Data Quality Working Group

In order to get an overview of the issues that have been discussed so far, I have summarized them a little and added some links and references. The interesting discussion is of course still going on and more and more people seem to be interested in these topics, but we have already gathered quite a lot of ideas. I think some topics need more input from the group, others could better be explored within smaller discussion or working groups.

We can use this short report as a base for further discussion - as you will see there are still some points that need a more general investigation, others might need concrete suggestions or a more focused discussion. If people are interested in working on a specific topic, please state this, others might join you and we could already concentrate on "producing" things. I am not indicating any directions or next steps here, it is up to you how to proceed.

Below you will find the topics which I have categorized into very broad categories, concerning geography, taxonomy, quality status of databases and some general issues.

Geography

Gazetteer

We have agreed that currently there is no useful gazetteer for marine locations available. Such a gazetteer should include all locations that are of interest for marine science. Surely some coastal (or adjacent) sites and locations should be included, such as capes, towns, islands, lagoons, river mouths or ports. These could be extracted from already existent terrestrial gazetteers and enriched with additional points.

Additionally there should be surely some underwater features included in the gazetteer, such as reefs, sandbanks, seamounts, abysses etc. However, such geographical features can barely be displayed as points due to their spatial extent. Maybe shapefiles with polygons or something like minimum/maximum extent of the feature can be a solution. There will always be an error or an uncertainty when referring to a "point" which has actually a certain extent itself (see presentations / papers by Arthur Chapman).

Another issue that has been raised was the use of different characters than those appearing in the Latin alphabet. Errors and different spellings might arise when non-Latin alphabets are being transcribed, causing inconsistencies. It might be helpful to include the "original" spelling, this in turn requires some technical modifications: databases have to support Unicode. There have been suggestions about solving this problem on a programming level rather than on the database level, but this would require an extra decoder / encoder on the client side.

Maps

The second geography-related problem concerns the absence of detailed maps. Generally there exist shapefiles for coastlines and bathymetry, but they are not detailed enough for many purposes. Also lagoons, coastal lakes and small islands are often not included in these maps.

There have been some suggestions on how to obtain or create those maps- we could either combine shapefiles that have been produced for special purposes by our institutions, but this might not be sufficient. For a start however those existent shapefiles might be helpful.

Another suggestion was to try and get shapefiles from official (governmental) agencies (e.g. shapefiles for the Water Framework Directive), this could be a start at least for the European Continent.

Standards

There was a suggestion of putting some emphasize on the utilisation of the *ISO 19115* standard also in the field of biodiversity and biogeography. We could extract relevant information from this standard and include it in a QA/QC "manual".

Climatology

During the workshop in Oostende Arthur had presented some methods of using environmental layers for modelling. For marine species we might have to distinguish between pelagic and benthic species and apply different models such as climatology or bathymetry, amongst others. There is a tool called *Aquamaps*[1], which models and displays species distributions, taking into account several parameters such as depth, SST, Salinity, Primary Production, Ice Edge Distance and Distance to Land. This tool might be a starting point to develop or adapt some methods to use environmental data for outlier checking.

Habitat Mapping

The idea of using a habitat classification as a tool for outlier checking is that species often occur in certain habitats such as sandy sediments or rocks and shouldn't occur in other habitats. However we have to check if the current attempts of defining habitat mapping standards are useful for our purposes. So far there are MESH[2] and

[1] http://filaman.ifm-geomar.de/tools/AquaMaps/HCMapSpeciesList.php

[2] http://www.searchmesh.net/

EUNIS[3] that are working on habitat classifications. We should check if we can help in developing such standards or modify them for our purposes. This is an area that requires some more investigation.

Species Distribution mapping

Another possibility of checking for outliers would be to check if a species occurs in an area from which it hadn't been reported so far. If a species endemic to the Atlantic Ocean it shouldn't occur in the Pacific Ocean. This requires a lot of work- compiling species distribution lists, and often knowledge about species is missing, but it could be achieved for some well known species and maybe gradually built for species in ERMS for example.

Similar is the idea that species often are restricted to a certain latitudal range. A tool that checks if a species that normally occurs around the equator shouldn't occur in the northern Atlantic, for example. Same as above, this requires knowledge about the species distribution and is only applicable for species that are indeed restricted to certain areas.

Species distribution mapping could also be applied for depth ranges. If a species is limited to benthos in shallow waters it should not occur in a deep-sea sample.

Outliers in time

Another possibility of outlier checking would be to include also the temporal aspect. Species get extinct from certain areas and get introduced to new areas. It would therefore be helpful to compile dates of first and last occurrence of a species in a certain area, if this information is available.

Taxonomy

Species Registers

There are various issues concerning taxonomy that have been raised during the discussion. The first requirement is the further development and extension of species checklists. The checklist of species available online should indicate the valid names and preferably all synonyms – valid names and synonyms clearly distinguished.
The checklists that are available (ERMS[4], Species2000[5], ITIS[6] etc.) are of course not complete yet. ERMS is for example lacking species from the Black Sea and

[3] http://eunis.eea.eu.int/index.jsp

Species2000 is until now missing many European species (but if I remember rightly ERMS will become a contributor to Species2000 soon). Consequently we will always encounter species that cannot be verified through any of those registers, and also the registers themselves contain errors both in spelling and in classification. There should therefore be attempts to expand the registers and quality-control them as well.

Tools for accessing Species Registers

As a next step there should to be tools to access these lists. The internet is a good source for verifying single species names, but often a whole list or even a database has to be verified. We need tools for interaction with those lists, that allow the submission of a list of names that are checked against a register and return all records that could not be found in the database. ITIS has already tools that allow this[7]. Another option would be to provide users with local copies of the registers. If these databases include some functionalities such as an update function that retrieves changed records from the central database, and tools to check the user's database against the register, users would certainly be encouraged to verify their species lists. An update function also ensures that always the latest version of the register is available locally. Classification changes, and often people check their databases once and from then on never again. Like this the will be outdated again after a few years.

Taxonomic Changes

Changes in taxonomy are another issue. It would certainly be helpful if the online systems provided also information about the classification changes that a taxon is undergoing. I have read somewhere that ITIS has developed a system of tracing classification changes but they haven't implemented it yet - this could be checked. Furthermore, sometimes different classifications for taxa exist, these could also be provided (if the system behind allows multiple classifications). The tools we intend to develop to assist in taxonomic quality control should not only check for spelling errors but also take into account classifications. Wrong classification into higher levels can influence biodiversity analyses as much as a wrong spelling of a species name.

Spelling / Sounding algorithms

The next point on the list is the validation of taxonomic names through tools that make use of sounding analyses, spelling similarities or fuzzy matching. There is a very interesting analysis of algorithms and tools described by Eduardo Dalcin in his

[4] http://www.marbef.org/data/erms.php
[5] http://www.species2000.org
[6] http://www.itis.usda.gov/
[7] http://www.itis.usda.gov/taxmatch_ftp.html

PhD thesis which is certainly worth reading[8]. We have discussed about algorithms like Soundex which compares names with similar soundings and like this can find spelling variations of the same word. Soundex was however developed to find similar names of people in the US. It might not work that well with taxonomic names which are of Latin or Greek origin and do not have a unique pronunciation. There are other improved algorithms such as *Phonix*; and there are algorithms that check for spelling similarities. ERMS employs a *Fuzzy Matching* method that checks for common spelling errors (such as for example the interchange of *s* and *c* or *t* and *th* etc.). All these options are worth exploring and quality control tools using them could surely help to reduce errors in databases.

Incomplete classifications

One more issue that comes across when dealing with taxonomic databases or species lists is the "incomplete", "uncertain" or new" identification of species that might look like *Abra cf. alba, Abra sp.1, Abra aff. alba, Abra nov. alba, Abra alba sp. nov., Abra sp. nov, Abra alba?, Abra ?alba, ?Abra alba* and so on. These cases have to be handled differently- some can be matched to a valid species, some only to a valid genus or even family. There was a suggestion to make this a separate section in a QA/QC "manual": how to deal with incomplete identifications.

Quality status indication in databases

One part of the discussion was dedicated to the issue of indicating the quality status of database records. Taxonomic data can be verified through different sources. One suggestion was a schema used at the plankton department of IBSS, Ukraine. They define different sources of taxonomic data, guide-books for the considered area, papers, monographs, etc., not testable data: protocols of samples analysis made by "unavailable" experts (not working any more) and testable data: protocols of samples analysis made by "available" experts (working here and now). Of these four categories a Quality Index is formed, indicating the reliability of the record. This schema also takes into account the number of different authors that report a taxon from an area. Thus, a record is less credible if occurs in three of the four categories but have the same author as a source all the time.

Another suggestion is to form categories of publications that are assigned a "quality status"- ranking primary sources such as taxonomic description in peer-reviewed journals highest and secondary sources such as field guides rather low. This can be refined by giving different weight to the authors of the publications, the finest step would be to assign the quality status directly to the author. Like this it can be easily checked by whom a species was reported, in what kind of publication, at what time

[8] Dalcin, E. Data Quality Concepts and Techniques Applied to Taxonomic Databases (2004). PhD Thesis, University of Southampton (http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf)

and how often. However, these attempts require much expertise that might go beyond that of data managers. When building such a system of quality rating we should take care that it remains "usable" for non-taxonomists.

Other issues

"QA / QC manual"

The idea came up of creating a "users guide" for quality control issues. Such a manual should not include long theoretical essays about data quality but have rather the form of a checklist of quality control issues that a data manager can use either as a reference to "tick off" the issues that should be considered when one wants to improve quality. Such a manual could include a list of plausibility checks for a database, a list of tools available on the internet (such as for example the *Georeferencing Calculator*[9] or the *spOutlier*[10] tool) or as free software. Also this manual can include short explanations about quality issues, standards, ongoing discussions or projects that deal with data quality.

Audit logs, validation logs and Documentation

An issue that has been mentioned a few times during the workshop is the proper documentation of databases, validation actions and the general maintenance of an audit log to register changes in a database. These issues are too often neglected even by data managers but can really help to improve quality. A proper documentation of a database can help to prevent entering data into wrong fields; a proper definition of allowed values for a field can prevent wrong values right at data entry. Validation checks and other actions in the database should always be documented to trace who modified data and for what reason. Like this, credibility of a record can be assessed. Setting up an automated audit trail can be done almost automatically in more sophisticated database management systems but can require lots of time and programming in less advanced systems. But even for systems like Access an audit trail is realizable. Such a code could also be made available for public in order to encourage people to document their changes.

Modification of existing tools from other fields of science

In terms of data quality we are not alone. Many other fields of science have already well functioning mechanisms and tools for quality control and quality assessment. We should therefore check if we can use or modify existing tools or methods from areas such as other oceanographic sciences (physical and chemical oceanography), terrestrial biology or other data managing disciplines, such as medicine, economy, or other natural sciences. This needs some research and hasn't been discussed so far

[9] http://elib.cs.berkeley.edu/manis/gc.html
[10] http://splink.cria.org.br/outlier?criaLANG=en

during the e-mail conversation but would certainly be helpful in order not to re-invent the wheel.