Project no. GOCE-CT-2003-505298
ALTER-Net

A Long-Term Biodiversity, Ecosystem and Awareness Research Network

Requirements for an IT Framework for ALTER-Net
Results from the Requirements Analysis conducted by Work Package I6

Katharina Schleidt; Mandy Lane; Herbert Schentz

WPI6_2005_01

Instrument: Network of Excellence
Thematic Priority: Global Change and Ecosystems (Sub-priority 1.1.6.3, Topic 6.3.III.1.1)
Due date of deliverable: 1st January 2005
Submission date: 16th March 2005
Start date of project: 1st April 2004
Duration: 5 years
Deliverable lead contractor: umweltbundesamt
Revision: 1.0

Requirements for an IT Framework for ALTER-Net Results from the Requirements Analysis conducted by Work Package I6

CONTENTS

1 Executive Summary

The goal of the ALTER-Net WP I6 is the creation of a framework for sharing data, information and software tools amongst the ALTER-Net partners in support of biodiversity research, policy and public understanding of science. In order to ascertain the requirements for such an information framework, we designed a questionnaire and circulated this among the network partners. This document summarises the results of this questionnaire, and further information gained from the subsequent workshop, held jointly with I3 (LTER networks) in March 2005, in Bratislava (see also workshop report).

Of the 24 network partners contacted about the questionnaire 11 returned filled out responses, 4 replied that their institutions do not work at this level with data and 9 did not respond despite repeated reminders, where we are still hoping that the lack of response was due to lack of time or problems reaching the responsible parties instead of lack of interest in the subject matter.

The need for such a network is high; the vast majority of the responders consider such a network necessary or desirable. The data available in the network would be required mainly for scientific use. The network partners wish access to raw or filtered/normalized data. The requirement for complete analysis is quite low.

When questioned on what types of analysis and modelling tools required, the strongest need was for publishing and reporting tools, followed by statistics and geostatistics tools. The need for sharing of modelling tools was less predominant, though this may be due to the specialized questions asked. Most of the software currently being used by the partners was classified as homemade, which indicates a good deal of experience in creating such tools within the institutes. There was also a strong interest shown in creating such tools together with network partners.

Most of the data available within the ALTER-Net network pertains to the terrestrial biosphere followed by the aquatic biosphere. This is partially due to the abundance of different life forms as well as the greater general interest in this area. While there is less data pertaining the atmosphere, hydrosphere and pedosphere available, the parameters documented in this area tend more to physical parameters which are easier to communicate and share. Most of the sites being observed tend to be of a point type, although many sites also monitor areas.

The prerequisites for the creation of a data exchange network are mostly present, with most institutes within the network storing their data centrally and willing to provide access to this data free of charge. Unfortunately, very few of the institutes currently offer their data online. This is partially due to a lack of standards and a clear concept of what networks provide the necessary functionality to make it cost-effective to participate. Discussions at the workshop revealed a lack of data management support at some of the partner institutes represented. Some of these had not felt able to complete the questionnaire. It will be important for I6 to find ways of engaging with these institutes to ensure their inclusion.

In order to determine what existing system to adopt or what technology to use to build a new system, following steps must be taken:

- Review existing technological developments
    - Ecogrid (SEEK, UK-EcoGrid, MyGrid, GBIF, TEMS/GTOS)
    - State-of-the-Art in Metadata Registries
    - State-of-the-Art in Grid Technology, especially the data exchange interfaces
    - State-of-the-Art in Ontology languages and tools
- Analyse what these systems can offer and what further developments would be required for ALTER-Net purposes.

- Determine which sites could best be involved in a prototype data grid.

Based on the questionnaire survey and workshop, we can conclude that there is a clear requirement for sharing data and tools amongst ALTER-Net partners. In addition, some of the necessary building blocks for an information framework are in place: many of these partners already have centrally stored data gathered according to standards, plus expertise in the development of sharable analytical tools. The challenge for ALTER-Net will be to maximise the use of these existing systems and at the same time accommodate those partners with less well developed information management facilities.

2 Introduction

In order to fulfil the I6 mission objective: *"To construct a framework within which can be built a system to manage biodiversity data, information and knowledge from the NoE, and to make them available to scientists, policy makers and the public"* the first questions we needed to ask the project partners were the following:

- What data is currently available in what form within your institution?
- What is the availability of the data existing within your institution?
- What data would you require access to from other institutions?
- What ways of access would you require for data from other institutions?
- What tools do you require to further process the data (your own and that made available to you from other institutions)

In order to gain an overview of the partners' requirements, we designed a questionnaire covering all of these topics and circulated this amongst the project partners. This paper is an analysis of the results of this questionnaire.
These results were subsequently discussed at the first I6 workshop, held jointly with WP I3 (LTER networks) in March 2005 in Bratislava (see also workshop report). The workshop provided useful feedback on the questionnaire responses, further insight into requirements and into the arrangements for data management and analysis at partner institutes. By pairing the partner requirements with our overview of existing initiatives towards networking ecological data (SEEK. GBIF, Ecoinformatics…) this has enabled I6 to define a more detailed workplan, incorporating as many partner requirements as possible whilst assuring accordance with emerging standards where possible.

3 Methodology

As many of the project partners are working in a wide area of fields (domains), we designed the questionnaire with individual columns for answers pertaining to the specific domain (Biosphere Terrestrial, Biosphere Aquatic, Atmosphere…). In the part pertaining to existing data, we further divided these domains into relevant sub domains (Biosphere Terrestrial: vertebrates, invertebrates, micro organisms… Atmosphere: air chemistry, meteorology; …). In this way it was possible for institutes to differentiate between the types of existing data reported as well as their requirements dependant on domain.
For most analysis purposes, we have been regarding each pair of institute and domain as an individual entry (CEH – Hydrosphere, INSU – Socioeconomic …). We have done this based on the assumption that different departments within the institution are involved with the individual domains, and thus these may be seen as individual units. In those situations where

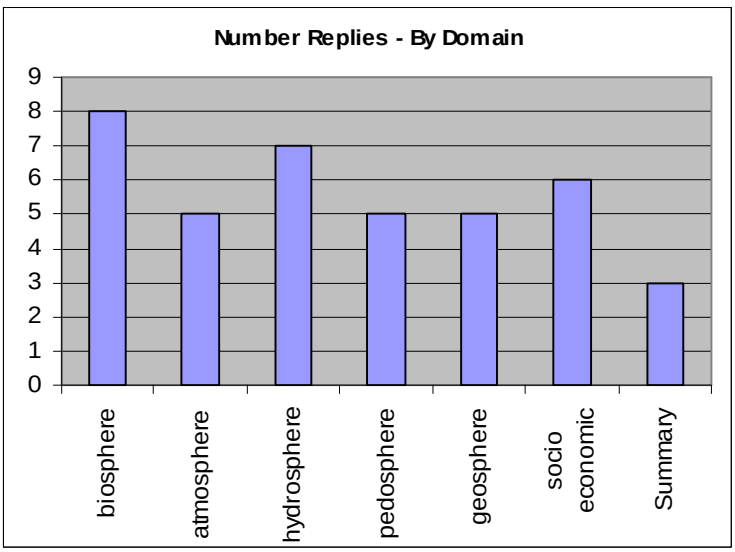
we looked at the data by domain, i.e. the number of replies by domain, it seemed to make more sense to just work with the total number of institutions responding.
As many institutions only responded for one or a few domains, the first step taken was to create one data source encompassing just the filled in columns from all responding institutions. As some institutions elected to only fill in the summary column whilst leaving the individual domain columns empty, it was decided to treat "summary" as an individual domain.
In order to better analyze the data, it was then transformed so that the sum of the answers to one question and domain were apparent. In cases where we specified explicit answers such as [YES|NO|CAN'T SAY] we created a sum of the number of responses for each allowed answer (3 answered YES, 1 NO, 2 CAN'T SAY). In cases where the answers were not explicitly specified (number of sites, list of parameters …) all responses were listed allowing us to total numeric values and gain an overview of free text responses.
The following analysis was based on this transformed data.

4 Questionnaire Response Statistics

Although we would have wished for a far wider response from the various ALTER-Net partners, we did receive replies from 11 of the 24 partner institutions we contacted. Four institutions replied that they couldn't answer our questionnaire for reasons such as

- they neither produce nor analyze data
- All data within the institution is so decentralized (on the pc of the scientist working with it) that they couldn't provide any information
- Aren't involved in I6

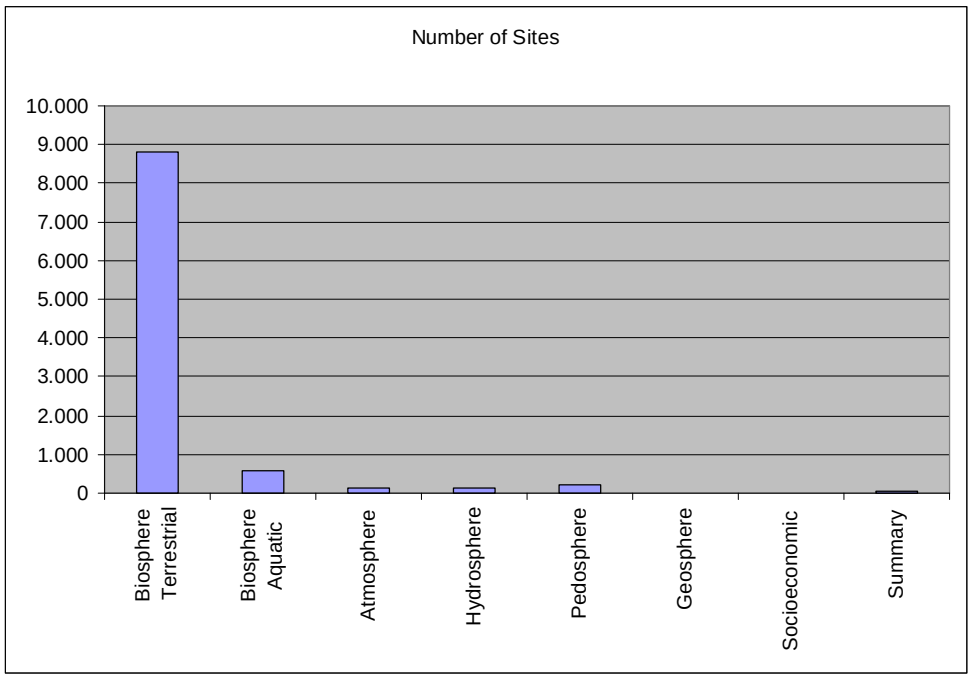
The remaining 9 Institutions returned no answer, where we are still hoping that the lack of response was due to lack of time or problems reaching the responsible parties instead of lack of interest in the subject matter.
So – of the 18 Institutions officially committed to working on I6, we got about a 60% response rate (less than we'd hoped for, but at least more than half!)
As most institutions only work in certain areas, one of the first questions we tried to answer is what the predominant domains are



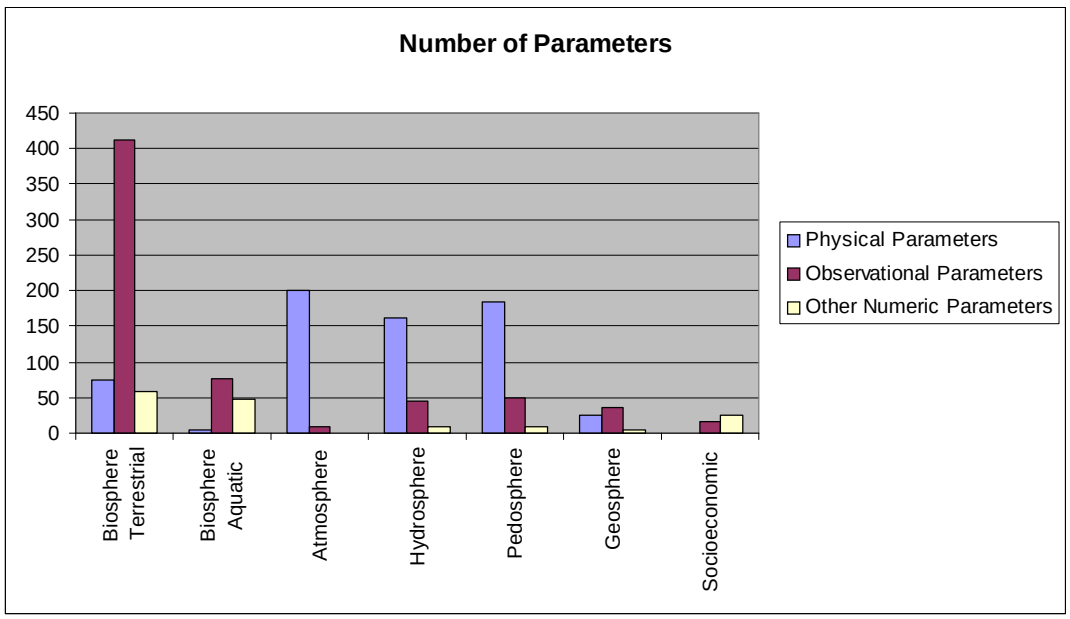
5 Existing Data

5.1 Sites and Measurements

Most of the existing data reported (number of sites) pertains to Biosphere Terrestrial with over 8000 sites, followed by Biosphere Aquatic, with a bit over 500.

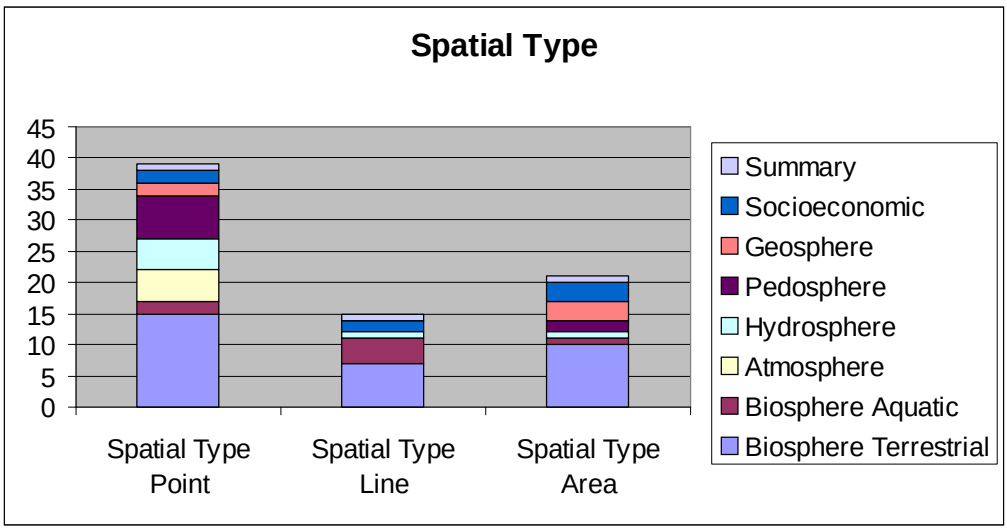


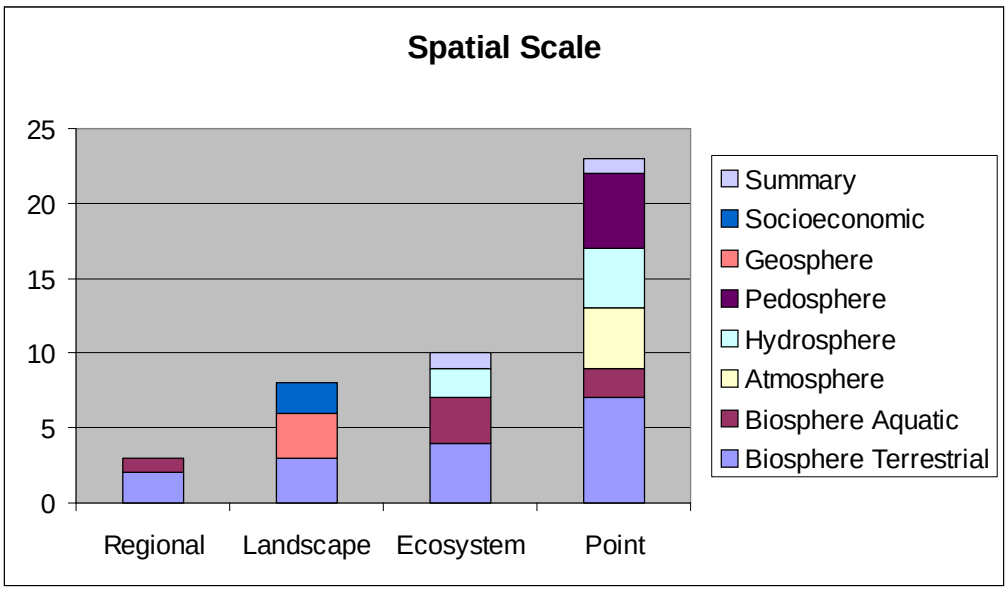
In the Terrestrial as well as the aquatic biosphere the number of observational parameters clearly dominates, although this value is strongly influenced by the large number of observational parameters pertaining to invertebrates reported by CEMA (310 of the 412 total reported). Physical parameters dominate in the areas Atmosphere, Hydrophere and Pedosphere.



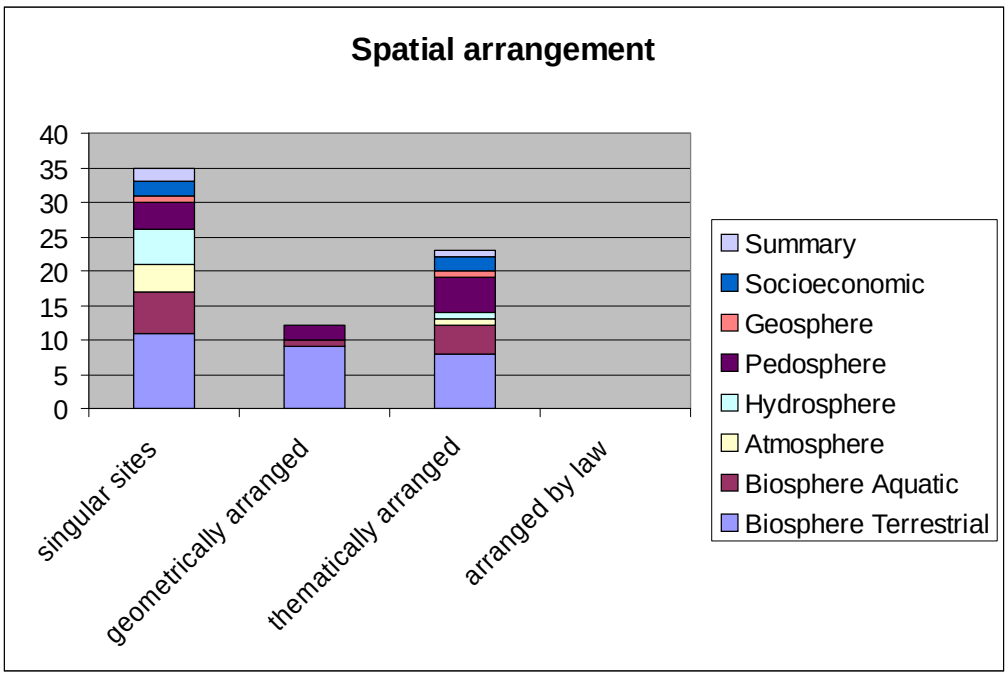
The following definitions for physical and observational parameters were used in conjunction with the questionnaire:
Physical Parameters: Number of parameters that are measured by a chemical or physical procedure.
Observational Parameters: Number of Parameters that are determined by scientific expertise.

Most of the data reported is of point type, followed by areas. A few partners reported data of line type. This is also reflected in the spatial scale of the sites, where the majority was again of type point, followed by Ecosystem and landscape, with regional only being reported for two terrestrial and one aquatic biosphere site.

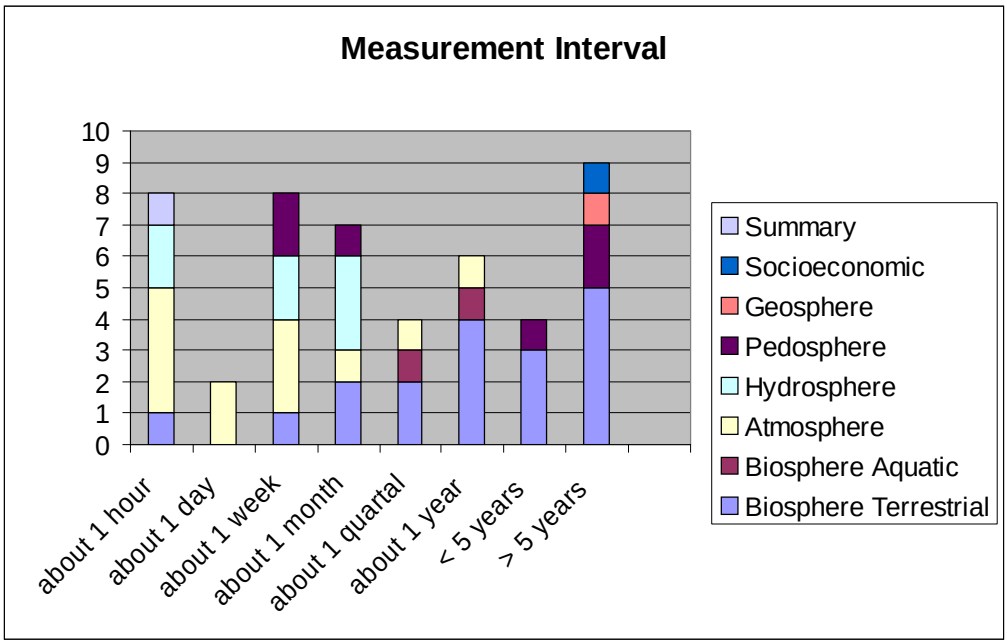




The spatial arrangement of sites tends to singular sites, although many partners have arranged their sites thematically. A few partners have arranged their sites geometrically, whereas no partners arrange their sites due to legal requirements.



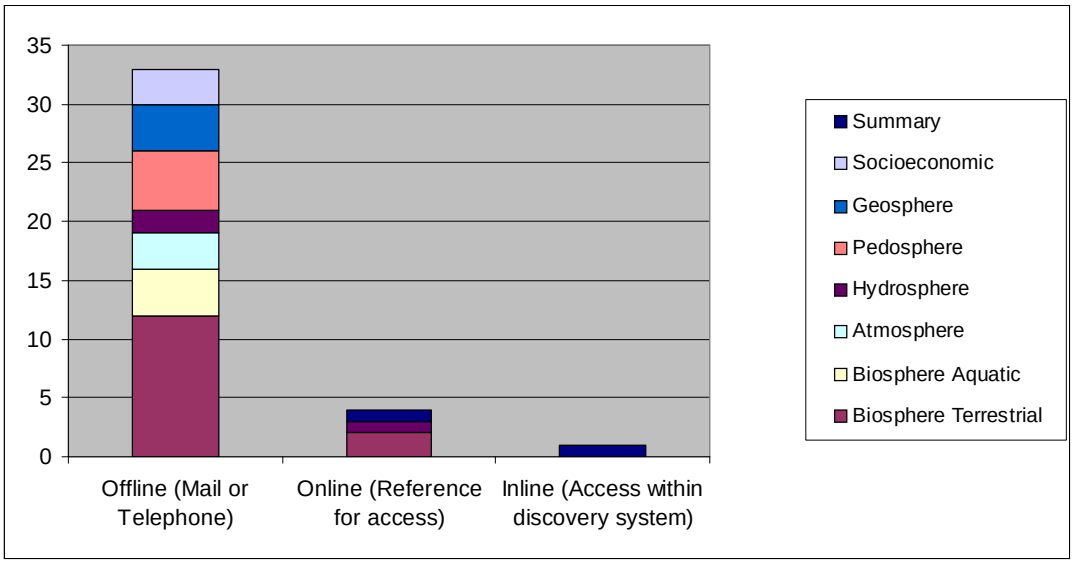
Measurement intervals reported varied by domain, with longer intervals (yearly or longer) more prevalent in terrestrial biosphere monitoring and pedosphere monitoring whereas short intervals (hourly to monthly) being predominant in atmospheric and hydrospheric measurements.



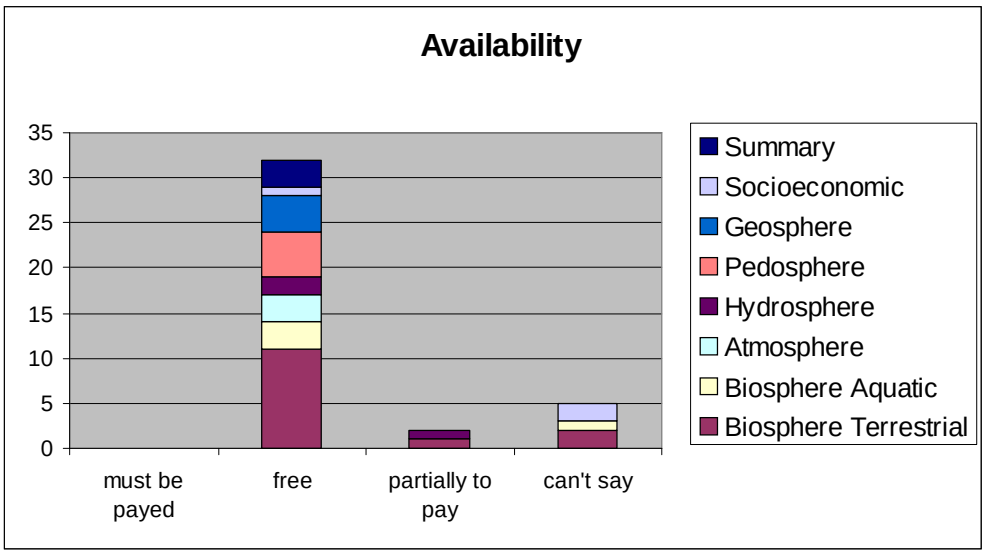
5.2 Accessability

At present, very few of the partners offer their data online, and very few are members of existing networks. Two partners are involved in the GBIF Biodiversity network, although only one is currently a data provider. Four partners are involved in the UN ECE Integrated Monitoring network. Two partners report data within the framework of the Waterframe

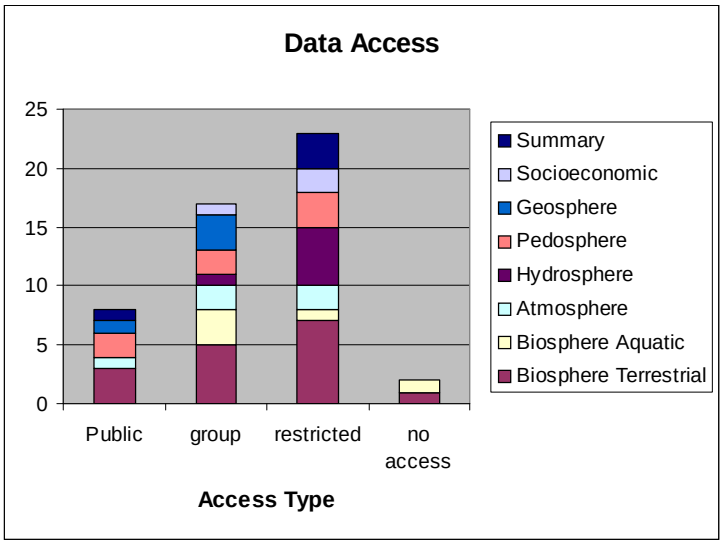
Directive. Seven of the eleven responding partners are involved in national networks and three of the partners are involved in other networks, mostly LTER networks



While a few of the partners charge for access of their data, the overwhelming majority is willing to share their data free of charge. This is especially important, as none of the partner institutions are willing to pay for the use of data from the network.

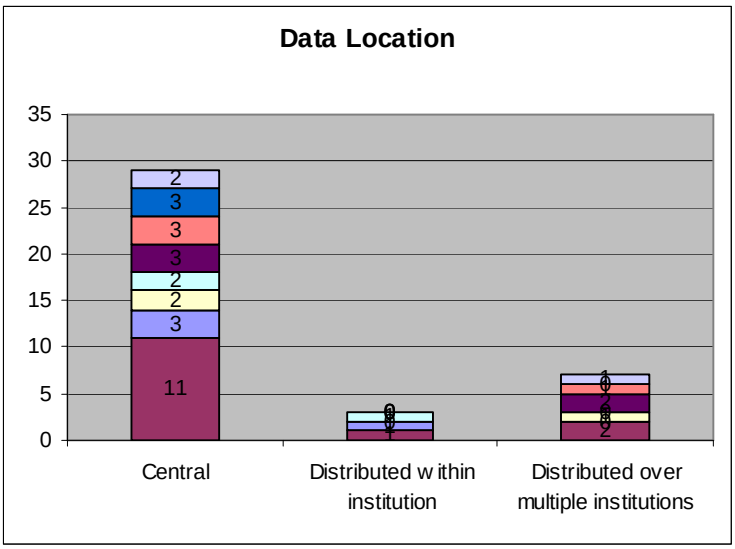


Currently most partners restrict the access to their data, with very few willing to provide their data to the general public.

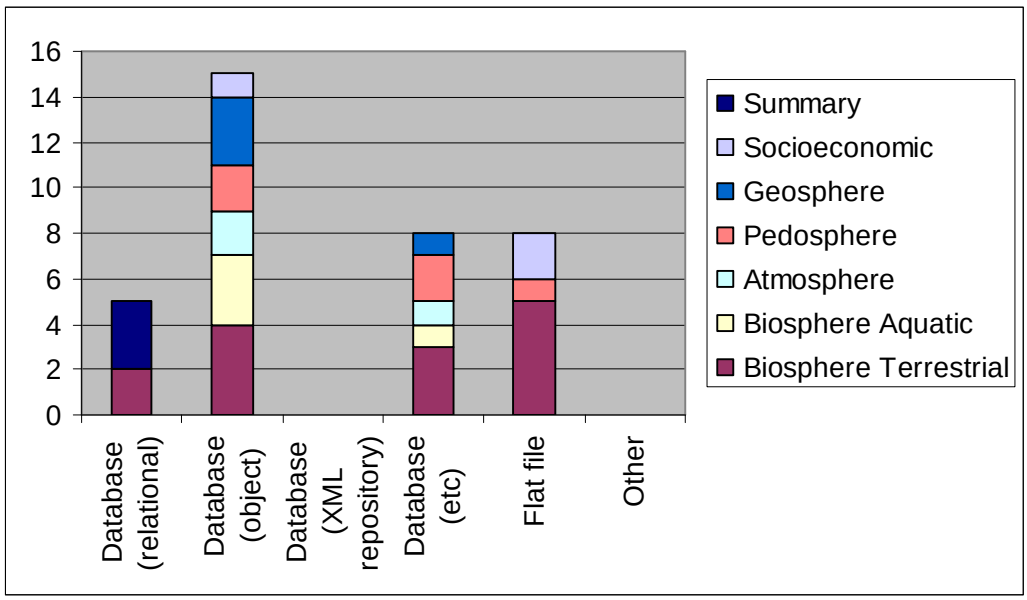


5.3 Data Storage

The fact that most of the partners responding to the questionnaire store their data within a central database is quite positive for our plans of networking the existing databases as it is much easier to tap central resources (one point of contact, one place to implement an interface) than those spread out over various locations, or even worse stored on the desktop pc of an individual scientist.

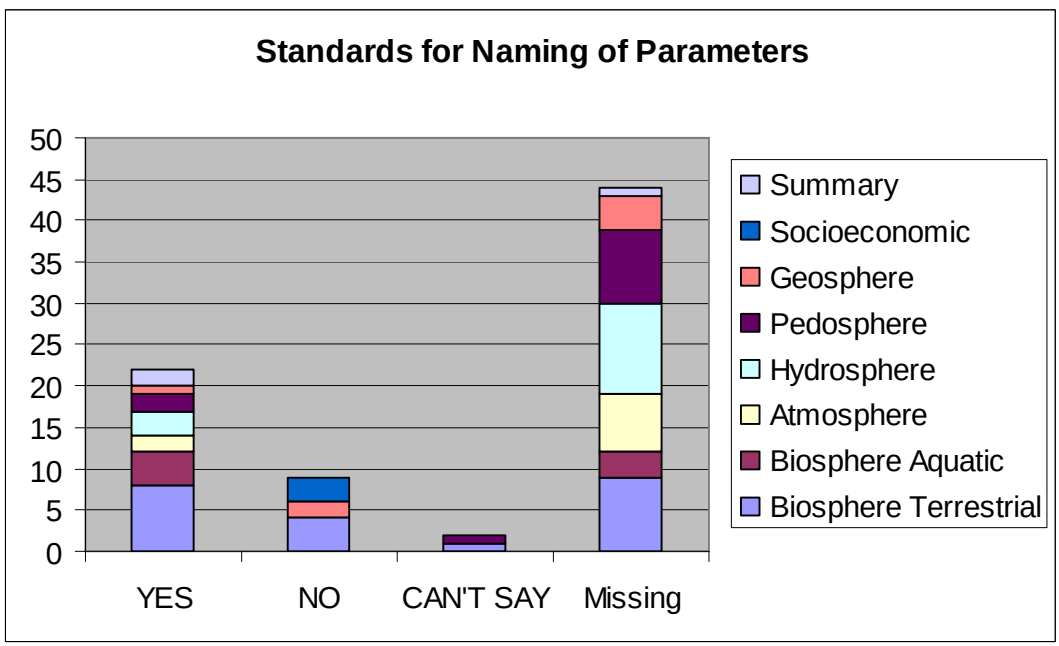


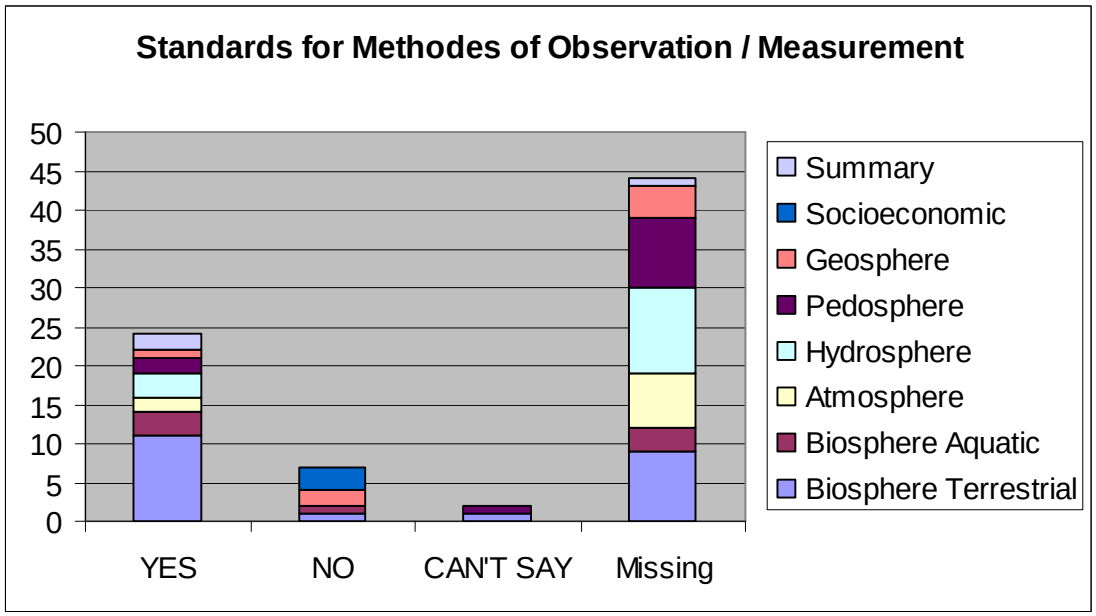
Most of the data available is stored in databases, but a few of the partners are still storing their data in flat files. The file based data must be transferred to a database in order to access it via a network, but this should not be very difficult as long as it is truly a flat data model.



5.4 Use of Standards

Whilst of those responding to the questionnaire regarding use of standards, in regard to naming of parameters as well as in methodology used for observation and measurements, predominantly adhere to existing standards, the majority of responders left the field empty, which leads to the suspicion that no standards are being used. This could lead to difficulties in data exchange, as standards are necessary in order to determine how the data was gained. It could also make it difficult to define a common ontology for the classification of the data, as classification requires common concepts.



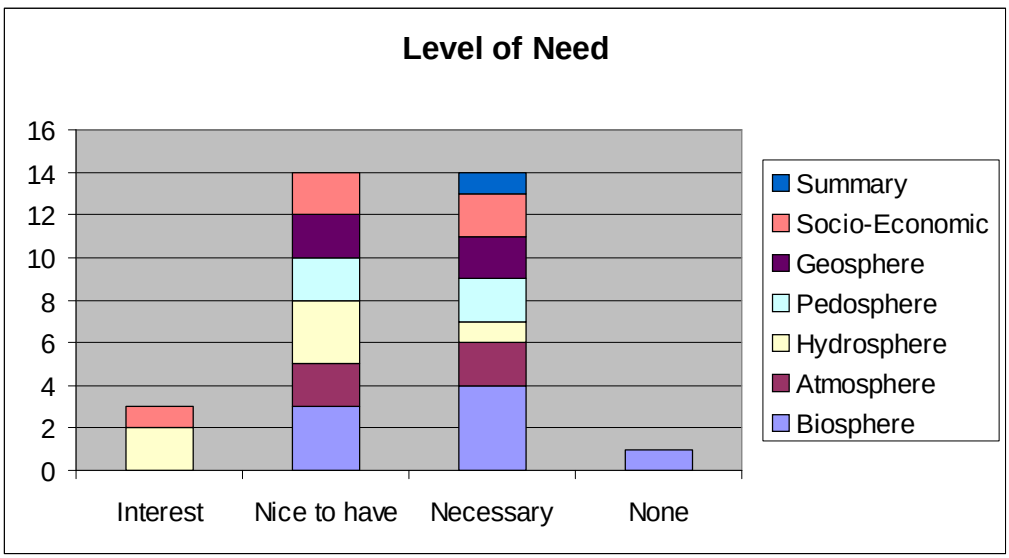


When questioned to the use of existing metadata standards (ISO 19115, EML, Reportnet, GML, ABCD, Darwin Core), one partner mentioned partly supporting ISO 19115 and one partly EML. As one partner is involved in the GBIF Network as a data node it can be assumed that this partner either supports ABCD or Darwin Core, but this was not explicitly mentioned.

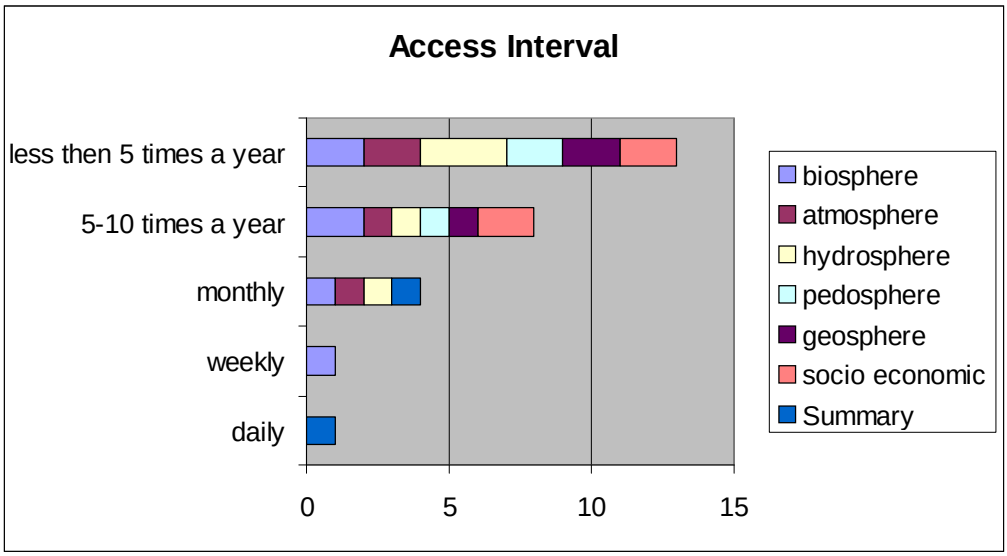
6 Network Requirements

6.1 Level of Need

The need for a system for networking data between the partners seems to be fairly strong, with most partners responding their level of need being between necessary and nice-to-have. Very few responded with level of need set to interest or none. This value must unfortunately be relativized by the fact that many partners did not respond at all to the questionnaire, where we are still hoping that the lack of response was due to lack of time or problems reaching the responsible parties instead of lack of interest in the subject matter. No partners were willing to pay for accessed data.

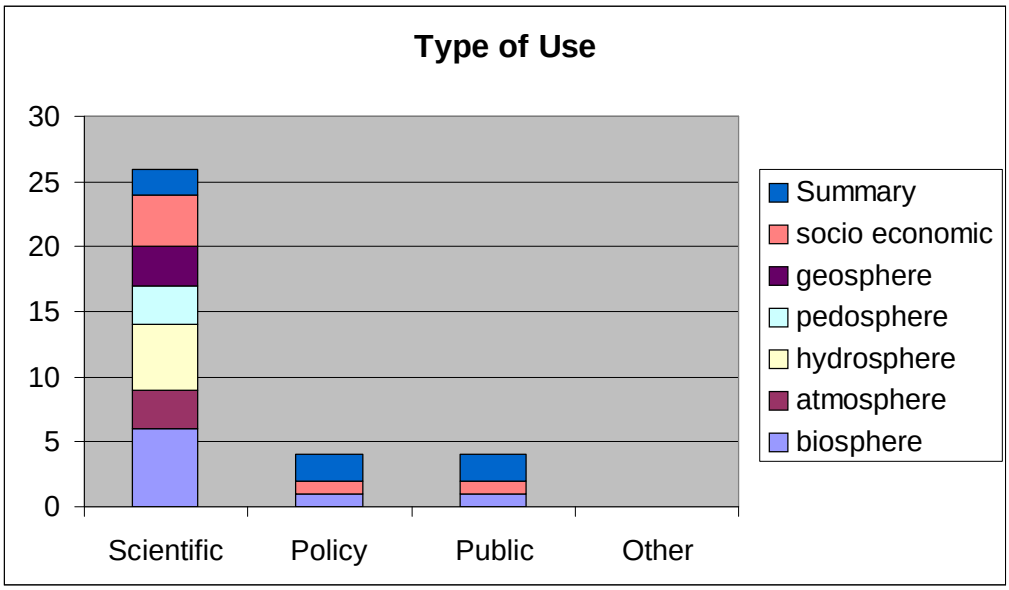


Data access would be required on a fairly sporadic basis, with the prognosed access intervals ranging from monthly to a few times a year. Very few partners expect to use data from partner institutes on a daily or weekly basis

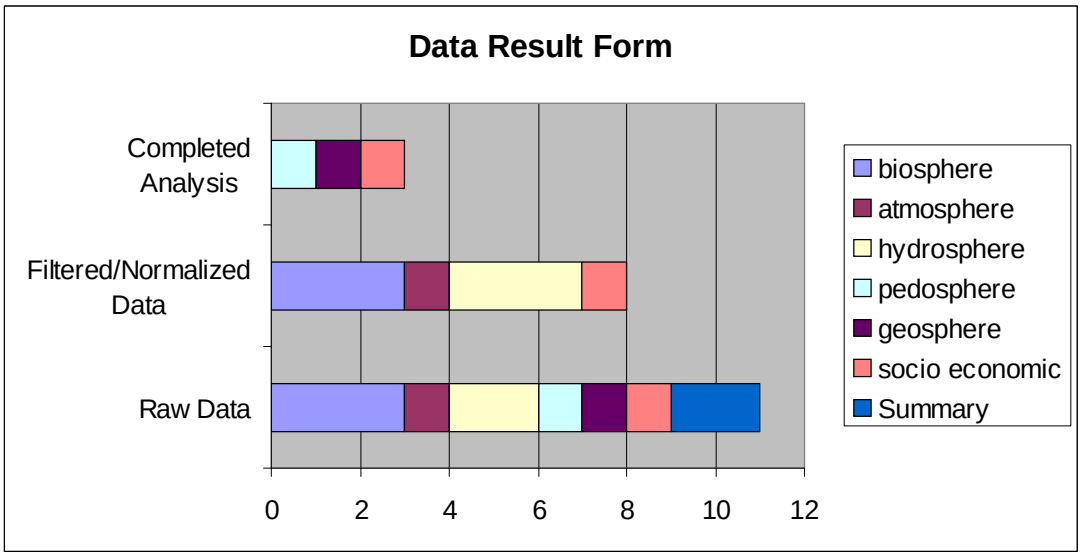


6.2 Requirements for Use

The main type of use for data reported by the partners is scientific, although a few do require data for policy use or for informing the general public.



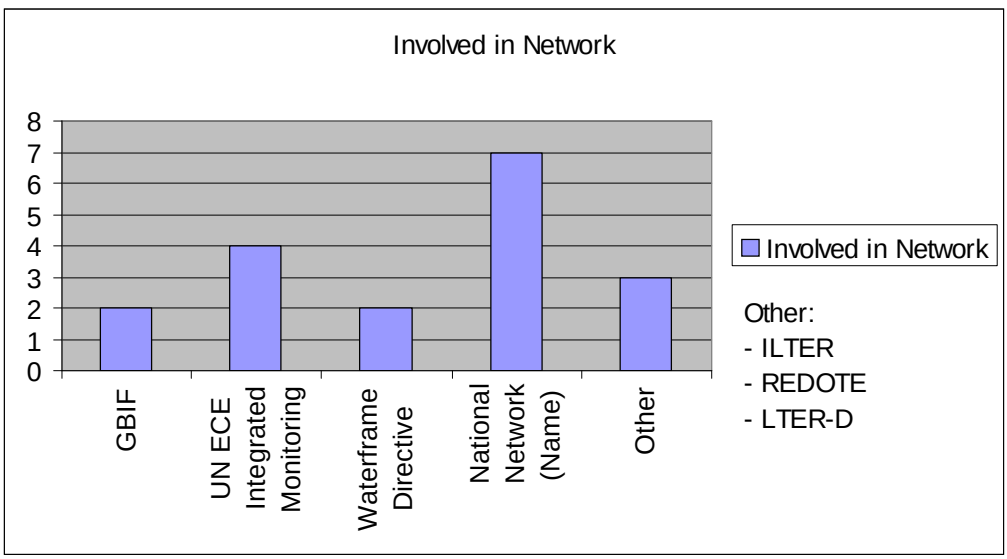
Mostly the partners wish for raw data, followed by filtered/normalized data. Very few partners express a wish for completed analysis data from partner institutions.



6.3 Participation in Networks

Although there is a strong interest in accessing the data resources of other institutions, this is often not possible as the necessary networks do not yet exist. Where they do exist, there are often difficulties due to lack of integration and standards. This seems to be a strong mandate to define standards in this field, and create a coordinated network for the sharing of relevant data.

While a majority of partners are involved in some sort of national network, the participation in international networks is still a bit rare. While 4 of the partners are involved in UN ECE Integrated Monitoring networks, only 2 are involved in GBIF (one only as a portal node and not providing data), 2 are involved in the Waterframe Directive, and 2 are involved in LTER networks.



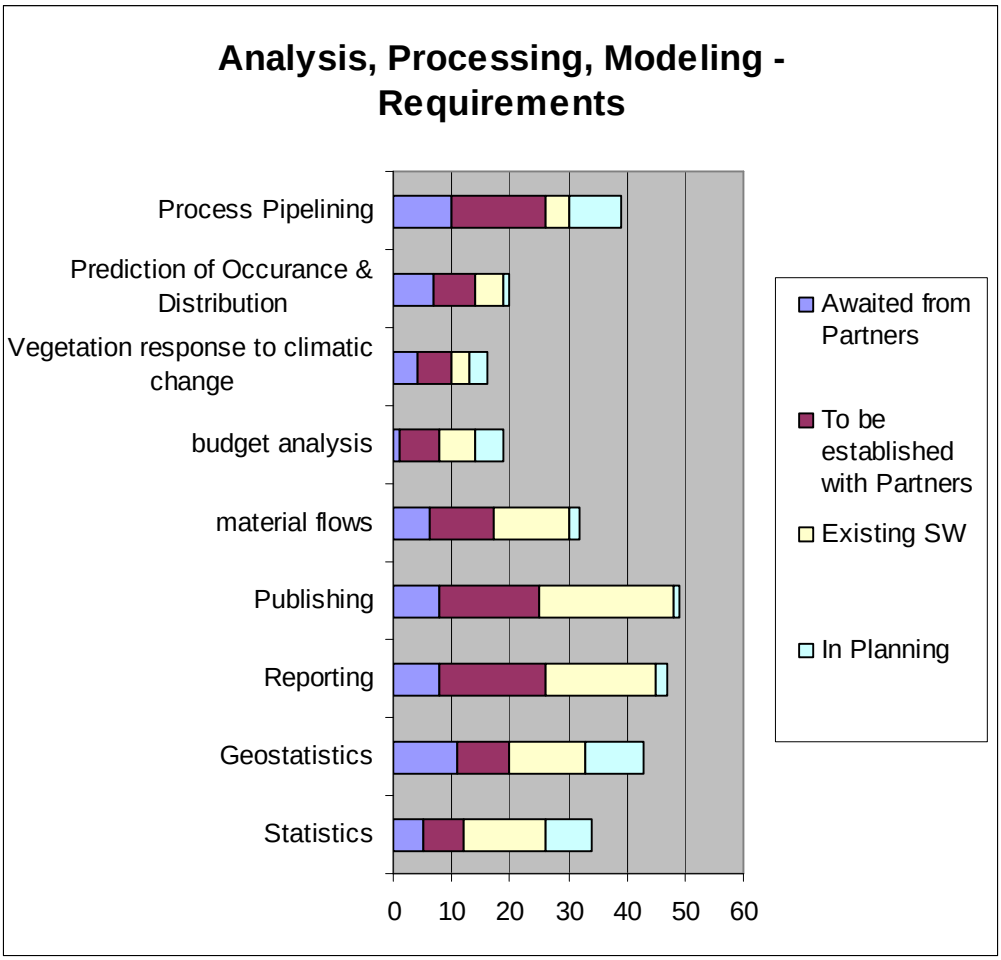
Concerning existing security restrictions, some of the partners currently limit access to registered or known users or only release data under licence following authorization. Some partners are also wary of providing all data online, as some of the data may be quite sensitive, for example when it pertains to locations of endangered species. Most partners would wish

the system to have some sort of authorization and authentication system, and also wish for mechanisms to assure proper acknowledgement of the data originators and licensing agreements.

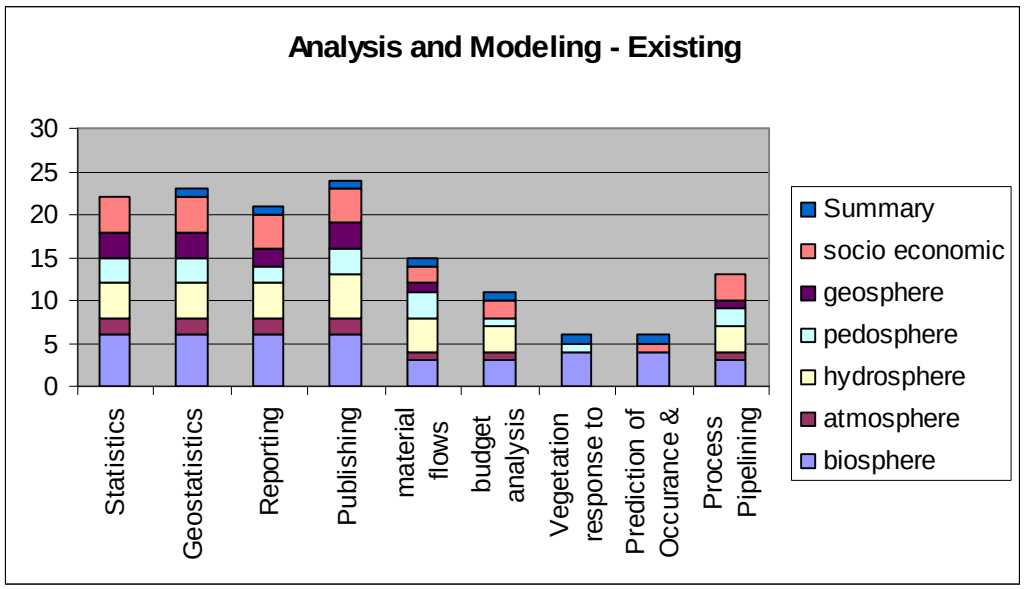
7 Analysis, Processing and Modelling Requirements

7.1 Types of Requirements

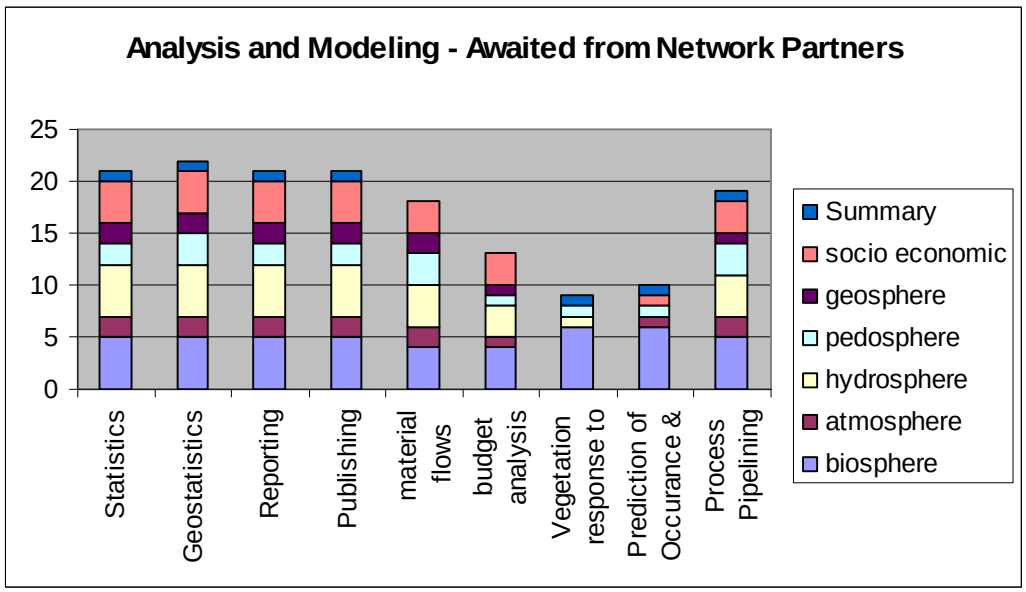
Starting with an aggregate overview of partners requirements for analysis, processing and modelling tools, one quickly sees what type of tools have what priority amongst the network partners. The strongest response pertained to publishing and reporting tools, closely followed by analysis tools such as statistics and geostatistics utilities.
A strong interest was shown pertaining to process pipelining, which would allow the integration of multiple processing steps into one workflow. Many partners showed interest in either establishing such tools with network partners or gaining access to existing systems.
Less interest was reported on the modelling tools suggested. This topic must be scrutinized at the I6 Workshop in order to ascertain if the lack of interest is to be generalized to modelling tools in general, or if the options we supplied within the questionnaire (Material Flows, Budget Analysis, Vegetation Response to Climatic Change, Prediction of Occurrence & Distribution) did not correlate with the users actual needs. The other possibility is that the responses were more spread out among the supplied answers, leading to a lower total. In general it seems unlikely that there isn't interest in this area, as there is a strong interest in process pipelining, which in turn supports modelling.



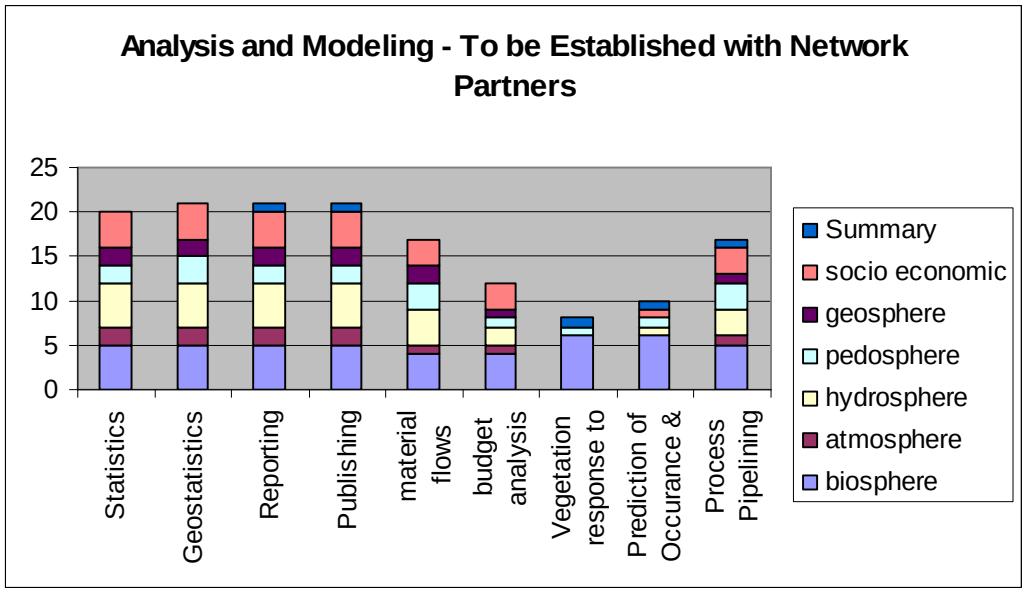
For more in depth perusal, we have added an analysis of the various types of analysis and modelling tools as they are required by domain. In the first graph, we show the currently existing software.



In the second graph, we have the same axis, but are showing the data representing tools awaited from network partners.

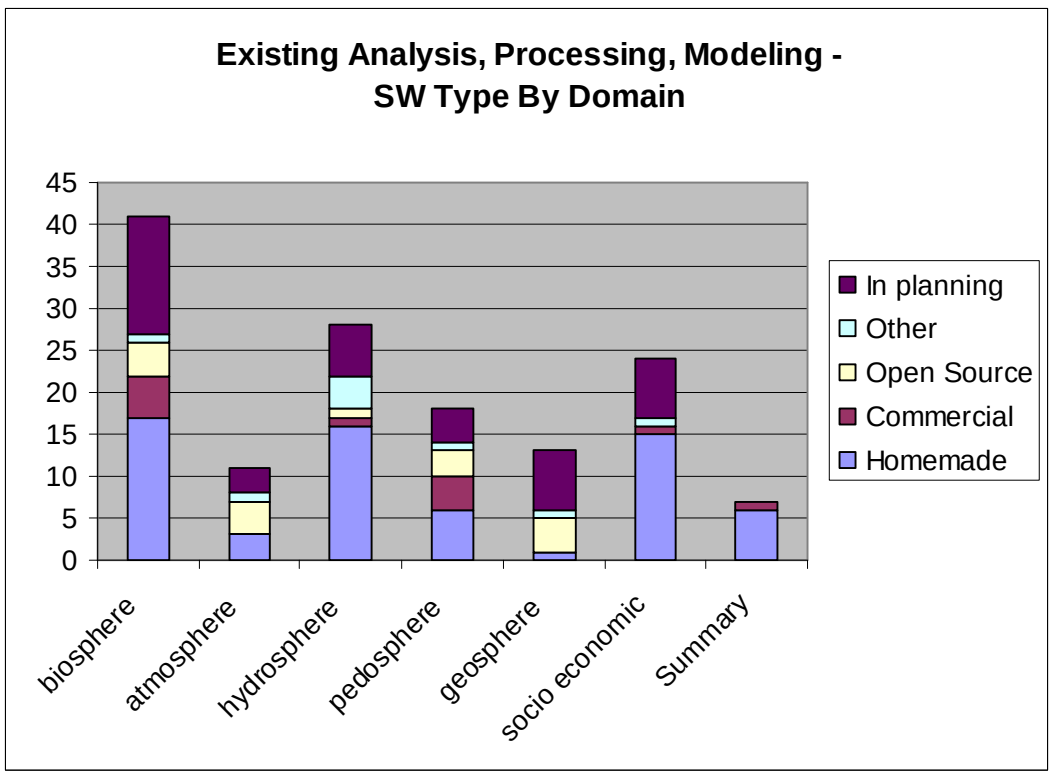


In the third graph, the values represented show the types of tools that the responders would like to actively establish with network partners.



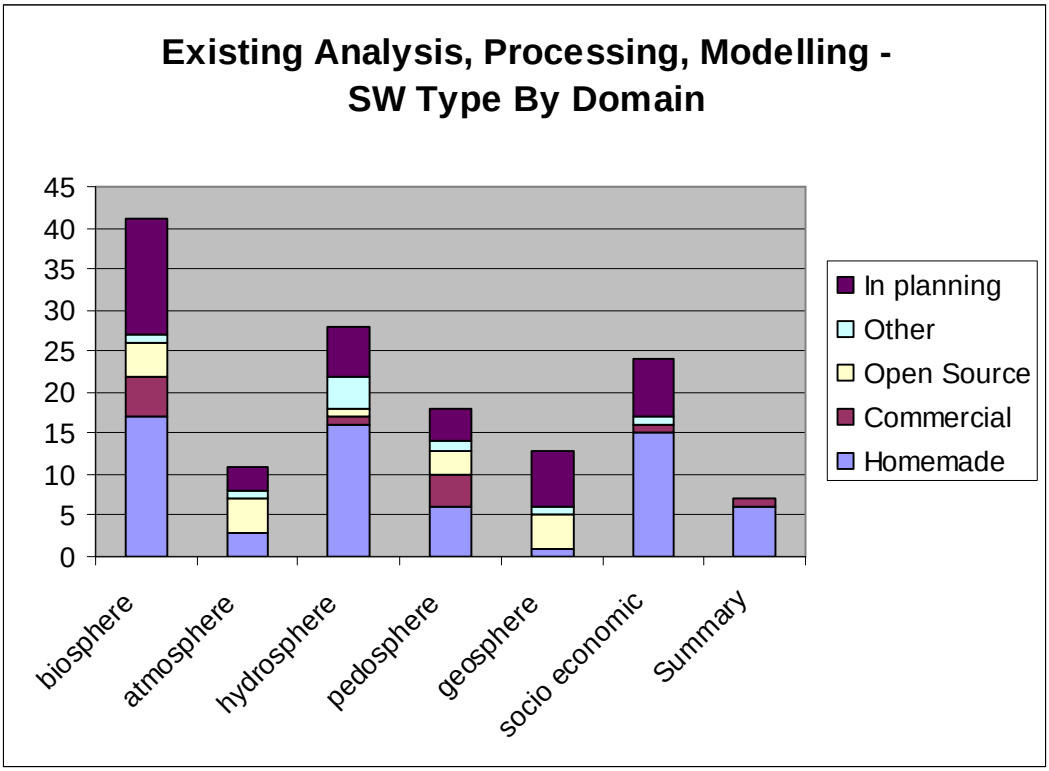
7.2 Sources of Software

The vast majority of tools used by the network partners are homemade, either created in house or contracted out to software companies. Very few partners use commercial tools, and when they do mostly in the area of statistics and geostatistics. Several partners use open source products, not only in the area of statistics but also for their publishing and reporting needs. Many partners have various analysis, processing and modelling tools in planning. Here the interest in statistics is again quite strong, but in this case the interest in establishing processing pipelining tools is equally strong.

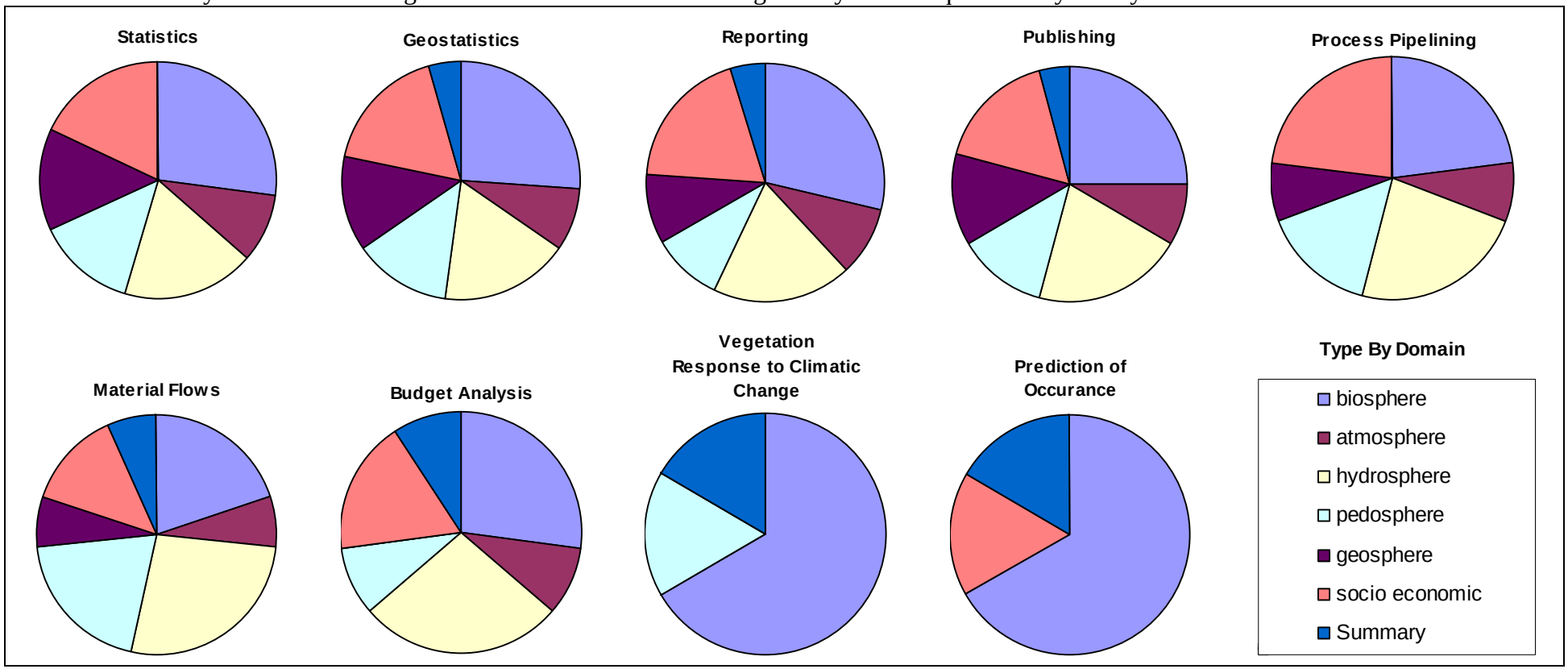


7.3 Requirements by Domain

If one interprets the results of the questionnaire pertaining to existing analysis, processing and modelling software by domain, one clearly sees that the majority of existing tools are utilized in studies of the biosphere, closely followed by hydrosphere and socioeconomic studies. The results for tools in planning are similar, again with a clear lead for biospheric tools, followed by those for socioeconomic research and those pertaining to the pedosphere.



Looking at types of analysis, processing and modelling software as they are required within the various domains, one sees a great similarity in the weighting of requirements by domain in the areas statistics, geostatistics, reporting, publishing and process pipelining. In the area of modelling tools, the predominance of biosphere scientists interested in tools for modeling vegetation response to climatic change as well as prediction of occurrence is fairly obvious. Modeling tools for material flows and budget analysis are required fairly evenly across the various domains.



8 Conclusion

Based on the results of the questionnaire and on discussions at the subsequent I6/I3 workshop, there was a clear message about the need for data sharing, data harmonisation and suitable software tools for biodiversity research and effective LTER networks. Work carried out under I3 endorsed this need, revealing for example how few sites monitored common variables. Information gained through I6 and I3 activities during this first year demonstrates the need for a network allowing scientists to share and integrate their data in a structured manner and also to share and collaborate on analysis and modelling tools.
As we do not have the budget to create our own system, and this probably wouldn't be such a good idea anyway, as there are already several systems being created, it would be advisable to adopt and maybe expand an existing system. In order to reach this goal following steps must be taken:

8.1 Analyze success of existing developments

Several attempts have been made to create this type of data sharing and modelling environment. A number of these attempt have been successful (GBIF, TEMS, …?), some are in progress (SEEK, UK Ecogrid, MyGrid, …). However, one often hears comments along the lines of "it's too much work to do the metadata annotation" or "it's too complicated to navigate amongst all the data". By discovering what difficulties have been faced by similar initiatives, we can try to tackle these issues early on in developments for ALTER-Net

8.2 Review existing networks

In order to gain an overview of existing networks and tools we must do further in depth research. As many initiatives do not supply much information on their websites, it would be advisable to create personal contacts with the key players in the various initiatives currently running. The following areas must be thoroughly examined:

8.2.1 Existing Ecogrids and Networks

- SEEK: determine the current status of the SEEK tools (Morpho, Kepler, Metacat, …) and their suitability for our requirements
- UK Ecogrid and MyGrid: determine status of these grid environments, their suitability, and their compatibility with other networks
- GBIF: currently restricted to collection and observation data. Much valuable information contained, how to integrate?
- TEMS/GTOS: as far as we have been able to ascertain, this network only provided metadata to existing data sources. It does not allow the user direct access to the data. However, one should clarify this fact and also inquire what the future plans for this network are

8.2.2 State of the Art in Metadata Registries

As the principles of metadata annotation and metadata registries are being used in other disciplines that ecology, it would be important to gather information pertaining to the state of the art in this area. Thus it will be possible to better evaluate existing systems.

8.2.3 State of the Art in GRID and Data-GRID Technology

Much work is being done in the field of Data-GRID technology. Of special interest for our purposes is the OGSA-DAI (Open Grid Services Architecture – Data Access Integration) system, which has been implemented based on the Globus Toolkit version 3 and is currently being ported to Globus Toolkit V4. This technology provides the mechanisms required for seamless integration of inhomogeneous decentrally located data sources.

8.2.4 State of the Art in Ontology Languages and Tools

Our current investigations have shown a strong interest in the use of ontologies for data discovery as well as for harmonizing semantics for data transport. Most initiatives in this area are using the W3C standard OWL, which has the advantage of newly integrating services (OWL-S) into the ontology. This functionality would facilitate our plans of integrating data with analysis and modelling tools.
In order to decide if this language fulfils our requirements, it is important that we evaluate the tools currently available to merge, manipulate and display ontologies.

8.3 Develop prototype systems

The I6/I3 joint workshop helped to define I6 'next steps' in developing an information framework for ALTER-Net. Two development steps emerged:

i) Development of an LTER sites and datasets meta-information system with associated core ontology as the first step towards an eventual information framework for ALTER-Net
ii) Development of a prototype information system using a case-study approach focussing on the effects of climate change on vegetation.

Both these activities are being built into the workplan for I6, and will involve working closely with other work packages and with partner institutes, their scientists and data managers.