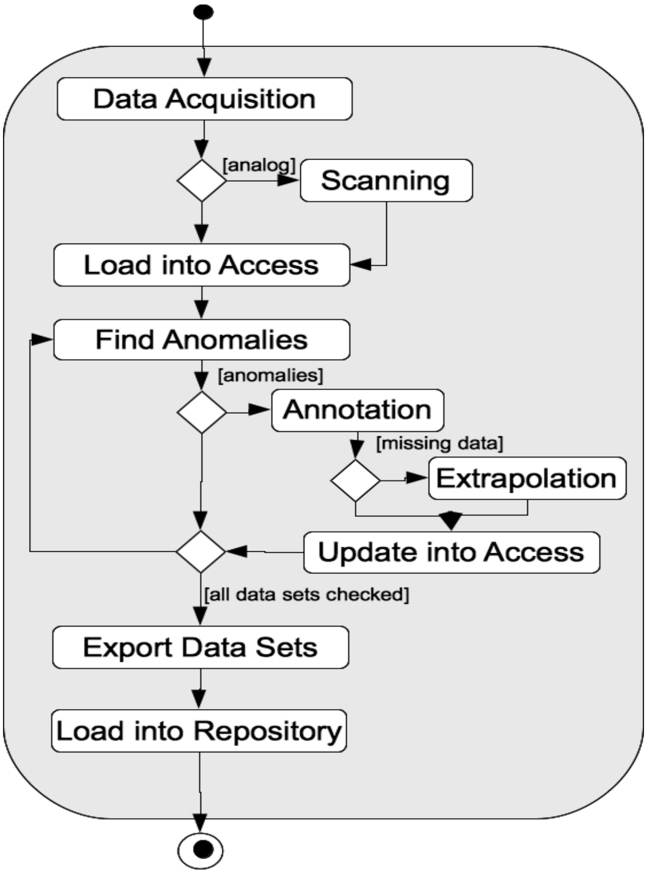


Figure 2: Activities flowchart

Activity	Tool
Data acquisition	Either manual (notes on predefined forms – numbers and multiple choice), or via PDA forms.
Scanning	Manual input of notes on computer.
Loading into access	Ad hoc scripts
Finding anomalies	1. data check out 2. import into excel 3. (possibly) some simple data derivation (e.g., bud-flower ratio for phenology time series) 3. visualization of a time series 4. Manual detection of anomalies in the time series (non monotonic evolution of the bud flower ratio)
Annotations	Ad hoc annotation in the time series
Extrapolation	A new data set is generated based on the primary data and the extrapolated data. This data is then stored into access as a new table. There is no lineage between the original table (primary data) and the table containing the extrapolated data (derived data).
Updating into access	Ad hoc scripts
Exporting data	Each derived table is exported as a CSV file
Loading into repository	CSV files are loaded as data set in the NERI repository (see 2.2 for details).

Figure 3: Tools used in Biobasis throughout the lifecycle

2.2.BioBasis Repository

The NERI repository is organized as a relational database. Its scheme is based on one table for each data set stored in the repository, and four metadata tables (shared across all data sets). The metadata tables are *Elements*, *Variables*, *ElementGroups* and *Programmes;* they are shown in Figure 4 below. Each row in the table *Elements* is the description of a data set. The table *Variables* describes the attributes in a data set. *ElementGroups* categorises Elements into groups; each group can be part of a Research Program described in the Programmes table.

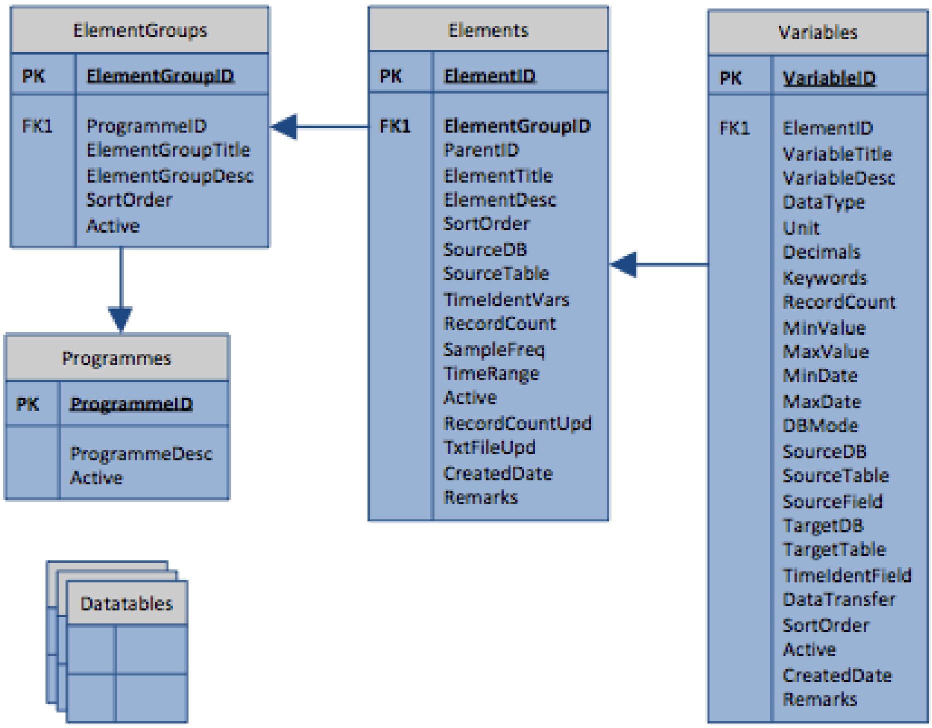


Figure 4: Meta data tables in NERI repository

2.2.1.Current Repository

The repository currently stores 117 data sets. They cover measurements about various Biobasis parameters such as plant phenology, bird counts, water temperature and snow depth collected at specific locations since 1995:

- 45 of the data sets have a global location and specify its measurement location in UTM (Universal Transverse Mercator coordinate system; coordinates (all of them in the Zackenberg area).
- All data sets have temporal information. 81 data sets are accurate to hour or minute. 27 are accurate to date and the rest are either accurate to month or year. No time zone information is given.
- Eight tables have UTM coordinates specified. They all omit UTM zone; it can only be derived from explicit information.
- The table attributes data types are set accordingly to content. The data types strings, float and int are the most used.
- The table with least attributes has 3 attributes. The largest one has 44 attributes. The mean number of attributes are ~13 attributes and median is 9 attributes.
- The data set tables store approximately six million facts. 23 tables have more than 100.000 facts. The mean and median are ~49.000 and ~3590 respectively.
- The data sets are regarded as immutable. Either an updated table are stored or a new data set is created. No history of updates is maintained.

- The data sets use 255.8MB of storage in the database.
- There are no foreign key constraints between data tables and metadata tables.
- It is impractical to query information across a large number of tables. The number of tables grows linearly with the number of data sets.
- The database is designed for manual insertion of information. e.g. metadata tables may have an "active" attribute that is of data type varchar and is initialized to 'x' if active.
- The data types used throughout the database are integer, float, datetime and varchar.
- The metadata tables *Elements* and *Variables* both store aggregated data from the data tables, with fields such as min/max and date, source, active, etc. i.e. these data can be easily retrieved from the database without retrieving the actual data table. These tables also contain information about origin, placement and dates of different events. (e.g. DataTransfer and CreatedDate).

Because of the flat table structure, it is very fast to load data and extract data from data sets[10]. On the downside, several metadata tables must be updated on insertion of a new data set. In addition, a lot of maintenance and integrity constraints are ignored in the database model. Instead, these constraints have to be manually enforced by the data manager. Also, the flat scheme with a table per data set prevents efficient querying over/across a large number of data tables.

2.2.2.Requirements not met by the Current Repository

Provenance. In an ideal world, ecologists can annotate and share their data seamlessly without thinking about where the data are stored and how they are maintained. The repository should store all information about a data set: its content, its derivation history, as well as annotations.
For example a data set that is converted from raw form into human readable form, it should be annotated with what kind of transformation has been applied, which previous/current data sets were involved in the transformations, possibly annotations about the transformations that have been applied (change time representation, origin) and the reasons for the changes.
Origin and history of data are the key element in assessing the soundness and quality of a data set. Provenance enables ecologists to understand the evolution of their data and can help them trace back the origin of an error or an interesting phenomenon.

Query capabilities. Also, it should be easy to find and query data in a repository. It should be possible to easily formulate queries that ask for spatial and temporal information and the associated data to it. More specifically, the following queries should be efficiently supported:

1. *Spatial queries:* Where is the node located that samples are coming from? What probes are also active in a given area around a specific data set?
2. Temporal queries: What is the status of a data set at a specific point in time? What nodes reported in a specific period? How many samples are collected in a period of hours?

[10] An obvious caveat concerns the format of the data – most measurements are float and care needs to be taken when casting floats across various formats (say from a file generated by a data logger to the Oracle float format). This problem has been relevant for Geobasis for example.

3. Spatio-temporal queries: Is there any correlation between node samples in a special area at a given period? For example, ten nodes are installed in Greenland that measure the snow depth. Each is manually sampled. It may be possible to detect outliers by looking at correlation over these data sets.

The DMU repository does not efficiently support any of these queries. To support more advance queries, a more elaborate database structure is required as well as tools that make it easy for ecologists to interact with their database. A priori, it seems that the functions that we describe above can be integrated on top of existing database management systems.

No usage control. The data sets stored in the repository are made available online via a web interface. These data sets are public domain and can be accessed and used without restrictions. Today, it is impossible for NERI to monitor who is actually using these data and how. The best the data manager can do is to monitor who is downloading the data sets – but while it is an interesting statistic it is not enough to enforce that credit is given to NERI for the utilization of the data sets. Conversely, if a data set is annotated or is invalidated (because an error has been detected in the cleaning process), then there is no systematic way to inform those who are using this data set that a new version is available.

3. Case Study: CO2 Flux Data at Abisko

The carbon dynamics of Arctic ecosystems is a very interesting topic of study in the context of the predicted climatic changes. Wetland and associated palsa mires are particularly relevant as they often contain large pools of carbon. In this section, we focus on the life cycle of the data collected at Stordalen mire (68° 21'N, 19° 02E) in northernmost Sweden, where measurements of CO2 (as well as methane CH4) flux measurements have been carried out since summer 2000 [12]. While similar measurements are collected at Zackenberg in the context of Geobasis[11], we focus on the measurements from Abisko because they are not yet part of a long term monitoring program, and thus representative of the measurements and observations obtained by researchers in the context of a given project. Because the carbon lifecycle is so important, generations of projects that federate measurement activities across many sites have supported continuous measurements for more than 10 years. Such projects include Carbo Euroflux (FP5 2000/2003 – including 26 eddy-flux sites throughout Europe[12]), NECC (Nordic Centre for Studies of Ecosystem Carbon Exchange and its Interactions with the Climate System – initiated by the Joined Committee of the Nordic Natural Science Research Councils (NOS-N), the Nordic Council of Ministers and the NorFA in the period 2003/2007 - including 26 eddy-flux sites with measurements in forests, agricultural land, wetlands and above lakes in the Nordic regions[13]), and most recently ICOS (FP7 infrastructure project with a preparation phase 2008-2011 and an operational phase up to 2031 on 30 ecosystem sites across Europe[14]).

[11] http://www.zackenberg.dk/monitoring/geobasis/
[12] http://www.bgc-jena.mpg.de/public/carboeur/projects/cef.html
[13] http://www.necc.nu/
[14] http://www.icos-infrastructure.eu/

3.1.1.Data Life Cycle

Figure 5 that describes the data life cycle for Biobasis also fairly accurately describes the data life cycle for the flux measurements at Abisko. On site data acquisition with couriering of data via removable media to a program manager that proceeds to data derivation followed by a cleaning process, and finally the resulting data product is published on a publically accessible web site.

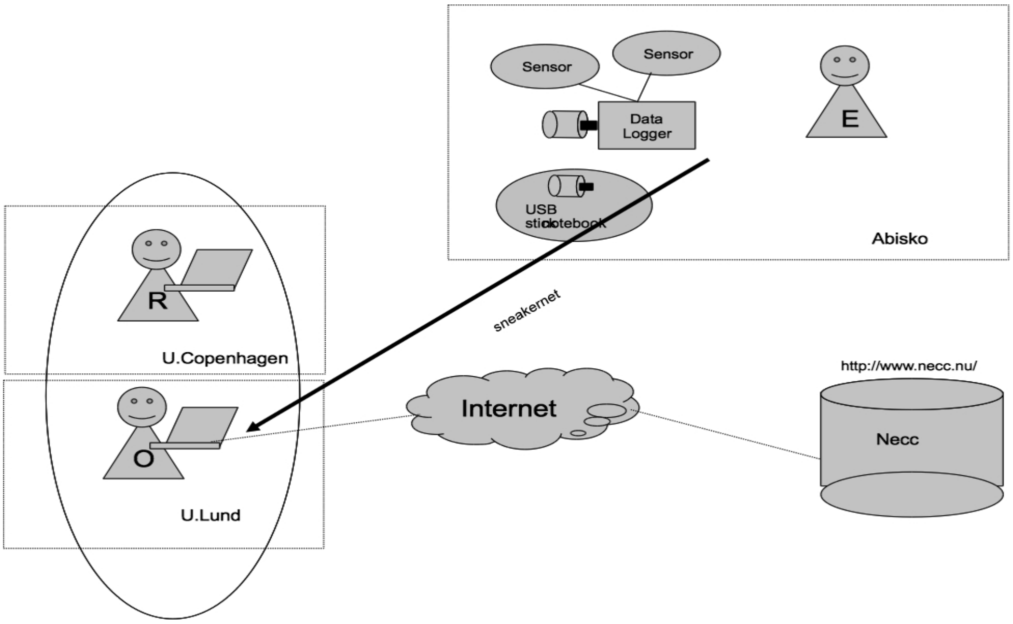


Figure 5: Data life cycle for CO2 flux data at Abisko.

More specifically, the data life cycle can be described as follows:

1. Data is automatically collected (in digital form) using data loggers connected to relevant sensors. There are two types of measurements involved in CO2 flux measurements: either eddy correlation measurements (with sensors that measure wind speed, temperature, CO2 and water concentration[15]) or chamber measurements (with a closed chamber where CO2 concentrations are measured and CO2 flux obtained by derivation). The measurements are stored on a data logger, nowadays a CR1000 from Campbell Scientific. Data is regularly extracted via USB key – some places the data loggers are available online (facilitating this transition to online data loggers is largely the goal of WP5). Note that these primary data are time series. Typically, the primary data can be captured at 10-20 Hz, which gives around 50-100 MB of data per day. The primary data is stored on files.
2. Methods have been developed and shared across the years with standard software emerging as efficient tools to derive flux data from the primary data. In 2006, Mauder et al. compared seven tools focused on eddy covariance measurements. Mikhail Mastepanov at U.Lund developed a visual tool focused on deriving and controlling flux data obtained from chamber measurements (see Figure 6). In the context of ICOS, a new generation of software tools is being developed, which is destined to become a standard in the

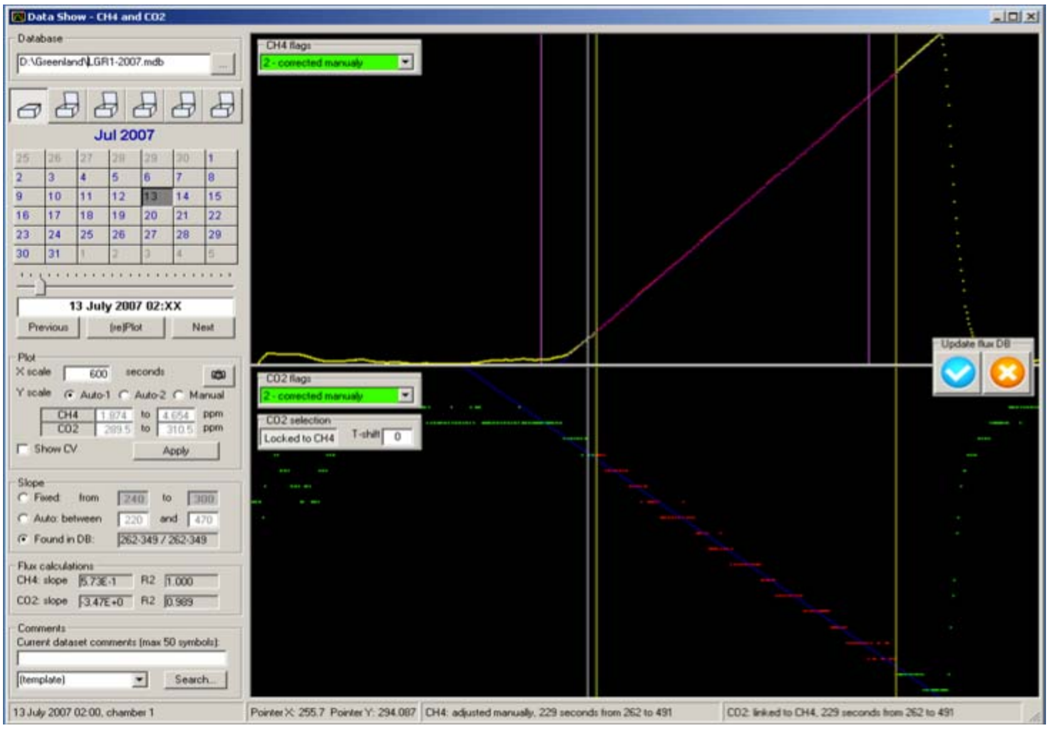
[15] http://en.wikipedia.org/wiki/Eddy_covariance

community. The derived times series corresponds to a flux per time unit – typically there is a data point per half hour. At this point, data is stored on files or on an Access database.

3. Once flux data is derived, data cleaning consists in finding and fixing anomalies in the resulting time series. This data cleaning process is performed on the derived data rather than the primary data because of the reduced size of the data set. It is interesting to note that the data sets are now managed with a WISKI system – basically a time series management system on top of which a set of tools have been defined, originally for water management[16]; the reason for adopting WISKI is twofold (a) ease of loading the primary data obtained from the data loggers into the repository and (b) ease of finding and fixing anomalies in the time series.
4. So far, not all the data sets obtained at Abisko for CO2 flux are available online. These measurements are not part of the data sets managed by the data manager associated with the Abisko Scientific Research Station. So, it is up to the researchers that collect the data to ensure that data are published. This is a significant overhead – mainly because there are no appropriate high level tools to support the data publication process (note that data was never loaded into a database throughout the life cycle defined above). Some data sets were published on the NECC web site – but this site is no longer live. Summaries of some data sets are published on the fluxnet web site[17]. ICOS should provide a long term support for data publication; however ICOS does not focus on making it easier for individual researchers (or research groups) to manage their data – e.g., there is no repository defined (yet). Note that CO2 flux data obtained at Zackenberg in the context of Geobasis have been stored until 2006 in the repository presented in Section 2 (this archival process stopped because of problems loading the data).

[16] http://www.kisters.eu/english/0/8FB26FB39796717FC125737F005CE7CB/$FILE/WISKI_Overview_US_letter_email.pdf

[17] http://www.fluxnet.ornl.gov/fluxnet/index.cfm



Data Show - CH4 and CO2

Database: D:\Greenland\LGR1-2007.mdb

13 July 2007 02:XX

Plot
X scale: 600 seconds
Y scale: ● Auto-1 ○ Auto-2 ○ Manual
CH4: 1.874 to 4.654 ppm
CO2: 289.5 to 310.5 ppm
□ Show CV

Slope
○ Fixed: from 240 to 300
○ Auto: between 220 and 470
● Found in DB: 262-349 / 262-349

Flux calculations
CH4: slope 5.73E-1 R2 1.000
CO2: slope -3.47E+0 R2 0.989

Comments
Current dataset comments (max 50 symbols):
(template)

CH4 flags: 2 - corrected manualy
CO2 flags: 2 - corrected manualy
CO2 selection: Locked to CH4 T-shift 0

13 July 2007 02:00, chamber 1
Pointer X: 255.7 Pointer Y: 294.087
CH4: adjusted manually, 229 seconds from 262 to 491
CO2: linked to CH4, 229 seconds from 262 to 491

Figure 6: Mikhail Mastepanov tool for visualizing chamber measurement and derived flux data

4. ScanDB Functional Requirements

4.1.Lessons learned from the case studies

The two case studies presented above are meant as points of references for (a) the INTERACT community whose needs we aim at investigating further, and (b) the design of ScanDB. The parameters that are monitored and the derivation and cleaning processes are significantly different. Let us first focus on these differences:

- In the Biobasis case study, the primary data that are collected are – once cleaned up – the data products that are published. In the CO2 flux case study, the primary data that is collected must be derived to obtain the data products that are published. This derivation process requires intimate knowledge gain in the field.
- While most of the Biobasis data is collected manually at a low resolution in space and time and for a limited period (Zackenberg is not open all year), the primary data is collected in the context of the CO2 case at a rate of 10-20 Hz all year long.
- Many of the data collected manually in Biobasis correspond to counts which are represented by integers. The data collected from the sensors involved in CO2 flux derivation is represented as floats – with a precision that varies from sensor to sensor. Note that sensor-based data collection in Biobasis will lead to similar concerns in terms of data representation and precision requirements.

Let us now turn to the commonalities. They are significant:

- All data both primary and derived are spatial time series. Whether it is possible to isolate a few operations on those time series that are recurrent throughout the derivation and cleaning processes relevant to the INTERACT community is an open question.
- Even if the primary and derived data sets that are considered are all spatial time series, the data management tools that are used by researchers range from plain files, to database systems (again ranging from Access to Oracle) and even elaborate tools based on a time series management system (WISKI). Today, the main barrier to using elaborate data management systems is the overhead of loading data into it. This is basically the first barrier when using a data management system. If we do not make it easier for researchers to load data into an advanced data management system, they will never use it.
- In both case studies, ecologists care about (a) the tools that they can use to speed up data cleaning/validation processes, and (b) the tools that they can use to make it easier to share their data, document the quality of the published data sets (provenance) and control how data sets are accessed and used.

4.2.Functional requirements

As a consequence, we can define the following goals for ScanDB V0:

1. Define a repository structure that leverages the spatio-temporal dimension of all data sets. The goal is to allow researchers to search and correlate data sets (and not just the meta data) and to provide support for provenance.
2. Define a tool that makes it easy to load data into such a repository.
3. Define a tool that makes it easy to find and fix anomalies on the time series stored in the repository.
4. This repository should be accessible by existing tools (for visualization or automatic validation).

References

[1] Jim Gray, David T. Liu, Maria A. Nieto-Santisteban, Alexander S. Szalay, Gerd Heber, and David DeWitt. Scientific Data Management in the Coming Decade. ACM SIGMOD Record. Volume 34 Issue 4, December 2005

[2] L.Osterweill, L.Clarke, A.Ellison, E.Boose, R.Podorozhny, A.Wise. Clear and Precise Specification of Ecological Data Management Processes and Datasets Provenance. IEEE Transactions on Automation Science and Engineering. Vol 7, No 1, January 2010.

[3] W.Michener, J.Brunt. Ecological Data: Design, Management and Processing. Wiley-Blackwell; 1 Edition. February 2000.

[4] Samantha Romanello, James Beach, Shawn Bowers, Matthew Jones, Bertram Ludäscher, William Michener, Deana Pennington, Arcot Rajasekar, Mark Schildhauer. Creating and Providing Data Management Services for the Biological and Ecological Sciences: Science Environment for Ecological Knowledge. International Conference on Scientific and Statistical Database Management (SSDBM). 2005.

[5] NOAA’s Environmental Data Management:
Integrating the Pieces. An Assessment of NOAA’s Environmental

Data and Information Management. March 2006

[6] J. Porter. A Brief History of Data Sharing in the US Long Term Ecological Research Network. Bulleting of the Ecological Society of America. January 2010.

[7] R.Ingersoll, T. Seastedt, M. Hartman. A Model Information Management System for Ecological Research. Computers in Biology. Vol 47, Nr 5. May 1997.

[8] Credit where Credit is Due. Nature Biotechnology. Vol 27, No 579. 2009.

[9] Bryn Nelson. Data Sharing: Empty Archives. Nature 461. 2009.

[10] W K Michener, James W Brunt, John J Helly, Thomas B Kirchner, Susan G Stafford. Nongeospatial metadata for the ecological sciences. Ecological Applications, 1997.

[11] Robert B. Cook, Richard J. Olson, Paul Kanciruk, and Leslie A. Hook. Best Practices for Preparing Ecological and Ground-Based Data Sets to Share and Archive. Environmental Sciences Division, Oak Ridge National Laboratory. October 2000

[12] Thomas Friborg, Torbjörn Johansson, Marcin Jackowicz-Korczynski, Torben R. Christensen and Patrick M. Crill. Palsa Mires – CO2 exchange from Stordalen mire. Proceedings of the PALSALARM symposium Abisko, Sweden 28–30 October 2009.

[13] M. Mauder, T. Foken, R. Clement, J. A. Elbers, W. Eugster, T. Grunwald, B. Heusinkveld, and O. Kolle. Quality control of CarboEurope flux data – Part 2: Inter-comparison of eddy-covariance software. Biogeosciences, 5, 451–462, 2008