A data manager's guide to marine taxonomic code lists

M. Kennedy and L. Bajona

Fisheries and Oceans Canada
Bedford Institute of Oceanography
Science Branch
Ecosystem Research Division
P.O. Box 1006
Dartmouth, Nova Scotia
B2Y 4A2

2009

Canadian Technical Report of Fisheries and Aquatic Sciences 2827

Fisheries and Oceans Canada | Pêches et Océans Canada

Canada

Canadian Technical Report of
Fisheries and Aquatic Sciences 2827

2009

A DATA MANAGER'S GUIDE TO MARINE TAXONOMIC CODE LISTS

by

Mary Kennedy and Lenore Bajona

Fisheries and Oceans Canada
Bedford Institute of Oceanography
Science Branch
Ecosystem Research Division
Dartmouth, NS
B2Y 4A2

Cat. No. Fs 97-6/2827E-PDF ISSN 0706-6457

Published by:
Fisheries and Oceans Canada
1 Challenger Drive
Dartmouth, Nova Scotia
B2Y 4A2

Correct citation for this publication:

Kennedy, Mary and Lenore Bajona. 2009. A data manager's guide to marine taxonomic code lists. Can. Tech. Rep. Fish. Aquat. Sci. 2827: iii + 23 p.

TABLE OF CONTENTS

List of Tables iv
List of Figures iv
Abstract 1
Résumé 1
Introduction – The Need for a Standard 3
Linking Species Lists 4
Proposal #1: Link multiple collections based upon scientific name and authorship 5
Proposal #2: Map individual collection species lists to a Master Standard Code List 6
Benefits of Linkage to a Common Master Standard Code List 7
Complications of Linkage to a Master Species Code List 9
Registers of Species 10
Quality Control of a Species List through the Utilization of Standard Master Lists and Registers of Species 12
BioChem : A Case Example 12
BioChem 13
Map and Assign Standard Codes to Species List Names 14
Summary 20
Appendix 1: Negative TSN Code Series 21
References 22

LIST OF TABLES

Table 1. A dataset species list often contains names of taxa not currently in the dataset. The numbers presented in this table are based upon database queries performed in March 2008 4
Table 2. Scientific name must include the authorship or else major mismatches may occur. The following table shows three Integrated Taxonomic Information System (ITIS) records with identical scientific names. 5
Table 3: Article 58 of the ICZN states that variant spellings of species-group names are deemed to be identical and lists fifteen accepted spelling variations. 6
Table 4. Pros and cons to using ITIS as the master standard code list. 7

LIST OF FIGURES

Figure 1. Sample locations for two DFO biological data collections. The map on the left show the geographical distribution of a subset of OBIS Canada collections and on the right a subset of BioChem database collections 3
Figure 2. IODE GE-BICH diagram depicting the relationship between various international initiatives 11
Figure 3: Flow chart showing the various tables used during the coding procedure. 13
Figure 4: Map showing a subset of BioChem sample locations. 14
Figure 5. A flow diagram showing the various pathways to follow upon receipt of a new species list in the attempt to assign a TSN. This entire procedure requires long term funding and support for local registers of species and for master lists such as ITIS. 19

ABSTRACT

Kennedy, Mary and Lenore Bajona. 2009. A data manager's guide to marine taxonomic code lists. Can. Tech. Rep. Fish. Aquat. Sci. 2827: iv + 23 p.

A comprehensive taxonomic list is required to efficiently share and integrate biological data by organism names. Beyond the frequent human error introduction of mixed cases and typos, there is the common occurrence of multiple names for the same organism (synonyms) and the same name applied to many different organisms (homonyms). A given dataset may refer to an old name that has been updated by the taxonomic experts and may also have a separate entry for the new currently accepted name. Users accessing the data may not be aware of the multiple names thus may only obtain a subset of the data they were looking for and likely need. Linking multiple datasets only increases the chances of missing relevant data. Sharing biological data over the web necessitates a decision on standardization of organism names. This report suggests methods to standardize taxonomic lists and develop species registers to provide quality control.

RÉSUMÉ

Kennedy, Mary, et Lenore Bajona. 2009. A data manager's guide to marine taxonomic code lists. Can. Tech. Rep. Fish. Aquat. Sci. 2927 : iv + 23 p.

Nous devons disposer d'une liste taxonomique exhaustive si nous voulons être en mesure de partager les données biologiques et de les associer correctement aux noms des organismes. Outre les erreurs fréquentes occasionnées par les fautes de casse et de frappe, il arrive souvent que plusieurs noms (synonymes) désignent un même organisme et qu'un même nom (homonyme) désigne plusieurs organismes très différents les uns des autres. Par ailleurs, un même ensemble de données peut être lié à un ancien nom qui a été changé. Après une mise à jour par les spécialistes en taxonomie, cet ensemble peut faire l'objet d'une rubrique distincte sous le nouveau nom. Les utilisateurs qui accèdent aux données peuvent ne pas connaître l'existence des noms multiples associés à une espèce. Si c'est le cas, ils risquent de n'obtenir qu'un sous-ensemble des données qu'ils cherchent et, par conséquent, de passer à côté des données dont ils ont besoin. Si on se contente de créer des liens entre des ensembles de données de sources multiples, on ne fait qu'augmenter le risque de passer à côté de données pertinentes. Le partage par Internet des données biologiques nécessite une normalisation des noms des organismes. Le présent rapport propose des méthodes pour la normalisation des listes taxonomiques et pour l'établissement de registres d'espèces, dans le but de permettre un contrôle de la qualité.

PREFACE

In 2006/07 the Canadian Department of Fisheries and Oceans National Science Data Management Committee (DFO-NSDMC) funded a project to improve DFO taxonomic data in the BioChem plankton database archive. The taxonomic code table in BioChem contained one field that was to contain a value to enable linkage to a standards database which at the current time is the Integrated Taxonomic Information System (ITIS). In order to perform this direct match these values must be positive. However, many of the values in BioChem had been assigned a negative value when a matching name was not found in ITIS. This project's aim was to standardize and document procedures to answer the question "What do we do when the Integrated Taxonomic Information System database does not contain an identified taxon from a data collection?"

In February of 2007 a meeting was held to discuss this project status and to exchange ideas between Canadian east and west coast taxonomists. This document contains the protocols suggested/adopted at this meeting and during ensuing discussions amongst data managers at the Bedford Institute of Oceanography. Although the procedures created during the initial project were aimed strictly at solving problems in one database, BioChem, the principles applied have since been adapted for use by other databases. These improved procedures have been documented and are the subject of this report.

A poster on "The need for a standard" was presented at the Ocean Biodiversity Informatics Conference held at the Bedford Institute of Oceanography in October of 2007 (DFO. 2008).

Acknowledgements
DFO-NSDMC for funding portions of this project; OBIS Canada; Claude Guay at Integrated Science Data Management (ISDM) for his assistance modifying the BioChem code table design; Guy Baillergeron at ITIS Canada; Moira Galbraith and Steve Romaine at the Institute of Ocean Sciences (IOS) for their feedback and all the suppliers of data to BioChem especially Isabelle St.Pierre and Caroline Lafleur at the Institut Maurice-Lamontagne (IML). We would also like to thank Laure Devine for her helpful comments and direction.

INTRODUCTION – THE NEED FOR A STANDARD

It is commonplace to have large biological databases that lump together datasets from many different collections. Plus, the ability now exists to link to additional sources of data via the internet. This conglomeration of data collections will make comparison of datasets frustrating unless decisions relating to standardization of the taxonomic naming conventions are initially adopted.

Data managers at the Bedford Institute of Oceanography (BIO) maintain many marine biological databases and contribute data collections to the Department of Fisheries and Oceans (DFO) regional and national archives, such as BioChem, as well as export subsets of these collections to the Ocean Biogeographic Information System (OBIS) via Canada's regional OBIS node. These data collections include a wide range of biological material ranging from bacteria to whales. The source of these collections is primarily from DFO programs but these data are supplemented by additional data collected in the Canadian region by other research programs. New and legacy data collections are being appended to these rapidly growing databases. Sampling gears used to catch specimens may range from Niskin bottles and plankton nets to benthic grabs and fishery trawls. Specimens caught span the entire marine water column – benthic, planktonic, mesopelagic, nektonic organisms, etc. The location of the sampling is also broad. Samples may be from inshore to offshore and not just local waters but also from Arctic, mid Atlantic, Caribbean, etc. (see Figure 1).

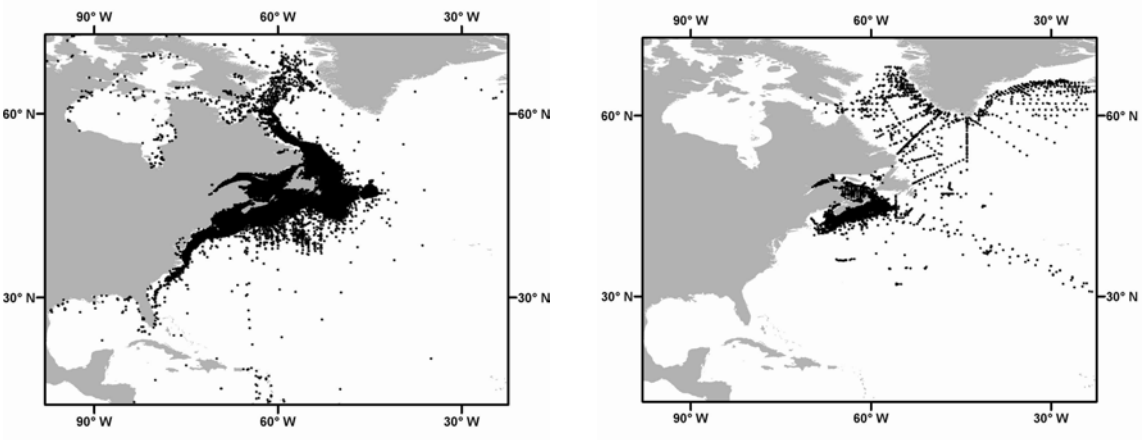

Figure 1. Sample locations for two DFO biological data collections. The map on the left show the geographical distribution of a subset of OBIS Canada collections and on the right a subset of BioChem database collections

The data contained in each of these datasets may vary widely. One dataset may contain catch statistics for commercial fish while another contains counts of cells per litre of water for phytoplankton. The common thread is that each dataset will also contain a list of biological names associated with the data. These lists of names may contain common names, taxonomic names at various ranks other than species, informal "convenience grouping" amalgamating distinct taxa, groupings of taxonomic names, in addition to valid species scientific names. Hereafter, these lists will be collectively referred to as the dataset "species list".

It is often the case that individual datasets have their own species list. These lists may contain more names than actually found in the dataset. Example, in March 2008, the DFO Industry Surveys Database (ISDB) species list contained 2204 entries, however the database only contained information associated with 835 of these names. The Northwest Atlantic Fisheries Organization (NAFO) database uses the Food and Agriculture Organization of the United Nations (FAO FIGIS) code list. This latter list contains more than 10,650 entries of which only 174 are utilized by NAFO and of these only 161 exist in the DFO collection. These counts as well as those for the DFO Maritimes Research Vessel Trawl Surveys (RV) collection and the East Coast North America Strategic Assessment Groundfish Atlas (ECNASAP) are listed in Table 1.

Dataset	Number of data records in dataset	Number of names in dataset species list	Number of names with data records in dataset
NAFO (uses FAO Codes)	1,025,027	174 (10,650)	161
ISDB	3,174,629	2,204	835
RV	155,239	2,248	472
BioChem	775,000+	3,884	3,884
ECNASAP	471,798	274	274

Table 1. A dataset species list often contains names of taxa not currently in the dataset. The numbers presented in this table are based upon database queries performed in March 2008.

Combining or linking data collections requires that the individual species lists be matched or "mapped". The question arises *"Is it sufficient to only map the subset of names that are associated with data?"* The answer is that it is advisable to initially invest the time to map the entire list to an adopted common standard list. A method to map the list will be outlined below. Once mapped to a common standard these lists can be easily linked.

LINKING SPECIES LISTS

There are two methods to link multiple species lists. The first method (Proposal #1) involves mapping names from one list to names in the second list. The second method (Proposal #2) involves "codes". The names in the individual species lists are mapped to a standard species code and then the lists are linked via the species code. The pros and cons to both approaches will be outlined below.

In biological databases that span multiple kingdoms one must expect homonyms (the same name used for different organisms or groups of organisms). According to Article 1.1.1 of the INTERNATIONAL CODE OF ZOOLOGICAL NOMENCLATURE online "Zoological nomenclature is independent of other systems of nomenclature in that the name of an animal taxon is not to be rejected merely because it is identical with the name of a taxon that is not animal".

One detail relevant to both mapping approaches is that many lists do not include the 'scientific authorship' associated with the scientific name. Blind mapping of scientific names without taking into account the authorship is not advised because of homonyms. For instance, one could easily map a taxon from one kingdom to the same name string in another kingdom or map a species to an invalid name (see Table 2).

Scientific Name	ITIS.TSN	Group
Animalia - Holopedium Zaddach, 1855 – valid	83956	Crustacean
Monera - Holopedium Lagerheim, 1883 – valid	818	Blue-green algae
Animalia - Phycinae – valid	131752	Insect – (flies)
Animalia - Phycinae – valid – phycine hakes	555704	Fish – (hake)
Animalia - Digenea Carus, 1863 – valid	55189	Flatworm
Animalia - Digenea Carus, 1863 – invalid	185496	
Plantae - Digenea C. Agardh – accepted	183223	Red algae

Table 2. Scientific name must include the authorship or else major mismatches may occur. The following table shows three Integrated Taxonomic Information System (ITIS) records with identical scientific names.

Historically, collections were caught and analysed for one particular scientist. This scientist knew his taxonomic group and did not need to include the authorship information in his species list. A common and generally limited set of reference books were used when the samples were analyzed. Now, problems arise when we try to combine data collections from different scientists who used different reference books. The best approach is to consult the original owner of the dataset and obtain their list of references then populate the missing "author" field. However, for many legacy data collections, this might not be possible, and if this field is populated by someone other than the original collector, this fact should be included with the metadata associated with the dataset.

PROPOSAL #1: LINK MULTIPLE COLLECTIONS BASED UPON SCIENTIFIC NAME AND AUTHORSHIP

In theory, one should be able to combine data collections by simply sorting the list on taxonomic name and authorship. In reality, unless one is combining collections from one source it is unlikely that this will be a simple task.

A few of the reasons why matching by name and authorship might fail include:

- Use of common names and/or ranks other than species instead of scientific names
- Different synonyms for same taxon
- Spelling variations such as *i* vs *ii* – see Table 3 (although spelling differences may be valid, they cannot be mapped if spelled differently)
- Variation in handling of diacritical marks (accents) in authorship: oe vs ö
- Various languages: 'and' (English) vs 'et' (French)
- Use of initials, abbreviations, parentheses, commas and spaces in authorship
- Authorship year variations
- Inconsistency in the use of terms such as 'variety', 'form' and 'subspecies'
- Grouping of taxa in one list vs individual species in another
- Grouping of several taxa together
- Lack of authorship

Extract from Article 58 International Code of Zoological Nomenclature (ICZN) (http://www.iczn.org/iczn/index.jsp.)

58.1. use of *ae oe* or *e* (e.g. *caeruleus, Coeruleus, ceruleus);*

58.2. use of *ei i* or *y* (e.g. *cheiropus Chiropus chyropus*);

58.3. use of *i* or *j* for the same Latin letter (e.g. *iavanus, javanus*; *maior, major*);

58.4. use of *u* or *v* for the same Latin letter (e.g. *neural, nevra*; *miluina, milvina*);

58.5. use of *c* or *k* for the same letter (e.g. *microdon, mikrodon*);

58.6. aspiration or non-aspiration of a consonant (e.g. *oxyrhynchus, oxyrynchus*);

58.7. use of a single or double consonant (e.g. *litoralis, littoralis*);

58.8. presence or absence of *c* before *t* (e.g. *auctumnalis, autumnalis*);

58.9. use of *f* or *ph* (e.g. *sulfurous, sulphureus*);

58.10. use of *ch* or *c* (e.g. *chloropterus, cloropterus*);

58.11. use of *th* or *t* (e.g. *thiara, tiara*; *clathratus, clatratus*);

58.12. use of different connecting vowels in compound words (e.g. *nigricinctus, nigrocinctus*);

58.13. transcription of the semivowel *i* as *y*, *ei*, *ej* or *ij* (e.g. *guianensis*, *guyanensis*)

58.14. use of *-i* or *-ii*, *-ae* or *-iae*, *-orum* or *-iorum*, *-arum* or *-iarum* as the ending in a genitive based on the name of a person or persons, or a place, host or other entity associated with the taxon, or between the elements of a compound species-group name (e.g. *smithi*, *smithii*; *patchae*, *patchiae*; *fasciventris*, *fasciiventris*);

58.15. presence or absence of *-i before a suffix or termination (e.g. timorensis, timoriensis; comstockana, comstockiana).*

Table 3: Article 58 of the ICZN states that variant spellings of species-group names are deemed to be identical and lists fifteen accepted spelling variations.

The authors feel that, because of the high likelihood of non-matches, the next proposal is preferable.

PROPOSAL #2: MAP INDIVIDUAL COLLECTION SPECIES LISTS TO A MASTER STANDARD CODE LIST

In theory this method should also be easy. Initially time would have to be invested to map the individual dataset species lists to the master list to obtain the standard code such as the ITIS TSN but once mapped, these lists could then be matched to other lists via their standard codes. The individual lists could retain their own spellings, language, etc.

A few of the reasons why matching by name and code might fail include:

- Requires adoption of a master standard code list
- Requires support and maintenance of master code list
- Requires procedures to handle names that are not in the master code list

The authors wish to elaborate on the benefits of this second proposal – linking via a unique numeric code identifier. Most of the species lists associated with DFO Maritimes datasets include a field for this code. Data managers in charge of these datasets have collectively adopted the Integrated Taxonomic Information System (ITIS) as the "standard" and the ITIS Taxonomic Serial Number (TSN) field as the "standard code".

ITIS is a leading international source for authoritative taxonomic information. Developed jointly by federal agencies in United States, Canada and Mexico, ITIS provides a central source (mirrored in several locations) to verify scientific names and to obtain full taxonomic classification for marine and terrestrial organisms (ITIS^{CA} 2007). The taxonomic database maintained by ITIS contains authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world. ITIS is usually updated on a monthly basis. As of March 14, 2008, ITIS contains 467,638 scientific names and 109,072 common names. (ITIS 2008)

BENEFITS OF LINKAGE TO A COMMON MASTER STANDARD CODE LIST

Utilizing the suggested method of mapping dataset species lists to an adopted master list in order to link lists has the added benefit that by adopting the master list names the taxonomic nomenclature can become standardized. Linkage to valid names and hierarchy is also available through the master list. This would solve the quality control issues which arise when large biological oceanographic databases and portals are queried to combine collections from multiple sources. Users of this data need to feel confident that:

a) The analysts assigned the correct name
b) The name is spelled correctly
c) The name is the most recent commonly accepted name.

Is ITIS the solution to sharing and linking marine biological data? There would be pros and cons to using any existing database as a standard and ITIS is not an exception (see Table 4). However, if the international community were to throw their support behind the concept of a standard exchange code many of the cons listed in Table 4 could be resolved.

Cons	Pros
Strong North American focus in content and spelling (although ITIS has world coverage for many taxonomic groups)	Standardized spelling of scientific name and authorship
Many records out of date and require revision/verification	Linkage to parent record/taxonomic hierarchy
Many marine taxa missing	Linkage to synonymy
Time delay when new names forwarded for addition to ITIS database	Linkage to "valid/accepted" name
Will not enable linkage for non taxonomic entries in species lists	Responsibility to update taxonomic changes (= maintenance of master list) and
	standardization of taxonomic nomenclature (ICZN)

Table 4. Pros and cons to using ITIS as the master standard code list.

Linkage to ITIS can provide more than standardized spelling of scientific name and authorship (Branton et al, 2007). Given the standard code, the ITIS TSN, it is also possible to extract enriched metadata from ITIS, i.e. the currently accepted scientific name, author, rank, synonym taxa, validity and hierarchy. This information does not have to be stored and maintained in the source database.

Datasets managed by DFO Maritimes span the period from the early 1900s to the present. Additional legacy data collections are being recovered and appended to these datasets. The fact that data sampling spans the last century presents an additional taxonomic nomenclature challenge. Classifications may have changed over this period! The issues of taxonomic changes (i.e. which species name is valid and which name isn't and which species name is a synonym of another name) are daunting to a data manager in charge of maintaining a local species list and who is most likely not a taxonomist. Linking to a standard master code enables the data manager to retain the original name and link to the currently accepted name. If the validity for a name changes then, theoretically, this will be updated in the master code table and will not need to be tracked by the originator. Leave the task of taxonomic updates to the managers of the "master standard list", i.e. ITIS. The master standard list will always be in a state of flux. It will require constant maintenance – it will always require updating of existing records and will require means to append new species.

The ITIS code is gaining acceptance as a leading international source for authoritative metadata and one of its database fields, the TSN, has been accepted as an international exchange code (IOC IODE GE-BICH-III/3. 2006). If huge data providers such as OBIS include the ITIS TSN field as one of their required data elements then data management will be facilitated. Linkage to taxonomic hierarchy and validity will be standardized. Data providers would be able to perform searches for specific organisms which would find and link the data for all synonym names for the specified organism. Data providers could perform more robust searches (both advanced and by taxon groups) from which users could be more confident that all data for the specified organism(s)/group has been obtained.

The initial task of mapping species lists to ITIS will not be easy and it will require collaboration between taxonomic experts, database and ITIS data managers. International adoption of the concept of an standard 'exchange' code is required. Once adopted, support must be provided to the managers of the master standard list if this list is to remain functional. These tasks are not trivial, but once done, future maintenance and querying of data would be easier and more accurate.

COMPLICATIONS OF LINKAGE TO A MASTER SPECIES CODE LIST

What happens when a species list name is not in the master list? One procedure is to assign a negative code to the standard code field. This method of code assignment was briefly described in a previous paper (Branton et al 2007) and will be discussed in more detail below.

Many data collections have species lists that contain a variety of names. These names may be:

1. Scientific names with authorship
2. Scientific names without authorship or variances in years
3. Common names
4. Non-taxonomic names (i.e. non-living, picoplankton, mesoplankton, nekton, etc)
5. Groupings of valid names

Category 1
Mapping of scientific names with authorship to a master list such as ITIS should be straight forward. However, the master list may not always be up to date and or may not contain an older invalid synonym if the valid name is in the database. How should these cases be handled? How can the names be appended to the master list?

Category 2
Mapping of scientific names without authorship should not be automatic. As mentioned above, homonyms do exist and a data manager should consult with a taxonomist prior to mapping a name without authorship to a name in the master list that does have an associated authorship. Similar consultation should occur prior to mapping names with authorship to names in the master list that do not currently have an associated authorship. Where should the fact that the names were forced matched be recorded? Year variances of one or two years can be assumed to be the same authorship.

Categories 3 and 4
Common names and other non-taxonomic names such as phytoplankton, jellyfish, krill, plastic, rocks, shells, etc cannot be directly mapped to a taxonomic standard such as ITIS. Another common example would be H4B and CYT. These terms refers to fish eggs from the groups hake/four-bearded rockling and cunner/yellowtail and are valid taxonomic identifications (Colton and Marak, 1969). How should these names be treated?

Category 5
Groupings of valid names require special handling. These names may individually have entries in the master list but the grouping in the species list is not taxonomically comprehensive in nature. Two species may have been grouped together but this group does not include all species within the genus level, eg. *Paracalanus*/*Clausocalanus*.

One way to answer these questions is to assign negative TSN value to any name not in the original master list, i.e. ITIS. The authors wish to manage these negative codes used by DFO Maritimes through the creation of a master Negative code table and would like to

suggest that this method be adopted by others. This table will be referred in the following as the Standard MasterNegativeTSNs list.

This MasterNegativeTSNs table would include the following fields:

- TSN (negative code value)
- name
- authority
- Reference information (including publication and page #)
- Comments
- MasterList_Comments (comments from ITIS re inclusion in master list)
- TSN_ITIS

The new field called 'TSN_ITIS' would include a link to the MasterList. Several examples are listed below to clarify the meaning of the TSN_ITIS.

- If a 'negative record name' was a group of two species then the TSN_ITIS would be the genus TSN value. (similar pattern of using next level if grouping at other taxonomic ranks)
- If the negative record name was krill then the TSN_ITIS would point to Euphausiacea.
- If the negative record name was a group of jellyfish, ctenophores and salps then the next highest common rank would be the kingdom Animalia and this TSN would be assigned.
- If the negative record name contained taxa from multiple kingdoms then the TSN would be set to 0 to indicate 'living'.
- If the negative record was plastic or rocks then the link could remain null to indicate that the name is 'non-living'.

Names that fall into Category 1 but not yet in the ITIS tables would follow the above procedure and a negative code would be assigned. The next step would be to forward the new name and any required metadata to the managers of the MasterList (ITIS) with the request that this name be considered for inclusion into the master list. If this name is accepted and appended to the master list then the record TSN_ITIS will be assigned the new non-negative code value for this name. This name may remain in the Master Negative TSN table for historical tracking purposes.

The assignment of negative TSN values is not a new concept. DFO Maritimes follows the basic prefixes assignment described by the World Ocean Database (WOD) (Appendix 1). .

REGISTERS OF SPECIES

A marine species register is more than a simple compilation of names from local collections. Compilation also requires an exhaustive search of the literature, the internet, sample collections, etc. Plus, names in the register must be verified by taxonomic experts. In theory, all species listed in an accepted register should be in the master species list.

The authors of this paper are DFO Maritimes data managers and as such are primarily interested in oceanic data collections and registers of marine species (RMS). DFO is in the process of establishing Canadian registers of marine species for our three oceans, the Atlantic, the Pacific and the Arctic oceans. Plans may expand to cover inland waters as well. These registers once created will feed into existing parent registers such as World Register of Marine Species (WoRMS).

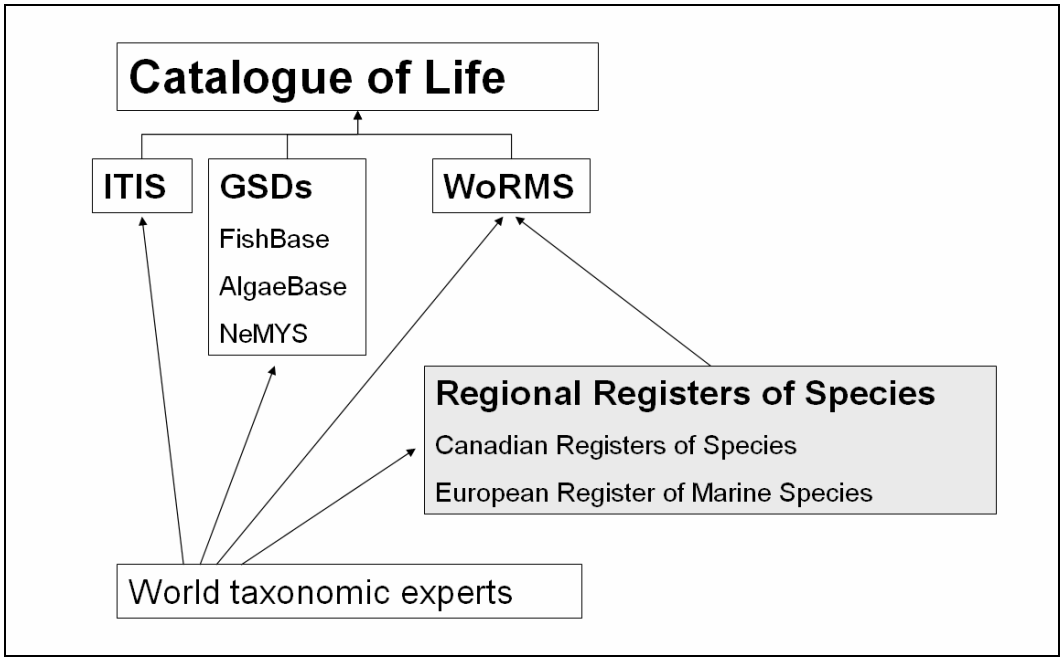

Figure 2. Diagram depicting the relationship between various international initiatives.

The aim of the WoRMS is to provide an authoritative list of names of all published marine species globally. Many of the Global Species Databases (GSDs) now maintained within the WoRMS system will automatically be contributed to the Catalogue of Life. The Catalogue of life is a collaborative venture between Species 2000 and ITIS, and is regarded by many as the prime supplier of information on taxonomy. WoRMS will serve as the taxonomic backbone of OBIS. Figure 2 above is a sketch showing the relationships between various international initiatives. This figure was redrawn from an IODE-GEBICH (Intergovernmental Oceanographic Data Exchange Group of Experts on Biological and Chemical Data management and Exchange Practices) informal presentation.

The comprehensive lists of regional species names, i.e. local registers, are required for biodiversity studies and presence verifications. Species registers should become the accepted regional source of verified marine taxonomic names and the taxonomic authority for names not in the master species list. These lists could be used to:

- determine if new species are being introduced
- determine if species are becoming extinct
- assist in taxonomic identifications
- assist in species mapping applications

Large collections, such as BioChem, may contain datasets gathered from many different sources. The level of expertise of the person performing the original identification could

range from student to an internationally renowned taxonomist. This information is rarely available or preserved.

Accepting the fact that we can't always check the analyst's credentials, perhaps we can verify that the identification is reasonable, i.e. verify that the taxon identified is known to occur in the area sampled. How do we do this?

In addition, a one to one match between a species list and the adopted standard master lists enables correction of spelling errors and will link older unaccepted names to newer valid names but this linkage does not authenticate the original analyst's identification. Is there another method that can be employed to aid in the quality control of our taxonomic nomenclature assignment?

One method accounting for both these problems is a register of species. The basic concept is to compare a dataset species list to a comprehensive list of all known species names from the sampling area. If the names in the species list are not contained in the local register then the original identification should be questioned. Perhaps an error was made, a new species has entered the local area, or a new species has been identified and the register should be updated.

QUALITY CONTROL OF A SPECIES LIST THROUGH THE UTILIZATION OF STANDARD MASTER LISTS AND REGISTERS OF SPECIES

BIOCHEM : A CASE EXAMPLE

The procedures required to quality control a species list can be divided into two basic steps:

1. Map the species list to standard codes and *clean* (revise) the species list's existing scientific names and authorities
2. Verify species presence in the species register that covers the sampling area and update the local marine species register.

The first procedure outlined below will describe how to append a **dataset species list** to the **collection species list** and how to assign the **master list standard code** values (Figure 3). In this example the Integrated Taxonomic Information System (ITIS) was adopted as and will be referred to as the **master list**. The **standard code** used to map the lists will be the ITIS TSN field. The dataset mapped to the master lists was a large marine plankton archive, BioChem. The method described below, although specific to one particular database, could easily be adapted for use with other marine collection species lists.

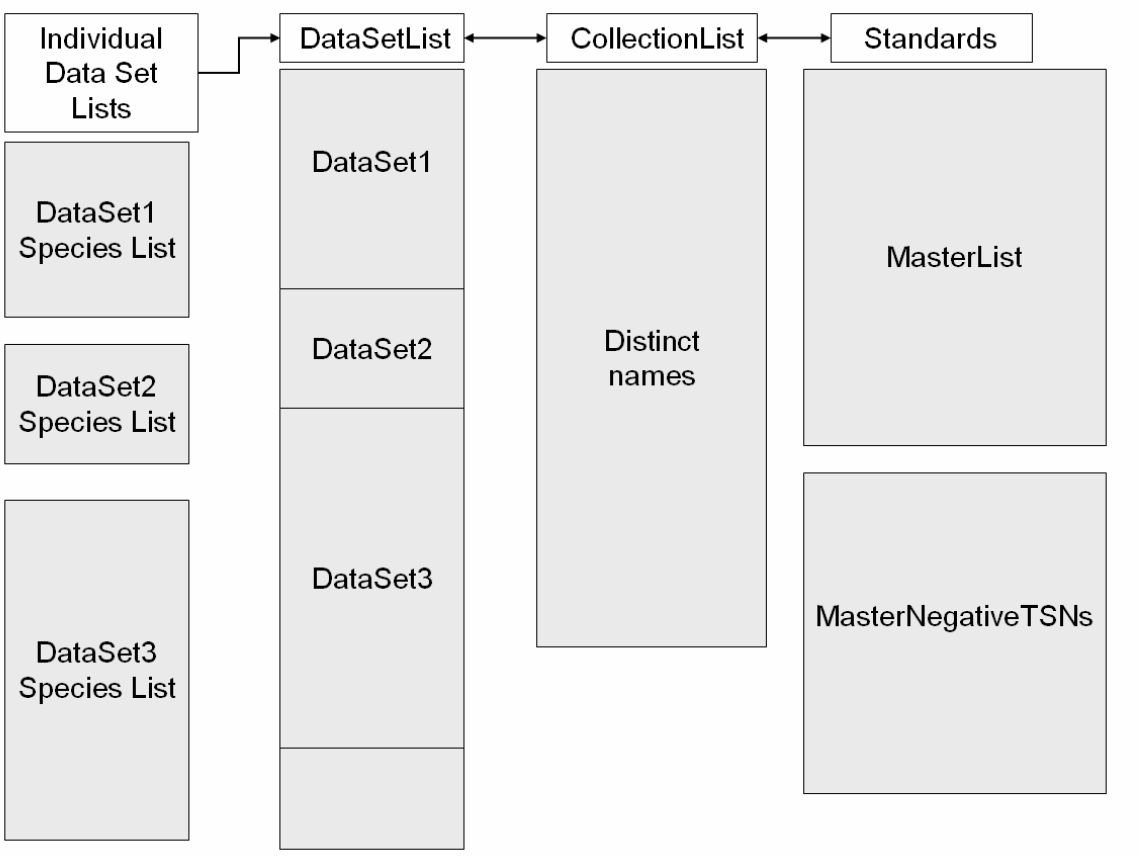

Figure 3: Flow chart showing the various tables used during the coding procedure.

The procedure described below is simplistic in nature – map a species list to the standard TSN codes and presto the list is "cleaned" – we now can link to the correct spelling and most recent taxonomic information. However, in practice the procedure is more complicated:

- Many existing species lists contain only the scientific name and do not include the scientific authority citation.
- ITIS, if adopted as the master list, is incomplete and many taxa in its list are currently under review.
- A one to one match between a species list and the ITIS database does not authenticate the original analyst's identification.
- An additional coding system is required for collection names that are non-taxonomic and not included in the master species list
- Many geographical areas do not currently have their own local species registers.

The procedures developed to handle these obstacles for the plankton archive database, BioChem, are described below.

BIOCHEM

BioChem is a national Department of Fisheries and Oceans (DFO) archive for plankton and chemical data (Gregory and Narayanan 2003). BioChem is a free access DFO national application Oracle database. Its purpose is as a repository for biological (Bio) and

chemical (Chem) marine environmental sample measurements. Scientific research missions originating from the various DFO research institutions are the primary source of information in the holdings.

BioChem's archived data cover Canada's 3 oceans, the Atlantic, the Pacific and the Arctic (Figure 4). This ever growing collection of datasets currently includes records that span the period from 1921 to the present (BioChem DFO 2006).

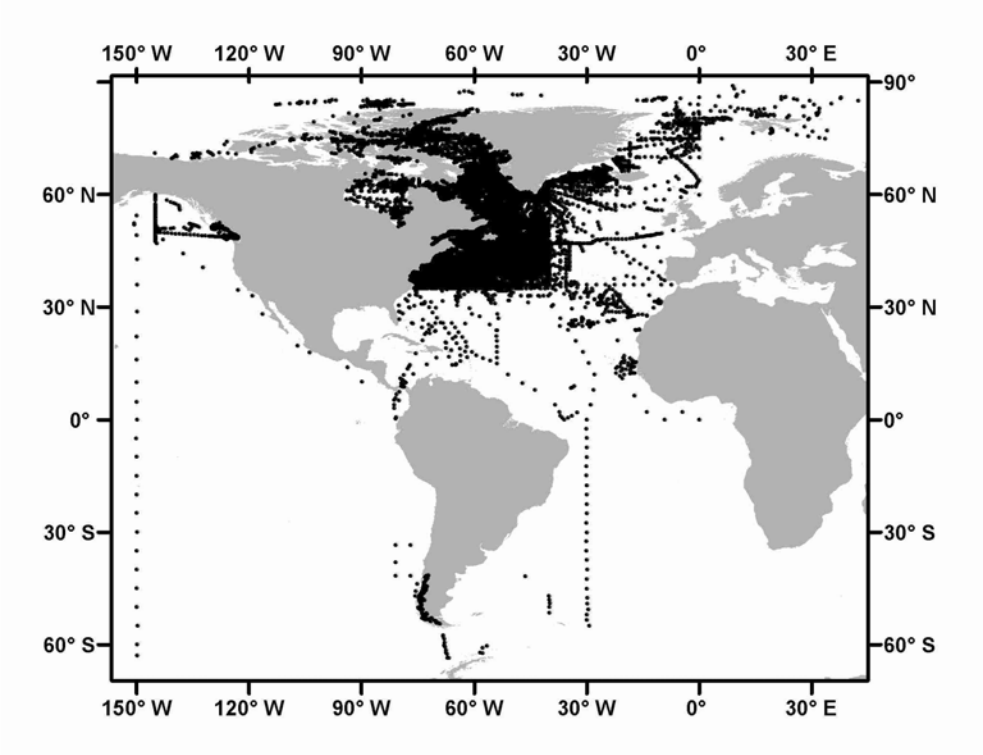

Figure 4: Map showing a subset of BioChem sample locations.

It is BioChem's taxonomic code table, i.e. its species list, that is the subject of the following exercise. This code table includes fields for the taxonomic name and scientific authority, comment fields, a local database code and the ITIS TSN code value.

MAP AND ASSIGN STANDARD CODES TO SPECIES LIST NAMES

The first step in the procedure is to design a collection list which will contain a list of distinct standardized names from a set of source datasets. This step is required since many 'species lists' contain separate records/codes for individual life history stages for one species. The aim of the following steps will be to append distinct names from the dataset species list to the collection list.
The CollectionList table will include the following fields:

- *Source_name (optional)*
- *Source_authority (optional)*
- Collection_SpeciesCode
- Collection_EditedName

- Collection_EditedAuthority
- TSN
- TSN_ITIS
- Comments

Before it can be determined if the species list has new names that need to be added to the collection list, the entire list must be "cleaned". In order to track changes in the source name to a cleaned/edited name, an intermediate level table was created. This table can house all records from a group of dataset species lists from different sources.

This intermediate table will be referred to as **DataSetLists** and the table design should include the following fields:

- Source_TaxonomicName
- Source_ScientificAuthority
- Source_DataSet - Identifier assigned to indicate the origin of the species list
- Source_SpeciesCode
- Collection_EditedName
- Collection_EditedAuthority
- Modifier
- LifeHistoryStage
- Rank
- Collection_SpeciesCode
- TSN_local
- TSN_ITIS
- Local_Register_Name

Dataset species lists, in addition to having multiple records due to the inclusion of life history stages, may have multiple records due to the association of a modifier with a particular name. Examples of modifiers are: sp.; spp.; aff.; ca 500µ; damaged; ?; etc.

The step by step procedure – Step 1

- Receive a dataset from a new source
- If the new dataset's species list contains names but does not contain the associated scientific authorities then contact the provider of the dataset. The following procedure requires the scientific name, authority name and year to be present whenever possible. The 'authorship' must be included in order to ensure proper matching to taxonomic nomenclature in the master list and in the local species registers.
- Load the new species list into the DataSetLists table filling the first four fields.
  - Source_TaxonomicName
  - Source_ScientificAuthority
  - Source_DataSet - Identifier assigned to indicate the origin of the species list
  - Source_SpeciesCode
- Clean/Edit/revise the records
  - Copy source_TaxonomicName and source_ScientificAuthority to Collection_EditedName and Collection_EditedAuthority fields

- Remove “sp.”, “spp”, “unidentified”,”damaged”, etc terms from edited_name field and record these terms in the Modifier field
- Remove “ssp.”, “var”, “form” terms from edited_name field and record these terms in the Rank field
- Remove any references to life history stages from edited _name and record these terms in the LifeHistoryStage field
- Remove any references to size from edited _name and record these terms in the modifier field, example *Thalassiosira* ca. 50µ
- Replace common names with scientific names, example replace copepods with Copepoda
- Contact the original supplier for clarification of unclear grouping names – example what is meant by term “other copepods”, “eggs”, etc
- Remove “cf”, “aff” terms from edited_name field and record these terms in the Modifier field
- Replace any abbreviations with the full name
- Correct any obvious typographical spelling errors, extra spaces, etc
- Check the authority name for punctuation and year matches
  - Standard formats: “name1, yyyy” or “(name, yyyy) name2 yyyy”
- Check source language
  - is ‘et’ used instead of ‘and’
  - check inclusion of accents

The step by step procedure – Step 2

Map[1] the edited names and authorities in the **DataSetLists** to identical names and authorities in the Collection List table

- Update the Collection_SpeciesCode field in the DataSetLists table where edited names and authorities match collection names and authorities
- Extract the subset of names with null Collection_SpeciesCodes (i.e. records where the datasetlist name does not have a match in the CollectionList). Append these new names and associated authorities to the CollectionList table.
- Remove selected records or add prefix of “IGNORE”. Species lists may contain records that indicate that a name was reassigned and that this dataset name record is no longer valid. These records do not need to be appended to the collection species list and therefore do not need to be verified. The choice is to either remove these records or to include a procedure that will skip these records if the first few characters in the name field equal “ignore”.

For new names in the **CollectionList** table, i.e. for records where the TSN fields are null:

- Copy Source_Name and Source_Authority entries to Collection_EditedName and Collection_EditedAuthority fields

[1] The method of mapping the supplied list to BioChem, to ITIS and/or the register has not yet been standardized. Mapping may be manual - use the ITIS on-line application to search for a name *or* ad hoc routines could be employed to search the databases for similar names.

- Map Collection_EditedName and authority entries to the MasterNegativeTSN table for known non-taxonomic names.
    - Update the TSN and the TSN_ITIS fields with the values from the MasterNegativeTSN table.
- Map Collection_EditedName and authority entries to the MasterList.
    - If name is in the MasterList update the TSN field
    - If the TSN value is positive then the TSN_ITIS field should be updated with the TSN value
    - If name is not in the MasterList, review and reformat name and reattempt mapping
    - If reformatted name is not in the MasterList append reformatted name to the MasterNegativeTSN table and update the TSN fields from the table.
- Link to the DataSetLists table and update the TSN fields.
- *Populate the individual Dataset Species List TSN fields (optional).*

The step by step procedure – Step 3 (Determine if new names are in the local register of marine species)

It is often difficult to authenticate the original specimen identification that resulted in the assignment of a name. It is possible, however, to check new names in a dataset list against a list of taxonomic names known to occur within a given area, i.e. verify that a new name exists in the local species register.

There may be many names in the supplied taxonomic code list that are not taxonomic in nature or they could be groupings of valid taxonomic names. Examples include phytoplankton, unarmoured dinoflagellates, algae, jellies, shrimp, H4B, CYT, "group of *Paracalanus*, *Clausocalanus* and *Pseudocalanus*". These kinds of entries would have been assigned a negative TSN_local value (see Appendix 1). It is questionable if it is worthwhile searching the local register for these groups.

The assignment of a -7000 series negative TSN code number simply implies that the name in question has not yet been reviewed for acceptance or rejection into the masterlist. These names could be present in the local species register.

Ideally the species register table includes the TSN as one of its required fields. *If this field is not present then the mapping from list to the register will have to match names and authorities.*

- If TSN is positive then map TSN values
    - If match update RegisterName in DataSetList table
    - If no match skip to Case 1 below
- If TSN in -7000 series then map edited name and authority to register name and authority
    - If match update Register Name in DataSetList table and skip to Case 2
    - If no match skip to Case 3

Case 1: For records in ITIS but not in Species Register

- Check the geographical area of the source sample – perhaps the sample was collected outside the bounding area for the register
    - If sampling area within register area
        - Review the literature and internet for additional authoritative information related to geographical distribution for this taxa and consult with taxonomic experts.
            - If it appears reasonable that the taxa could have been caught in the region update the local species register (see below) and update Register Name in DataSetList table
            - If it appears unreasonable then contact the original supplier and request additional info related to the analyst's level of expertise. Were they an expert taxonomist, a student, … What reference material was used to identify the individual specimen? (reference book and page number used to identify the specimen)
                - Revise the taxonomic name and repeat the procedure

                  -or-
                - Assign "unverified" to RegisterName field

    - If sampling area outside register area
        - Literature and internet search for existence of new registers and/or contact taxonomic expert(s) for group(s) in question
            - If identification confirmed
                - Update the Register (see below)
                - Update Register Name in DataSetList table with either the register's name or the Taxonomic authority's name
                - If no register available for geographical area then commence compiling list of names for this area
            - If identification not confirmed then contact the original supplier and request additional info related to the analyst's level of expertise. Were they an expert taxonomist, a student,… What reference material was used to identify the individual? (reference book and page number used to identify the individuals)
                - Revise the taxonomic name

                  -or-
                - Assign "unverified" to RegisterName field

Case 2: For records with names in Register but not in ITIS (-7000 series records)

Extract relevant fields from register to satisfy ITIS requirements for data submission as outlined on their web page.

Case 3: For -7000 series records not in ITIS or Register

Check geographical area for sample and bounding area for various registers

Literature and internet search for existence of new registers and/or contact taxonomic expert(s) for group(s) in question

- If identification confirmed
  - If no register available for geographical area then contact the regional data manager or consult WoRMS
  - Update the Register (see below)
  - Update Register Name with the register's name
- If identification not confirmed then contact the original supplier and request additional info related to the analyst's level of expertise. Were they an expert taxonomist, a student,… What reference material was used to identify the individual?
  - Revise the taxonomic name and repeat the procedure

-or-

  - Assign a -1000 series TSN_local value and assign "unverified" to RegisterName field

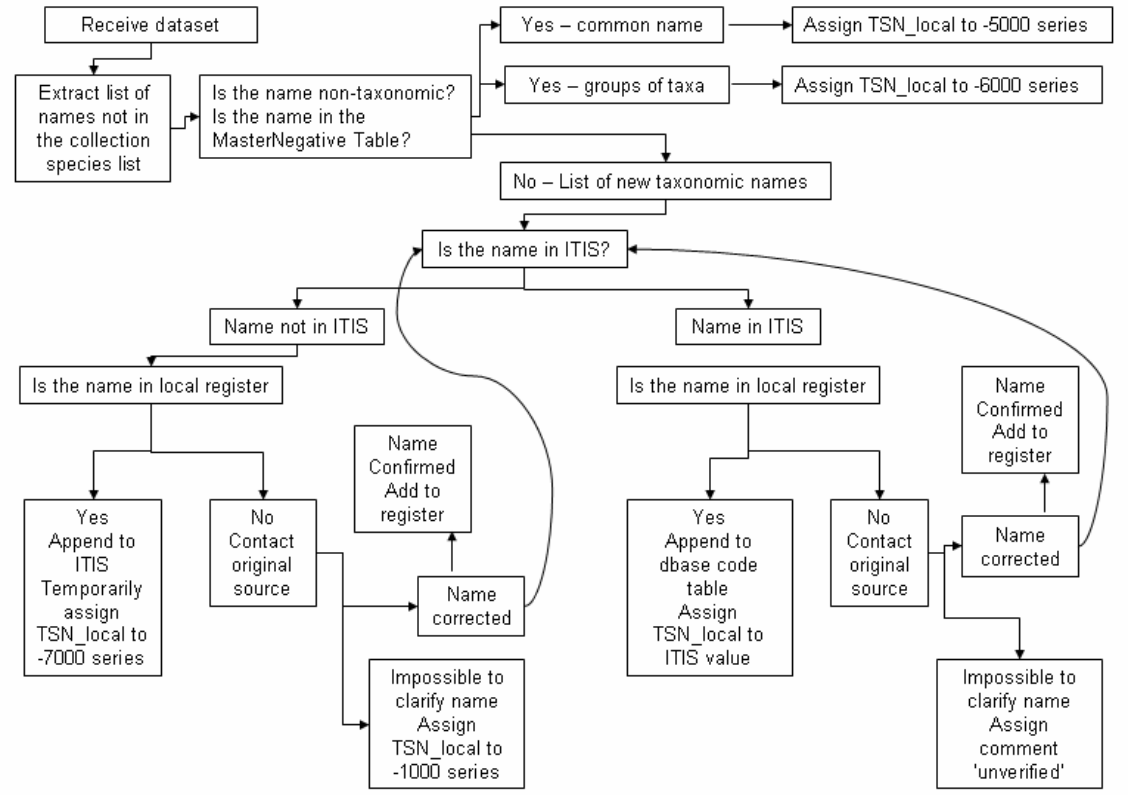

Figure 5. A flow diagram showing the various pathways to follow upon receipt of a new species list in the attempt to assign a TSN. This entire procedure requires long term funding and support for local registers of species and for master lists such as ITIS.

SUMMARY

For taxa in the collection species list that have a positive TSN, linkage exists to the currently accepted name and hierarchy for the source name. For names with negative TSNs this linkage to validity and synonyms is not available. However, the TSN_ITIS enables linkage to the master list and a partial hierarchical tree should be available depending on the amount of information recorded in the Master Negative Table.

Fundamental to the above procedures are the existence of standard codes for names (a master taxonomic code list and a master non-taxonomic code list) and local species registers. Both must be accepted and maintained. Without adequate support, financial and taxonomic expertise, these databases will quickly become dated.

APPENDIX 1: NEGATIVE TSN CODE SERIES

There are records in taxonomic code lists that are not true taxonomic names. These include groups of valid species, common names, non-living material as well as unverified/questionable names. These records may never be assigned an ITIS code. Procedures to handle these exceptions are outlined below.

The National Oceanographic Data Center's World Ocean Database (WOD) has adopted a procedure to assign negative values to their non-ITIS taxa[2]. BioChem has adopted their procedures and the following definitions for negative series TSNs are valid for both databases. BioChem has added an additional series for 'non-living' names.

NOTE: within the –ve TSN series DFO Maritimes codes do not equal those of WOD. These are non-intelligent and unlimited numerical codes. The codes are referred to as the –#000s series however codes within the series are not limited to 4 digits - the series designator is the first negative digit.

-9000s
These -9 codes are for name descriptions which are non-living items. Example: sand, rocks, plastic, garbage, metal, glass, bone, wood.

-7000s
These -7 codes are for taxa names which were still in the ITIS review process at the creation of WOD01. Upon completion of the review, most will be assigned an official ITIS TSN which will replace the temporary -7 code. Those which fail to meet the ITIS review criteria will be re-assigned a -1 code value.

-6000s
These -6 codes are for taxa names which contain two or more taxonomic groups, making it difficult to assign a single ITIS taxonomic code which preserves the original meaning. Example: "Salps and Doliolids", both legitimate by themselves, but combined they cannot be matched to a single ITIS TSN.

-5000s
These -5 codes are for taxa names which are non-taxonomic, or cover too many taxonomic groups to assign a single ITIS taxonomic code which preserves the original meaning. Example: shrimp, worms, plankton.

-1000s
These -1 codes are for taxa names which were submitted to ITIS and failed to meet the ITIS review criteria. The descriptions are either non-taxonomic, non-existent, or misspelled beyond recognition.

---

2 Extracted from World Ocean Database (WOD) code table definitions. See tax_####:http://www.nodc.noaa.gov/OCL/WOD01/code01.html

REFERENCES

BioChem (DFO 2006). Department of Fisheries and Oceans Canada, 2006. BioChem: database of biological and chemical oceanographic data.
URL http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/Biochem_e.htm Version 8 (2005).

Branton, Robert, Lenore Bajona, Shelley Bond, Mary Kennedy, Daniel Ricard and Lou Van Guelpen. MS 2007. Methods for Standardizing, Validating and Enriching Taxonomic Metadata. NAFO SCR Doc. 07/08, Serial No. N5349, 8p.

Chapman, A.D. 2005. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
URL http://www.gbif.org/prog/digit/data_quality

Colton, J.B. and R. Marak, 1969. Guide for identifying the common fish eggs and larvae of continental shelf waters, Cape Sable to Block Island. Woods Hole Laboratory Reference #69-9, Bureau of Commercial Fisheries, Woods Hole, MA

Conkright, M. E., J. I. Antonov, O. Baranova, T. P. Boyer, H. E. Garcia, R. Gelfeld, D. Johnson, R. A. Locarnini, P. P. Murphy, T. D. O'Brien, I. Smolyar, C. Stephens, 2002: World Ocean Database 2001, Volume 1: Introduction. S. Levitus, Ed., NOAA Atlas NESDIS 42, U. S. Government Printing Office, Wash. D. C., 167 pp.
URL http://www.nodc.noaa.gov/OCL/WOD01/code01.html

Costello, M.J., K. Stocks, Y. Zhang, J.F. Grassle, D.G. Fautin. 2007. About the Ocean Biogeographic Information System.
URL http://www.iobis.org/about/

DFO. 2008. Proceedings of a conference on ocean biodiversity informatics; 2-4 October 2007. DFO Can. Sci. Advis. Sec. Proceed. Ser. 2008/024.
URL http://www.dfo-mpo.gc.ca/CSAS/Csas/Publications/Pro-CR/2008/2008_024_e.pdf

ECNASAP (DFO Maritimes, 1994).
East Coast of North America Strategic Assessment.
URL
http://geodiscover.cgdi.ca/gdp/search?action=fullMetadata&entryType=productCollection&entryId=10982&entryLang=en

ISDB DFO Maritimes (2006).
Industry Surveys Database.
URL
http://geodiscover.cgdi.ca/gdp/search?action=fullMetadata&entryType=productCollection&entryId=31837&entryLang=en

DFO Maritimes Nafo Landings Database.
URL http://www.nafo.int/

DFO Maritimes Research Vessel Trawl Surveys (RV)

Food and Agricultural Organization (2006).
FAO FIGIS database Species List.
URL http://www.fao.org/fishery/collection/asfis/1

Gregory, Doug and Savi Narayanan. 2003. BioChem: A National Archive for Marine Biology and Chemistry Data. AZMP/PMZA Bulletin 3: 11-13.
URL: http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/azmp-pmza/documents/docs/bulletin_3_2003.pdf

ITIS (2008). Integrated Taxonomic Information System.
URL http://www.itis.gov/about_itis.html

ITIS *CA (2007) Integrated Taxonomic Information System. Canada
URL http://www.cbif.gc.ca/pls/itisca/taxaget?p_ifx=cbif

OBIS (2006).
Ocean Biogeographic Information System.
URL http://www.iobis.org/