Northwest Atlantic Fisheries Organization

Serial No. N5349 **NAFO SCR Doc. 07/ 08**

SCIENTIFIC COUNCIL MEETING – JUNE 2007

Methods for Standardizing, Validating and Enriching Taxonomic Metadata

by

Robert Branton[1]
Science Branch, Population Ecology Division, DFO, BIO
1 Challenger Drive, P.O. Box 106, Dartmouth NS, B2Y 4A2

Lenore Bajona
Science Branch, Ocean Sciences Division, DFO, BIO
1 Challenger Drive, P.O. Box 106, Dartmouth NS, B2Y 4A2

Shelley Bond
Science Branch, Population Ecology Division, DFO, BIO
1 Challenger Drive, P.O. Box 106, Dartmouth NS, B2Y 4A2

Mary Kennedy
Science Branch, Ecosystem Research Division, DFO, BIO
1 Challenger Drive, P.O. Box 106, Dartmouth NS, B2Y 4A2

Daniel Ricard
Department of Biology, Dalhousie University
1355 Oxford St., Halifax NS, B3H 4J1

Lou Van Guelpen
Atlantic Reference Centre, Huntsman Marine Science Centre
St. Andrews NB, E5B 2L7

Abstract

Taxonomies are always changing; hence providing reliable taxonomic metadata is a difficult task, especially when working with multiple databases containing data on hundreds of organisms. In this paper we describe how taxonomists at the Atlantic Reference Centre of the Huntsman Marine Science Centre in St. Andrews, New Brunswick and scientific data managers at the Bedford Institute of Oceanography in Dartmouth, Nova Scotia prepare organism names in biological databases for global internet access via portals such as the Ocean Biogeographic Information System and the Global Biodiversity Facility.

[1] Corresponding author
BrantonB@mar.dfo-mpo.gc.ca
902-426-3537

Introduction

The Ocean Biogeographic Information System (OBIS, 2006) is the main data portal of the Census of Marine Life (CoML, 2006) as well as main contributor of marine species data to the Global Biodiversity Facility (GBIF, 2006). Taxonomic metadata available on such portals must be consistent across the data sources.

Metadata give information about data to enable analysts or end-users to understand them. Biological data include the name(s) of the organism(s) in question either in the body of a database, in its title, or in its descriptive text. In each case the name comprises what we call taxonomic metadata. Simply providing the common name(s) for species, particularly when data are to be combined with those from other sources, is generally unacceptable. Data providers must give scientific names and authorities for the organisms at hand. Furthermore, the names must be currently accepted by taxonomic specialists. Taxonomic metadata also can be expected to include taxonomic hierarchies. Taxonomies are always changing; hence providing reliable taxonomic metadata is a difficult task, especially when working with multiple databases containing data on hundreds of organisms.

Numerous species lists and authoritative repositories exist.

> In Atlantic Canada, the Atlantic Reference Centre (ARC) of the Huntsman Marine Science Centre in St. Andrews, NB maintains a research collection of organisms collected in the Northwest Atlantic. This collection and the taxonomic expertise of the ARC staff provide a strong regional basis for authoritative information about species names and taxonomic classification. In addition, the ARC maintains a series of Registers of Marine Species that list the scientific names of species found in different regions of the Northwest Atlantic, namely the Bay of Fundy, the Gulf of Maine, the Canadian Atlantic and the Northwest Atlantic (ARC, 2004; Table 1). Work in progress includes validation of scientific names and coordination with the European Register to create a North Atlantic Register of Marine Species.
>
> A leading international source for authoritative taxonomic metadata is the Integrated Taxonomic Information System (ITIS, 2006). This system provides a central location to verify scientific names and to obtain full taxonomic classification for marine and terrestrial organisms. The system was designed and implemented after the 1992 Rio de Janeiro Earth Summit and the subsequent ratification of the Convention on Biological Diversity (United Nations, 1992). The taxonomic database maintained by ITIS contains authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world. The latest version of this database that we have was released on October 19th 2006 and contains 424,130 accepted scientific names (ITIS, 2006). Its main aim is to support various biodiversity initiatives and to facilitate the exchange and sharing of reliable biodiversity metadata.
>
> Another international source of standard taxonomic metadata is the Food and Agricultural Organization (FAO) list of 10,650 species for its FIGIS database (Fisheries Global Information System, 2006). This list covers commercially exploited species and contains scientific species names as well as common names in English, French and Spanish. The FIGIS database is the focal point for the collection, management and dissemination of capture and production data from various regional fisheries bodies (e.g. NAFO - Northwest Atlantic Fisheries Organization).

In this paper we describe how taxonomists at the ARC and scientific data managers at the Bedford Institute of Oceanography (BIO) in Dartmouth, Nova Scotia prepare organism names in biological databases for global internet access via portals such as OBIS and GBIF.

Materials and Methods

The taxonomic metadata of four databases from the Department of Fisheries and Oceans (DFO) were examined and harmonized with the contents of ITIS and of the FAO list of species of commercial importance. These databases include:

1. DFO BioChem database (DFO, 2006)
2. DFO Maritimes copy of the NAFO landings database (NAFO, 2006)
3. The East Coast of North America Strategic Assessment (ECNASAP) (DFO Maritimes, 1994)
4. DFO Maritimes Industry Survey database (DFO Maritimes, 2006)

A summary of record and name counts by database are given in Table 2.

The ITIS and FAO species lists were obtained from their respective websites and loaded into an Oracle database. Procedural Language/ Structured Query Language (PL/SQL) programs were developed to perform the following basic steps:

For each given name and authority where available:

- Find matching name(s) and associated ITIS Taxonomic Serial Number(s) (TSN)
- For each TSN found, find currently accepted TSN
- For the accepted TSN, find the accepted (scientific) name, author and taxonomic hierarchy
- If no TSN found then mark for manual handling

For each accepted name (if no accepted name, then use given name):

- Find matching FAO abbreviation (3a_code, e.g. HAD for haddock)
- Find matching English and French common names

For a given name the accepted scientific name and author, TSN, and taxonomic hierarchy comprise the “enriched” taxonomic metadata for that name.

A flow chart for this program is given in Figure 1.

Results

In most cases the PL/SQL procedure succeeded in providing accepted names and hierarchies. A summary of accepted names by taxonomic level and database is given in Table 3 and sample of the resulting enriched taxonomic metadata is given in Table 4.

In some cases in which the authority name was not provided our PL/SQL returned multiple TSNs; in other cases no TSNs were returned. Where multiple TSNs were returned data managers worked with ARC staff to determine which TSN to select. A genus name for example may appear in two different kingdoms, in which case we simply had to choose between kingdoms. Where no TSN was returned because the given name was obviously misspelled we worked with the data provider to determine the correct spelling and hence the appropriate TSN. Where the problem could not be resolved, a negative TSN was assigned under the following scheme:

–7000	name appears to be valid, but is not currently in ITIS
–6000	a group of two or more valid ITIS names
–5000	taxonomic slang or an assemblage of groups
–1000	non-taxonomic or misspelled beyond recognition

These negative TSNs are used to locally manage and track outstanding problems. DFO database managers are working with ITIS Canada, to expedite entry of the -7000 group of names and authorities into the ITIS database. Our scheme of negative TSNs is very similar to that used by the World Ocean Data Center (World Ocean Database 2001).

Discussion

Providing end users with reliable taxonomic names, hierarchy information, FAO abbreviations and common names clearly facilitates biodiversity analyses by enabling drilldown and rollup across multiple data sources. Providing taxonomic hierarchies, in particular, improves understanding of the taxonomic complexity and resolution of a given database.

The task of finding accepted names and hierarchies was greatly facilitated by the PL/SQL procedures. Problems typically encountered were:

- A name was to the taxonomic rank that a scientist, technician or observer could identify the particular specimen; hence a name could be at any taxonomic rank.
- A name was not yet in ITIS.
- A name was not a true scientific name(s) (e.g. Invertebrata, Algae), but was easily assignable to phylum or kingdom.
- A name returned multiple ITIS TSNs in different taxa (e.g. when a scientific name did not include the authority). Identical names could have different ITIS credibility ratings.
- Manual intervention sometimes was required to resolve discrepancies.

Issues raised by the validation and the enrichment of taxonomic metadata are shared by all laboratories conducting fisheries research. While the usefulness of these various systems is clear, it is important that they be integrated into a consistent framework. Local species lists, for example, must contain not only the species names, but also species authorities. With this information we can enhance the PL/SQL procedures to automatically manage both names and authorities, thus reducing the need for manual intervention.

Taxonomies will always change, but local species lists do not need to be kept updated with changing taxonomy if they include ITIS TSNs. Use of our PL/SQL procedures will return updated taxonomic information. To perform this protocol we maintain a local copy of the ITIS database, frequently updating it to ensure that names, authorities and hierarchies are current. To streamline this process we are investigating use of PL/SQL Web Services Access to directly query the ITIS website as an alternative to maintaining a downloaded copy of the ITIS database.

References

ARC (2004).
Marine Species Registers for the North Atlantic Ocean.
URL http://www.marinebiodiversity.ca/nonNARMS/

CoML (2006).
Census of Marine Life.
URL http://www.coml.org/coml.htm

Biochem (DFO 2006).
Database of biological and chemical oceanographic data.
URL http://www.meds-sdmm.dfo-mpo.gc.ca/biochem/Biochem_e.htm

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

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

Food and Agricultural Organization (2006).
FAO FIGIS database Species List.
URL http://www.fao.org/figis/servlet/static?dom=root&xml=species/species_search.xml

GBIF (2006).
Global Biodiversity Information Facility.
URL http://www.gbif.org

ITIS (2006).
Integrated Taxonomic Information System.
URL http://www.itis.usda.gov/

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

NAFO (2006).
DFO Maritimes Nafo Landings Database.
URL http://www.nafo.int/

United Nations (1992).
Convention on Biological Diversity.
URL http://www.biodiv.org/

World Ocean Data Center (World Ocean Database 2001).
Code List files for World Ocean Database 2001.
URL http://www.nodc.noaa.gov/OC5/WOD01/code01.html

Table 1 – ARC Registers of Marine Species.

Register	Names
Bay of Fundy	2,680
Gulf of Maine	3,300
Canadian Atlantic (CARMS)	5,120
NW Atlantic (NWARMS)	6,120

Table 2 – Summary of record and name counts by database.

Database	Catch Records	Names
1. Biochem	750,000+	3,427
2. NAFO Landings	1,025,027	159
3. ECNASAP	471,798	274
4. Industry Survey	3,089,458	791

Table 3 – Summary of accepted names by taxonomic level and database.

	Species	Genus	Family	Order	Class	Phylum	Kingdom	Unassigned*	Total
1. Biochem	1,864	771	226	97	60	34	0	375	3,427
2. NAFO Landings	123	18	6	4	3	3	2	0	159
3. ECNASAP	271	3	0	0	0	0	0	0	274
4. Industry Survey	464	89	83	24	20	8	0	103	791
* given names currently not automatically assigned via ITIS, hence manual intervention required									

Table 4 – Sample of resulting enriched taxonomic metadata.

Following are examples of taxonomic metadata typically given by the data providers …

GIVEN SCIENT NAME	ECHINODERMATA P.	ASTEROIDEA S.C.	GADIFORMES	RAJIDAE F.	SEBASTES SP.	LYCODES ESMARKI	GADUS MORHUA
GIVEN COMMON NAME	SPINY SKINNED ANIMALS	ASTEROIDEA S.C.	HAKE (NS)	SKATES (NS)	REDFISH UNSEPARATED	VACHON'S EELPOUT	COD(ATLANTIC)
GIVEN SPEC CODE	6000	6100	18	211	23	643	10

Following are corresponding examples of taxonomic metadata obtained from the Integrated Taxonomic Information System (ITIS) …

TSN FOR GIVEN NAME	156857	156862	164665	160845	166705	165287	164712
SCIENT NAME UPDATED	N	N	N	N	N	Y	N
ACCEPTED TSN	156857	156862	164665	160845	166705	630982	164712
ACCEPTED SCIENT NAME	Echinodermata	Asteroidea	Gadiformes	Rajidae	Sebastes	Lycodes esmarkii	Gadus morhua
ACCEPTED AUTHOR	Klein, 1734	de Blainville, 1830		Blainville, 1816	Cuvier, 1829	Collett, 1875	Linnaeus, 1758
ACCEPTED RANK	Phylum	Class	Order	Family	Genus	Species	Species
KINGDOM	Animalia	Animalia	Animalia	Animalia	Animalia	Animalia	Animalia
PHYLUM	Echinodermata	Echinodermata	Chordata	Chordata	Chordata	Chordata	Chordata
CLASS		Asteroidea	Actinopterygii	Chondrichthyes	Actinopterygii	Actinopterygii	Actinopterygii
ORDER			Gadiformes	Rajiformes	Scorpaeniformes	Perciformes	Gadiformes
FAMILY				Rajidae	Scorpaenidae	Zoarcidae	Gadidae
GENUS					Sebastes	Lycodes	Gadus
SPECIES						esmarkii	morhua

Following are corresponding examples of taxonomic metadata obtained from the Food and Agriculture Organization (FAO) …

FAO CODE	ECH	STF	GAD	RAJ	RED		COD
FAO E COMMON NAME	Echinoderms	Starfishes nei	Gadiformes nei	Rays and skates nei	Atlantic redfishes nei		Atlantic cod
FAO F COMMON NAME	Oursins, bèches-de-mer	Astéridés nca	Gadiformes nca	Rajidés nca	Sébastes de l'Atlantique nca		Morue de l'Atlantique

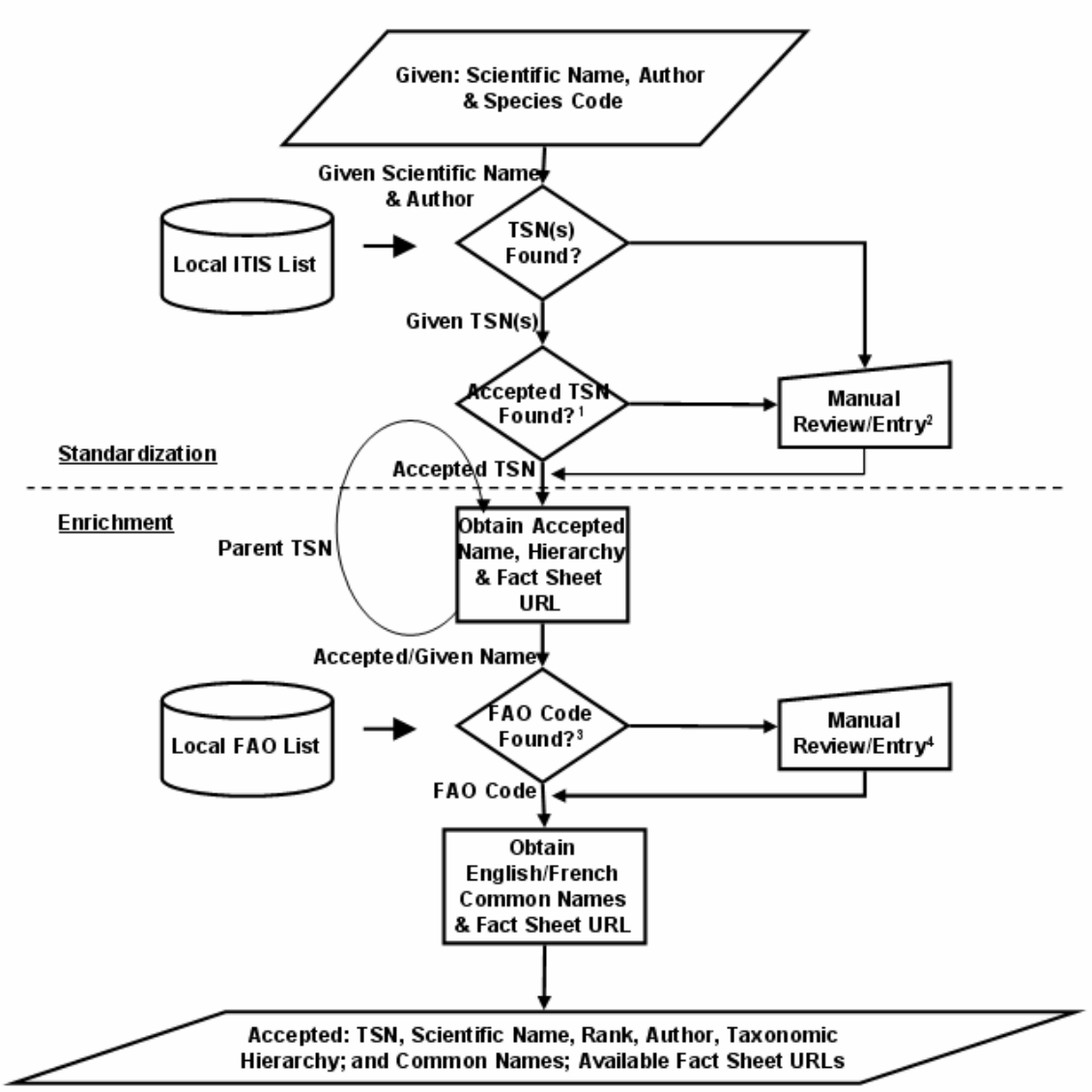

Figure 1 – Flow chart of PL/SQL program used.