Available online at www.sciencedirect.com

SciVerse ScienceDirect

Polar Science 6 (2012) 97–103

NiPR National Institute of Polar Research

http://ees.elsevier.com/polar/

The implementation of initial data populations of environmental data and creation of a primary working database

D. Fleischer [a], M. Bölter [b,*], R. Möller [b]

[a] *Helmholtz-Zentrum for Ocean Research in Kiel (GEOMAR), 24148 Kiel, Germany*
[b] *Institute for Polar Ecology, University of Kiel, Wischhofstr. 1-3, 24148 Kiel, Germany*

Received 20 May 2011; revised 4 October 2011; accepted 24 January 2012
Available online 10 February 2012

Abstract

Biological and environmental changes are creating a growing demand for historical and global data sets. Comparing up-to-date ecological and biological findings with historical statements has become a major part of scientific work in the field of ecology. This evaluation and comparison procedure is very time-consuming while the availability of raw data is very low. Comparisons between original findings – if available – require a lot of work from print publication to digitalization or transformation to appropriate data formats. The effective use of working capacity is a general issue and has become important, should the use of information technologies be invoked to minimize time-wasting copy and paste operations.

In this paper we aim to present a working repository for terrestrial biological data. The implementation of this type of data repository will provide various services to participating scientists as long as the final aim is the publication of these repositories. Furthermore, the security and long-term availability of environmental data is an issue of increasing importance to the scientific community. Unrepeatable sampling events and any data thus obtained are precious in time series analysis. For this reason, a well-structured storage of data is necessary for easy accessibility, retrieval and comparability. This is an important issue for the community of environmental scientists. The need to construct and implement repositories should prevail against all hitches and we are therefore describing our on-going task with the primary population of this kind of data repository. A biological and ecological information system is a matter of public interest and should also be a key issue for ecologists.

Keywords: Environmental data; Data management; Working group repository

1. Introduction (primary biological and ecological data)

Collecting environmental data is a basic aspect of the field of ecology during lab and field experiments. Limitations in both, time and sample size sometimes make these analyses unique, while field observations are often unrepeatable and thus most valuable. Nevertheless, the day-to-day handling of these data is far from subject to cautious management. There is no difference between projects with only one or two participants, such as a Masters or a PhD thesis and wide-scale research projects taking place on a global scale.

Long-term availability and reusability are not usually considered or implemented from proposal to project finalization. The reusability factor is underestimated in most projects or programs. An example

* Corresponding author.
E-mail address: mboelter@ipoe.uni-kiel.de (M. Bölter).

1873-9652/$ - see front matter © 2012 Elsevier B.V. and NIPR. All rights reserved.
doi:10.1016/j.polar.2012.01.001

can be seen in the following project developments. In the mid-1980s, NASA taped over 200,000 previously used master tapes involving high-resolution records from spacecraft such as Landsat satellites and Apollo 11, due to a shortfall in supplies of long-life tape. The Nimbus II satellite launched in 1966 soared over the Earth in a polar orbit every 108 min. The data resolution was higher than the processing capacities of that time and data were stored on analogue tapes.

In February 2010 NASA researchers recovered the oldest and most detailed NASA image on global heat radiation from the analogue tapes (Pringle, 2010). Such pictures brought pressure to bear on investigations into sea ice reduction due to climate change, but data recovery from archaic storage systems needed a collaborative approach from various scientific disciplines. This example drastically illustrates the serious problems arising at the end of projects when there is a shortage of money and time for all participants, and shows that data management played a secondary role in this project.

A future oriented data convention on these projects has generally been prevented by the termination of reports and the need for the presentation of results or findings at conferences or in journals. The worst-case scenario is that project employees need to change positions after the project has ended and all the knowledge of spread sheets and the quality of raw data has been lost. Disastrous data loss is exacerbated by a side effect in scientific publishing leading to additional knowledge loss.

It is much easier to publish an article with significant and positive information. Experimental designs resulting in non-significant results are scarcely published, even though the information is equally important to the scientific community. This problem was first identified by Sterling (1959). He reviewed four psychology journals and found that 95% of the articles reported statistically significant ('positive') results. Sterling updated this study 40 years later, and Dickersin (1990) reported that there had only been minor changes in the situation. This is a widely known problem and there is also proof of publication bias in clinical studies (Simes, 1986). The problem has been identified in life sciences, but it is known in all scientific fields. Nevertheless, the performance of experiments or field methods with insufficient results is still valuable to the scientific community. These experiments do not need to be carried out in the same way again.

The World Medical Association makes the following recommendations on this issue in its statement "Principles for all medical research" (Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects, 2008): "Authors, editors and publishers all have ethical obligations with regard to the publication of the results of research. Authors have a duty to make publicly available the results of their research on human subjects and are accountable for the completeness and accuracy of their reports. They should adhere to accepted guidelines for ethical reporting. Negative and inconclusive as well as positive results should be published or otherwise made publicly available…" There is no reason for any divergences from this statement in other scientific fields. So it is essential to store research data in a data repository right away without any evaluation or selection that might influence the final empirical results. Therefore, all ecological data collections need special means of storage in order to ensure that the data are useful and ready for analysis.

The analytical tools required for immediate presentation or inspections in long-term databases are essential for later use and comparisons with other data sources. Sampling of environmental data not only consists of collecting data but also comprises considerable qualitative information and often-subjective impressions of the localities. Methods of storing these impressions are photos of landscapes and sites showing the environment and providing non-verbal information. However, the problem arises of how digital data can be combined with images or other non-verbal information. A combination of both methodologies is therefore deemed necessary and will be presented.

Individually collected data sets are usually only available in local spread sheet files. These files are not generally available to the public. This may not only create a doubling of research, but it could also hamper the essential exchange of original data between research institutions. In order to obtain the best data accuracy, environmental research has to be based on interdisciplinary cooperation. Piwowar et al. (2011) claimed that the sharing and archiving of data was a good investment and that this was also true of environmental data. Institutional databases or repositories of scientific data could close this gap by at least making their metadata public (see Fig. 1). Harvesting techniques can collect this meta-information and spread them over a huge number of portals and search engines. The final consequence would be individual data publication in a world data centre, thus providing all of the repository aspects and benefits to the scientific community. This may be combined with publication in the data journal 'Earth System Science Data' for citability (Fig. 2). Data archives are focused on the

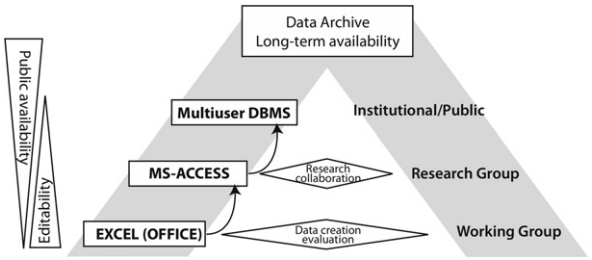

Fig. 1. Setting up an institutional repository has to take into account the individual preferences of the scientific employees. While data capture at the creation level is hard to achieve the data creator and evaluator have to be allowed to choose their preferred tools and techniques to achieve the best data quality. The research collaboration level needs to implement the best possible interoperability with the evaluation and creation level.

long-term availability and usability of data with strong Internet linkage for public visibility. In the data-curating continuum (Treloar and Harboe-Ree, 2008) Treolar and colleagues published the fact that the transition of data in the public domain was accompanied by migration procedures (Treloar et al. (2007) – see Fig. 2) and most scientists need assistance during this procedure. Fleischer and Jannschk (2011) published this bottleneck as one of the future drawbacks towards a real time access to scientific data.

2. Cooperation and data sharing

The Internet enables scientists all over the world to work together on a daily basis. Sending data files back and forth by e-mail makes cooperation easier than ever. Understanding such files and knowing the meaning of each entry can be a time-consuming process. This procedure, however, contains an underestimated source of error. Interim results of the completely wrong file may involve the recipient partner in anything from hours to days of work. This problem is only a mouse click away and may cause delay and unnecessary work. Well-described data sets with meta-information on

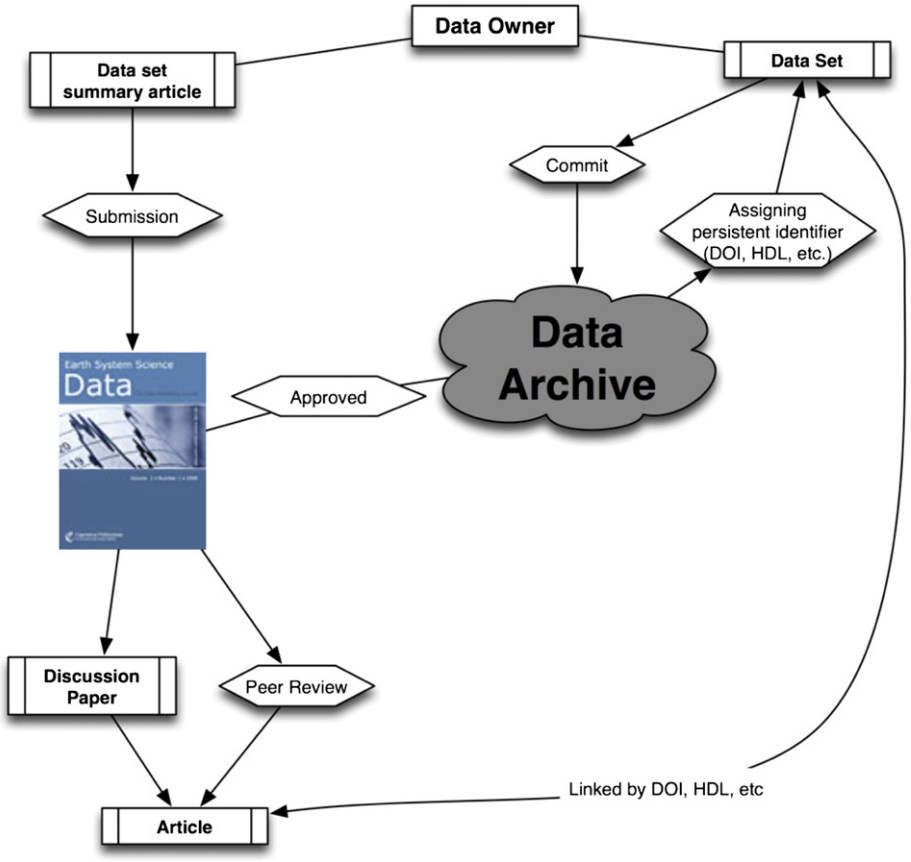

Fig. 2. Citability of data is a major issue due to scientific credit for the data provider. The data journal 'Earth System Science Data' (ESSD) has been established just recently to provide a solution to this issue. ESSD requires an article like description of the data set and a persistent identifier (DOI or HDL) of the digital data set stored in an ESSD approved data archive.

their reusability should prevent this problem, but their transferability is less reliable.

In the days before the Internet became available, scientific publications were the usual way to show methods and results or findings to the scientific community. Today, commercial publications are not evaluated as the right way to reach the broadest possible audience. The 'Open Access' movement is the current philosophy of how to open up scientific knowledge to humankind, and access to raw data is the next necessary step to a sustainable knowledge system for future generations.

Cooperation at data level is the basis of interdisciplinary science and has been identified as a good return of investment by Piwowar et al. (2011). Central data storage allows the use of intelligent algorithms on large data banks (Fiedler et al., 2006a, b). The European Network of Excellence MarBEF accomplished a macro-zoobenthos database in Theme 1 to investigate biodiversity patterns on a continental scale (Somerfield et al., 2009). The creation of this database (Vanden Berghe et al., 2009) first made it possible to run an analysis above the regional scale and test theoretical hypotheses with empirical data. Webb et al. (2009) used this opportunity to test the applicability of macro-ecological methods to the marine environment on this 'macroben' database MarBEF (http://www.marbef.org/).

Within the Institute for Polar Ecology we went through an evaluation phase to find the most appropriate solution to our observation data collected over the last 30 years (items such as microbial soil communities, marine macro-zoobenthos community data and planktonic community data). This evaluation included the possibility of a real data publication in a data archive and an additional description in an article in the Open Access Journal, Earth System Science Data (ESSD) (http://earth-system-science-data.net/) (Fig. 2). The participation in MarBEF and other collaborative projects sensitized us to the need for an institutional solution. The compilation of the different data set for the MarBEF Theme 1 made it perfectly clear that unformatted data required a tremendous effort to achieve a data set usable for comparative analysis. Within a research institution this investment of personnel costs is unsustainable. Therefore we started with a controlled Excel format based on agreement such as the taxonomic project (World Register of Marine Species) and the amount of meta-information necessary. The second step will be the compilation of these working group based data sheets in a more structured manner, such as MS Access (in order to obtain the best possible interoperability within the MS Excel primary level tool).

3. Technologies available

The variety of applicable technologies is huge. In some cases, choice depends on special features, which make the decision quite simple, but in other cases it is a subjective decision. Spread sheets are well known to people and therefore preferred during the data acquisition stage, as long as selectivity and deductive abilities are not required. As soon as multi-user access is required, it is absolutely necessary to use Database Management Systems (DBMS). Several different systems are commercially available, as well as open source software tools. These solutions require detailed knowledge of data modelling and interface development. The advantage of a self-made solution is that users can be convinced to use the system through customization. There are some drawbacks, such as the implementation of the necessary interoperability features to achieve public awareness of a repository. All this developmental work can be very time-consuming and expensive, while data libraries are available. An individual institutional repository is not big enough to be recognized by large harvesting projects or implementation of the techniques required may be too expensive and unnecessary within the institute itself.

The main feature of an integrated environmental information system is timesaving. Structured data storage enables much more sophisticated and less time-consuming analysis procedures. It is also possible to apply new analysis procedures that are not included in an MS Excel data set. Scientists from marine fields willing to transfer their data into data archives such as Fishbase (http://www.fishbase.org), OBIS (http://iobis.org) or PANGEA (http://www.pangea.de) are able to do this with a single mouse click. The proposed information system will be able to create all kinds of transfer formats, such as Darwin Core (http://rs.tdwg.org/), OpenDirectory (http://www.dmoz.org/Science/) by providing further links to data stores, or the DiGIR (http://digir.sourceforge.net/) provider, which itself relates to the Darwin Core project or other generally usable XML formats. This simplification of everyday routine will free employees from wasting time on everyday routines and support international collaboration.

Apart from this in-house solution, the use of a public data center like the World Data Centre for Marine Environmental Sciences PANGAEA in Bremerhaven, Germany (http://www.pangea.de) has

entered the realms of possibility. This data center provides a huge variety of support for scientists who are willing to upload their data into PANGAEA and out of the system. The PANGAEA data warehouse recombines data that has already been stored in PANGAEA. A data warehouse is basically a delivery system with query optimization, memory and access structures. The warehouse is made to execute queries as fast as possible. Usually, these techniques are used to combine huge data sets from large supermarket chains, in order to run statistical data mining algorithms, for instance.

The PANGAEA database contains billions of values. All kinds of marine environmental sciences and other designated World Data Centers are available to take care of data quality and retrievability (http://earth-system-science-data.net). The OAIster-harvesting database hosted by the world's largest library cooperative OCLC (http://www.oclc.org/oaister/) is a metadata database harvested from PANGAEA and other data centers to provide a central meta-information service pointing to the true data storing databases. Further databases, especially for soils with applications to soil monitoring programs are well presented by Slavecz et al. (2006).

4. Data integration

The estimated impact of experimental results and findings require surveying activities. With its approach to experimental studies or surveys, the field of ecology is able to analyse large-scale environments. Physiologists create data based on metabolic rates and energy investments in a species. On the other hand, taxonomy and community analysis create a lot of data on population levels. These data can be combined to check the overall result or finding: Can a population be supported by its environmental conditions and its metabolic rates? By setting up information systems, data cooperation between working groups and participants will become easier. The security of primary scientific data in terms of theft protection is only one side of a much larger coin. The protection of data from loss has become more important.

The ethical expectations and concerns of individual scientist suspicious of sharing a real time data repository with other scientists need to be cleared up. The incorporation of a complex user access control system will, in any case, assure the security of raw data. The fact that scientists will become available from the data they have collected, instead of from the publications they have written, will be a completely new approach, one that will enhance scientific cooperation.

The development of this kind of information system will combine technologies, which have previously been isolated. It will integrate archiving, the analysis and publication of biological data as an aid to environmental scientists. On the other hand, this system will use technologies from Databases, Content-Management-Systems, Knowledge Bases and Artificial Intelligence in a single system and reveal a guideline for the architecture and construction of other systems of the same kind. In addition to this scheme, the modelling of the incorporation of improved interactivity between the user and the information system will be practicable. The system will display the correct information to the correct user at the correct time. This will involve the ability to adapt data presentation, and in these terms, it will be necessary to include user references, needs and purposes in the system. This has already been commenced in the Codesign approach and needs further practical implementation.

5. Our methodological approach

The actual data sets in the Institute for Polar Ecology are stored as EXCEL files, as they are still being processed and under evaluation regarding any further steps that might need to be taken. We are still at a much earlier stage than the sophisticated databases presented above. It is our primary aim to gather a combined store of data in order to compare their contents and make additions through qualitative descriptions of soil properties (e.g. grain size, soil colour etc.), which can be added and re-evaluated at any time. This means that any work on the original files and its completion is important. It also allows for reassessment of databases and data treatments. Furthermore, combinations of data for various applications, as well the search for individual properties, numbers or sites may be performed. The implementation of links to image files has already been performed.

The use of imaging techniques is becoming quite popular in the field of ecology, as it provides further information on sites or organisms. The use of well-established sampling methods, in combination with imaging, is creating a large amount of data. Photos or videos, in combination with primary digital data, are not usually stored side by side. For this reason, pictures are considered completely separately, and not as a combined set of information providing both data and images. If the metadata are inappropriate, the

connection between the two will become undiscoverable and soon the possibility of combing these sources of knowledge will be lost to either the scientific community or the public. It is not unusual for some data spread sheets never to be used again because nobody is able to physically retrieve primary data. This may be true of all kinds of larger and more expensive projects.

Thus it has become possible to access data, either through the original digital data of the samples or sites, or via the photo of a particular site. Either way, the latter offers a particular perspective enabling sites firstly to be compared via their visual description, and secondly by going into their description through physical, chemical or biological measurements. Maps from GIS sources may complete these measurements. The creation of the IPOE Data Repository is based on advantages during the analysis and interpretation of scientific results or findings, together with the combination of imaging and measurements.

The proposed methodology, ranging from primary storage in data spread sheets and connections with photo libraries, presents a way of realizing a vision of field impressions before they enter digitalized data storage. The raw data may be used further during publishing efforts, but are presented as a gateway for further users, in order to visualize environmental conditions and interaction between digitalized information in the form of numbers and analogue information in pictures. The availability of such comparative data, through their addition to existing databases or references is an attempt to achieve further knowledge of sites that are not accessible to all research groups. They may thus serve as an interface between such groups for ecological work in Arctic Science, especially to those, which are not directly linked to large consortia like LTER Studies.

Fig. 2 shows the strategy we are following. After evaluation procedures, the data will be stored within a medium-sized database tool, such as MS ACCESS. In the final stages, while editing and checking will become of minor importance, a more sophisticated database (DB2 or Postgres) may become more feasible. For this step, the cost benefit ratio will have to be evaluated, since the separation of database and front-end will come with new development investments for data display and any other factors that may prove necessary.

Acknowledgement

The authors wish to express their thanks to Mrs. P. Johnson for her intensive work during proofreading of the manuscript.

References

Declaration of Helsinki Ethical Principles for Medical Research Involving Human Subjects. 2008. In: Medical Journal World, 54 (Dec. 2008), 122–125.

Dickersin, K., 1990. The existence of publication bias and risk factors for its occurrence. J. Am. Med. Assoc. 263, 1385–1389.

Fiedler, G., Czerniak, A., Fleischer, D., Rumohr, H., Spindler, M., Thalheim, B., 2006a. Content Warehouses. Rep. No. 605, Institute für Informatik und Praktische Mathematik, Christian Albrechts-Universität Kiel, p. 71.

Fiedler, G., Thalheim, B., Fleischer, D., Rumohr, H., 2006b. Data Mining in Biological Data for BiOkIS, vol. 1. Knowledge Media Technologies. First International Core-to-Core Workshop, p. 10.

Fleischer, D, Jannschk, K., 2011. A Path to filled Archives Nature Geoscience 4, 575–576 http://dx.doi.org/10.1038/ngeo1248.

Piwowar, H.A., Vision, T.J., Whitlock, M.C., Whitlock, M.C., 2011. Data archiving is a good investment. Nature 473, 285. http://dx.doi.org/10.1038/473285a.

Pringle, H., 2010. NASA dives into its past to retrieve vintage satellite data. Science 327, 1322–1323.

Simes, R.J., 1986. Publication bias: the case for an international registry of clinical trials. J. Clin. Oncol. 4, 1529–1541.

Slavecz, A., Terzis, A., Ozer, S., Musaloiu, E.R., Cogan, J., Small, S., Burns, R., Gray, J., Szalay, A., 2006. Life under your feet: an end-to-end soil ecology sensor network, database, web server, and analysis service. Microsoft Technical Rep 90, 1–16.

Somerfield, P.J., Arvanitidis, C., Vanden Berghe, E., Van Avesaath, P., Hummel, H., Heip, C.H.R., 2009. MarBEF databases, and the legacy of John Gray. Mar. Ecol. Prog. Ser. 382, 221–224.

Sterling, T.D., 1959. Publication decisions and their possible effects on interference drawn from tests of significance or vice versa. J. Am. Stat. Assoc. 54, 30–34.

Treloar, A., Harboe-Ree, C., 2008. Data management and the curation continuum: how the Monash experience is informing repository relationships. Proc VALA Conf. http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf.

Treloar, A., Groenewegen, D., Harboe-Ree, C., 2007. The data curation continuum – managing data objects in institutional repositories. D-Lib Magazine 13 (9/10). http://dx.doi.org/10.1045/september2007-treloar.

Vanden Berghe, E., Claus, S., Appeltans, W., Faulwetter, S., Arvanitidis, C., Somerfield, P.J., Aleffi, I.F., Amouroux, J.M., Anisimova, N., Bachelet, G., Cochrane, S.J., Costello, M.J., Craeymeersch, J., Dahle, S., Degraer, S., Denisenko, S., Dounas, C., Duineveld, G., Emblow, C., Escaravage, V., Fabri, M.C., Fleischer, D., Gremare, A., Herrmann, M., Hummel, H., Karakassis, I., Kedra, M., Kendall, M.A., Kingston, P., Kotwicki, L., Labrune, C., Laudien, J., Nevrova, E.L., Occhipinti-Ambrogi, A., Olsgard, F., Palerud, R., Petrov, A., Rachor, E., Revkov, N., Rumohr, H., Sarda, R., Sistermans, W.C.H., Speybroeck, J., Janas, U., Van Hoey, G., Vincx, M., Whomersley, P., Willems, W., Wlodarska-Kowalczuk, M., Zenetos, A., Zettler, M.L., Heip, 2009. Macro-Ben integrated database on benthic invertebrates of European continental shelves: a tool for large-scale analysis across Europe. Mar. Ecol. Prog. Ser. 382, 225–238.

Webb, T.J., Aleffi, I.F., Amouroux, J.M., Bachelet, G., Degraer, S., Dounas, C., Fleischer, D., Gremare, A., Herrmann, M., Hummel, H., Karakassis, I., Kedra, M., Kendall, M.A., Kotwicki, L., Labrune, C., Nevrova, E.L., Occhipinti-Ambrogi, A., Petrov, A., Revkov, N.K., Sarda, R., Simboura, N., Speybroeck, J., Van Hoey, G., Vincx, M., Whomersley, P., Willems, W., Wlodarska-Kowalczuk, M., 2009. Macroecology of the European soft sediment benthos: insights from the MacroBen database. Mar. Ecol. Prog. Ser. 382, 287–296.