ENBI – Work Package 11

1st Workshop on "Multi-lingual Access to European Biodiversity Sites"

Institute of Marine Research, Kiel, Germany

06. – 07. October 2003

Workshop Report

Introduction: The first workshop in relation to the ENBI-project and work package 11 ("Multi-lingual Access to European Biodiversity Sites") was conducted as scheduled in the beginning of October 2003 (6.-7. Oct. 2003) at the Institute for Marine Research (IfM) in Kiel. The major purpose of this workshop was to introduce the translation partner, to establish the translation consortium for work package 11 , to make translation partners familiar with the ENBI project in general and specifically with the tasks of work package 11, to introduce tools and techniques necessary to provide multi-lingual access to biodiversity information in the Internet and to train partner on the encoding procedure of the specialized dictionaries. Partner contracts were negotiated and finally settled at the workshop. A detailed workshop agenda is attached as ANNEX 1 to this report, a list of participants is attached as ANNEX 2.

Results: All workshop participants introduced each other with a short statement on their institutions and their background for being a partner in Work Package 11 in the ENBI project. Some partners had already successfully co-operated with the FishBase project in the past in relation to translation matters (e.g. Afonso Marques, LMG / IMAR, has coordinated the translation of the FishBase 2000 Manual into Portuguese). Rainer Froese gave an interesting introductory presentation on FishBase with special consideration on multilingual aspects titled:

"FishBase as a Model for a Multilingual Database."

FishBase was introduced as a model for a multilingual database in the Internet. The strategies which will be developed for the translation of this database can serve as a template for the translation of any other database in the Internet. As a first task for translation several lists were submitted to the participants for word-by-word translation (static translation) in order to make the search page of FishBase quickly available in different languages. The dead-line for this task was set to 31.10.2003. In order of priority the following lists were submitted:

1. Labels list
2. Country list
3. Ecosystem list
4. Order list

Many aspects were discussed on the translation of e.g. family names. Specifically in connection to FishBase as a model for a multilingual database, the discussion concerned the following issues:

- How to proceed with missing order common names or a family common names in a given language? The consequence would be, that this family will not appear at all in the list with translated family names (thus consequently not be displayed to visiting experts who may have suggestions for a better/correct translation which is not a desired situation). To avoid this, it was suggested to try first to find an existing name, exploring sources as "official lists", field guides, aquarium guides, and the Internet. It was agreed, to mention the source for a common name in an extra column in the translated lists (e.g. paper, book, URL) to allow to follow up for a quality check.
- As a second option, it was suggested to look for the species common names of the main genera and select the one used mostly, especially for species of different genera. As a third option, it was suggested to translate the English term, and if the first three options fail, to translate the scientific name into the given language.

- In addition it was mentioned, that for 74 families no English terms are available in FishBase yet. It was suggested to contact Jo Nelson in order to check if he has advanced a work planned for the possible 4th edition of Fishes of the World. With a new family created it was decided to choose the English name, and, if not available, the scientific name instead until a translation is available.

- How to deal with the fact, that in the literature families appear which are listed in FishBase as Subfamilies? At this point, it was suggested to provide the translation as well, if applicable or not can be decided later.

- Several common names for a family. If several common names are used for one family, the following suggestions shall be applied as a guideline for translation:
- "and": used when the family/order contains species with different common names.
- e.g. Eagle rays and Manta rays: some fishes in this family are called Eagle Ray, others are called Manta Rays.
- "or": used when a family has different common names which are in fact synonyms.
- e.g. Racehorses or Pigfishes: the same species in this family is sometimes called racehorse and sometimes pigfish.
- ",": used when more than two names are known for that particular family. It can be used in combination with "and" and "or".
-
- Rainer Froese specifically has asked not to apply "&" and brackets "()" in any case.

- How to proceed with the translation of countries and Islands? The standard to use for country names is the ISO 3166-1 short name which should be available in each language in the national standardisation agency, at least to check the correct name and spelling. However it was noted to be careful with country names which were recently changed, like Burma in Myanmar, east

European countries, Pacific island states like Nauru. The second part of this standard (3166-2) could be consulted for archipelagos or islands that are also administrative entities. The 3rd part (3166-3) contains less than ten previous country names like Czechoslovakia which need to be translated as well.

- The English names start almost consistently with a capital initial. There is probably no specific reason for this, and it is suggested to handle this in accordance with the specific rules of a given language (e.g. in German the adjectives start with a small letter).

- Abbreviations shall not be applied for technical reasons, e.g. do not use "St." for "Saint.

- While working with the FishBase lists, English typing errors should be highlighted and passed with the translated lists to the FishBase team for correction.

- For the meaning of some words it was suggested to use following standard reference: *Roger Lincoln, Geoff Boxhall and Paul Clark (1998) – A dictionary of ecology, evolution and systematics. Cambridge Press.*
- N. Bailey will ask the authors for permission to use their content for the translation in FishBase.

- Another issue was raised on common species name: the fact that, on the species level, one species will often have more than one “main” common name. There are e.g. official common names set by law for trade purposes for most commercial species but there is no national list of official unique common names. As this matter (translation on the species level) at present is not going to be addressed in the translation project and FishBase has a strategy to deal with this issue, there is no further need for discussion on this issue.

Bernd Ueberschär gave a presentation entitled:

"ENBI - European Network for Biodiversity Information and Work package 11: - Multi-lingual Access to European Biodiversity Sites."

The ENBI project was introduced in general and the specific tasks and the time schedule for Work package 11 was presented (date of commencement of WP 11 is the 1st of march 2003, the project will officially ends at the 28th of February 2005). The presentation is available for download at the website www.enbi.linguaweb.org.

Bernd Ueberschär gave a presentation entitled:

"Technical instructions: tools and guidance how to translate. Introducing tools and techniques necessary to provide multi-lingual access to biodiversity information in the Internet."

Specifically the EU-Systran Service was introduced and some tools suggested to support the translation tasks, e.g. EURODICATOM, the European Commissions multilingual term bank. The Systran translation engine was presented as a Personal Computer Edition (Systran Premium) for "hands on" sessions and the machine

translation abilities tested by the workshop participants with some free text paragraphs taken from FishBase in their respective language.

In relation to the machine translation, many questions were raised, most of them concerning the differences in language structure and the difficulties with the results which are caused by those language-specific features when a machine translation was applied:

Some specific issues and problems identified in relation to machine translation:

- With the standard vocabulary (standard dictionaries), Systran translates words and terms often in a different sense, such as "Stocks" or "Order" (tested in German). Only the implementation of specialised dictionaries and the introduction of "Domains" will assure a qualified translation.

- Using Systran from e.g. English to Italian the translation engine does not recognise the difference between a verb and a noun in some cases. For example the word ‘fish’ is both a verb and a noun in English whilst in Italian there are two discrete words. Systran will preferably translate ‘fish’ into a verb. Similar problems were found with other languages. One solution will be to implement the translation for fixed phrases into the dictionary (e.g. "fish stocks"), this will enable Systran engine to apply the proper translation after analysing the context. Another measure to solve these problems is to simplify the English source text in such a manner that allows Systran to recognise such differences. Guidelines for this simplification were already drawn from Systran´s recommendations on "How to prepare an English text for a qualified translation" and were passed to the FishBase team (for those partners who are interested in these recommendations, see ANNXEX 4 of this report).

- It was argued, that the machine translation of the FishBase "Diagnosis Field" can easily lead to errors and/or incorrect interpretation of the numbers of e.g. nº spiny rays, nº line lateral line scales may result in wrong translation of the numbers and their position resp., this issue will probably be improved under consideration of the measures as mentioned in the paragraph above.

- A problem for the machine translation of English into Dutch seems to be the obviously incomplete implementation of Dutch grammar rules into the translation engine. Dutch grammar is sometimes completely different compared to the English grammar and at present this is not yet reflected with the translation engine of Systran for English to Dutch. In a lot of cases, the meaning of the sentence will be lost and this is resulting in ‘rubbish’. However, it can be expected that the rules, grammar and basic dictionaries are being improved continuously by Systran, thus a better translation result can be expected in the future (this is valid for all languages). In addition, as an instant measure, a more standardised and simplified English source text will help to improve the translation as well.

- In general, the EU-Systran service is not yet ready (by the date of this workshop) to be used for online translation of websites, thus as an intermediate solution, the Systran web service will be used as a demo for the translation of free text fields such as Species Summary in FishBase. Ongoing

effort to negotiate with the EU-Translation service on the implementation of the website translation is underway, initiated by the Project Coordinator Bernd Ueberschär.

- All workshop participants were asked to suggest other databases and glossaries on Biodiversity in the Internet for translation (translation of glossaries will not produce a new definition, rather than a plain translation of the given text, mostly presumed to be in English) , following the translation strategies which are applied to FishBase. The Italian partner suggested e.g. EUNIS, European Nature Information System
- http://eunis.eea.eu.int/eunis/index.jsp, CLEMAM: Checklist of European Marine Mollusca http://www.somali.asso.fr/clemam/index.clemam.html
- MarLIN: The Marine Life Information Network for Britain and Ireland, with special attention on two fields: (i) Species information pages and (ii) Habitats (Biotope) Information Pages: http://www.marlin.ac.uk/
- CephBase: http://www.cephbase.utmb.edu/, Crustacea Glossary of the Natural History Museum of Los Angeles County, http://crustacea.nhm.org/glossary/

Date of next Workshop:
All workshop participants agreed to have rather an interim workshop instead of a final workshop at the end of the project. It is suggested that an interim workshop (August/September 2004 in Kiel) will facilitate the translation work a lot and will give enough room for discussions on open issues for the translation tasks before the termination of the project. In case the workshop budget allows a workshop at the end of the project the final workshop scheduled for January/February 2005 will be conducted as well.

ANNEX 1:

ENBI – Work Package 11

1st Workshop on "Multi-lingual Access to European Biodiversity Sites"

Institute of Marine Research, Kiel, Germany

06. – 07. October 2003

Workshop Agenda:

Monday 06.10. 2003:

- **10:00am:** Welcome and Opening Remarks, Coffee, Refreshments, Tea and Cookies available
- **10:15am**: Introduction of Workshop Participants (~ 5 min each)
- **10: 30am**: Presentation **Rainer Froese**: FishBase as a Model for a Multilingual Database.
- **11:15am:** Discussion
- **11:30am:** Presentation **Bernd Ueberschär:** -General Introduction-
- ENBI - European Network for Biodiversity Information and Work package 11: - Multilingual Access to European Biodiversity Sites
- **12:15am:** Discussion
- **12:30pm:** Lunch (Kantine Landeshaus, ~10 min. Walk Outside)
- **13:30pm:** Presentation **Bernd Ueberschär**: Technical instructions: tools and guidance how to translate. Introducing tools and techniques necessary to provide multilingual access to biodiversity information in the Internet.
- **14:15pm:** Discussion
- **15:00pm:** Coffee Break
- **16:15:** Open for Discussion,
- **17:00:** Contract Issues, Billing.
- **17:30:** Return to Hotel (Collective Transport, for those who want to change the Dress)
- **19:15:** Depart from Hotel to Dinner Location (Restaurant Louf, close to the Place for Lunch) Collective Transport.

Tuesday, 06.10.2003:

- **9:00am – 12:00am:** Open for Discussions and Strategies and Exercises on Translation, Tools and Measures (Extraction Tools, Systran, Filter Tools etc.). Coffee, Refreshments, Tea and Cookies available.
- **12:00am – 12:30pm**: Processing of Reimbursement Forms
- **12:30pm:** Lunch (Kantine Landeshaus, ~10 min. Walk Outside)
- **13:30pm:** Continue Discussions on Translation Strategies and Exercises
- **14: 30pm:** Workshop Outcomes, Recommendations and Final Thoughts
- **16:00pm:** Workshop officially closed.
- **15:00pm:** If desired, open for further Discussions, Library Visit, Contractual Issues

ANNEX 2:

List of Workshop Participants

NicolasBailly, MNHN - Muséum national d'histoire naturelle, (French)
57 rue Cuvier 75231 PARIS CEDEX 05
E-mail: bailly@cimrs1.mnhn.fr

Margaret Eleftheriou, IMBC - Institute of Marine Biology of Crete (Greek)
P.O.Box 2214 Heraklion Crete,
E-mail: margaret@imbc.gr

Bernd Ueberschär, IfM - Institute for Marine Research, Kiel (Germany)
Düsternbrooker Weg 20, D - 24105 Kiel, E-mail: bueberschaer@ifm.uni-kiel.de,

Rainer Froese, IfM - Institute for Marine Research, Kiel (Germany)
Düsternbrooker Weg 20, D - 24105 Kiel, E-mail: rfroese@ifm.uni-kiel.de

Afonso Marques, LMG / IMAR - Laboratório Marítimo da Guia Faculdade de Ciências de Lisboa, Estrada do Guincho, 2750 Cascais
E-mail: ammarques@fc.ul.pt

Carlo Froglia, CNR Istituto di Scienze Marine Sezione Pesca Marittima (Italy)
Lgo Fiera della Pesca, Dr. I-60125 ANCONA, ITALY E-mail: c.froglia@ismar.cnr.it

Elisabetta Betulla Morello, CNR Istituto di Scienze Marine Sezione Pesca Marittima (Italy), Lgo Fiera della Pesca, Dr. I-60125 ANCONA, ITALY E-mail: b.morello@ismar.cnr.it

Gert Boden, Africa Museum, Department of Zoology Fish Section Leuvensesteenweg 13, B-3080 TERVUREN BELGIUM (M.Sc.) E-mail: boden@africamuseum.be

Wijnand Heitmans, ETI - Expert Center for Taxonomic Identification (Netherlands), University of Amsterdam, Mauritskade 61, 1092 AD Amsterdam, E-mail: wijnand@eti.uva.nl

ANNEX 3

Specific comments of the Workshop participants (not edited). However, many comments and suggestions were extracted for the section general results of this report.

Margaret Eleftheriou, IMBC - Institute of Marine Biology of Crete (Greek)
As far as my comments on the Workshop are concerned:
The workshop was extremely useful in that we all participated in using the Systran machine translation as a group which covered several native languages. This was particularly useful and saved a good deal of time, because we were able to have first-hand, native speaker feedback about the translations. I think I can safely say that all group members were initially not very happy with the quality of the translations, but that this reaction changed to a certain extent as the meeting went on and discussions took place as to why there were such imperfections in the machine translations.
We agreed as a group that it was not simply a question of the length and complex grammatical structure of the sentence that threw up problems. We did in fact agree to recommend that all sentences submitted for machine translation should be kept as short as possible, as the machine translation was noticeably better with short structures.
However, it was also clear that some other language function impaired the efficiency of the translation. Again, after some discussion, one of the failings was identified as coming from the dual or even triple grammatical function of certain words, the most important of which is, crucially for FISHBASE, "fish".
In English, the term fish can be used as a noun, a verb and an adjective. As a result , the machine cannot differentiate between the different functions and produces, occasionally gibberish.
Another major problem comes from the use of the phrasal verb in English, which has a preposition as an integral part of the meaning, though not of the form. That is to say, "look up", "look into" "look out" all have specific meanings dependent on the preposition. However, the machine frequently treats the preposition as a preposition and not as part of the verb, and the result is also occasionally gibberish. (There is a slightly similar situation in German, but in that language, much more logical, the preposition becomes part of the verb and therefore there is not the same opportunity for confusion.
There is also considerable confusion coming from the complex verbal structure of the present tense in English, with its three separate forms (they fish, they are fishing, do they fish?).
It was clear to me as a language specialist that all of these identified difficulties come from the fact that the matrix language is English. I do not think that similar translation difficulties would be encountered if French, German, Spanish of Italian were used as the matrix. However, since I am not a native speaker of these other languages, perhaps I am merely seeing those difficulties which exist in my native language.

NicolasBailly, MNHN - Muséum national d'histoire naturelle, (French)
also for A. MARQUES
During the morning of the 7th, A. MARQUES and N. BAILLY worked together on the CD-ROM provided by B. Überschär.
1. Static translation files
1.1. Orders
No specific discussion during the workshop.

I was remembering that the common names of orders were in previous versions of the glossary. Actually, they are in the fields English / FrenchDef / PortugueseDef / SpanishDef (table Glossary) that give definition of the family common names. From the glossary it is also possible to extract the French, Portuguese, and Spanish family names that are entered (table GlossaryFamilies in the ENBI_WP11_v1_031016.mdb).

Other questions:
- Are the English FishBase family names in the table Families (field CommonName) consistent with those of the Glossary?
- Are the order common names consistent between the records of the family of the same order in the fields English/XXXDef?
- Are the English FishBase order names in the table Orders consistent with those of the Glossary?
- Is there a mechanism in data encoding insuring that when a record is added or modified in one the fields concerned above, there is a check in the other related tables?
- Is there a simple way to extract from the Glossary the list of the order common names for each of the 3 languages? The problem here is to make a link with the scientific name that should be extracted from the English definition field.
- Layout: the capital initial is inconsistently used in the different fields TermXXX.

1.2. Families

The discussion has been on the following issues:
- Missing family common names in a given language?

We discussed the possibility to use the English name or the scientific name instead, but we agreed better to make an effort here during this project to choose a common name.
Nevertheless, it does not solve the problem when a new family is created. The time that the different translators send a common name in their own language, we have to choose a name to be put in the translated text.
During the workshop we decided to choose the English name, and if does not exist the scientific name. I think it is not good. In almost all languages, we have the possibility like in French to use the termination "idés" instead of "idae", or in English "ids". The problem here is that we did not see that a "real" common name is missing. If it is possible to mark it, let's use this solution rather than the English or scientific names. But if it not possible to mark a missing real common name, I rather prefer to use the scientific name alone or followed in parentheses by the English name, if possible, both in italics in the translated text (but is it possible?).
Here is a summary of the discussion how to proceed to choose a common name. It can be used either for the order common names.
First, try to find an existing name, especially through "official lists", field guides, aquarium guides, and Internet.
Second, if possible, look for the species common names of the main genera and select the one used mostly, especially for species of different genera.
Third, translate the English term.
Fourth, translate or "Frenchise, Portuguesise, Spanishise" the scientific name (like in French labridés for Labridae).
Fifth, still 74 families miss English terms. What to do with them? Maybe contact Jo Nelson to check if he has advanced a work planned for the possible 4th edition of Fishes of the World.

- Subfamilies.

There are 518 families in the table Families of FishBase, but the file prepared by B. Überschär proposes 554 terms. It seems that the added names are very often used as subfamilies in FishBase, where they can be found as families in other works, especially those not following the Catalog of Fishes. How do we deal with that problem? With a subsequent one that may appear, where we may have to give slight different names for the family and the nominal subfamily in the Glossary. Should we stand now for the FishBase strictly (518). Or should we try also to give names for subfamilies? It is not a problem for translation I suppose, although ambiguities may appear during the translation if we don't know if the target of the text is the family or the subfamily (I am not sure if it happens really).

- Several common names for a family.

We discussed about the few inconsistencies and signification of the different layouts used when several common names are used for one family. It is not clear if a comma means synonyms, or the list of names used for different parts of the family (often slight different subfamilies or genera). Should be clarified and explicated, especially, should we add names here (it has been done in the German example). [Note that since the workshop, it has been asked not to used & and ().]
But more important, and it has not been discussed, we have to make a file translating only word by word, because in a text, only one of the names will be used, not all. We did not discuss that during the workshop. Or will it be automatically extracted (is it possible?). Should be discussed and clarified.

Note that:
- There are some control characters in the English names in the file given on the CD-ROM.
- The English names start almost consistently with a capital initial, even when it is an adjective. It is note the case for German. What should we do?

Other questions:
- Is there a simple way to extract the list of the family common names from the Glossary for each of the 3 languages where some exist already? The problem here is to make a link with the scientific name that should be extracted from the English definition field.
- See above related question for Orders.

1.3. Countries, Islands

The standard to use for country names is the ISO 3166-1 short name (two or three name must be simplified from this standard). It should be available in each language in the national standardisation agency, at least to check the correct name and spelling (be careful of name recent changes, like Burma in Myanmar, east European countries, pacific island states like Nauru and so on).
The second part of this standard (3166-2) gives the administrative entities of the country up to 2 levels if I remember (which corresponds to Région and Département levels in France). I think it is less useful, because the names are given only in the national language. Nevertheless, could be consulted for archipelagos or islands that are also administrative entities.
The 3rd part (3166-3) contains less than ten old country names like Czechoslovakia that we may need to translate as well.

Other questions:

- In the excel file delivered for the workshop, some entities are reported like Alaska (USA) where it is only Alaska in the field Paese of the table CountryRef. I am not sure that it should be added. To be discussed and clarified.
Here we have a problem between translation and signification of the word. We cannot have at the same time a field for a translation that gives the political entity from which it depends.
Also, in main texts we won't find Alaska (USA), but Alaska alone. I understand that there is a problem here for island with the same English name, that may have different translation (but is it true ?).
So from FishBase CountryRef table, Systran will not translate "Alaska (USA)" as a whole, but "Alaska" and "USA" separately.
- That leads to another question: what about the abbreviations like USA, UK, EU, etc.? To be discussed.

1.4. Ecosystems
No issue discussed.

Other question:
- What about the abbreviation "St" for "Saint"? I am not in favour to use the abbreviation because of sorting problems? Will Portuguese use constantly São and Spanish San?

1.5. Labels
Explanation by Eli on the technical aspect of the translation of the labels.

Other question:
- Some of the terms can be found in the Glossary. The glossary can be imported in the Systran dictionary, at least the relevant fields (to help the work for French, Portuguese, Spanish). Note that there is difficulty with suffixes in the Glossary beginning by a dash "-" (three terms) when using Excel, because the copy/paste from Access to Excel transforms these terms in Excel function. An apostrophe has to be added before the "-" in Excel before the importation from Systran.

2. Dynamic translation
Discussion on the quality of the tests without dedicated dictionaries.
Discussion on the necessity to correct and standardised the comments field in FishBase before extracting the words/sentences to be translated.
Avoid to discuss about the glossary work, i.e., we are not here to give definitions from our own, but focus on the translation of the existing one.

Other remarks
- After some test with Systran, one of the main problems is the loose use of articles "a" and "the" in English. Especially when the style is telegraphic. The results are quite bad in French, and maybe in other language, when these type of articles are required apart from locutions like "un peu de coton".
- We did not discuss if we want to add html tags in dictionaries. I assume that Systran can deal with them properly. But what about the extraction of terms from FishBase, is it not a problem?

Elisabetta Betulla Morello, CNR Istituto di Scienze Marine Sezione Pesca Marittima (Italy),

1. Manual translation

· No real problems were envisaged with respect to manual translation of static fields such as labels ecc., if not that in some cases it is virtually impossible obtain succinct translations into Italian language.
· Family and order common names have no real meaning in Italian language, these fields will, thus, be translated directly from Latin, as is usually done.
· Problems could be encountered when translating common species names. It often happens that one species will have more than one “main” common name (e.g.: Dicentrarchus labrax is translated to ‘Brazino’ or ‘Spigola’ both having equal importance) or, vice versa, that more than one species will be grouped under the same common name (e.g. the official trade name of Euthynnus lineatus, Thunnus maccoyii, Thunnus albacares, Thunnus obesus and Thunnus tonggol is ‘Tonno’ for all). Adriatic and Thyrrhenian coasts often use completely different common names. There are official common names set by law for trade purposes for most commercial species but there is no national list of official unique common names. A strategy needs to be drawn up.

2. Machine translation

· The main downfall of machine translation using Systran from English to Italian is that the system does not recognise the difference between a verb and a noun in some cases. For example the word ‘fish’ is both a verb and a noun in English whilst in Italian there are two discreet words. Systran will preferably translate ‘fish’ into a verb. Another example is water. Over and above this, serious problems were encountered with the basic grammar and sentence construction of the resulting translation from English to Italian, sometimes completely unintelligible.
To solve these problems, the English text forming the template for machine translation should be simplified in such a manner as to allow Systran to recognise such differences. Guidelines for this simplification should be given to the FishBase team from each partner according to each different language, thus, leading to an overall simplification strategy. Furthermore, to partly overcome this problem, groups of words or phrases should be translated and included in the thematic multilingual dictionaries.
In order to coin such strategies, the partners should receive chunks of FishBase text (possibly the same for all partners) to use as templates for translation trials with Systran.

3. Glossary

· A strategy for the future compilation of glossaries should be drawn up. The main problem involving definitions of specific words is that the translation should not be literal from English, but should have a real meaning in the product language within each specific area of the overall subject (e.g. ecology vs. genetics).
· A template against which such definitions should be standardised is to be identified. The following text was suggested:

Lincoln, R., Boxshall, G. and Clark, P. 1998. A dictionary of ecology, evolution and systematics. 2nd edition. Cambridge University Press: Cambridge. 361 pp.

4. Suggested websites upon which to test translation model

· EUNIS, European Nature Information System

http://eunis.eea.eu.int/eunis/index.jsp
Extremely important database with information regarding Species, Habitats and Sites compiled in the framework of NATURA2000 (EU Habitats and Birds Directives), as well as a glossary of terms, all in English. Most important of all is the EUNIS Habitat types classification ("a comprehensive pan-European system to facilitate the harmonised description and collection of data across Europe through the use of criteria for habitat identification; it covers all types of habitats from natural to artificial, from terrestrial to freshwater and marine").

· CLEMAM: ChecklList of European Marine Mollusca (http://www.somali.asso.fr/clemam/index.clemam.html)
Translation of this website is suggested by the Italian partner. There is a very large "community" of 'amatorial' malacologists in Italy which publish important papers on molluscan taxonomy and distribution and which would greatly benefit from an Italian version of this website, being it an extremely important source of up-to-date molluscan nomenclature. Translation of the website should be simple mainly due to the prevalence of static terms, but fields regarding the distribution of Molluscan species may benefit from appropriate machine translation.

· Other suggested websites:
- MarLIN: The Marine Life Information Network for Britain and Ireland, with special attention on two fields: (i) Species information pages and (ii) Habitats (Biotope) Information Pages: http://www.marlin.ac.uk/
Despite not being "European"……
- CephBase: http://www.cephbase.utmb.edu/
- Crustacea Glossary of the Natural History Museum of Los Angeles County: http://crustacea.nhm.org/glossary/

Gert Boden, Africa Museum, Department of Zoology Fish Section Leuvensesteenweg 13, B-3080 TERVUREN BELGIUM (Dutch)
A. Static Translation.

FishBase will be used as a model for the 'Multilingual Translation'. First task is to make the translation for the Search-page of FishBase. This will be done by tables which will be send in excel-format. For every word, a translation will be given. The deadline for this task is 31/10/2003. In order of priority we have:

1. labels list
2. country list
3. ecosystem list
4. order list
5. family list

A rough translation of these lists can be given on 31/10/2003. For the country list and ecosystem list we can use a geographical reference. For the order list and family list, we need a good reference. But a lot of families have no common name in Dutch. We foresee no real problems for fish families distributed in Europe or fishes used in the Aquarium trade. But for the remaining families, especially deep-sea fishes, we have no common name. At this moment we can leave them blank; so that we can do some more research in specialized references, before naming them. Therefore we have to

do some more research work after 31/10/2003 in order to provide the best possible translation for some families.

Other problems:
- words with the same singular and plural form in English; while in Dutch they will have a different singular and plural form.
e.g. 'fish'; in Dutch this can be 'vis' (singular) or 'vissen' (plural).
- combination of words; sometimes a combination of words has to be used to give a better translation; sometimes two words in a language translates as only one word in another language.
e.g. 1 'dorsal'; in Dutch translated as 'dorsaal'.
'fin'; in Dutch translated as 'vin'.
'dorsal fin'; in Dutch translated as 'dorsale vin'.
e.g. 2 'anal'; in Dutch translated as 'anaal'.
'fin'; in Dutch translated as 'vin'.
'anal fin'; in Dutch translated as 'anaalvin' (in one word).
I think these problems can be solved as follows: we do a first translation, then we can look at the result it will give, and then do a second translation for those problems which appear in the result.

Question: What is the convention for the signs 'and', 'or' and ',' in the family and order lists?
- 'and': used when the family/order contains species with different common names.
e.g. Eagle rays and Manta rays: some fishes in this family are called Eagle Ray, others are called Manta Rays.
The Dutch translation for 'and' is 'en'.
- 'or': used when a family has different common names which are in fact synonyms.
e.g. Racehorses or Pigfishes: the same species in this family is sometimes called racehorse and sometimes pigfish.
The Dutch translation for 'or' is 'of'. Attention, because this word 'of' has another meaning when it is an English word.
- ',': used when more than two names are known for that certain family. It can be used in combination with 'and' and 'or'.

For the meaning of some words we can use following standard reference:
Roger Lincoln, Geoff Boxhall and Paul Clark (1998) – A dictionary of ecology, evolution and systematics. Cambridge Press.

B. Machine Translation.

The Machine Translation will be done by SYSTRAN. The Dutch translation is however not available on the CD version, and therefore we had to check it on the internet version (In the meantime, I believe this problem is solved). To compare the translated text with the original English text, we had first to find a good way to do this. We had to do some copy/paste and to change the view first, before we were able to compare both texts. With the CD version this will be a lot easier.

Other remarks on the Machine Translation.
1. The dictionary used in SYSTRAN seems to be OK. Sometimes it translates it to a wrong word with a slightly different meaning, but this can mainly be solved to give a higher priority for words with a 'biodiversity' meaning.

2. The real problem for the translation into Dutch seems to be the grammar. Sometimes this is completely different with the English grammar. In a lot of cases, the meaning of the sentence will be lost and this is resulting in 'rubbish'. So, SYSTRAN has to improve a lot on the grammatical side of the translation.
For the translation of the 'distribution', this problem can be more or less negotiated. But the grammar of a sentence is important in the 'biology' field and certainly in the 'diagnosis' field. In these cases we will have misinterpretations and it could give a wrong diagnosis of the fish. To solve this problem, we will have to make some appointments for these two fields. This can be done to standardize the English in these fields in FishBase and the Machine Translation will give less 'errors'.

ANNEX 4

Preparing English Text for MT

SYSTRAN has formulated six lists of rules which are designed to aid in the preparation of English text for machine translation.

- **Rules for the Use of Articles**
- **Rules for Lists**
- **Rules for Phrase Structure**
- **Rules for Punctuation**
- **Rules for Formatting**
- **Other Rules**

A) Rules for the Use of Articles

1. Use articles to reduce the ambiguity caused by homographs.
Instead of: *empty file*
Use: *empty the file*
Or: *the empty file*

2. Use articles to clarify the function of the present and past participles.
Instead of: *moving car*
Use: *the moving car*
Or: *moving the car*

3. Use articles and punctuation to clarify the part of speech of a word.
Instead of: *Check the lighting, electrical, and navigation systems*.
Use: *Check the lighting, the electrical, and the navigation systems.*
Or: *Check the lighting, and the electrical and navigation systems*.

B) Rules for Lists

The use of the article is especially useful in the identification of list items. Also, it aids greatly in translation to have each item of the list stated in the same manner.

1. Use articles at the beginning of each item in a list.
Instead of*: the brake and tail lights*
Use: *the brake light and the tail lights*

2. Use articles to show that only a certain item in a list is modified.
Instead of: *the brake pedal and accelerator*
Use: *the brake pedal and the accelerator*

3. Write list items as clauses or full sentences whenever possible; don't mix sentences, words, and phrases.
Instead of: *Rotate the wheels, lubricate, clean head.*
Use: *Rotate the wheels, lubricate the joints, and clean the head.*
Or*: Rotate the wheels. Lubricate the joints. Clean the head.*

C) Rules for Phrase Structure

1. Repeat nouns or noun phrases instead of using pronouns; avoid particularly the pronoun "it"
Instead of: *Wash the car, clean the windshield, and then wax it*.
Use: *Wash the car, clean the windshield, and then wax the car.*

2. Put phrases as close as possible to the nouns that they modify.
Instead of: *Engine cover for sale by elderly gentleman with a few bolts missing.*
Use: *Engine cover with a few bolts missing for sale by elderly gentleman.*

3. Use the word "to" or an auxiliary verb to indicate the infinitive form of a verb and to distinguish it from a finite form of the verb.
Instead of: *Rotate the wheel to clean and then lubricate the head.*
Use: *Rotate the wheel to clean it and to lubricate the head.*
Or: *Rotate the wheel to clean it, then lubricate the head.*

4. Arrange sentences to minimize ambiguity.
Instead of: *Remove the bolts holding the assembly with the left hand.*
Use: *Remove the bolts, which hold the assembly, with the left hand.*
Or: *To remove the bolts, hold the assembly with the left hand.*

5. Don't leave out subordinate clause markers (that, which, who, etc.).
Instead of*: Make sure you select the proper tool.*
Use: *Make sure that you select the proper tool.*

C) Rules for Punctuation

Punctuation is essential for dividing sentences into their logical components, allowing for their correct interpretation.

1. Separate two main clauses with a comma followed by "and," or with a semicolon.
Instead of: *Check the figures, verify the test results.*
Use: *Check the figures, and verify the test results.*
Or: *Check the figures; verify the test results.*

2. Set off subordinate phrases and subordinate clauses with commas.
Instead of: *After you have checked the lights, the brakes, and the steering make a report.*
Use: *After you have checked the lights, the brakes, and the steering, make a report.*

3. Put commas after all prepositional phrases that begin sentences.
Instead of: *During the landing personnel should remain seated.*
Use: *During the landing, personnel should remain seated.*

4. Use commas to set off embedded clauses.
Instead of: *Wolf's analysis (which supports this conclusion) is scholarly and detailed.*
Use: *Wolf's analysis, which supports this conclusion, is scholarly* and

detailed.

5. Use parentheses where the material enclosed in the parentheses does not have a close logical relationship to the sentence.
Instead of*: Burke's discovery, see page 23, supports this conclusion.*
Use: *Burke's discovery (see page 23) supports this conclusion.*

6. Limit the use of the slash.
Instead of: *at the beginning/end*
Use: *at the beginning or at the end*

7. Hyphenate phrases that modify other words or phrases.
Instead of: *man eating shark*
Use: *man-eating shark*

8. Do not insert hyphens to break words that fall at the end of a line; do not hyphenate, or use "soft" hyphens instead.
Instead of: *If you hyphenate the words that fall at the end of a line to try to get an even right margin, then the program will look for each half of the word in the SYSTRAN dictionaries or UDs. Depending on the word fragments, it will either list both parts of the word as separate Not-Found Words or assign incorrect or partial definitions to the word.*

Use: *If you hyphenate the words that fall at the end of a line to try to get an even right margin, then the program will look for each half of the word in the SYSTRAN dictionaries or UDs. Depending on the word fragments, it will either list both parts of the word as separate Not-Found Words or assign incorrect or partial definitions to the word.*

9. Avoid the use of a dash as a punctuation mark, use other punctuation instead.
Instead of: *Because the data were incorrectly analyzed – the reason for which will be discussed later – the wrong conclusions were drawn.*
Use: *Because the data were incorrectly analyzed (the reason for which will be discussed later), the wrong conclusions were drawn.*
Or: *Because the data were incorrectly analyzed, the reason for which will be discussed later, the wrong conclusions were drawn.*

D) Rules for Formatting

SYSTRAN requires certain indicators in order to identify the ends of sentences and of paragraphs. If these indicators are not in place in the corpus to be translated, the program may not be able to determine this information and, as such, it will not provide accurate translation.

1. Use two spaces at the end of all sentences and after a colon.
Instead of: *These are the critical areas: development, production and marketing.*
Use: *These are the critical areas: development, production and marketing.*

2. Use one space after abbreviations, commas, and semicolons.
Instead of: *Mr. Smythe*
Use: *Mr. Smythe*

3. Use the word wrap or the soft return feature of your word processor for all sentences within a paragraph, instead of inserting hard returns.

4. Use a hard return at the end of each paragraph.

5. Use two hard returns after all titles and headings, unless they end in a punctuation mark.

6. Use indents, tables, and tabs instead of many spaces.

E) Other Rules

1. Use abbreviations consistently.
Instead of*: a 3 min. min.*
Use: *a 3 min. minimum.*
Or: *a 3 minute min.*

2. The use of different fonts helps to draw the eye, but the program has no way of identifying fonts. Use other methods to set off text for machine translation.
Set off names, such as key names, icon names, and functions; by punctuation, usage, or case rather than by font alone.
Instead of: *Press enter.*
Use: *Press "ENTER."*
Or: *Press the ENTER key.*

3. Single words or acronyms that are not to be translated should be preceded by a period.
Instead of: *I work for the CORE.*
Use: *I work for the .CORE.*