Oceanography Committee **ICES CM 2002/C:12**
Ref. ACE

Report of the

ICES-IOC Study Group on the Development of Marine Data Exchange Systems Using XML

Helsinki, Finland
15–16 April 2002

International Council for the Exploration of the Sea

Conseil International pour l'Exploration de la Mer

Palægade 2–4 DK–1261 Copenhagen K Denmark

TABLE OF CONTENTS

1 OPENING OF THE MEETING

The meeting was opened by R. Gelfeld (Co-Chair) and was hosted by the Finnish Institute of Marine Research, Helsinki (FIMR), Finland. P. Alenius welcomed the participants on behalf on FIMR R. Olsonen outlined the local arrangements.

Members of the Study Group present were: P. Alenius (Finland), T. Carval (France), D. Collins (USA), J. Gagnon (Canada), R. Gelfeld (USA, Co-Chair), A. Isenor (Canada, Co-Chair), P. Knight (UK), R. Lowry (UK), N. Mikhailov (Russia), F. Nast (Germany), R. Olsonen (Finland), G. Reed (IOC Consultant), L. Rickards (UK), H. Sagen (Norway), J. Szaron (Sweden), and E. Vanden Berghe (Belgium). The ICES Secretariat was represented by H. Dooley, ICES Science Coordinator/Oceanographer. Absent members included: P. Haaring (Netherlands), C. Haenen (Netherlands), T. O'Brien (US), B. Pelchat (Canada), and R. Starek (US). Meeting participants not official members of the Group included M. Fichaut (France), R. Hietala (Finland), K. Manni (Finland) and K. Tikka (Finland). A complete list of names, addresses and contact points of participants can be found in Annex 1.

2 ADOPTION OF THE AGENDA

The agenda (Annex 2) for the SG meeting was adopted as a resolution at the 89th Statutory Meeting (C.Res. 2001/2C11). Note that this is the first meeting of the Group.

3 WHY MIGHT XML BE IMPORTANT?

R. Gelfeld introduced the discussion noting that many have exchanged data in the past, but formats often change between publication of data sets (e.g., the World Ocean Database, WOD. For a complete list of acronyms, see Annex 3). Such changes often necessitate rewriting software to handle the data, which is troublesome for users. This is where XML could help and prevent problems in the future. As well, the meeting will provide a forum for sharing knowledge and experiences related to XML developments. However, standards need to be developed. Obviously this will take some time, but a start can be made within this meeting and intersessional work. Hopefully, the meeting would result in the initial development of a plan to guide an investigation into how this technology might best be used in an oceanographic context.

The Group was reminded of the Terms of Reference that provides direction for the activities. The Terms of Reference are:

An ICES-IOC Study Group on the Development of Marine Data Exchange Systems using XML (SGXML) (Co-Chairs: R. Gelfeld, U.S.A. and A. Isenor, Canada) will be established and will meet in Helsinki, Finland on 15 April 2002 to:

a) develop a framework and methodology for the use of XML in marine data exchange in close consultation with IOC and the Marine XML Consortium;

b) develop a Workplan that within 4 years will lead to published protocols for XML use in the marine community;

c) explore how to best define XML tags and structures so that many ocean data types can be represented using a common set of tags and structures.

d) test and refine these common tags and structures using designated case studies i.e.;

   i) Point (physical/chemical) data (profile, underway, water sample).

   ii) Metadata (cruise information, building from the ROSCOP/Cruise Summary Report).

   iii) Marine Biology data (integrated tows (e.g., zooplankton-phytoplankton tows), demonstrate the use of taxonomy);

Several members of the Group noted the potential flexibility of XML. This flexibility could make software development easier. Grammar is recognised as an important element in the development. It was noted that XML is only part of the integrated technology for data management. XML provides the syntax to describe hierarchy, but for applications, semantics are very important and often more difficult than syntax. The brick concept, developed by R. Keeley, was thought by some to be a good approach, but others felt that it was necessary to start with a data model and common dictionary (i.e., only use one code for temperature). The size of XML files was also noted to be a problem as was the time required for parsing the XML files.

From an IOC/IODE perspective, the requirement was to design a framework for an XML structure that data centres can use. Flow of data between centres is very important. If an acceptable structure is produced, this will solve many problems.

The Group recognised that many international groups were looking toward this Study Group to suggest the appropriate direction for XML (e.g., IOC, JCOMM, ICES, etc.).

4 OVERVIEW OF MARINE RELATED XML PROJECTS AND INVESTIGATIONS

A series of short presentations were given to illustrate XML developments in the various organisations represented. These presentations, which predominantly served to initiate a series of discussions, are summarised below (summaries are delimited by the bold text). Slide presentations are provided in Annex 4.

The **IOC use of XML** was noted to be in metadata management. The Marine Environmental Data Information Referral Catalogue (MEDI) is the IOC metadata system. It began in 1979 and several versions have been published in the IOC manuals and guides series. In 1996, IODE 15 requested that an electronic system be developed. MEDI uses the NASA Global Change Master Directory (GCMD)/Directory Interchange Format (DIF) structure.

MEDI provides a mechanism to upload entries to global directories. The software incorporates a Java database management system with client server software and connection to the Internet. XML is also used. There is GIS functionality and an Adobe SVG plug in. At present this works on Unix systems and PCs. It is not yet available for Macs.

There is a quasi-common DTD between GCMD and MEDI, although there are some differences. More functionality exists in MEDI for marine metadata thus there are some extra fields. A copy of the DTD for MEDI is available (see: ioc.unesco.org:8080/medi/Welcome.jsp - this needs a password for access). The GCMD keyword validations are used within MEDI. The lists are maintained and updated by GCMD. Updates can be suggested.

On behalf of **AODC**, G. Reed presented XML activities at the Australian Data Centre. Within AODC, XML is used primarily to manage data within the Centre, as opposed to an exchange mechanism with other Centres. The structures developed at AODC include quality control parameters as well as data. AODC has also extended spatial descriptions to include points, lines and polygons. The temporal space is instant, continuous and periodical. All of this is encoded as data objects and is currently being used internally by the Royal Australian Navy for profile, meteorological, seabed composition, transparency and bioluminescence data. AODC uses MEDI for metadata.

At AODC, the XML technology has simplified and accelerated quality control procedures and has improved data processing. It has simplified storage by having all raw and edited data along with metadata in one file. A DTD and documentation are available.

A description of the **Marine XML Consortium** was given. This is an effort originally proposed by the Chair of IODE. At IODE XVI, the potential of XML for a transfer mechanism was recognised and IOC participation was recommended. The current situation has a Consortium XML office located at IOC. A proposal has also been submitted to the EU (during the WGMDM meeting we were informed that the proposal had been approved at about 65% of the proposed funding level).

The proposal is composed of several work packages. The overall aim of the proposal is to develop a prototype of a marine markup language. Present partners include: HR Wallingford, UK MIC, 7seas, NERSC, CLCRL, RIKZ, VLIZ, SCO (Social Change Online), SMHI (EuroGOOS) and IODE.

A **Dublin Core** presentation dealt with its potential application to ocean metadata. Librarians seeking interoperability between metadata on collections initially created Dublin Core. The Core represents a consensus of metadata elements. Extended to a set of attributes, it has resulted in an ISO standard. The Core set is now 15 elements. The elements are multi-lingual and could be useful for something like cruise reports. The Core also allows for extension into a particular field of study (see www.dublincore.org).

A recent WMO meeting on Integrated Data Management, ISO 19115 on Geographic Metadata was also noted (Geneva Nov 2001, www.wmo.ch/web/www/WDM/reports/ET-IDM-2001.rtf). An overview of elements is presented in their report. This effort should be followed. The WMO requirements are considered to conform to Dublin Core and the draft ISO standard geographic metadata (19115) could be applied to WMO requirements. The WMO team is expected to review the proposed core metadata standard with respect to existing datasets, develop a draft list of keywords to describe WMO datasets, and develop a list of proposed extensions needed for ISO code lists.

The Group recognised that one advantage of WMO is the established communications system. However, to use the communication system you have to use WMO codes. This does offer application in an operational environment and is

useful when moving smaller amounts of data in the communication system. There are obvious links to JCOMM. A WMO paper describing links between systems (XML, BUFR, etc.) was also noted.

Work at the **Finnish Environmental Administration (FEA)** was also presented. In this effort, co-operating partners wanted to receive environmental datasets from multiple sources without fixing the format of the data being delivered. Here, the data transfer is not in XML. Instead, the format description is in XML (with defined elements). The data are delivered in free format (not XML). The collaborators developed their own tag names for the metadata descriptions. For more details, see www.ymparisto.fi

A description of the **FIMR cruise planning system** used operationally on board Finnish Research vessels was also given. The system has evolved since 1981, and was initially ASCII based. However, changing needs have resulted in version support problems. XML may help the problems by offering easier version control because of extensibility. In this environment, new ideas can be added, cruise files can be validated and one has easy presentation of information using XSL. Within this system, the most difficult and perhaps the most important issue is the data model.

The Group also considered **metadata standards**. The basic questions of metadata deal with who, what, when, where, how, and why. Existing standards focus on a group of data as a dataset, not the individual observations. Existing standards include:

- GCMD Data Interchange Format (DIF) – This has 'valids' such as geographic location, platform types, sensors. Many items are lumped such as address for an Institute. Here, lumped is used to describe the granularity of the data representation. Lumped indicates a collection of data. For example, address may be a single field. It has a well-defined DTD and is the basis for MEDI.
- US FGDC Content Standard for Digital Geospatial Metadata (CSDGM) – This has no controlled vocabulary but allows selection of thesauri to control the vocabulary. This is a "splitter" system and has a growing population of users. For example, each piece of an address has its own element. Address is split into its parts.
- ISO 19115/TC211 – Not yet accepted as a standard. This has a well-defined DTD but no controlled vocabulary. It has conformance levels to identify the amount of detail included. Level 1 is the "light" amount while Level 2 contains verbose descriptions.
- MARC 21 (MAchine Readable Cataloguing record) – from Library of Congress. www.loc.gov/marc/marcsgml.html. This is an XML/SGML work in progress. The Library of Congress controls the vocabulary. This is a mix of splitter and lumped.
- Other systems include:
  - GML
  - OCLC – Co-operative Online Resource Catalogue
  - Dublin Core
  - European cataloguing standards

All Geospatial Standards converge around the ISO Metadata standard. It was suggested that SGXML utilise common features in the metadata standards as we work towards standards.

The Group noted that splits and lumps are related to local vs. global views. For example, in a local domain a system only need consider one address format. A global approach to something like address would require splitting. Such split and lump issues make mapping in one direction easier than the other (split to lump is easier). It was also noted that the granularity in the Keeley bricks is related to splits and lumps.

A presentation on the use of **XML for oceanographic data exchange** considered the requirements for such a system. For example, development of a document type definition (DTD) and standardised syntax for common content (code tables). With these, we would be able to exchange data knowing the details of the content.

Such DTD development would incorporate mandatory, optional and undeclared variables. Flexibility would be provided by including all three-element types.

The mandatory elements would define the elements without which the document would be useless. These elements would cover both structural integrity and content integrity (e.g., make sure fields like latitude and longitude are included). There would be extreme control through markup in some cases (e.g., positions and times).

For optional elements, these would be described as desirable but not essential. This would introduce elements that are only required for certain types of data (e.g., for zooplankton an optional might be net mesh size while for a CTD this is irrelevant).

The undeclared elements would allow for necessary anarchy. These elements could have any content and structure but may only be useful with consenting data exchanges.

The DTD development is an exercise of modelling oceanographic data and such modelling has been done before under IOC (e.g., GF3). Another useful aspect of GF3 was the code tables (see GF3 published as Manuals and Guides No 17). However, there are much more extensive dictionaries now - BODC, Pangaea (world data center for marine environmental data, AWI).

One thing that is currently missing is accessibility. Code tables need to be mounted on web servers as universally available resources (BODC available as relational tables). In this case, the URI needs to be incorporated into an XML document and we need to standardise the code table syntax.

There are also problems with code table maintenance. Maintenance is an open-ended commitment. In the past, others have made local extensions and submitted to BODC and these have been incorporated (when appropriate). Maintenance is possible through mutual co-operation.

The Russian presentation noted that the **Russian NODC (RNODC)** has been using XML for about 4 years. The RNODC are using XML for managing metadata within an interagency meta database.

At present, the RNODC has cruise metadata served over the web - www.oceaninfo.ru. The GTS data are available at www.meteo.ru:8080/gisserver. They are using XML for the metadata with a reference to the data file name. Mikhailov stressed that the syntax is the easy part, while semantics are the difficult part.

One possible approach would be to use XML for the distribution of data obtained from a distributed ocean data system. The users want integrated and co-ordinated data products (JCOMM requirements also). This is the general thinking behind MedBlackDODS. Here, there are four regional centres each with its own area of responsibility and data sets. The National centres connect to regional centres through the distributed model. The user would like to define a request irrespective of where data are located

The main tasks for such a project could be described as:

- A common data model
- A common dictionary and codes
- XML application
- Global navigation
- Software for integration and services
- Middleware - needs to be defined

This was considered as a demonstrator project at IODE XVI. Here, the connection between data sets would use XML files. The ESIMO (Black Sea experiment) connects an experimental observing network in the Black Sea, Russian NODC, AARL St Petersburg, Vladivostock, etc. This is a pilot project 2001/2002. There may also be GIS connections in this project.

The important aspect for XML transfer is both the semantics and the syntax. The semantics are in the DTD. Here, the important issues are a common parameter dictionary, a common data model, common codes and a request model. The syntax is important for the DOM. Issues here are related to tags, attributes, entity references, processing instructions, character data sections and document type data.

The common data model is the key, and the bricks are an important component of the model. Related to this are various data relationships (internal and external), normalisation, common dictionary development, common code tables and general data model description. Then, the description could be mapped to an XML syntax and thus produce a marine XML DTD.

In France, **IFREMER** is currently using XML for metadata, data and for web services. In terms of metadata, current investigations are focused on Argo, GIS, ROSCOP and catalogues. For data, it is expected that XML will replace NMEA frames on French research vessels (NMEA is a real-time data system on oceanographic vessels. It displays time, location, etc. and can be combined with actual oceanographic measurements). For web services, the SISMER portal will display information as if it is a central system when it is in fact a distributed system.

Regarding Argo, some would like to use XML to distribute metadata and technical data. At the moment, this distribution is using netCDF for the profile and trajectory data. The Argo data are in the Coriolis database with other data (e.g., XBT data and maybe other SOOP data).

In Canada, XML work has evolved since beginning in 1999. Efforts have concentrated on applying the Keeley bricks to XML. The Keeley bricks really represent small packets, or atomic units of data or metadata. The bricks, initially in the domain of data types dealt with by MEDS (mainly physical data, GTSPP, waves, etc.), have undergone substantial changes over the past 12 months. However, the idea of the packaging of data into these atomic units remains.

A point data example was presented. It was noted that some bricks (e.g., location variables) occurred in various places to satisfy the need for flexibility in the point data type. The flexibility allows point data to have the dependent variable as the x, y, z, or t dimension. In this way, point data includes such data types as underway, profiles, current meter data and profiling current meter data (shipboard or moored).

5 GENERAL APPROACHES TO XML

A. Isenor introduced this item, noting that it was intended to provide the Study Group with a foundation on "XML modelling". The goals were:

- to provide the Group with a familiar starting point for XML initiatives.
- to establish the types of tags one might use in an XML structure, and also the structure itself.
- to obtain an understanding of when to use attributes and tags within an XML structure.

With regards to structure vs. content, a simple relational structure and corresponding XML structure was presented. It was suggested that a particular structure should not be modelled, but rather a general structure.

Another potential issue deals with content globalisation. An example would be date, where there exist many different ways to express the date. Should we be specifying a format, which is then globalisation? Or, perhaps a localisation? In the localisation model the format is not defined, but rather the local specification of the format is registered and software converts to the global standard. The Group considered globalisation to be a better solution.

The presentation then examined the attribute vs. element question. There are several facts related to attributes that can be stated:

- You cannot have an attribute name without defining a value
- No two attributes can have the same name in a single tag
- There is no order dependence to attributes
- You cannot expand attributes to contain other attributes within the same document
- You can limit values for attributes to predefined lists

General Discussion:

It was noted that life can be made harder with the bricks. For example, within BODC the water bottle tables hold data for 4000–5000 parameters (2 million rows), but not every sample has data for every parameter. Most users want a spreadsheet format, so how should the output be displayed to a user? XML will ignore things that are not required. If a parameter code is accessible by an API (i.e., the parameter code is not within the content), then one can tell the application “I'm only interested in a few parameters”. How should the parameter code be delivered? If not as an element, then it should be as an attribute. But Lowry was not happy with using attributes as parameter codes. The alternative is to have the parameter code as content, but then it cannot be addressed directly.

However, others in the Group considered the value to be important and didn't think it made a big difference if it is element or attribute content. In both cases value is important.

Some Group members thought that the bricks are the basis for moving forward. However, the Group needs to think about global bricks, not local bricks.

It was noted that most discussion has dealt with the DTD, while we should be considering schemas rather than DTD. However, some wondered if schemas have been ratified by W3C yet? (XML Schema was approved as a W3C Recommendation on 2 May 2001) Dates represent a good example of where the control offered by the schema would be useful. Positions also need clear definition. For both date and position, the use of existing metadata standards was recognised as very important. SGXML should investigate date, time, latitude, and longitude to see the most appropriate standard to use?

Regarding access to parameter codes, it was noted that the BODC parameter dictionary could be moved to XML from the relational database. The Group thought that producing a common structure for parameter dictionaries may be useful. However, content may have already been defined in GF3. The bricks introduce structure and mix content. We need structure-less agreement and a simple set of content.

Some thought that there was a lot to be gained from looking at the GF3 header/data model, for how the data are stored. This could be a case where something very simple gives you 90% of the answer. When solutions try to cover 100%, the often expend large resources in software development.

The Group recognised that the next stage of advancement for parameter dictionaries is mapping on the web. Many thought taking GF3 and putting it into XML was also a good idea. This would describe the content and the structure of the data. It was suggested that the GF3 to XML mapping exercise be completed in the evening.

A general overview and questions that resulted from the discussion may be summarised as:

- Detailed view - presentations
- Grammar issue
- Is XML a viable mechanism?
- XML is syntax - -semantics is a problem
- Parameter dictionaries - we need agreement
- What are the vendors doing?
- File size issue - cannot put everything inside XML. Simply not feasible.
- Will XML help with version control?
- Validating parser
- Lumps and splits

Day 2: General Discussion Continues

A. Isenor began the discussion by reviewing yesterday's activities with the hope of setting the stage for progress. The results of the evening GF3 investigation noted that the elements resulting from the GF3 series header should be considered useful (Annex 5). However, the elements from the GF3 data cycles section were weak.

In terms of offerings, the Group recognised that XML provides existing tools (one data file has the potential to satisfy multiple clients) which may solve problems related to diverse topics, like displaying cruise reports on web pages or providing CSV Argo metadata. However, the Group needed to refocus and consider the Terms of Reference.

In terms of positives to build from, the Group displayed some common thinking on metadata. As well, there is a Canadian point data commitment and SGXML should assist and benefit from this effort. This effort will use the information gained here to adjust the Keeley bricks.

The Group recognised that the ICES CSR was one form of metadata. We should also remember MEDI and EDMED. The XML solution should try to establish one system that can present the information in the form of a CSR, MEDI or EDMED.

D. Collins then gave a presentation on metadata. Collins suggested the Group utilise tools that already exist such as FGDC. We also need to split specific pieces of information rather than lump this information together. FGDC standard has 10 sub-sections that can be repeated anywhere within previous ones (e.g., look at contact info (pseudo-DTD)). This

allows the definition of each piece of information in a dataset in a detailed or overview form. The overview could refer to publications.

The tag explanations available in FGDC should be examined with the results of the GF3 investigation. Pieces of the FGDC could translate to the GF3 items and may also apply to CSRs. We can get a considerable tag base from FDGC. It was noted that most people prefer to just provide the overview information.

The Group recognised XML as a component of data distribution. However, there are other components like the parameter dictionary. XML is only for the future distribution of data.

The Group was also given a description of the Russian NODC experimentation. The original data centre has its own data model and the inhouse data model is not reconstructed in the XML solution. Here, the DOM is used for exchange. At the minimum this would be on a cruise level. The DOM will allow connection with Oracle after some modification. The DOM connects and acts as a buffer between data and Internet space. A navigator is also required for content and conditions on each server. This also combines the information received from the distributed sources. There are lots of examples of integrators. We should be trying to make a DTD for oceanographic data (e.g., profile) then develop a DOM model. This year, RNODC will connect 3 organisations in Russia: GTS data, cruise data and underway time series.

R. Lowry proceeded to give the group a description of the BODC parameter database. He proposed that the list of fields be extracted from Oracle to XML. Files are already available over the web as CSV files. There is also an Access version but it is not up-to-date. Instructions are also available on the BODC website (www.bodc.ac.uk). It was noted that ICES uses the codes but ignores the units.

Some thought that for applications on the web, we need calculated parameters (e.g., mean temperatures, concentrations, etc.). However, this is a dangerous path because there are many possible calculated parameters. Do we need mean and standard deviation for all parameters? As well, there is the problem of defining what is meant by mean (surface, bottom, population, sample, etc.). This leads to an infinite number of parameters. Although we need a way to describe derived parameters, it is unlikely to be within the dictionary.

The Group thought that it would be useful to provide XML access to parameter dictionaries. On the local node, data would be output to a particular form and the mapping converts one to other. Storing locally we would use the local system. However, outside the local environment the data would be delivered in an agreed list. This is fairly simple for temperature, salinity and nutrients. The problem is for things like plankton where there are many codes for species.

It was noted that much information about species is not in the BODC dictionary. We could map the BODC codes to ITIS but ITIS is also limited. However, it was noted that the online parameter dictionary would be a resource. Individual organisations may not use the codes internally, but will use the parameter dictionary in comparisons.

It was recognised that the definition of tags, etc. was important. However, the importance of the dictionary stems from the distributed data system approach. If such a system sends out a request for data (e.g., using spatial-temporal-parameter query) and collects together data from different databases, and then pulls all collected data together, it would presently result in many different names for the same parameter.

Regarding tags, the Group agreed to use lowercase text, with an underscore to represent a blank.

D. Collins then lead the group in flip-chart definition of an XML structure for parameter dictionaries. The resulting strawman for this neutral dictionary structure is below. Note that:

- red text indicates example content
- square brackets [] indicate number of occurrences
- blue text indicates comments that refer to the BODC application of the structure

```
<dictionary_entry> [1]
   <dictionary_term>
      <dictionary_entry_type>Parameter</dictionary_entry_type>
      <definition> ([1,n] e.g., 26 entries for BODC chlorophyll)
         <instance> (description 1 in BODC dictionary) </instance> [1]
         <definition_owner>BODC</definition_owner> [1]
         <creation_date> </creation_date> [1]
         <change_date> </change_date> [1]
         <methodology> (description 2 in BODC dictionary)</methodology>[1]
         <unit_of_measurement> </unit_of_measurement> [1]
         <min_value> </min_value> [1]
         <max_value> </max_value> [1]
         <null_representation></null_representation> [1]
         <short_name></short_name> [1]
         <accuracy></accuracy> [1]
         <authority_citation>BODC Data Dictionary</authority_citation> [1]
         <codeset> [1,n]
            <codeset_name>BODC Data Dictionary</codeset_name>
            <code>CHPL0492</code>
            <codeset_owner>BODC</codeset_owner>
         </codeset>
      </definition>
      <synonym> [0,n]
         <synonym_instance>klorofylli</synonym_instance> [1]
         <synonym_owner>FIMR</synonym_owner> [1]
      </synonym>
   </dictionary_term>
</dictionary_entry>
```

This structure represents a step to get individuals away from their own codes and to begin thinking about other code sets. It was recognised that spelling may be a problem for some biology.

A further discussion on details related to the BODC dictionary details ensued. There are 3 fields for describing parameters:

The 1st is 4 bytes - (this as a starter for the above structure)

The 2nd is all 8 bytes

The 3rd is the 8 bytes plus a qualifier for the method

This is designed to be rigorous on codes not text. The 8 byte code is unique.

It was noted that in the above XML, we are trying to make the structure more like a dictionary that people will understand. Who outside of BODC will look up CPHL for example? No one. But people might look up chlorophyll. As well, the synonym allows the mapping between codes.

It was then noted that the system was not designed to produce an international (standard) version for a parameter dictionary. Rather, all dictionaries were considered equal and this would simply allow a mapping between the different dictionaries.

The Group recognised that this XML structure may be useful for other codes such as ship codes, country codes, project codes, etc.

6 ACTION ITEMS RESULTING FROM THE MEETING

The meeting structure was predominantly a series of discussions. Considering this, it was considered better to summarise the Action Items in a single section as during the discussion the items were only loosely defined. Here, the Action Items are presented under four topic headings.

6.1 Parameter Dictionary

Action 1: D. Collins will provide the definitions for the above elements and tags.

Action 2: A total of 11 internal dictionaries will be mapped to the XML structure defined in this document. The mappings will be conducted by: K. Manni, R. Gelfeld, A. Isenor, N. Mikhailov, G. Reed, F. Nast, J. Szaron, P. Alenius, R. Lowry, J. Gagnon, T. Carval.

Action 3: E. Vanden Berghe will provide a DTD for the above structure.

Action 4: P. Alenius will provide a schema for the above structure.

6.2 Point Data Investigation

Action 5: A. Isenor will investigate applying Keeley bricks to point data as a test case.

Action 6: E. Vanden Berghe will provide biological and taxonomic input to the Keeley bricks. This will probably result in the taxonomic brick being completely reformed.

Action 7: The draft point data definition should be commented on by others in the group. Please send comments to A. Isenor who will produce version two of the point data structure by July. Identified reviewers were P. Alenius, F. Nast, T. Carval, K. Manni, E. Vanden Berghe, and R. Lowry. Then at the next meeting we will discuss how to take the point data structure further.

6.3 Metadata Investigation

Action 8: N. Mikhailov will attempt to construct a general metadata model including the definition of EDMED, MEDI, CSR, etc***. Mikhailov will look to the GETADE work with EDMED/ROSCOP and version descriptions. The hope is that next year there will be something tangible to work with. CSR is only one visualisation of the metadata.***

Action 9: Reviewers of this general metadata model were identified as: P. Alenius, R. Gelfeld, D. Collins.

Action 10: The Group will attempt a mapping between MEDI, EDMED, and CSR and produce a description as for dictionaries. Metadata elements for MEDI are well established. After mapping, design new tags for non-mappable fields. Supply the results of the mapping exercise to A. Isenor for incorporation into Keeley bricks and point data structure.

Action 11: Reviewers were identified to review the mapping: H. Dooley, A. Isenor, F. Nast, P. Alenius, and D. Collins. The review should be complete by July. Then A. Isenor will incorporate into point data structure by September.

6.4 Other Items

Action 12: G. Reed will establish a SGXML communication site under the marine XML community portal.

7 ANY OTHER BUSINESS

G. Reed noted that IOC have registered marinexml.net, and would be willing to establish a community portal for marine XML discussion if required. The Study Group thanked G. Reed for this offer and accepted it.

The Group noted that there may be input to the Terms of Reference from the IOC/IODE GETADE group that will be meeting later this week.

It was again noted that an essential component of the development is a general data model containing a common structure. The Group should be instigating development of a common data model on the basis of the full application

domain (marine data in general). Russia will be leading this issue at a national level, to produce a general data model as the basis of a DTD for the widest possible domain.

The proposed 2002/2003 Terms of Reference for the Study Group were briefly discussed and are presented here in Annex 6.

8 MEETING CLOSURE

It was suggested that the next meeting be held in April 2003, at SMHI, Sweden. R. Gelfeld thanked the Swedish members for volunteering to host the next meeting. He then closed the meeting by thanking the Finnish hosts and all of those who had participated.

ANNEX 1: NAMES AND ADDRESSES

Names, addresses and contact points of participants.

Alenius, Pekka,
Finnish Institute of Marine Research,
P.O. Box 33, (Lyypekinkuja 3),
00931 Helsinki,
Finland
Tel (operator): +358 9 613 941
Tel (direct): +358 9 613 94439
Fax: +358 0 61394494
E-mail: pekka.alenius@fimr.fi
Web page: http://www2.fimr.fi/ or www.fimr.fi

Carval, Thierry,
Institut Francais pour le Recherche et
l'Exploitation de la Mer (IFREMER),
Center de Brest,
Departement IDM,
BP 70,
29280 Plouzane
France
Tel: 33–2-98–22–4597
E-Mail: theirry.carval@ifremer.fr

Collins, Donald W.,
U.S. National Oceanographic Data Center
1315 East West Highway, 4th Floor,
Silver Spring MD, 20910,
USA
Tel: +1 301 713 3275 extn 179
Fax: +1 301 713 3302
E-mail: donald.collins@noaa.gov

Dawson, Garry,
Maritime Environment Information Center,
UK Hydrographic Office,
Admiralty Way, Taunton,
Somerset TA1 2DN,
UK
Tel: +44 1823 337900 extn 3225
Fax: +44 1823 284077
E-mail: garry.dawson@ukho.gov.uk
Web page: http://www.hydro.gov.uk/

Dooley, Harry,
ICES Oceanographer,
International Council for the Exploration of the Sea (ICES),
Palaegade 2-4,
1261 Copenhagen K,
Denmark
Tel (operator): +45 33 154225
Tel (direct): +45 33 152677 (tone) 210
Fax: +45 33 934215
E-mail: harry@ices.dk
Web page: http://www.ices.dk

Fichaut, Michele,
Institut Francais pour le Recherche et
l'Exploitation de la Mer (IFREMER),
Center de Brest,
Departement IDM,
BP 70,
29280 Plouzane
France
Tel: 33–2-98–22–6663
E-Mail: michele.fichaut@ifremer.fr

Gagnon, Jean,
Marine Environmental Data Service (MEDS),
Department of Fisheries and Oceans,
200 Kent Street, 12th Floor,
Ottawa, Ontario K1A OE6,
Canada
Tel: +1 613 990-0260
Fax: +1 613 993-4658
E-mail: GagnonJ@dfo-mpo.gc.ca
Web page: http://www.meds-sdmm.dfo-mpo.gc.ca/

Gelfeld, Robert D., (Co-Chair)
U.S. National Oceanographic Data Center/
World Data Center - A Oceanography,
1315 East West Highway, 4th Floor,
Silver Spring MD, 20910-3282,
USA
Tel: +1 301 713 3295 extn 179
Fax: +1 301 713 3303
E-mail: rgelfeld@nodc.noaa.gov
Web page: http://www.nodc.noaa.gov

Hietula, Riikka,
Finnish Institute of Marine Research
P.O. Box 33, Fin-00931,
Helsinki,
Finland
E-Mail: riikka.hietula@fimr.fi

Isenor, Anthony, (Co-Chair)
Bedford Institute of Oceanography,
P.O. Box 1006,
Dartmouth,
Nova Scotia B2Y 4A2,
Canada
Tel: 902 426 4960
Fax: 902 426 7827
E-mail: isenora@mar.dfo-mpo.gc.ca
Web page:
http://www.mar.dfo-mpo.gc.ca/science/ocean/welcome.html

Knight, Phil
Proudman Oceanographic Laboratory,
Bidston Observatory, Prenton,
Merseyside, CH43 7RA,
United Kingdom
Tel: +44 151 653 1556
E-mail: pjk@pol.ac.uk

Lowry, Roy,
British Oceanographic Data Center,
Proudman Oceanographic Laboratory,
Bidston Observatory, Prenton,
Merseyside, CH43 7RA,
United Kingdom
Tel: +44 151 653 1519
E-mail: rkl@bodc.ac.uk
Web page: http://www.bodc.ac.uk

Manni, Kati
Finnish Environment Institute,
Data and Information Centre
P.O. Box 140,
FIN-00251 Helsinki
Finland
Tel: 09–40300698
Fax 09–40300691
E-Mail: kati.manni@ymparisto.fi

Mikhailov, Nicolay,
Russian National Oceanographic Data Centre
6 Korolev St.,
Obninsk, Kaluga Region,
Russian Federation, 249020
Tel: 7–084–397–49–07
Fax: 7–095–255–22–25
E-Mail: nodc@meteo.ru

Nast, Friedrich,
Deutsches Ozeanographisches Datenzentrum (DOD),
Bundesamt für Seeschiffahrt und Hydrographie
Bernhard-Nocht-Str. 78
D-20359 Hamburg,
Germany
Tel: +49- (0) 40 - 3190–3530
Fax: +49- (0) 40 - 3190–5000
E-mail: friedrich.nast@bsh.de
Web page:
http://http://www.bsh.de/Oceanography/DOD/htm

Olsonen, Riitta,
Finnish Institute of Marine Research,
P.O. Box 33, (Lyypekinkuja 3),
00931 Helsinki,
Finland
E-mail: riitta.olsonen@fimr.fi
Web page: http://www2.fimr.fi/ or www.fimr.fi

Reed, Greg,
Intergovernmental Oceanographic Commission (IOC),
1 Rue Miollis
75732 Paris Cedex 15
France
Tel: 01 45 68 3960
E-Mail: g.reed@unesco.org

Rickards, Lesley
British Oceanographic Data Center,
Proudman Oceanographic Laboratory,
Bidston Observatory, Prenton,
Merseyside, CH43 7RA,
United Kingdom
Tel: +44 151 653 1514
Fax: +44 151 652 3950
E-mail: ljr@bodc.ac.uk
Web page: http://www.bodc.ac.uk;
http://www.oceannet.org

Sagen, Helge,
Institute of Marine Research
Norwegian Marine Data Centre
Nordnesgt 50,
5817, Bergen
Norway
Tel: 47 55 23 8500
E-Mail: helge.sagen@imr.no

Szaron, Jan,
Swedish Meteorological and Hydrological Institute,
Oceanographic Services,
Nya Varvet 31,
SE - 426 71 Vastra Frolunda,
Sweden
Tel: +46 (0)31 751 8971
Fax: +46 (0)31 751 8980
E-mail: jan.szaron@smhi.se
Web page: http://www.smhi.se

Tikka, Kimmo
Finnish Institute of Marine Research,
P.O. Box 33, (Lyypekinkuja 3),
00931 Helsinki,
Finland
E-mail: kimmo.tikka@fimr.fi
Web page: http://www2.fimr.fi/ or www.fimr.fi

Vanden Berghe, Edward,
Manager, Flanders Marine Data and Information Centre
Flanders Marine Institute
Vismijn, Pakhuizen 45–52,
B-8400 Ostend,
Belgium
Tel: +32 59 342130
Fax: +32 59 342131
E-Mail wardvdb@vliz.be
Web Page: http://www.vliz.be

ANNEX 2: 2001/2002 TERMS OF REFERENCE

2C11 An **ICES-IOC Study Group on the Development of Marine Data Exchange Systems using XML** [SGXML] (Co-Chairs: R. Gelfeld, U.S.A. and A. Isenor, Canada) will be established and will meet in Helsinki, Finland on 15–16 April 2002 to:

a) develop a framework and methodology for the use of XML in marine data exchange in close consultation with IOC and the Marine XML Consortium;

b) develop a work plan that within 4 years will lead to published protocols for XML use in the marine community;

c) explore how to best define XML tags and structures so that many ocean data types can be represented using a common set of tags and structures.

d) test and refine these common tags and structures using designated case studies i.e.;

   i) Point (physical/chemical) data (profile, underway, water sample),

   ii) Metadata (cruise information, building from the ROSCOP/Cruise Summary Report),

   iii) Marine Biology data (integrated tows (e.g., zooplankton-phytoplankton tows), demonstrate the use of taxonomy).

SGXML will report by 1 June 2002 for the attention of the Oceanography Committee and ACE.

ANNEX 3: LIST OF ACRONYMS AND TERMS

Acronym or Term	Description
AODC	Australian Oceanographic Data Centre
Argo	The Array for Real-time Geostrophic Oceanography
BODC	British Oceanographic Data Centre
BUFR	Binary Universal Format Representation
CSDGM	Content Standard for Digital Geospatial Metadata
CSR	Cruise Summary Report
CSV	Comma Separated Variable format
DIF	Directory Interchange Format
DOM	Document Object Model
DTD	Document Type Definition
EDMED	European Directory of Marine Environmental Data
FEA	Finnish Environmental Administration
FGDC	Federal Geographic Data Committee (USA)
GCMD	Global Change Master Directory
GETADE	IOC's Group of Experts on the Technical Aspects of Data Exchange
GF3	General Format 3
GTS	Global Telecommunications System
GTSPP	Global Temperature-Salinity Profile Program
ICES	International Council for the Exploration of the Sea
IOC	Intergovernmental Oceanographic Commission
IODE	International Oceanographic Data and Information Exchange
ISO	International Standards Organisation
ITIS	Integrated Taxonomic Information System
JCOMM	Joint Commission on Oceanography and Marine Meteorology
JGOFS	Joint Global Ocean Flux Study
MEDI	IOC Marine Environmental Data Information Referral Catalogue system
MARC	MAchine Readable Cataloguing record
MEDS	Marine Environmental Data Services - Canada
NODC	U.S. National Oceanographic Data Centre
OCL	Ocean Climate Laboratory/U.S. NODC
ODAS	Ocean Data Assimilation System
RNODC	Russian National Oceanographic Data Centre
ROSCOP	Report of Observations/Samples Collected by Oceanographic Programmes (now CSR)
SGXML	ICES/IOC Study Group on the Development of Marine Data Exchange Systems using XML
SISMER	French National Oceanographic Data Centre
SOOP	Ship of Opportunity Program
SQL	Structured Query Language
TOR	Term of Reference
WDCA	World Data Centre for Oceanography/Silver Spring
WGMDM	Working Group on Marine Data Management
WMO	World Meteorological Organisation
WOD	World Ocean Database
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language

ANNEX 4: PRESENTATION SUMMARIES

Content List

Development and Use of Marine XML within the Australian Oceanographic Data Centre to Encapsulate Marine Data

Belinda Ronai, Paul Sliogeris, Matthew de Plater, Krystyna Jankowska
Australian Oceanographic Data Centre.
Maritime Headquarters, Wylde St, Potts Point, NSW, 2011, Australia.
http://www.aodc.gov.au

Abstract

The Australian Oceanographic Data Centre (AODC) archives, quality controls and disseminates a large variety of marine data. The extensible Markup Language (XML) has been tailored to encode marine data and provides a clearly defined way to structure, describe and interchange marine data and has been established as the internal data format standard for the AODC. The development, advantages, disadvantages and uses of Marine XML within the AODC have been outlined.

Development and Use of Marine XML within AODC

Design Goals

- To provide a means of encoding all types of marine data for storage and portability.
- To establish a basis for developing a streamlined data management system within the AODC by creating interoperable software, based upon Marine XML encoded data, utilising XML technologies.
- To create easy to understand, self describing encoding of marine data.

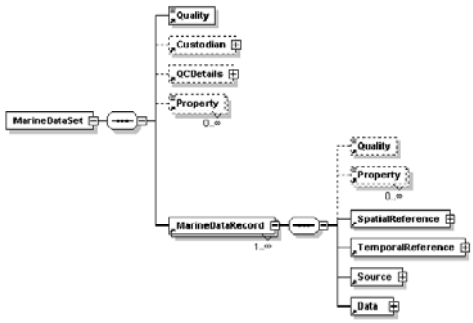
Design Goals

- To develop a means to track all quality control processes and edits and operators within the one XML file.
- To reformat data using XSLT into other marine formats to meet national and international data exchange obligations.
- To display data using XSLT straight from XML without reformatting data into graphics files.
- To integrate metadata within the marine data itself.

Representing Marine Data

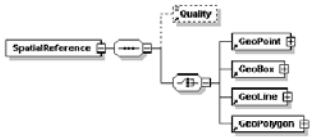
- The data structure needs to be able to handle varying temporal and geographical extents.
  - Ability to specify temporal data as an instant in time or a period (continuous or in-continuous).
  - Ability to specify geographical extents as a single point, line, box, or polygon region.
- The structure needs to handle both multi-dimensional and multi-parameter data.
- Ability to track any changes on the data, provide details on quality control processes and quality flag systems used.

Marine XML Structure



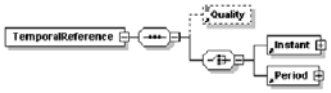
Spatial Extent

- Single point – SST measurement
- Line data – side scan sonar
- Polygon – gridded climatological data
- Box (3D polygon) – climatological data at standard depths



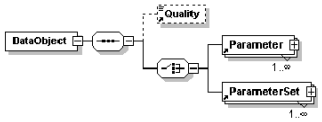
Temporal Extent

- Instant – SST observation
- Continuous period – ADCP or ARGO float
- Periodical period – climatological data, e.g. monthly



Encoding Marine Data

- Each Data element encapsulates a series of *DataObjects* elements
  - measured data
  - parameter measured
  - sensor used



Encoding Marine Data

- Example: Simple point data using *DataObject* and *Parameter* elements:

```
<DataObject index="0“  type=“Ancilliary" numberOfParameterSets="1" reject="false">
  <Parameter index="0" name="Bathymetry" units="Metres">
    <Value>100.0</Value>
  </Parameter>
</DataObject>
```

Encoding Marine Data

- Example: Point data with quality flag using *DataObject* and *Parameter* elements:

```
<                ="0"      ="XBT"     ="Primary"                          ="2"      ="false">
  <                ="0"      ="Water Temperature"      ="Degrees Celcius">
    <              ="XBT"      ="Mk12 T-10" />
    <      >23.5</      >
  </           >
  <                ="1"      ="Quality Control Flag"      "">
    <      >1</      >
  </           >
</           >
```

Current Status

- The Marine XML is currently being used within the AODC to manage various types of RAN collected marine environmental data.
  - temperature-depth profiles
  - sound velocity-depth profiles
  - meteorological data
  - seabed composition data
  - water transparency data
  - bioluminescence data

Current Status

- Quality Control
  - MarineQC software has been developed to quality control the data stored in Marine XML.
  - Data is visually inspected and all edits performed using the software are reflected in the Marine XML file.

Current Status

- Metadata management
  - The IOC MEDI system is used to generate DIF metadata records from marine data encoded in Marine XML.

Conclusions

- The development of MarineQC, AODC quality control software, and metadata system is based around data encoded in Marine XML.
- Simplified and accelerated the quality control procedures within the AODC, improved data processing capabilities.
- Simplified the storage of marine data by having all of the data, raw and edited, along with metadata within the one XML file for the whole of its life.

Marine XML Consortium

Background

- At IODE-XVI (2000), the chairman (Ben Searle) introduced his proposal for development of a Marine XML Specification through the creation of an international consortium.
- There would be a membership fee and a number of different levels of participation. The membership fee will assist in defraying the costs of supporting a consortium

Background

- The IODE Committee acknowledged the importance of XML and recognized the need for IODE to be closely involved in the development of a Marine XML
- Recommended the participation of IOC, through IODE, in the development of a Marine XML as part of a consortium of interested groups

Background

- One of the medium term objectives of the IODE Group of Experts on Technical Aspects of Data Exchange (GETADE) is to:
  - develop a marine XML as a mechanism to facilitate format and platform independent information, metadata and data exchange

Current Situation

- As from January 2002 the IOC Secretariat has been responsible for hosting the Marine XML project office.
- IOC has registered Internet domain name:
  MarineXML.net
- MarineXML Project proposal submitted to EU for funding
  - Decision on approval expected April 2002

Project Overview

- To demonstrate that XML technology can be used to develop a framework that improves the interoperability of data for the marine community and specifically in support of marine observing systems
- To develop a prototype of an XML-based Marine Mark-up Language (MML)

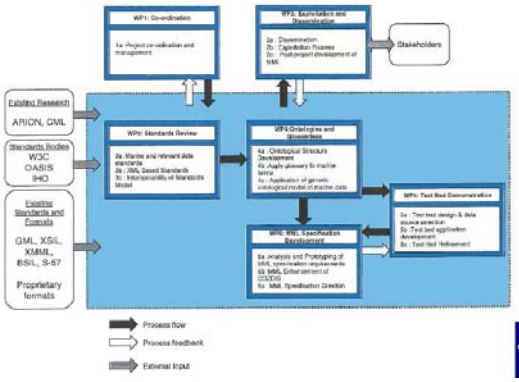
Project Overview

- MarineXML is to be developed in partnership with international agencies, government departments and organisations responsible for data standards to ensure that the research meets the needs of key stakeholders with interests in global ocean observing systems.

Project Objectives

- To produce a prototype marine data ontology framework for interoperability
- To produce working demonstrations of the data interoperability framework
- To develop a prototype MML specification
- To advance the standardisation of a Marine Mark-up Language

Project Work Plan



Consortium Participants

- HRW HR Wallingford UK (Coordinator)
- UKMIC UK Marine Information Council UK
- 7CS SevenCs DE
- NERSC Nansen Environmental and Remote Sensing Centre NO
- CLRC Central Laboratory of the Research Council UK
- RIKZ Rijkswaterstaat NL
- VLIZ Flemish Marine Institute BE
- SCO Social-change On-line AU
- SMHI Swedish Metrological and Hydrological Institute (EuroGOOS) SE
- IOC/IODE International Oceanographic Data & Information Exchange Committee INT

Project Outcomes

- The MarineXML Project will not result in the creation of a full MML specification
- The project will address the underlying framework issues of interoperability between existing and emerging standards
- It will provide a technical basis for the development of full specification

G. Reed – IOC Use of XML in Metadata Management

The Marine Environmental Data Information Referral Catalogue (MEDI) is a directory system for datasets, data catalogues and data inventories within the framework of the Intergovernmental Oceanographic Commission's International Oceanographic Data and Information Exchange (IODE) programme. It has been set up to ensure the widest possible coverage of data holdings and included a review of existing national and international data directory systems as well as implications of inter-operability with similar systems within other international organisations.

The database structure for MEDI has been based on the Global Change Master Directory (GCMD) DIF structure developed by NASA. Metadata descriptions can be easliy transfered from MEDI to global data directories. In addition, major global data collection programs, such as GOOS, can use MEDI to describe their datasets, with MEDI providing the mechanism to upload metadata to global directories.

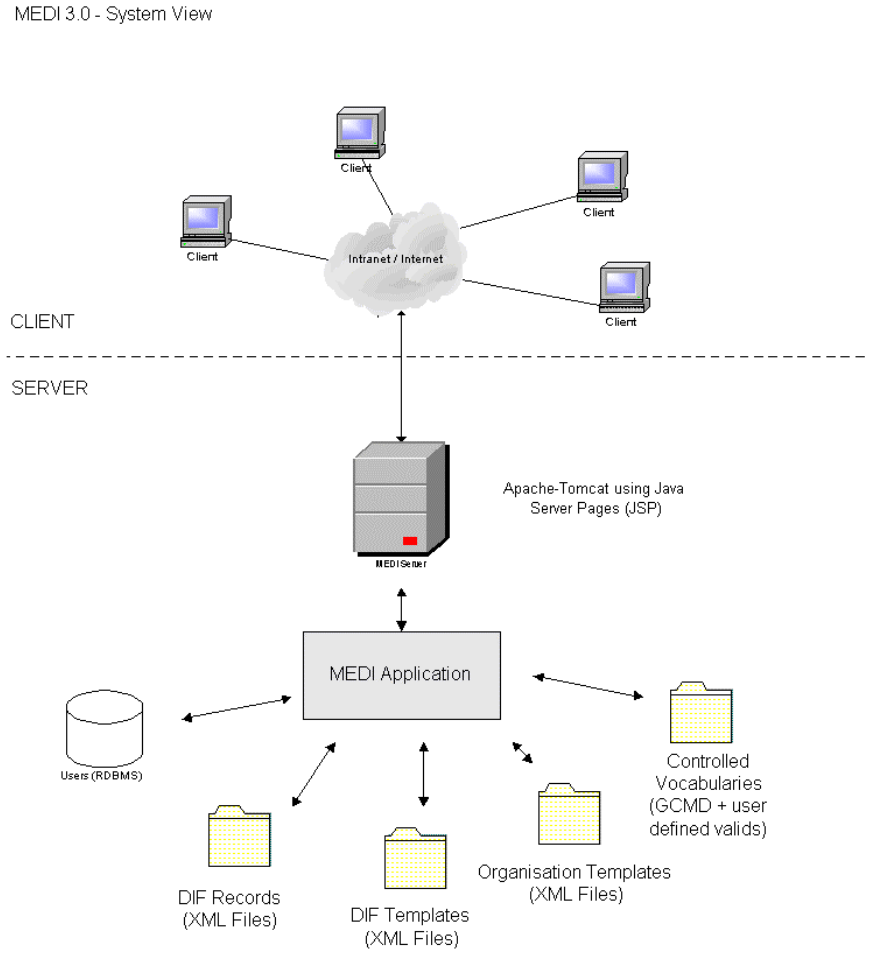
The MEDI software operates in a client-server configuration. Clients on a local network can access the software while keeping the database centralised. The software can also be run on a server that is URL addressable on the internet and consequently allow web access to the system.

Some of the features of the MEDI software include:

- Use of controlled vocabulary. GCMD keyword valids are used for parameters, locations, platforms, instruments, data centres, projects, URL content type
- Use of standard formats for spatial and temporal coverages
- Use of a spatial interface to search by geographic area
- Ability to represent the spatial coverage of the dataset as a rectangle, polygon, line, single point or multiple points
- Use of XML to validate and transfer metadata
- Use of Scalable Vector Graphics to dynamically generate graphics from the data

The MEDI software tool is browser-driven, thus allowing users to connect to the internet, if required, to search for marine-related metadata. The software can also be used locally, either as a stand-alone system or on a local network.

The current version includes the MEDI server that operates as a service under Apache Tomcat 3.2.3 using HTML, JSP and servlets to render functionality. Apache Tomcat operates on Windows, UNIX and LINUX platforms. MEDI uses standard HTTP protocol, hence can be accessed via the internet or intranet. Metadata records are stored as DIF-XML files and data can be imported and exported using standard ZIP formats. The GIS functionality is delivered using SVG (Adobe SVG plug-in 3.0). All text is displayed via a translation table that allows multi-lingual functionality. The current distribution size is 9.93MB and the software can be downloaded from the IOC web site. The software has been successfully tested using Windows and Unix operating systems. The SVG viewer software is currently only supported by Internet Explorer and does not function correctly with Netscape.
On-line help files, with examples, are available to assist the user.



Records can be imported and exported in XML format. The MEDI Document Type Definition (DTD) is used to define the rules and relationships between elements in an XML document used for transferring data and is compatible with the GCMD DTD providing ease of transfer of metadata records between the two directories.

Dublin Core Metadata

The Dublin Core Metadata Initiative is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies.

Dublin Core Metadata

2002-03-06

A proposal for a European network of Dublin Core implementers has been submitted to the European Commission. There are currently 53 members and a further 20 organizations that have expressed interest.

Dublin Core Metadata

An XML namespace XML-NAMES is a collection of names identified by a URI reference that are used in XML documents as element types and attribute names.

The use of XML namespaces to uniquely identify metadata terms allows those terms to be unambiguously used across applications, promoting the possibility of shared semantics.

Dublin Core Metadata

Each Dublin Core element is defined using a set of ten attributes from the ISO/IEC 11179 standard for the description of data elements.

Each element is optional and may be repeated. Each element also has a limited set of qualifiers, attributes that may be used to further refine (not extend) the meaning of the element.

Dublin Core Metadata

Each Dublin Core element is defined using a set of ten attributes from the ISO/IEC 11179 standard for the description of data elements.

Each element is optional and may be repeated. Each element also has a limited set of qualifiers, attributes that may be used to further refine (not extend) the meaning of the element.

Dublin Core Metadata

The Dublin Core standard comprises fifteen elements:

Title	Description
Contributor	Format
Source	Coverage
Creator	Publisher
Date	Identifier
Language	Rights
Subject	Relation
Type	

Dublin Core Metadata

The DCMI **Type Vocabulary** provides a general, cross-domain list of approved terms that may be used as values for the Resource Type element to identify the genre of a resource.

Ex.

Name: Dataset
Label: Dataset
Definition: A dataset is information encoded in a defined structure (for example, lists, tables, and databases), intended to be useful for direct machine processing.

Dublin Core Metadata

The Dublin Core Element Set was originally developed in English, but versions are being created in many other languages, including Finnish, Norwegian, Thai, Japanese, French, Portuguese, German, Greek, Indonesian, and Spanish.

Dublin Core Metadata

This model allows different communities to use the DC elements for core descriptive information which will be usable across the Internet, while allowing domain specific additions which make sense within a more limited arena.

WMO MEETING OF THE EXPERT TEAM ON INTEGRATED DATA MANAGEMENT
GENEVA, 5 - 8 NOVEMBER 2001

- www.wmo.ch/web/www/WDM/reports/ET-IDM-2001.rtf
- The experts determined that the elements needed to meet WMO requirements could be considered to conform to Dublin Core and that the draft ISO standard Geographic Metadata (19115) could be applied to WMO requirements.
- An overview of the ISO and WMO core elements and their corresponding names within the draft ISO standard is provided in the report of the meeting.

WMO MEETING OF THE EXPERT TEAM ON INTEGRATED DATA MANAGEMENT
GENEVA, 5 - 8 NOVEMBER 2001

The draft ISO standard 19115 is based on a number of subsidiary standards defining Geographic terms and processes starting at ISO 19000.

This set of standards provides an enormous range of definitions and specifications of metadata elements, provides a schema and establishes a common set of metadata terminology, definitions, and extension procedures.

WMO MEETING OF THE EXPERT TEAM ON INTEGRATED DATA MANAGEMENT
GENEVA, 5 - 8 NOVEMBER 2001

The ISO 19115 specifies a process (in ISO 19115 Annex C) where a community can adopt parts of the standard which it feels relevant (including the “Core Elements”) and also extend the elements, keywords and code table instances to suite that community.

WMO MEETING OF THE EXPERT TEAM ON INTEGRATED DATA MANAGEMENT
GENEVA, 5 - 8 NOVEMBER 2001

The expert team, recognising that a restrictive definition of a dataset could not be applied to the widely varied requirements of all WMO Programmes, agreed on a minimal and flexible working definition of a dataset as follows:

A dataset is a collection of information (data, products, etc.) that the owner considers as a unit.

Each dataset would have one and only one metadata description.

WMO MEETING OF THE EXPERT TEAM ON INTEGRATED DATA MANAGEMENT
GENEVA, 5 - 8 NOVEMBER 2001

•The team will review the proposed core metadata standard with respect to existing datasets, develop a draft list of keywords to describe WMO datasets, develop list of proposed extensions needed for ISO code lists, and propose an XML schema and examples.

Comments

•XML has the potential to greatly simplify the exchange of data and metadata between the oceanographic community and users of oceanographic data.

•Standard, freely available software, such as newer web browsers, can be used to display limited amounts of XML data clearly and simply.

•More complex software can be used to search XML data files for particular items of information, and to process these in various ways.

K. Manni (Finland) – Skye Finnish Environmental Administration (FEA)

THE MODE OF DATA TRANSFER OF THE FINNISH ENVIRONMENT ADMINISTRATION

The (Finnish) Environmental Administration (VYH) has created a VYH-mode of data transfer for environ-mental data transferring. The basic idea of this mode of data transfer is that the deliverer of the information describes the data and the form (structure) of it according to the directions stated in the Internet pages of Finnish environment administration (http://www.ymparisto.fi/eng/orginfo/database/vyhform/index.htm). Then the structure of the information (data) itself to be delivered can be chosen quite freely. The description of the delivered data is submitted in a separate description file or at the beginning of the transfer file.

Description file
- In XML form
- Describes the data and the data file structure
- Used tags are described on the FEI-web site

Data file
- Contains the data itself
- Data is in different blocks
- Free structure of the file
- Column delimiters or fixed width columns
- Order of the columns can be freely chosen

Example of a description (just the beginning of it):

```
<DESCRIPTION>
<HEADER>
 <DATE>10.9.1999</DATE>
 <SENDER>Lassi Lähettäjä</SENDER>
 <ORGANIZATION_S>Vesikonsultit oy</ORGANIZATION_S>
 <E-MAIL_S>lassi.lahettaja@vko.fi</E-MAIL_S>
 <RECEIVER>Kati Manni</RECEIVER>
 <ORGANIZATION_R>Finnish Environment Institute</ORGANIZATION_R>
 <E-MAIL_R>kati.manni@vyh.fi</E-MAIL_R>
</HEADER>
<FILE>
 <DATAFILENAME>pivet.dat</DATAFILENAME>
 <FILEDESCRIPTION>Vedenlaatutietoja</FILEDESCRIPTION>
<DATABLOCK>
 <TARGET>Notto</TARGET>
 <COLUMNSEPARATOR>;</COLUMNSEPARATOR>
 <NUMBEROFDATAROWS> 3 </NUMBEROFDATAROWS>
 <NUMBEROFCOLUMNS> 8 </NUMBEROFCOLUMNS>
<COLUMN>
 <NAME>NottoNro</NAME>
 <COLUMNDESCRIPTION>Näytteenoton numero</COLUMNDESCRIPTION>
</COLUMN>
<COLUMN>
 <NAME>Koordsto</NAME>
 <COLUMNDESCRIPTION>Koordinaatisto</COLUMNDESCRIPTION>
</COLUMN>
<COLUMN>
 <NAME>KoordPohj</NAME>
 <COLUMNDESCRIPTION>Pohjoiskoordinaatti/Leveys(latitude)</COLUMNDESCRIPTION>
</COLUMN>
<COLUMN>
 <NAME>KoordIta</NAME>
 <COLUMNDESCRIPTION>Itäkoordinaatti/Pituus(longitude)</COLUMNDESCRIPTION>
</COLUMN>
<COLUMN>
 <NAME>Nimi</NAME>
 <VALUEBEGINANDENDMARKER>"</VALUEBEGINANDENDMARKER>
```

```
 <COLUMNDESCRIPTION>Paikan nimi</COLUMNDESCRIPTION>
</COLUMN>
<COLUMN>
 <NAME>Nottolaitos</NAME>
 <COLUMNDESCRIPTION>Näytteenottolaitos</COLUMNDESCRIPTION>
</COLUMN>
<COLUMN>
 <NAME>Aika</NAME>
 <COLUMNDESCRIPTION>Näytteenottoaika</COLUMNDESCRIPTION>
</COLUMN>
</DATABLOCK>
```

Example of the data:

```
1;YK;6234565;3845678;"Kukkivajärvi";32;02.08.1998 12:30;"Levälauttoja";
2;YK;6234565;3845678;"Kukkivajärvi";32;15.01.1999 14:00;;
1;"DEPTHT";21.0;
1;"SDT";1.0;
2;"DEPTHT";20.0;
2;"THICKI";0.4;
2;"THICKS";0.3;
1;1;2.0;;"SLE";"1439_001";"Samea näyte";
1;2;0.0;10.0;"SLE";"1439_001";"Samea näyte";
1;3;5.0;;"H";"1439_001";;
1;4;20;;;"1439_001";;
2;1;0.0;10.0;;;"Levähippuja";
2;2;5.0;;;;;
2;3;10.0;;;;;
1;1;416;32;;2.5;0.1;
1;1;231;32;L;2.0;0.2;
1;1;321;32;;25;5;
1;1;216;32;;0.123;0.010;
1;1;456;32;;1.5;0.2;
1;2;416;32;;3.2;0.1;
1;2;231;32;;4.0;0.2;
1;3;216;32;;1.002;0.010;
1;4;456;32;L;0.5;0.2;
2;1;416;32;;3.4;0.1;
2;1;231;32;L;2.0;0.2;
2;2;416;32;;4.6;0.1;
2;2;231;32;L;2.0;0.2;
2;3;416;32;;10.9;0.1;
```

Metadata Standards Overview

Donald W. Collins
US NODC
ICES-IOC MarineXML Study Group
FIMR Helsinki

Basic Elements of Metadata

- Existing standards focus on a group of data as a dataset, not the individual data observations
  - Who
  - What
  - Where
  - When
  - How
  - Why

Metadata Standards 1

- **US NASA GCMD Data Interchange Format (DIF)**
  - widely used format and 'valids' = controlled vocabularies
    - valids are well defined and maintained by NASA/GCMD
  - well defined DTD
  - "lumper" system = multiple pieces of information may be "lumped" into one element
    - Address is most notable element demonstrating this characteristic
  - basis for IOC MEDI tools

Metadata Standards 2

- **US FGDC Content Standard for Digital Geospatial Metadata (CSDGM)**
  - growing population of format users (no statistics about DIF or FGDC use in the US)
  - well defined DTD
  - NO controlled vocabulary
    - allows users to select one or more 'thesauri' for vocabulary control
  - "splitter" system = each piece of information has its own element (and possibly valid domain)

Metadata Standards 3

- **ISO 19115 /TC211**
  - *NOT YET ESTABLISHED OR ACCEPTED!*
  - Well defined DTDs
    - Conformance level 1 and level 2
  - NO controlled vocabulary
    - allows users to select one or more 'thesauri' for vocabulary control
  - "splitter" system = more individual elements than US FGDC

Metadata Standards 4

- **MARC 21 (MAchine Readable Cataloging record)**
  - Widely used cataloging standard for all types of materials
  - XML/SGML tags are a work in progress…
    - http://www.loc.gov/marc/marcsgml.html
  - US Library of Congress controlled vocabulary
  - "hybrid" system = some types of information are 'lumped', some are 'split'

Metadata Standards 5

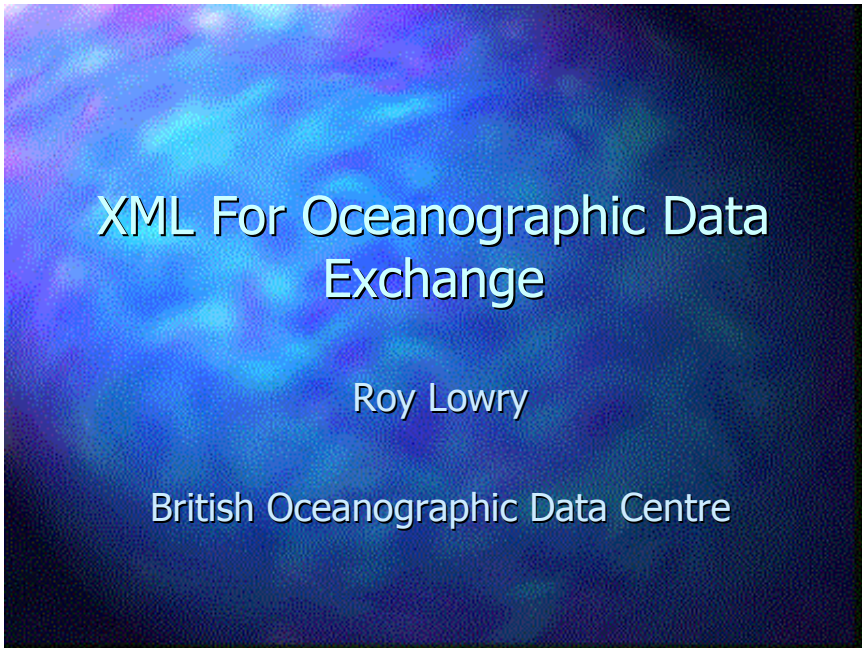
- **Other Standards...**
  - GML = GIS Markup Language
  - OCLC Cooperative Online Resource Catalog (CoRC) = expresses MARC records with XML
  - Dublin Core = well-defined DTD
  - European cataloging standards
    - MARC 21 (US, Canada, UK)
    - Przewodnik Bibliograficzny
- Crosswalks exist and Z39.50 provides a protocol to access across standards
- **ALL Geospatial Standards to "converge" around the ISO Metadata standard**

Crosswalks

- Examples
  - crosswalk from MARC to FGDC
    - http://www.alexandria.ucsb.edu/public-documents/metadata/marc2fgdc.html
  - crosswalk from FGDC to MARC
    - http://www.alexandria.ucsb.edu/public-documents/metadata/fgdc2marc.html
  - crosswalk between FGDC and DIF
    - http://gcmd.gsfc.nasa.gov/Aboutus/standards/fgdc_to_dif_7.0.html

Metadata, XML and Us

- Consider and utilize the common features in the metadata standards as we work toward data standards
- Be wary of making too many splits or too many lumps
- Look at an example of a metadata DTD...



XML 4 Xchange

- What to we need to do to set up successful XML-based oceanographic data exchange?
  - Develop a Document Type Definition (DTD)
  - Develop standardised syntax for common content (code tables)
  - Armed with these we should all know what each other is talking about

DTD Development

- DTD design can incorporate mandatory, optional and undeclared elements
- Applications processing XML ignore what they don't need
- By including all three element types we can design documents with flexible but controlled information content

Mandatory Elements

- These are elements without which the document would be useless
- Cover both structural integrity and content integrity
- Declared as mandatory and forced to have content
- Extreme control through mark-up in some cases e.g. positions and times

Optional Elements

- May be elements that are desirable but not essential
- May be elements that are only required for certain types of data
- Declared as optional and allowed to be empty

Undeclared Elements

- Necessary anarchy
- May have any content and structure (a wall for bricks?)
- Only useful to consenting data exchangers
- DTD elements with the content specified as 'ANY'

DTD Development

- DTD development is an exercise in modelling oceanographic data
- Oceanographic data modelling has been done before under the auspices of IOC
- This spawned GF3 (nice analysis, shame about the syntax)
- Why not develop an XML DTD based on GF3?

Code Tables

- GF3 had code tables
- With one exception, the GF3 code tables could be resurrected as a resource
- The exception is the parameter code table
  - Much more extensive dictionaries now available (BODC, Pangaea)
  - To start again would be madness

Code Tables

- Code tables need to be mounted on Web servers as universally available resources
- URI needs to be incorporated into the XML document
- Code table syntax needs to be standardised
- Obvious way to do this is to set up XML documents conforming to agreed DTDs

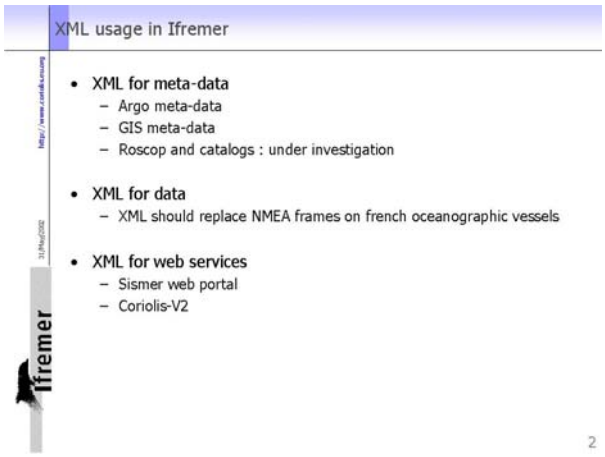
Code Tables

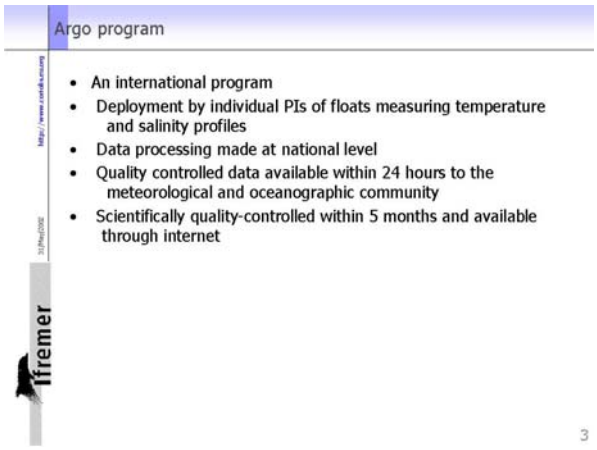
- Problems of maintenance (I'm not daft enough to take on an open-ended commitment)
- Solution may lie in local implementations agreed between consenting exchangers implemented through internal document type declarations

Conclusion

- Is the approach of DTD and XML code table document development considered worthwhile?
- If so, how shall we take it forward?

T. Carval (France) – XML at IFREMER



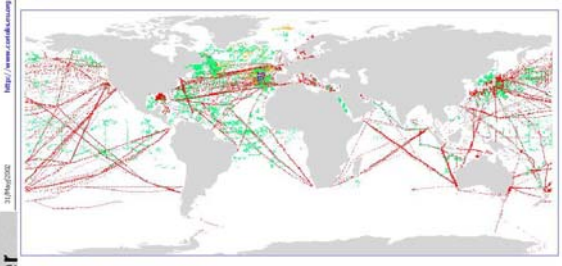


Argo 2001, worldwide coverage

- 491 floats
- 12 517 profiles

4

Coriolis data service 2001, worldwide coverage



Coriolis-data service 2001
- 165 190 profiles
- 5 millions of temperature measurements
- 0.6 millions of salinity measurement

5

Data management concepts

- Two Data Distribution Media
  - GTS for real-time modeling community
  - Internet for Oceanographers and modelers
- Unique data format for each distribution medium
- Standardized quality control procedures
- All data are public

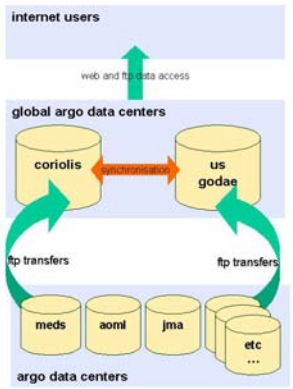
6

Argo data management actors

- PI's (principal investigators)
  The scientists who deploy the floats, then carry out delayed mode QC and return data to National Centres within 5 months of observations.
- National Centres
  The data centres who collect, qualify, process and distribute the float data they are responsible for. Data are distributed to PIs and the GTS within 24 hours of the float surfacing. They also send the data to the Global Data Centres.
- Global Data Centres (GDAC)
  Two central points of Argo data distribution on Internet for all the float data located in Coriolis/Ifremer/France and US GODAE/FNMOC/USA. Coordination between these centres occurs daily.

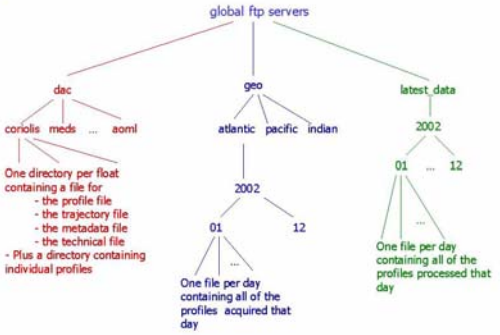
7

Argo data exchange



8

Global FTP servers



9

Global FTP servers

- **Data transfer to the global FTP sites :**
  - Each national center has an account on the global data servers
  - National centers transfer a new profile , to both global servers, in netCDF, simultaneously to the transfer to GTS
  - For new floats Metadata must be transfered first

10

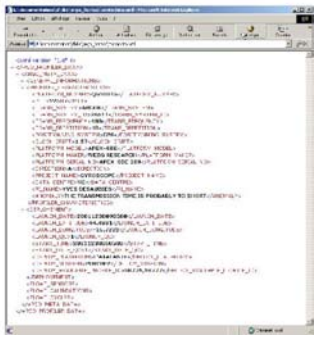
Argo data format

- 4 formats
  - Metadata : XML ascii format readable for users and computers
  - Technical : XML ascii format readable for users and computers
  - Profile : netCDF binary format
    Contains both original data acquired by the float and the best available profile together with quality flags.
    It also contains a subset of metadata necessary for users to manipulate the files
  - Trajectory : netCDF binary format
    Contains both original data acquired by the float and the best available profile together with quality flags.
    It also contains a subset of metadata necessary for users to manipulate

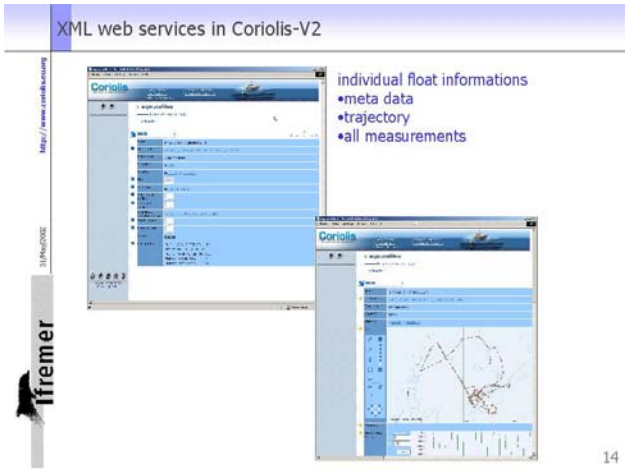
11

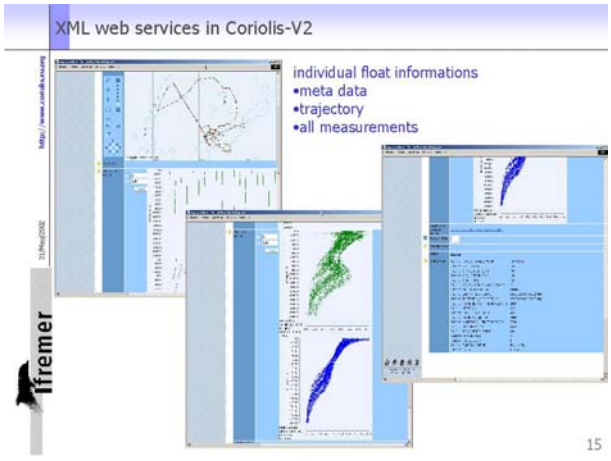
Argo meta-data proposal

- The proposal will probably not be accepted in the near future : users are reluctant
  - The content is not easily readable without a browser
  - Netcdf is popular
  - Ascii tables are popular
- A review by SGXML is necessary for a middle term acceptance

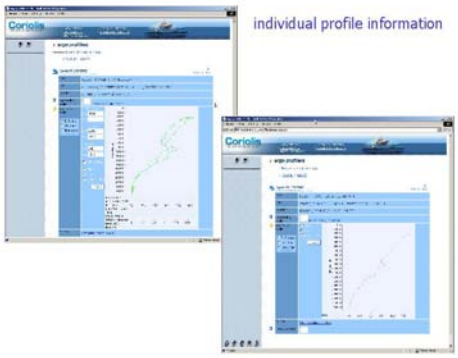



XML web services in Coriolis-V2
interactive maps
argo float directory
13


XML web services in Coriolis-V2
individual float informations
•meta data
•trajectory
•all measurements
14


XML web services in Coriolis-V2
individual float informations
•meta data
•trajectory
•all measurements
15

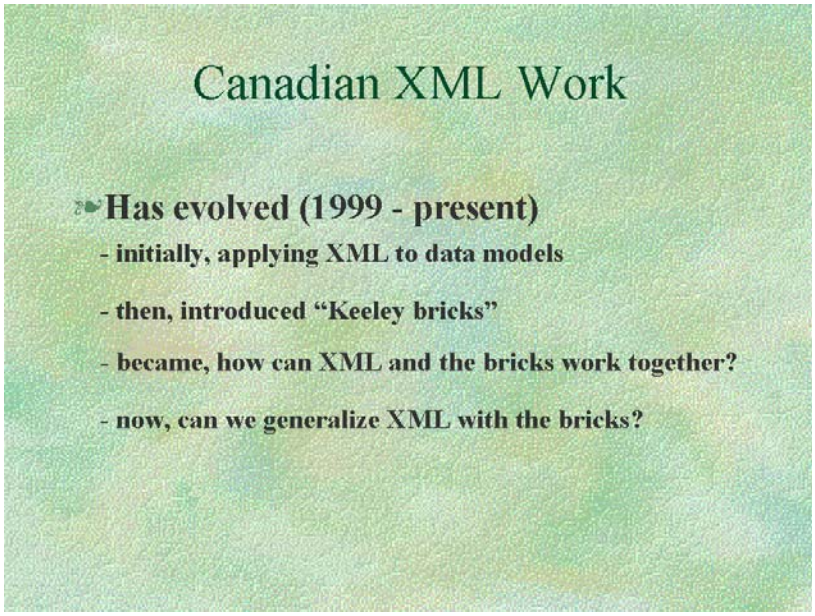
XML web services in Coriolis-V2

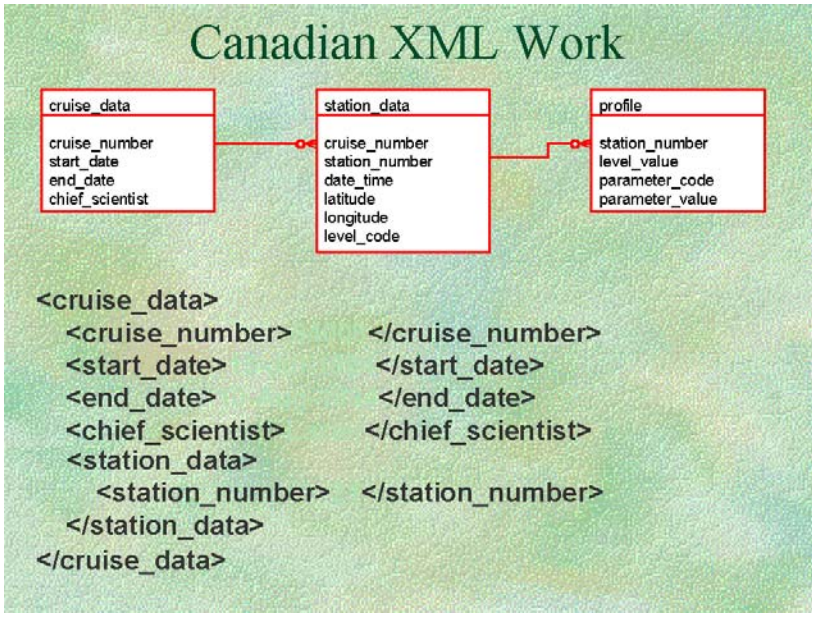


individual profile information

16

A. Isenor – Canadian XML Work





Canadian XML Work

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE cruise_data SYSTEM "concept.dtd">
<cruise_data>
        <cruise_number>       </cruise_number>
        <start_date>           </start_date>
        <end_date>             </end_date>
        <chief_scientist>      </chief_scientist>
        <station_data>
                <station_number>   </station_number>
                <date_time>            </date_time>
                <latitude>   </latitude>
                <longitude>    </longitude>
                <level_code>       </level_code>
                <profile>
                        <level_value>    </level_value>
                        <parameter_code>       </parameter_code>
                        <parameter_value>      </parameter_value>
                </profile>
        </station_data>
</cruise_data>
```

Ideas went to MEDS - Gagnon and Keeley

Canadian XML Work

- **Keeley Question? - Have you ever thought about bricks?**
  - **an atomic unit of data or metadata**
  - **A fundamental unit**
  - **A natural packaging**
  - **Consider "units" and "data point"**

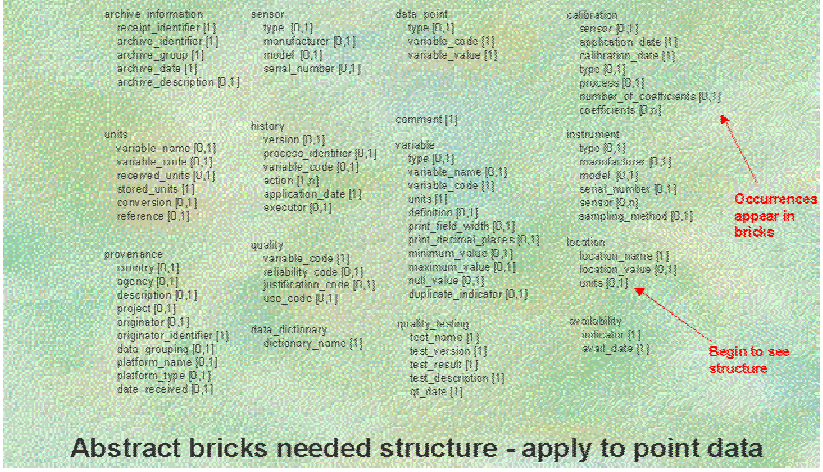
Canadian XML Work

units
- variable_name
- variable_code
- received_units
- stored_units
- conversion
- reference

data_point
- type
- variable_code
- variable_value

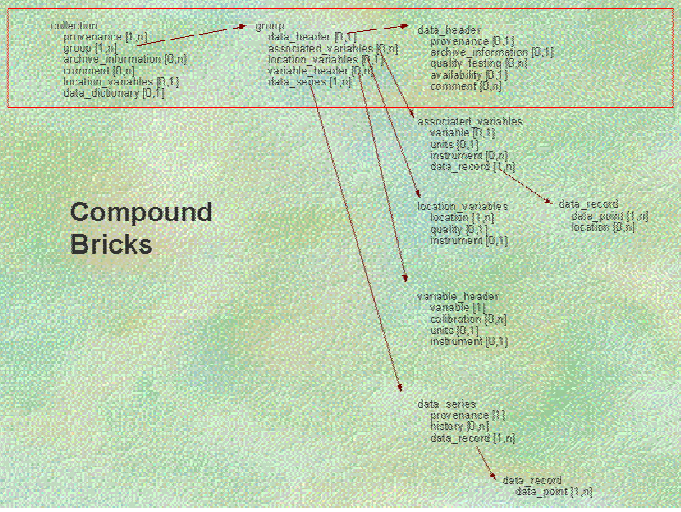
Canadian XML Work

15 of 22



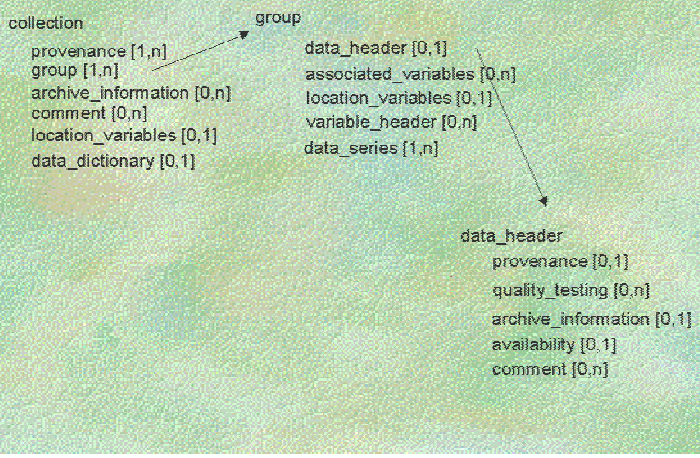
Abstract bricks needed structure - apply to point data

Canadian XML Work

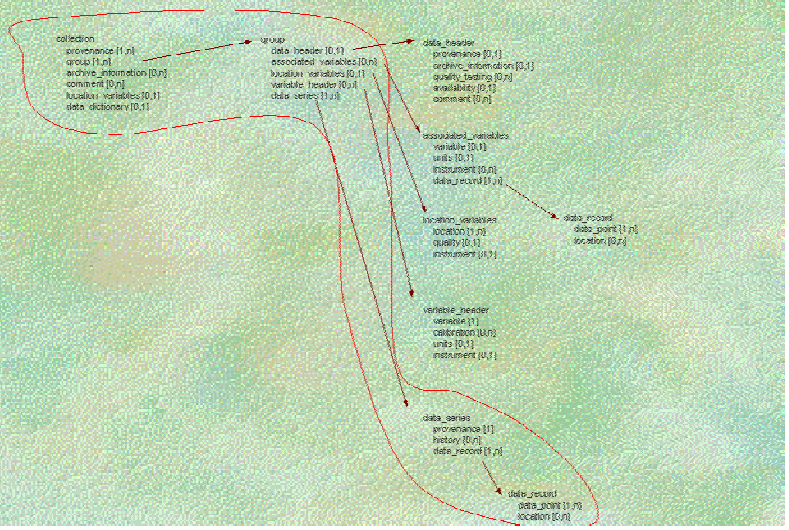


Compound Bricks

Canadian XML Work



Canadian XML Work



Canadian XML Work

```
<collection>
        <provenance>
                <originator_identifier>96006</Originator_identifier>
        </provenance>
        <group>
                <data_series>
                        <provenance>
                                <originator_identifier>1</Originator_identifier>
                        </provenance>
                        <data_record>
                                <data_point>
                                        <variable_code>TEMP</variable_code>
                                        <variable_value>2.3</variable_value>
                                </data_point>
                        </data_record>
                </data_series>
        </group>
</collection>
```

Canadian XML Work

- Generalize Keeley Bricks and XML?
  - Consider "Happy Birthday XML" (XML-Journal)
    - Asked 11 experts for the "gift"

⇒ Charles Goldfarb (The Father of XML)
use XML as a standardized means of modelling and representing information in all forms

⇒ Rick Jelliffe (Standards expert)
make local specs, not global

⇒ Jon Bosak (XML WG Leader [Sun Microsystems])
cooperation on vocabulary definition

Thank you.

ANNEX 5: GF3 MAPPING EXERCISE

Element Name	Sub elements	Description
series_header		
	country_code	
	country_name	
	institute_code	
	institute_name	
	creation_date	date and time are lumped
	data_centre_id	
	platform_type	
	specific_platform_code	
	platform_name	
	cruise_id	
	start_date	
	end_date	
	spatial_extent	probably a brick on its own (x,y)
	spatial_error	variation in position at the time of measurement – not the accuracy of the measurement
	bottom_depth	is there error on each element on x,y,z,t?
	minimum_observation_depth	
	maximum_observation_depth	if a point, set these to be the same
	originators_id	
	dictionary_reference	
	delimiter	
data_cycle_definition		
	parameter_code [1,n]	
data_cycle		
	data_record	text, with delimiters
spatial_extent		try to grab this from GML

ANNEX 6: 2002/2003 TERMS OF REFERENCE FOR SGXML

TOR 1) Create, evaluate and discuss intersessional work on SGXML parameter dictionary including the population of the dictionary for distribution via a defined XML structure.

The XML web distribution of the parameter dictionaries should be completed and the usefulness of the exercise for cross mapping of parameter dictionaries needs to be assessed. The applicability of the XML structure for other dictionaries should also be determined.

TOR 2) Evaluate and discuss intersessional work on point data structure. Evaluate the usefulness of the generalised Keeley brick approach with application to various point data types.

The generalised point data structure needs to be critically evaluated from the perspective of the international data centres. The applicability of the abstract Keeley bricks needs to be evaluated.

TOR 3) Report on the investigation into other available existing standards (e.g., geographers through the Open GIS consortium, taxonomy, ISO standards, metadata standards (MEDI, GFDC, EDMED, etc), utilising what has already been built.

The metadata problem is common to many organisations and considerable effort has been made by these other organisations. The usefulness of these efforts needs to be evaluated within the context of ocean data transfer.

TOR 4) Evaluate and discuss intersessional work on metadata. Evaluate the usefulness of linkages to other metadata standards and on the implications of a generalised metadata model to existing models.

Progress on the generalisation of the metadata model needs to be evaluated. The generalised model needs to be considered within the context of existing models.