Dealing with the challenges of presenting taxonomic data online: An introduction to PLANKTON*NET@AWI

Alexandra Kraberg, Friedrich Buchholz and Karen H. Wiltshire

Alfred Wegener Institute for Polar and Marine Research, Biological Station Helgoland,
Kurpromenade 201, 27498 Helgoland
E-mail: akraberg@awi-bremerhaven.de

Abstract

Phytoplankton taxonomy requires comparison of large volumes of information including images of taxa from different geographical areas. The internet should be ideally suited for this task. However, despite its advantages compared with traditional dissemination methods and the huge array of different online taxonomic resources, it lacks the evaluation and validation mechanisms of traditional resources, and 'ground rules' for the treatment of taxonomic data have not yet been established. PLANKTON*NET@AWI contains more than a thousand plankton images from the North Sea and different collections from all over the world. The database can be searched alphabetically or via collections. Each record can be viewed as a standardized data sheet with images and taxonomic descriptions. Comment functions are also provided but their administration has yet to be discussed. Images from different collections can be compared, facilitating the detection of taxonomic inconsistencies and geographic variations in morphology. PLANKTON*NET is a collaborative project with partners at Roscoff and in Woods Hole, but the individual sites are not yet networked. We are currently exploring mechanisms for future database formats and ways of networking existing resources to maximize the benefits for taxonomic research. Our favoured approach will be to follow the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as an application-independent interoperability framework based on XML technology, so that the integrity of local taxonomic initiatives can be maintained, while sharing content but this will also require discussion within the wider scientific community.

Keywords: Taxonomic database; long-term data; data analysis; OAI.

Introduction

Global change phenomena have already been shown to affect biological communities. Conclusions as to what exactly the long-term consequences of global change events on communities in different types of ecosystems might be vary widely, but there is some agreement that these impacts will be profound. One reason for the difficulties in analyzing and predicting long-term trends is that such analyses critically depend on reliable biodiversity data about the ecosystem under study. However, such baseline taxonomic information is often lacking and how environmental changes will affect species, communities and the stability of the affected ecosystems in general is therefore

difficult to predict (Danielsen, 1997; Fjeldså and Lovett, 1997; Gray, 2001; Piraino *et al.*, 2002).

Filling gaps in the knowledge about the number of species within a given geographic area would normally be the domain of trained, specialized taxonomists. However, there is today only a relatively small number of such specialists, facing millions of unknown taxa particularly in marine systems. In addition, the traditional publication process is very complex and expensive excluding many potentially interested parties from quick access to new information that could be vital for the design of new scientific programmes (Agosti and Johnson, 2002; Godfray, 2002).

Trained taxonomists are becoming rare for many species groups and many taxonomic collections that have been built over years or even decades are not maintained beyond the retirement of the scientists who assembled them. This together with slow and patchy access to existing and emerging information will lead to the loss of vital information and duplication of research efforts. This is happening at a time where financial resources are often stretched and where concurrently there is a renewed need for taxonomic expertise as entirely new habitats *e.g.* in the deep-sea are discovered (Morin and Fox, 2004). Online resources, if properly organized and co-ordinated could be one way out of that dilemma (Costello, 2003). This paper presents PLANKTON*NET@AWI a resource that is developing a concept for not only site design but also participation of the scientific community to provide a widely accepted, easy to implement and cost effective format that can deal with the problems described above. Once fully implemented PLANKTON*NET@AWI can complement and enhance traditional data dissemination methods.

Methodology and Data Acquisition

The database set-up is based on mySQL/ PHP which is open source software and easy to install and run. Further modules are based on XML and we are trying to establish a pilot for links using open archives initiative meta harvesting protocols (OAI-MHP) with our existing partners at the Station Biologique de Roscoff, who are hosting a second PLANKTON*NET site with their own data.

Data acquisition for PLANTKON*NET is essentially a collaborative process and the PLANKTON*NET developers and administrators are aiming at developing a resource that can act as a custodian for the data of different data providers, where on behalf of the data provider, PLANKTON*NET is handling all data processing and their addition to the database as a collection of images and taxonomic descriptions. In return these data providers will have a say in the further development of this resource. Individual contributors' records are clearly identifiable in the database and each contributor can provide additional information for the design of a customized introduction page that describes the institute at which the data were collected, provides geographical information about the sampling origin etc. The contributor wherever possible, provides image information in .tif format. These are then processed and will be used as jpg throughout the site and as .tif for download versions of all of the images. Where

necessary, PLANKTON*NET staff will also process original image material such as slides and photographs for the contributors, whenever they lack the resources to do so.

PLANKTON*NET aims and set-up

A major aim of PLANKTON*NET is to network biodiversity resources but at the same time to allow individual sites the freedom to develop modules and tools according to their own needs. In the case of the Biologische Anstalt Helgoland for instance (which is part of the Alfred-Wegener-Institute for Polar and Marine research), there is a strong emphasis on researcher and student training and this will be reflected in the analysis tools already planned for PLANKTON*NET@AWI.

At the heart of PLANKTON*NET lies its classification system, linked to the Taxonomic name server (TNS), originally developed at uBio, Woods Hole. The underlying databases of the TNS indexing system, contains approximately 1.7 million taxon names, synonyms and vernacular names in addition to valid scientific names, allowing quick and even more importantly complete assemblage of the taxonomic and other literature about a given taxon.

The taxon information within PLANTKON*NET is organized around collections (Fig. 1 and 2) each of which has been provided by one collaborator and is easily identifiable throughout the PLANKTON*NET site as originating from that collaborator. From a list with all collections in the database, a list with the genera present in the individual collections can be accessed.

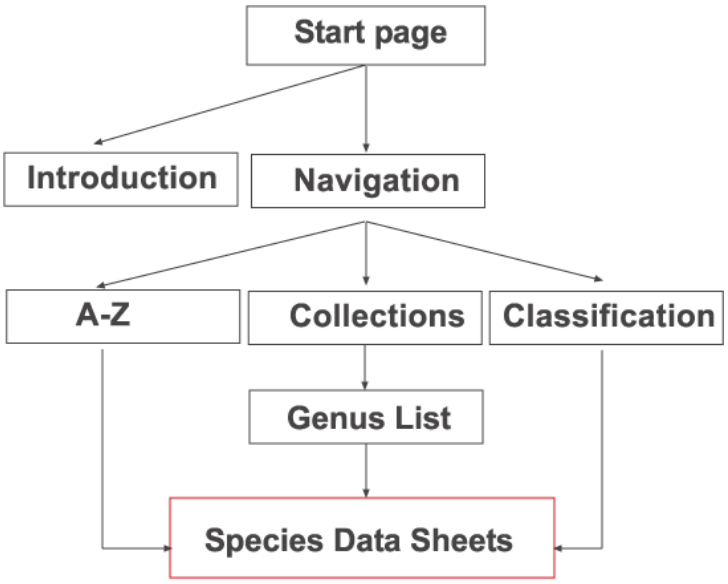

*Fig. 1. The PLANKTON*NET structure: Site Map with components and navigation of the PLANKTON*NET@AWI taxonomic web resource.*

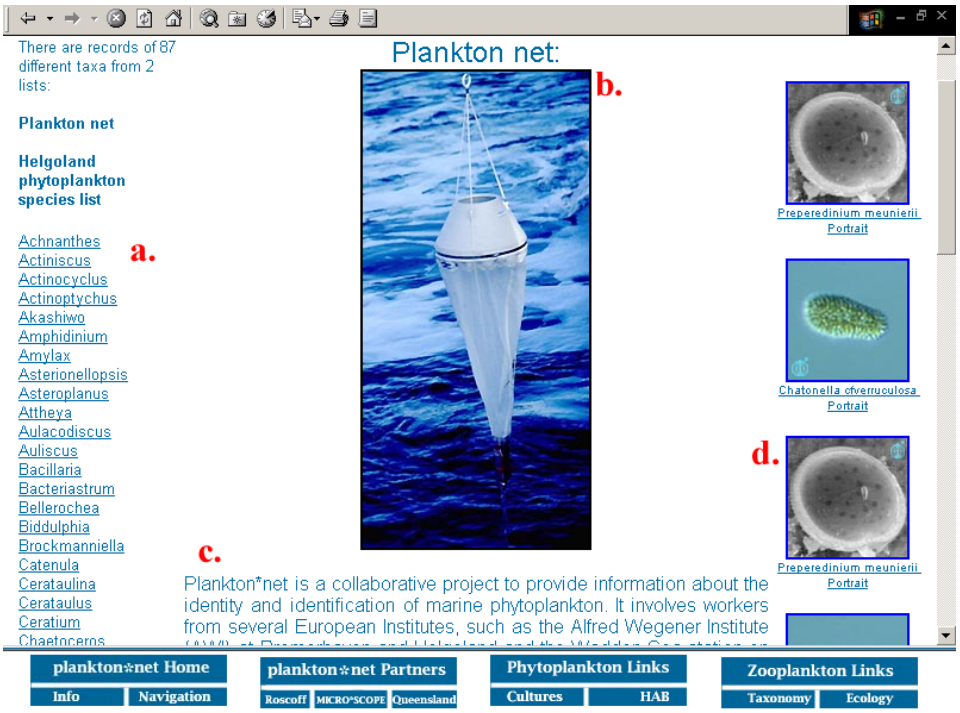

Fig. 2. A start page for an individual collection (accessed from the collections list): a list of all genera present in the collection; b. image or logo chosen by the data provider; c. introductory text for the providing institute and the image collection; d. examples of images contained in the collection.

From these lists the user can navigate to individual species sheets (Fig. 3), containing enlarged images, information about image origin and short species descriptions.

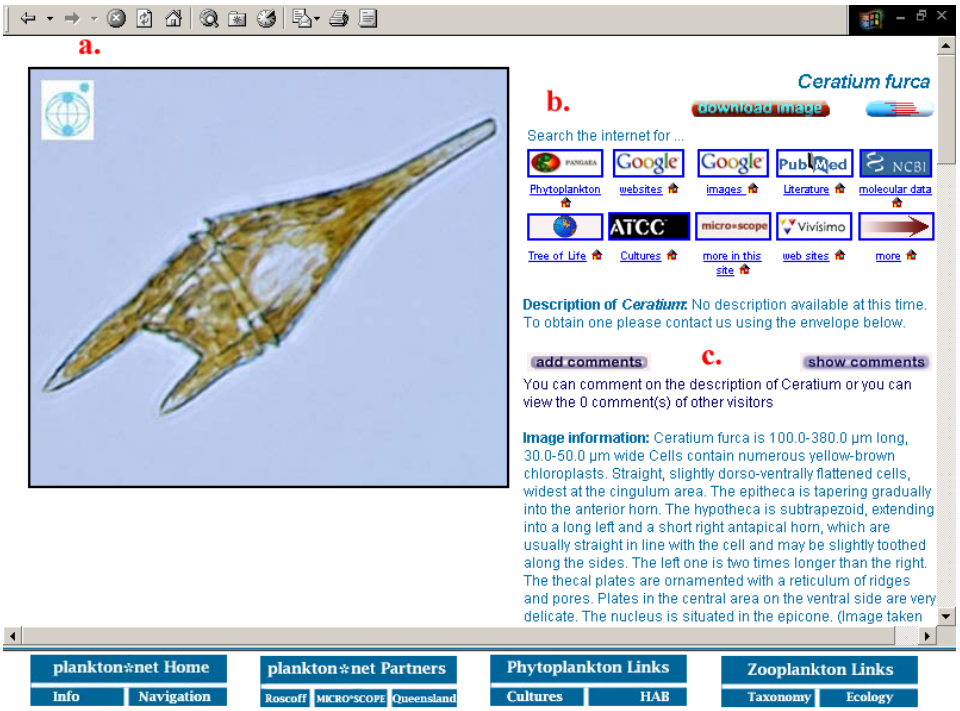

Fig. 3. Example of the individual species pages, as accessed from A-Z index and collections pages. The main components are: a. Large image of the organism; b. sets of outlinks with deep links to external resources containing information about the organism; c. image information and species descriptions.

These species sheets also contain a series of outlinks that connect to external online resources with information about the organism in question. These are deep links, *i.e.* they do not connect to a homepage, that has then to be navigated, but directly to the relevant information within these external websites. The outlinks have been customized so that, for instance species pages about Harmful algal bloom (HAB) species contain outlinks to HAB relevant rather than general phytoplankton resources. In addition, each species page contains a link to a high resolution download version of the image (in .tif format). These can be downloaded free of charge unless they are intended for commercial use. In addition, the bottom section of the species sheet contains a collation of all images of the same species and genus within PLANKTON@NET (Fig. 4), this simple set-up already allows limited comparisons of geographical images and importantly also facilitates the detection of possible species identification errors.

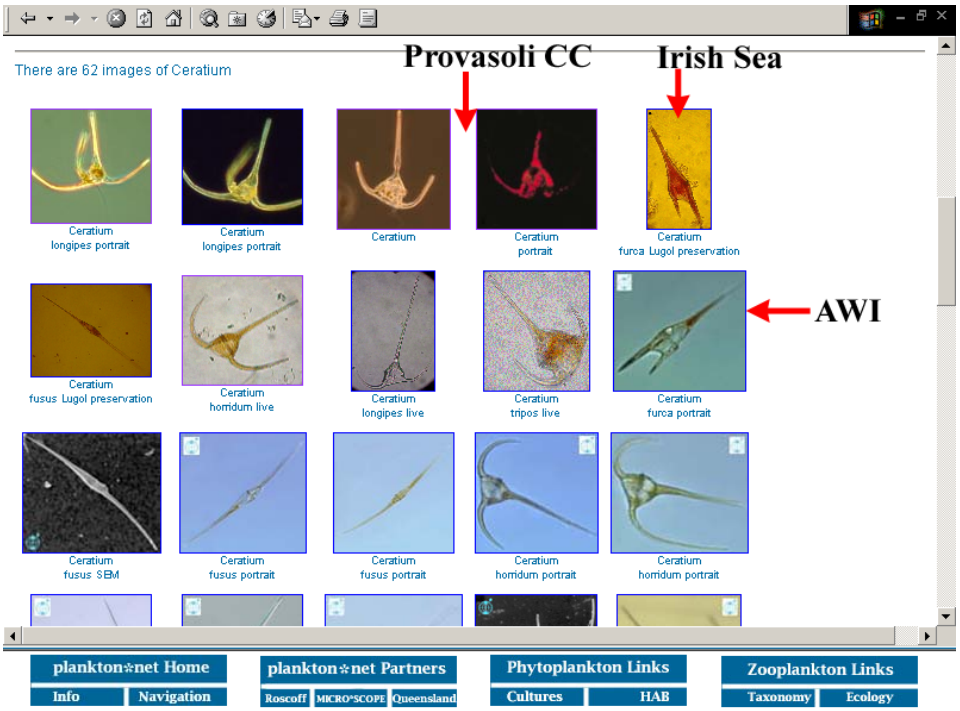

Fig. 4. The lower part of the individual species sheet, showing all images for this species present in the database. The example given here includes images from 3 different geographical sources.

While the above features are served out of PHP scripts/ MySQL database queries, there are also a number of static features in PLANKTON*NET, for instance a link list with links to culture collections as well as further online databases, taxonomic resources and identification aids. These static resources can of course be customized to the requirements of individual PLANKTON*NET sites.

Discussion and outlook

At present the emphasis of PLANKTON*NET is on the display and the collation of data. However we think that it will be vital to extend this basic format to be more conducive to data analysis. The challenge now is therefore to increase functionality so that the resource can be used by as wide a range of user groups as possible while maintaining and enhancing its user friendliness without disrupting site organization. Three stages are planned in the further development of PLANKTON*NET. First of all the current site structure will be modified to increase search functionality at species level. Although at present individual data sheets appear with a species heading, the site still essentially operates at genus level, which becomes apparent in Fig. 4. Such an increase in functionality could mean splitting the taxonomic descriptions into several separate fields to allow more refined searches. A key development will be to work on achieving

interoperability between PLANKTON*NET nodes and with already existing resources. The third step will be the addition of modules such as a library of taxonomic keys. These will be xml based matrix keys. By applying the emerging TDWG standard (Taxonomic Databases Working Group) for representation of descriptive data, (structure of descriptive data, SDD), PLANKTON*NET@AWI will be able to ensure that information is usable with different key production software tools, again allowing flexibility for local initiatives.

An important point to stress is that this resource is not in competition to existing resources. On the contrary, it is vital to maintain dialogue with these initiatives to maintain interoperability between resources. But with its clear focus on plankton taxonomy and in addition the analysis and interpretation of plankton data, PLANKTON*NET@AWI can provide a future essential tool for phyto- and zooplankton taxonomists world wide and can be used as a standalone tool or in conjunction with other existing resources, so that all the information necessary for gathering reliable biodiversity information, can be easily assembled and used.

The main underlying principle is to produce a resource that can be used on a par with traditional paper publications. This requires several prerequisites. Firstly the site structure has to be such that it is transparent, *i.e.* with all information necessary to judge the quality of the information presented. Secondly the format has to be supported and recognized as a valid format by the scientific community. It is for this reason that the PLANKTON*NET project is placing considerable emphasis on collaboration and communication with external experts and data providers to ensure that this resource will be of the utmost relevance to scientists in general, not only taxonomists and that importantly it is maintained in the long term (despite the financial pressures placed on most of the participating institutes). It is therefore vital that PLANKTON*NET is inclusive at every point of its development and is seen to encourage participation of additional data providers in the future. Providing a clearly and logically structured coherent site structure will aid this process.

The PLANKTON*NET concept allows scientists holding taxonomic collections, to have them preserved with minimum investment, while still having a say in the treatment of their taxonomic material within the database. With an increasing number of contributors this will become not only a means of preserving taxonomic information and the integrity of individual taxonomic collections but also, and importantly, a means of quality control of the data sets within PLANKTON*NET. This will allow the production of a reliable resource without the need for restrictive and, above all, time and money consuming editorial structures, but yet trusted in the same way that traditional paper based publications are. This resource is in no way meant to replace conventional resources but to complement them. The challenge of this and other future resources will be to achieve this without restricting one of the greatest strengths of the worldwide web: the ease and speed of the cost effective transfer of scientific information worldwide.

Acknowledgements

The authors, on behalf of the PLANKTON*NET community, wish to thank the developers of the original software, first at the University of Sidney and now at the Marine Biological Laboratory at Woods Hole. The actual data providers also deserve our sincerest thanks for the taxonomic information provided as well as for their comments and suggestions.

References

Agosti D. and N.F. Johnson. 2002. Taxonomists need better access to published data. Nature 417:222.

Costello M.J. 2003. Role of on-line species information systems in taxonomy and biodiversity. In: Heip C.H.R. (ed) Biodiversity of coastal marine ecosystems patterns and processes: a Euroconference, Corinth, Greece, 05-10 May, 2001, p58-59.

Danielsen F. 1997. Stable environments and fragile communities: does history determine the resilience of avian rainforest communities to habitat degradation? Biodiversity and Conservation 6:421-431.

Fjeldså J., J.C. Lovett. 1997. Biodiversity and environmental stability. Biodiversity and Conservation 6:315-323.

Godfray C.H.J. 2002. Challenges for taxonomy. Nature 417:17-19.

Gray J.S. 2001. Marine diversity: the paradigms in patterns of species richness examined. Scientia Marina 65:41-56.

Morin P.J., J.W. Fox. 2004. Diversity in the deep blue sea. Nature 429:813-814.

Piraino S., G. Fanelli, F. Boero. 2002. Variability of species'roles in marine communities: change of paradigms for conservation priorities. Marine Biology 140:1067-1074.