Vanden Berghe E., M. Brown, M.J. Costello, C. Heip, S. Levitus and P. Pissierssens (Eds). 2004. p. 15-24
Proceedings of 'The Colour of Ocean Data' Symposium, Brussels, 25-27 November 2002
IOC Workshop Report 188 (UNESCO, Paris). x + 308 pp.
– also published as VLIZ Special Publication 16

Integration of environmental datasets, formats and software in the IODE Resource Kit

Murray Brown

Phoenix Training Consultants
834 Elysian Fields Ave., New Orleans, Louisiana 70117, USA
E-mail: murraybr@bellsouth.net

Abstract

The Intergovernmental Oceanographic Commission's (IOC) International Oceanographic Data and Information Exchange Program (IODE) has developed an extensive suite of Web-based training materials called the IODE Resource Kit. A major section, entitled 'Data Analysis and Products,' has been designed to introduce data management trainees to global and regional environmental datasets, the principal formats used for their storage and distribution, and public-domain software for marine data quality-control and analysis. A principal aim of this section has been to achieve integration between the major datasets selected for IODE training, using existing pathways between them afforded by specific format compatibilities. It is now possible to identify a suite of (mainly) public-domain software programs that not only provide bridging functions between various databases, but also perform many analysis and quality-control functions. These interconnections are functionally illustrated by three functional schematics used in the IODE training curriculum. Every dataset used in the IODE workshops can be located on the diagrams, and easy paths can be traced to any desired software program. A set of 50 illustrated 'Roadmap Tutorials' has been developed to give step-by-step directions on the processes involved, including raw data entry, spreadsheet and relational database manipulations, grid-and-contour methods, and finally multi-parameter synthesis in Geographic Information System (GIS) applications. At the present time, most major formats and datasets of principal interest to hydrographers and chemical oceanographers are included in the schematics and in the IODE curriculum.

Keywords: Marine data management; Training.

Introduction

Since the late 1990s the Intergovernmental Oceanographic Commission has sponsored an international training program in marine database management (funded principally by Flanders and Belgium), later expanded to include marine information management. Focused primarily on developing countries, the program has based its training curriculum on the IODE Resource Kit (Reed, 2004), which contains extensive fundamental documentation and very practical workshop Tutorials. This paper discusses the central core paradigm for training in marine data analysis and data products, one of the three main components of the data portion of the Kit.

Earlier experience with the OceanPC Program in the early 1990s had indicated to the IOC trainers that the mere availability of data and software resources – even when a modest degree

of integration between the units was available – did not lead to development of national data collections or of especially noteworthy product development among the former trainees. It was apparent that a very robust practical and theoretical structure of relationships among the various datasets, formats and software programs must be developed, and that the training to be based on this structure must vividly demonstrate the possible connections and processes through practical demonstrations of them in 'real-world' situations.

Toward this end, the IOC trainers have canvassed the Internet to catalog the best ocean data sources, their formats, and the freeware that can be used to edit, display and analyze them. Special attention has been paid to finding or writing format conversion programs that accomplish connections between important resources. Using a list of principal datasets, formats and software, the trainers have developed integration diagrams that show explicitly the connections between them, and have written a complete suite of training Tutorials that show how to move among and between the resources. This suite is actually a step-by-step manual for the creation of a national ocean data collection for Namibia, and for the creation of important data products from that collection.

OceanPC Program

The predecessor program to the Data Analysis and Products section of the Kit was the OceanPC training activity sponsored by the IOC in the early 1990s. Based heavily on the many modular programs for data management and quality control written by Harry Dooly at ICES, the system of software could accommodate the World Ocean Database files, and contained very good quality-control utilities. It lacked sophisticated data analysis capabilities, and there were some issues of user-friendliness that caused concern, particularly in the area of data editing. Basically, it consisted of a host of small programs that 'communicated' with each other through two common formats: the ICES standard profile format, and the ICES spreadsheet format. The only other major 'external' program that the system supported was SURFER, through the export of simple XYZ flat files for gridding. Although the ICES programs could achieve some degree of data sub-selection, by space and time coordinates, this process was not straightforward and somewhat inflexible. It was felt then that the time had come to consider full relational database management technology, which was not at the heart of the ICES programs.

Planning the 'Next Generation'

In early 1995, the IOC convened a meeting of the OceanPC trainers to discuss possible future directions for marine data management training, with a special view toward the software problems described above. Two diverging philosophies emerged, named by their proponents The Central Engine (CE), and the Daisy-Chain (DC) Models.

Central Engine Model

The main feature of the CE model is that it requires a massive, all-encompassing software program (the 'Engine') that is compatible with all principal formats and can perform all desired QC and analysis work, in addition to exporting data and products in all import formats and other

necessary formats (e.g. GIS images and shapefiles, spreadsheets). This model, shown in Fig. 1, suffers from the problems associated with planning and supporting the creation of the major software required, and the prospect that the main program would probably be obsolete as soon as it was published. Quite frankly, it was discarded as impractical.

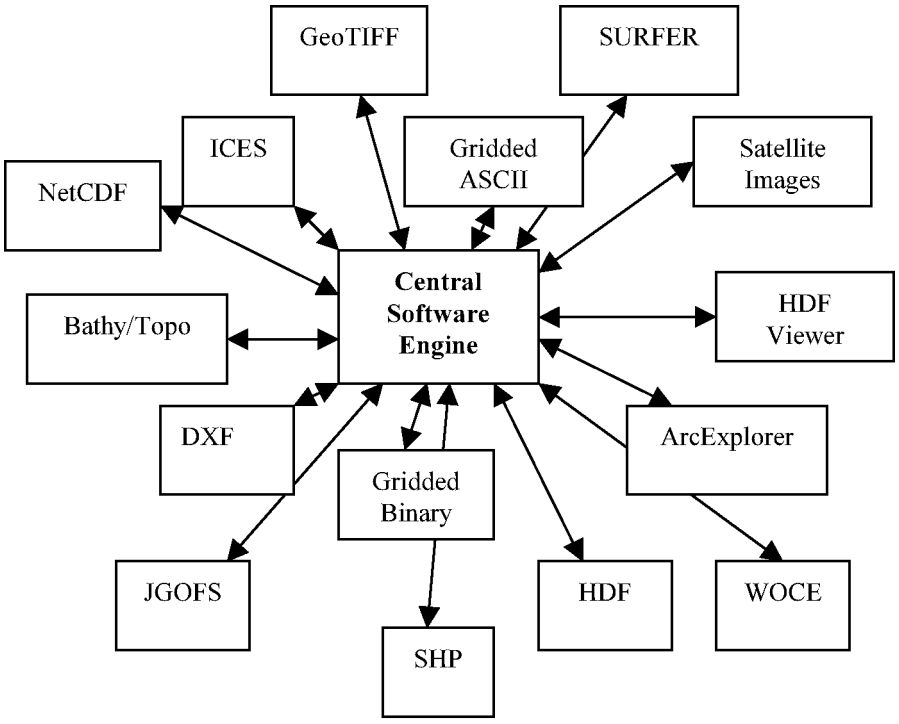


Fig. 1. Schematic representation of the Central Engine Model. The set of formats, databases and programs shown here is just representational, intended to show the profusion of connections required; it is not a faithful catalog of the resources considered at the time.

Daisy-Chain Model

The main feature of the Daisy-Chain Model is that it does not require any specific new software development, but instead relies on connections between existing application programs and on the future evolution of format compatibilities between these programs and major formats used in data publications. Its advantages lie in the fact that software development (except in the case of small format converter programs) is usually handled by other parties. Its disadvantages lie in the need for constant monitoring of the available software and formats in current use, in order to remain abreast of current developments.

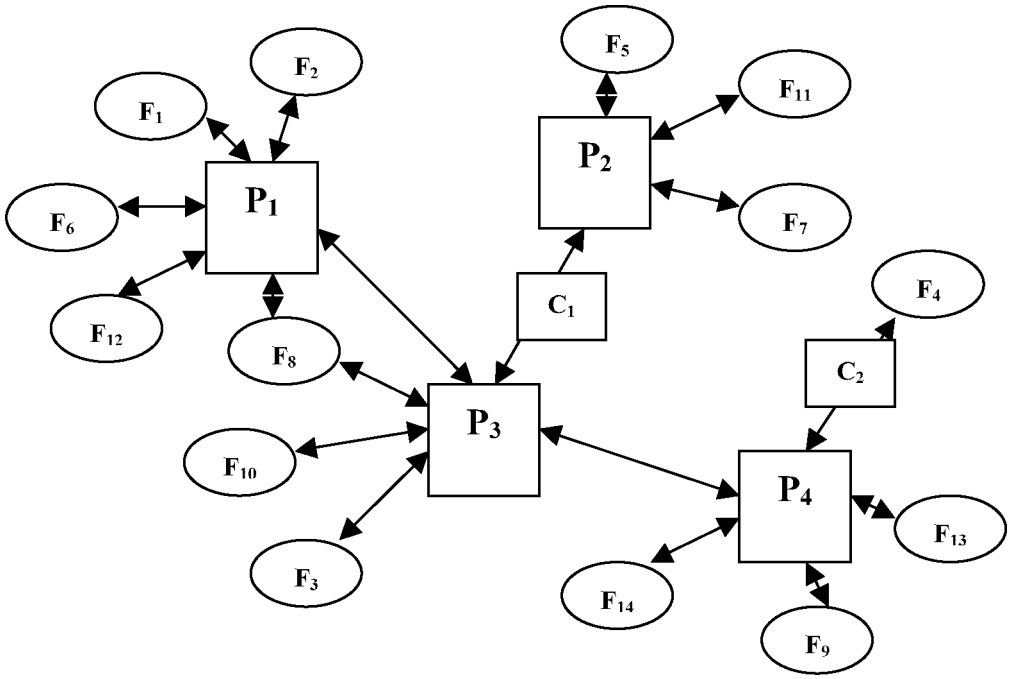


Fig. 2. Schematic representation of the Daisy-Chain Model. The 'system' is composed of various software programs (P), each with their own format (F) compatibilities, sometimes aided by special format converter programs (C).

Technology advances

Both because there was no budget at hand to continue the OceanPC development project, and because extremely rapid developments in Internet technology and the availability of new software seemed to require a strategic delay, nothing was done about system upgrade immediately after the 1995 meeting. A veritable explosion in the databases and software programs available to the ocean community occurred in the brief period after the introduction of Internet browsers (1994-6). Instead of searching high and low for resources, the IOC trainers (one of them an original OceanPC instructor) had the unexpected pleasure of wading through dozens of candidates.

New software included Ocean Data View (http://www.awi-bremerhaven.de/GEO/ODV/), an ocean database synthesis, quality-control and analysis program published by Reiner Schlitzer; Java OceanAtlas (http://odf.ucsd.edu/joa/jsindex.html), a similar program published by John Osborne; HDF Browser (http://www.intersys21.com/html1/hdf.html), a program for creating and viewing HDF files; HDFView (http://hdf.ncsa.uiuc.edu/hdf-java-html/hdfview/index.html), a program for viewing and exporting data matrices from HDF files; ArcExplorer (http://www.esri.com/company/free.html), a freeware GIS browser; and various upgrades of the popular SURFER gridding and contouring program (http://www.goldensoftware.com/demo.shtml). In addition a major portion of the ICES DOS-based software was upgraded to Windows format.

New database publications available either on-line or in CD-ROM format included the World Ocean Database 2001 (http://www.nodc.noaa.gov/OC5/SELECT/dbsearch/dbsearch.html); World Ocean Atlas 2001 (http://www.nodc.noaa.gov/OC5/WOA01/pr_woa01.html); satellite sea-surface temperature analyses from US NASA (e.g. POET [http://seablade.jpl.nasa.gov/poet/]) in global and regionally-subset form; the comprehensive WOCE dataset (http://www.nodc.noaa.gov/woce_v3/); numerous global relief datasets (gridded binary and ASCII) from the US NGDC (http://www.ngdc.noaa.gov); and very large sets of climate-related datasets from US NASA (http://redhook.gsfc.nasa.gov/~imswww/pub/imswelcome/plain.html).

It is difficult to list similarly any specific developments in the area of formats during this period, other than the usual committee-based development of new grand formats for international use by everybody, which activity is usually on-going, but has no practical impact. The mid- to late-1990s were marked, however, by a decrease in the publication of ad-hoc, new formats invented by CD publishers (e.g. the TOGA CD-ROM's ASCII chunk files and the WOCE Version 2 winds in binary grids), in favor of existing standard formats. HDF and NetCDF emerged as the formats of choice for satellite images and gridded climatological data. The ICES standard profile and spreadsheet formats dimmed in interest as users struggled with their eccentricities. The wild popularity of the World Ocean Database publications, however, did not focus interest on their format due to the ability of Ocean Data View to read the files directly, thus sparing the user the need to learn the ins and outs of stratified formats.

The bottom line to the technology explosion of the late 1990s was that the rapid availability of all these resources, amid the steady trend toward a smaller family of formats, simply presented itself as the solution to the problem. A ready-made Daisy-Chain Model of resources integration just appeared on the horizon, and it was greeted, adopted and fostered by the IOC trainers.

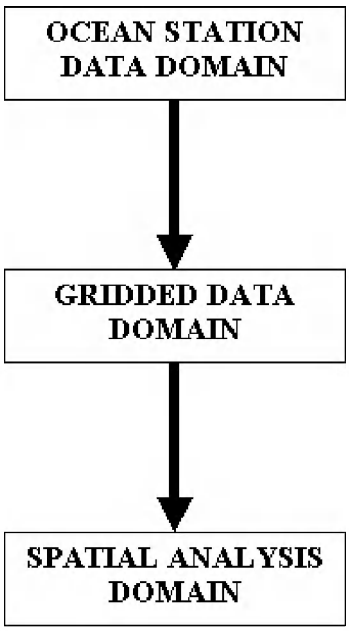
Emergence of the Resource Kit

The earliest use of a set of integrated resources was the IOC-supported Black Sea CD, used for marine data training in Sofia, Bulgaria, in 1998. It consisted of a simple one-page HTML interface to a set of programs, datasets and format descriptions. An integration diagram that showed the relationships between the formats was presented, but it was not included on the CD-ROM.

This first appearance of the integrated approach was followed by the use of CD-ROMs of increasing sophistications during the ODINEA and IOCINCWIO Marine Data Management Training programs for eastern African countries in the 1998-2000 period. As the pedagogic skills of the trainers increased, the total number of software programs decreased (from a high of over 40 to about 20 today), list of emphasized formats decreased (from 22 to 8), but the principal datasets used for training have remained constant at about 20. The early integration diagram has evolved into a set of three 'domain diagrams' in a separate section on Data Integration in the Data Analysis and Products part of the Kit. Beginning in 2001, manuals for each workshop have been prepared, primarily in view of the extreme size of the Kit (currently over 10,000 files). As well, tutorials have been included in the Kit, showing students how to navigate through the integration diagrams.

Integration diagrams

Although the Resource Kit's basic model for Data Analysis and Products is the Daisy-Chain Model, as described above, the many possible combinations and connections has made it necessary to break up the model into three 'domains,' as shown on the right. There are obvious areas of overlap, requiring a few programs or formats to appear on more than one diagram, but the division seems to successfully separate quite different areas of work into more-easily understood relationships. The domains portray the data analysis process as a sequential progression from 'data,' through gridding and contouring, to spatial analysis (GIS applications).



The Ocean Station Data Domain consists of the datasets, formats and software concerned with ocean survey data. It emphasizes the methods to create a national data collection from published and unpublished sources, and the methods to create 'data products' for analysis by other means (e.g. gridding and contouring in the Gridded Data Domain).

The Gridded Data Domain consists of the resources concerned with gridded data, and with Level 3 satellite images, which are – in a sense – visual representations of data grids on the earth surface. This domain also includes the mechanisms for creating special formats which can be used in the next domain.

The Spatial Analysis Domain consists of resources concerned with GIS analysis.

Ocean Station Data Domain

Two main 'routes' are emphasized in this domain: the pathway from raw data to Ocean Data View (or digital data capture), and the pathway from global archives to ODV (Fig. 3). The latter is easily navigated, due to ODVs built-in compatibility with principal formats. The former requires careful tutelage, due to the many different data reporting systems found in old hard-copy data or spreadsheets. Particular attention is paid to the 'units problem.' The role if ACCESS, as surrogate for all relational database management systems (RDMS), is entirely optional, because ODV currently performs such a wide range of data management functions that we have not found it necessary to use RDMS technology for our work. We do, however, provide training on methods to migrate data between ODV and a typical RDMS program if that is needed.

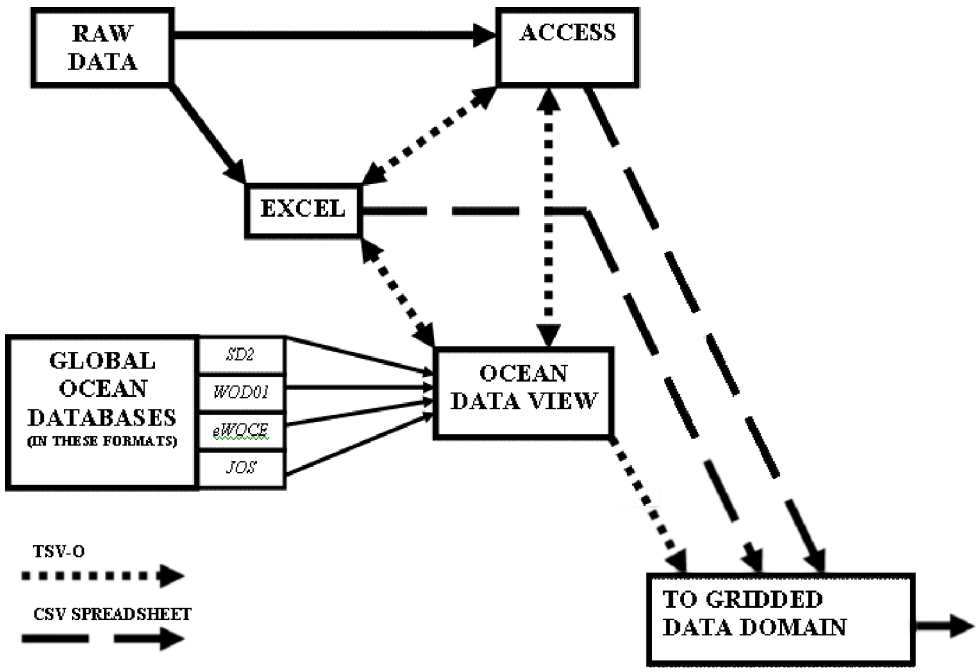


Fig. 3. Schematic diagram of the Ocean Station Data Domain. Training tutorials in the Kit emphasize the route from Raw Data to EXCEL to Ocean Data View, as well as direct data import to ODV from Global Databases and export from ODV of spreadsheet data for use in the Gridded Data Domain.

Gridded Data Domain

The Gridded Data Domain includes both data matrices obtained by gridding ocean station data and satellite images mapped to the earth's surface. Due to the limitation that the entire IOC training curriculum is focused on the ultimate use of ArcExplorer, which does not support image projection changes, the image mapping must be Cartesian (also called equirectangular or equatorial cylindrical equidistant). This Domain includes the enormous archives of HDF and NetCDF images and climatological means available from numerous online archives. Quite frankly, however, one of the most valuable formats in this Domain is the simple XYZ format used for transfer between programs. Data in XYZ format are ubiquitous on the Internet.

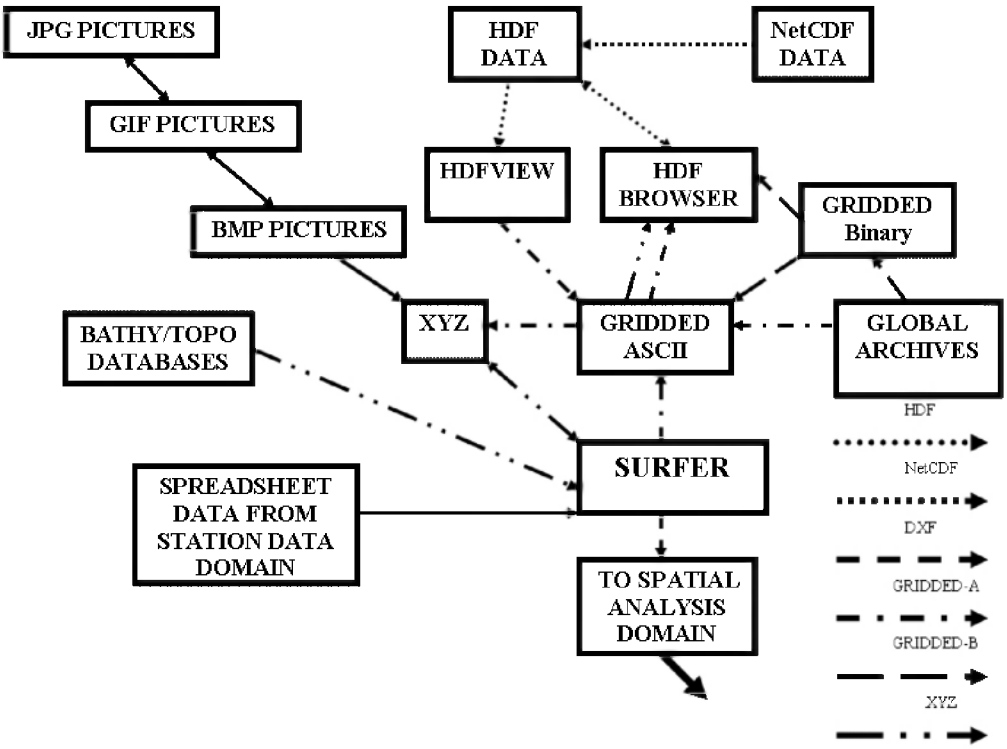


Fig. 4. Schematic diagram of the Gridded Data Domain. Tutorials in the Kit emphasize numerous interconnections between the many formats in this Domain. Some small format conversion programs are not shown here to simplify the diagram. SURFER has found extensive use as a format converter, in addition to its powerful grid-and-contour functions.

Spatial Analysis Domain

This Domain is concerned primarily with the synthesis of 'overlay' maps that combine data products and images from the Gridded Domain and from many global data resources. Students who have access to complete GIS systems can increase the scope of this Domain to include the use of images and vector files in other projections or coordinate systems. Students with only the resources provided in the Kit are limited to Cartesian display (e.g. X and Y in degrees, without projection). A perhaps surprising feature is the extensive use of the DXF format, an older workhorse interchange format in GIS work that is still widely supported.

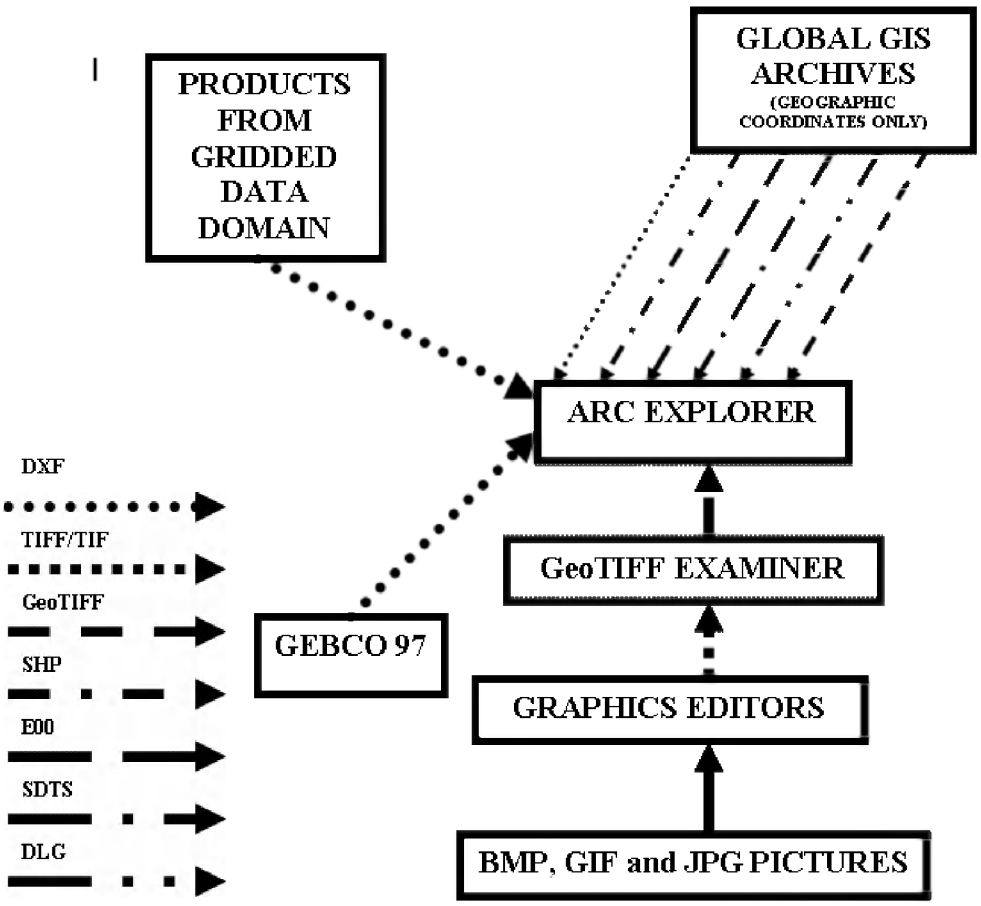


Fig. 5. Schematic diagram of the Spatial Analysis Domain. At this level the data products developed in the previous two Domains are combined with many different types of global resource data files in appropriate formats. Of special interest is the incorporation of GeoTIF images which often allows the use of actual photographic images for large-scale work (e.g. local analyses at ICZM levels of resolution).

Conclusions

The IODE Resource Kit has developed a methodology for the integration of a wide range of ocean data, data formats and applications software, using a set of three 'Domain' diagrams to illustrate the relationships. The marine data management training program of the IODE relies heavily on these diagrams to point students to documentation for the elements of the integration scheme, and to provide the big picture for the Roadmap Tutorials contained in the Kit.

Acknowledgements

This work of developing the IODE Resource Kit has been supported by the Intergovernmental Oceanographic Commission, the governments of Flanders and Sweden, and by details of employees from the governments of Australia and the United States. Special thanks are given to Mr. Peter Pissierssens who knows how to lead a team by letting them work in peace. For the development of the integration scheme, much is owed to the original OceanPC team, Mr. John Withrow, Dr Douglas McLain and Dr Harry Dooley.

References

IODE Resource Kit, n.d. http://www.oceanteacher.org. Intergovernmental Oceanographic Commission of UNESCO. Paris.

Reed G. 2004. OceanTeacher: building capacity in oceanographic data and information management. p. 39-44. In: Proceedings of 'The Colour of Ocean Data' Symposium, Brussels, 25-27 November 2002. Vanden Berghe E., M. Brown, M.J. Costello, C. Heip, S. Levitus and P. Pissierssens (Eds). IOC Workshop Reports 188 (UNESCO, Paris). x + 308 pp. – also published as VLIZ Special Publication 16.