ENBI – Work Package 11

2nd Workshop on "Multi-lingual Access to European Biodiversity Sites"

Institute of Marine Research, Kiel, Germany

20. – 21. September 2004

Workshop Report

Introduction:
Following an agreement of the translation consortium at the first workshop, a second interim workshop on "Multi-lingual Access to European Biodiversity Sites" was scheduled for August/September 2004 and successfully conducted from the 20.–21. September 2004 at the Leibniz-Institute of Marine Research, Kiel, Germany, with translation partners from six European countries. The major purpose of this workshop was a rather pragmatic approach to machine translation, taking advantage from the subsequently available online access to the EC-MT-Systran® translation system. Major issues were (a) to test the effects of source text variation on the translation results, (b) to compile guidance for source text in various languages (c) to monitor the effects of the user dictionaries which are being implemented into the EU-MC Systran® system and (d) to compile rules and recommendations for the treatment of terms for the user dictionaries.

A detailed workshop agenda is attached as ANNEX 5 to this report, a list of participants is attached as ANNEX 6.

Since two new participants were invited to attend the workshop (from Greece and Sweden) 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. Although Swedish is at present not among the target languages of the ENBI-project, the Swedish participant was invited because the Swedish FishBase team has expressed their great interest to implement machine translation for FishBase for English to Swedish as soon as possible.

Following introduction of the participants, the workshop started on Monday morning with a presentation of **Bernd Ueberschär** on the progress in ENBI and specifically WP 11 since the last workshop in Oktober 2003, entitled:

- ***General Introduction on the progress of tasks in ENBI WP 11 since the last workshop: - "Multi-lingual Access to European Biodiversity Sites-."***

The work package leader informed the translation partner about the progress of the translation project and reported on the following major issues of interest for the translation consortium in work package 11:

Three major dictionaries for the EU-MT-translation system, extracted from FishBase with the following topics were compiled and submitted for translation: "Biology", "Distribution" and "Morphology". The dictionaries contain each 300 – 1200 terms (+ ca. 20,000 species names) which are not included in the standard dictionaries for Systran. Portuguese and German translation are almost complete, other languages are under construction. A revision of the translated lists under consideration of the results of this second workshop will be necessary. This tasks will be completed until October 2004.

A trial version of EU-Systran for website translation "on the fly" is available now in the Internet and under evaluation for WP-11 (https://mt.cec.eu.int/ecmt/Login.do, restricted access). This trial includes the consecutive integration of custom-made dictionaries (produced from WP 11). The system is expected to be open for public use (other biodiversity than FishBase websites may use the service) from December 2004.

A beta version of machine translation using Systran´s free webservice is still shown in FishBase and open for the public (www.fishbase.org). A graph was shown, which demonstrates that the multilingual access to FishBase since November 2003 has caused a pronounced increase in Hits to FishBase from developing countries.

Four, partially multilingual, public domain glossaries in the Internet, related to the subject biodiversity, were "deep-linked" to FishBase. Search in 4 additional Internet glossaries beyond the genuine glossary of FishBase is feasible. Tapping of other glossaries is an ongoing effort until February 2005. Manual translation of the FishBase glossary of terms into German (WP 11 leader) is under construction, decisions on (manual or MT)translation into other languages are pending (subject at the workshop in September 2004).

A contract is in place (in accordance with the contractual agreements of WP 11 in ENBI) with FIN (FishBase Information and Research Group, Inc.); the specific purpose of this contract is to establish multi-lingual access to common species names in the Internet in 8 languages (English, German, Dutch, Spanish, Portuguese, French, Greek, Italian). The FAO lists of common names (containing also common species names beyond finfish) will be entered into the Species 2000 system, organized and funded by WP 11 and members of the FishBase Team, (WorldFish) start August 2004. This task will be completed in February 2005.

An article about machine translation matter was compiled on request for the first ENBI-Newsletter: Still a Challenge: Machine translation (MT) in the 21st century (10 pages). The article is available as downloadable file from the CIRCA site.

The WP-11 leader has attended the semi-annually ENBI-Steering Committee Meeting in March 2004 in Chania, Crete.

ENBI-WP 11 Website was continuously updated (responsibility coordinator of WP-11). http://www.enbi.linguaweb.org/

Bernd Ueberschär gave a second presentation entitled:

Experience with Manual and Machine Translation. Issues related to Dictionary compilation. Introducing Guidance for Source Text. Introducing Experience and Techniques how to translate. Status of the Cooperation with the EU-Translation Department

This rather technical talk informed the translation partner about the latest progress in relation to the cooperation between the EC-MT-System and the ENBI-project. Guidelines, which were compiled for source text and special dictionaries (joint effort from WP 11 and the EC-MT department (MT team, European Commission Directorate-General for Translation, Unit D.03 - Multilingualism and Terminology Coordination) were introduced and discussed. The technical talk was also intended as preparation for the following pragmatic exercises with the EU-MT System (effects of verifying source text on machine translation "on the fly", paragraphs with free text and websites)..

The FishBase coordinator **Rainer Froese** gave a presentation on the topic:

Fishbase and Non-European Languages, Future Plans.

In his presentation, Rainer Froese put emphasis on the fact, that not only translation is an important issue for information systems, but also the option to enter common names in the search routine in other scripts than Roman. FishBase offers common names in several scripts and obviously attracts now many more users from countries with other scripts than Roman. Rainer Froese presented a comparison on usage of FishBase in 2001 and in 2004 for countries which are using other scripts than Roman (e.g. China, Japan, Saudi-Arabia, South-Korea, Russia etc.). Many of those countries appear the first time in the statistics in 2004, and some of them moved much closer to the line of mean usage. One criterion to implement a new script in FishBase is the number of common names available in those script (minimum is supposed to be 100 names?). Some aspects on the maintenance of manual translation in FishBase were discussed with Rainer Froese. Options were discussed for facilitating future 'small' translation jobs, such as necessary whenever a title, label, foot note or choice field changes in any of the translated pages. Basically FishBase will give the translators online access to the translation table and have a semi-automatic email system that alerts translators whenever a term is added to the translation table.

Workshop Outcome: Some important outcome from discussions and exercises around source text, dictionaries and machine translation is summarized in the following section, follow-up comments from workshop participants were considered. The final notes are prepared in close cooperation with Cameron Ross, the responsible partner for WP-11 affairs in the EC-MT-System.

- Telegraphic style ruins the machine translation. A major finding of the exercises in the workshop was, that the way FishBase texts are written, a subject or verb is often missing, causes the program to misinterpret the context and/or the information. It is suggested to revise the free text in FishBase for more complete sentences. This task will be followed up by the WP 11 coordinator. Some more details about this issue are given in ANNEX 7.

- Context sensitivity is another major problem for translation. Terms have different meaning when appearing in another context. That is the reason, why categories are an important and powerful feature in machine translation technology. For the information systems which are treated as trials in ENBI-WP 11 special categories will host the translated list of terms (e.g. Fisheries, Environment, Biology) the translation engine will be advised to access those resources first for proper translation of the specific words when they appear in the related context (e.g. "stock" has a different meaning in relation to the category "Fisheries" compared to the category "Business").

- Since the English language is often not sufficiently precise, it is obviously required to replace English source text to facilitate machine translation. Some words have two meanings in English but not in other languages. An example is “to feed” which means both “to offer food to somebody else” and “to eat”. This results in very ambiguous translations. The easiest thing to do would be to substitute the problem term with a simpler one (or: rather more precise one) in the English source text (there is no option for an English-English dictionary), in this case feed replaced by "eat" would result into a precise translation. Some of those problems in that context are liable to be associated with the lack of subjects in sentences. "Feeds" is an example of homography, where "Feeds on XXX" could indicate a noun or verb.

- How to apply context-related translation. As mentioned, many words have a different meaning depending of the context in which they are used. Often the translator does not know what translation will apply to an individual word which appears in the translation list without any context. Thus, at least for the lists of words from FishBase it makes sense, to check the context in which those word appear. This can be done by either with the "search" function in the word-document which was delivered along with the lists and which contains all free text paragraphs from FishBase, or with e.g. the Internet search engine "Google" (other search systems may be applied as well). Just type the word in combination with FishBase and the result will be the species summary, mostly at the first position on the lists of search results, where the word can be considered in the real context. This is also a useful procedure for unknown words. In summary, when proposing a translation, it should be taken into account whether a proposed translation will work in all circumstances in the texts.

- The character of a word is an important information for the translation engine. It is helpful if the translator deliver, along with the translation, the respective character of the term, such as noun, proper noun, adjective, verb. This attribute helps the technical team of the EC-MT department with the (manual) encoding process for the user dictionaries.

- Sometimes the English language uses two or more words for a term, which is in another language only one word. On the other hand, English terms could not be translated in only one word, because there’s no proper word for it, e.g. Great Britain (2 words) – in Dutch: Groot-Brittannië (1 word). The same applies e.g. to “raker” in “gill-raker” or "mid-water" (gill-raker and mid-water should be one entry). Expressions (combination of several words) are certainly welcome as it gives the computer a clue as to how to translate a word in context. Another example in French is "rendre" -> "make" but "rendre un avis" -> " give an opinion". ". Basically, wherever we consider a group of words as being one semantic concept, then we have to keep them together (noun noun, verb object, verb preposition object, etc.). One other example would be "to feed on". In general, expression coding is quite powerful and good results are possible.

- How to treat Family names. Some families do not show English translations in FishBase (however, I pushed an effort from Rainer Froese and Joseph S. Nelson to add more English names for families in FishBase) only Latin names are given which

cannot be translated. In those cases it is advisable to keep the Latin name in. Also some English family names have not yet corresponding translation into other languages. In that case, keep the Latin name and add the common family name in English in parenthesis. This is acceptable for the translation engine.

- "Latinized" words in other languages than English. Not all "latinized" words can be translated. In English, e.g. "Cnidarians" (from the Latin word Cnidaria), have no matching word in other languages, e.g. Dutch has no "latinized" words. In that case the Latin word has to be applied as translation. However it is finally in the responsibility of the translator how to manage this item in his/her language, with a recommendation that the common name is given in the translation.

- Translation of common names. Many common names in English have a corresponding common name in other languages. However, sometimes the translation might be misleading. On the other hand, translation is supposed to be as complete as possible, many user may not be able to understand any English common name. So, it would be useful to show th latin name plus the sounding common name in parentheses. The following rules have to be obeyed: If we enter "Crangon" = "Crangon (Sandgarnele)" then that translation will always appear. But we could solve the problem in the source text: the first time we use the Latin term in English, and we put the English common name in parentheses too. Then in the terminology file, we have to indicate the English common name plus its translations. Example: In the terminology file: "tursiops truncatus = tursiops truncatus" + "bottlenose dolphin = dauphin souffleur". In the text, first to mention: "tursiops truncatus (bottlenose dolphin)" which should be translated as "tursiops truncatus (dauphin souffleur)"; thereafter just "tursiops truncatus" which will be translated simply as "tursiops truncatus".

- Capital letters: for the sake of automatic coding, it's better to use lower case for entries unless they are really proper nouns or the translation in the target language requires it (e.g. German). If a word can appear as lower or upper case, just enter the lower case. For the automatic dictionary, entering only a word in upper case instructs the machine to match only an upper case version of the word in the source text. If entering "Football" -> "Rugby", that entry would not be matched if the source text contained only "football". If entering "football -> rugby" in the dictionary, it should work for "football" or "Football".

- Abbreviations can certainly harm the translation if they're not recognised by the system: the "." can be interpreted as an end of sentence. So it would be helpful to include abbreviations which are used frequently. Abbreviations are entered with there final dot when necessary, like aff., sp. Ref., because it may confuse the machine, e.g.: “(Ref. 2834) Status of protection …” is translated in “(statut Réf. 2834) de protection …”.

- Some words are doubled with singular and plural: do we have to enter both forms of terms, the singular and plural (once needed in relation to the source text)? Automatic coding is expected to guess that the word is a noun, recognise the word in the plural in English (assume ....s) and then apply an appropriate plural ending in the translation language. This is less of a problem than correctly recognising "amphihaline" because in the absence of other information, the system will assume that an entry is a noun, and since most English words add -s in the plural, the plural form is easy to recognise. On the translation side, pluralisation is not so easy and there is no guarantee that the correct ending will be applied, but it should be obvious it's a plural. It does not help to enter both singular and plural forms It was tried in test, and it didn't help: the plural form seemed to be overruled by the singular form. So just put the singular.

- In the distribution file it is needed that combination of country names such as New Caledonia are kept together as an expression, instead of having New translated in one line and Caledonia in the next one. WP 11 will distribute a corrected file "Distribution" to the translation consortium.

- Misspellings in the English version of the word lists from FishBase. Request to translator: please gather words which are misspelled and send the list to the WP-11 coordinator (Bernd Ueberschär) The files will be collected and passed finally to the FishBase-team for correction in the source text in FishBase.

- Some English common names refer to groups of species from several families and there may be no corresponding term in other languages: this may render translation impossible. An example is Basslets.

- Other information systems shall be considered for translation trials. Since the major output of WP-11 is the compilation of a technical paper on translation techniques and strategies for biodiversity information systems, one other system than FishBase shall be included into the translation trials. This may help to gather even more broader exercise and would certainly enhance the final output of WP-11. Thus, other sites for translation on top of FishBase were discussed in the workshop. The CLEMAM website, proposed by the Italian partner in the previous workshop, was considered as a possibility as well as OBIS, the manager of the latter expressed high interest for translation to Rainer Froese at a recent meeting.

- Do bold letters have an impact on translation? Words in bold shouldn't pose a problem. However, the formatting is not always respected in the translation.

- Options for terms to be added to the given list: If the translation partner consider one term not explained enough in one translated form, you may add a term to the dictionary, example "Aral" appears in the dictionary and may relate to "Aral pubfish" in Fishbase, in other positions it appears as Aral sea (more often). So, just add the English term "Aral Sea" and the translation in your language. Please mark the added terms in your dictionary!

- The adjectives are to be entered with masculine gender, we assume that the EC-MT will automatically make the changes, or should we give indications. For instance, does the system know if "amphihaline species" will be translated in "espèce amphihaline" when we have entered amphihaline 'amphihalin' as the masculine only? It is correct to a noun, adjective, verb, it should always be entered in its base form - so "amphihalin" in French for the example. Automatic coding is then supposed to guess that the word is an adjective in English and apply the correct ending in French when the adjective is associated with a feminine or plural noun. So just leave it in the basic form. Of course, the guessing is limited and doesn't always resolve the part of speech correctly - in a test, "amphihaline" was assumed to be a noun: "amphihaline species" became "espèce d'amphihalin". So if an adjective like that is is generally associated with a particular noun, best to include the phrase e.g. "amphihaline species" -> "espèce amphihaline" to make sure the correct translation.

- The machine translation interface of the EC-MT-Department is now open for testing at the following address: http://mt.cec.eu.int/ecmt (restricted access). This very useful option was intensively applied during the workshop. As a workshop follow-up, all translation partner were asked to continue testing of source text translation applying the EC-MT-system for individual results in each language. To enable participants to work with this system, the WP-11 leader has send a request to the EC-MT-Department to allow access to the system for all translation partner. Please note, that,

at present the current interface does not offer domains for online translation. However, the Partner in the EC-MT-Department will be looking at ways of combining online translation with specific terminology, but for the moment it's necessary to extract text and run in batch mode in order to access domains (which is important in order to monitor the effect of the increasing number of translated terms in the dictionary resources).
Following the latest news from the EC-MT-Department, partner form the ENBI-Project get the permission now to access the EC-MT translation system for further trials from their own Computer. To register for the MT service, please go to the new interface of the MT-Service at: https://mt.cec.eu.int/ecmt. The user will be asked to create a password and to give an e-mail address which (since the service is not for private use) must be in the public domain - institute addresses are fine.

Other Items: Medium to long-term support for manual translation.

- To maintain the accuracy and completeness of static translation components within an information system, translation of static terms (labels etc.) requires a long-term commitment of translation partners in the future beyond the limited period of this specific project. Thus, the translation partner of the ENBI-project were asked to consider a medium to long-term commitment which secures the update of added items, at least for the FishBase system. The Italian and Portuguese partners declared already their support and are willing to grant full cooperation for future static translation tasks. Partner for other languages are welcome. To support this cooperation and to make it easy to update the concerned systems it is suggested to develop an automatic tool to warn translators (persons) when there are new items to translate. Nicolas Bailley is willing to initiate this technical feature together with the FishBase team in November 2004.

ANNEX 3

Specific comments of the Workshop participants (not edited).

1. Gert Boden
Koninklijk Museum voor Midden-Afrika (KMMA)
Tervuren, Belgium

Comments / Questions / Recommendations
Translation to Dutch

A. Static Translations.

FishBase is used a a model for the 'Multilingual Translation'. For the static translation some tables were sent in excel-format. Translation for these tables is finished.

However, changes can be made in those static pages and the static translation has to be maintained. Volunteers can have full access to the tables through RDE. An interface to do the RDE has to be set up. Most likely, these translations can be done by persons who have a long time connection with FishBase.

B. Machine Translations.

The machine translation will be done by SYSTRAN. The purpose for this project is to compile a good strategy for translations of biodiversity websites.

Biology, distribution and morphology fields are all translated by SYSTRAN in FishBase. An excel-table for each of these fields was sent to translate words not recognized by the dictionary of SYSTRAN. In order to get a good translation we have to assign also a category to these words.

Two major problems are linked with machine translation:
- context sensibility. This is the main problem for translation. Words have another meaning when placed in another context. Therefore a category 'fisheries' will be used in the future for translation of specific fisheries words. More in detail a category (noun/adjective/verb) to all words will be given.
- telegraphic style of sentences. Using a telegraphic style will ruin the machine translation. Because of a missing verb or subject, it will not translated properly by machine translation. Therefore it is recommended to use full sentences with a verb.

Other problems with machine translation:
- Sometimes the English language uses two or more words for a term, which is in another language only one word. On the other hand, English terms could not be translated in only one word, because there's no proper word for it.
e.g. Great Britain (2 words) – in Dutch: Groot-Brittannië (1 word).
- Names of families. Some families have only Latin names and can not be translated. Therefore it is better to have the Latin names. Also some English family names can not be translated. A solution is to use the Latin name and add the common family name between parenthesis.
- Latinized words. Not all latinized words can be translated. In English you can say 'Cnidarians' (comes from the Latin word Cnidaria), but in Dutch there are no latinized words. Therefore it is better to use the Latin word.
- Do we translate common names in the machine translation? Eventually we can use the Latin name and add more information for it, e.g. '*Crangon* (a small crab)'
- Sometimes a combination of words is necessary for a good translation. These are sometimes not considered or overlooked.

In the lists are sometimes also errors for the English word. A list of these errors can be made and given to Bernd Ueberschär. He will send this list to the FishBase Team in Los Baños.

C. The use of translations in FishBase.

FishBase now uses different scripts and also more common names in those scripts. A lot of internet users (China,...) do not know the Roman script and the scientific names. They will not use FishBase if there are no other scripts available.
What is the criterium to have other scripts in FishBase?
- the common names of all fishes of the country have to be in FishBase.

D. What about future plans with SYSTRAN?

SYSTRAN will be used by FishBase for the next 5 years as a free service, because of the ENBI-project. If FishBase provides more new words to SYSTRAN, maybe it will stay a free service.
The influence of the translated FishBase pages in now visible in the number of FishBase hits. FishBase gets more hits from the developing countries since the pages are translated.

E. Translation of the glossary in FishBase.

Have we to translate the glossary in FishBase? There are now also physical en chemical terms in the glossary. So, there are many terms in different categories, which we are not familiar with. The scope of the glossary has to be discussed then, rather than do the translation of it.

F. What to do?

1. Translation of the lists for biology, distribution and morphology.
2. Testing of the translation as a follow-up for the dictionnary implementation.
3. Make rules for a good translation.
4. Give specific features for your own language.

2. Nicolas Bailly (France)

During the 1st workshop in October 2003, we decided to have an intermediary workshop preferably to a final one.
Organisation
This workshop was very well organised like the last one, the provided facilities fitted perfectly the needs of this type of workshop.
Agenda
We followed all the agenda more or less in the hours indicated.
The agenda was very useful:
1) to have a common view of the advancement of the workpackage as a whole,
2) to precise details on the construction of the MS-Excel files to be translated,
3) to test the tool that will be used (ecmt),
4) to take some decisions on some final work to do.
Comments
Dictionaries
Some issues were raised during the translation of the files and decisions taken.
Some words are doubled with uppercase and lowercase initials: use only the word with lowercase.
Some words are doubled with singular and plural: use only the word as singular.
Some groups names are designated either by their Latin name, by their Latinised common name (like caproid instead of Caproidae), or by their common name. It was let to the responsibility of the translator how to manage, with a strong recommendation that the common name is given in the translation. For the French translation, I have decided to use always the corresponding name in French, because it is more precise in many cases, using one form for another possibly being misleading in some cases. But I will try to add common names in parentheses as often as possible., but always add the common names, possibly in parentheses.
The same decisions for other categories, like Acanthocephala in Acanthocephala (acanthocéphales) and acanthocephalan in acanthocéphale (Acanthocephala). The problem here is to know if the singular with be automatically as plural, for instance, if we enter only caproid → caproidé (sanglier); is caproids will be automatically translated into caproidés (sangliers), or if we have to enter the translation for caproids as well?
The adjectives are to be entered with masculine gender, we assume that the ecmt will automatically make the changes (?), or should we give indications. For instance, does the system know if

"amphihaline species" will be translated in "espèce amphihaline" when we have entered amphihaline → amphihalin as the masculine only?
In the distribution file, before going further, it is needed that the names with double words like New Caledonia are kept together, instead of having New translated in one hand, and Caledonia in another one.
Misspellings in English version
Gather them and send them to Bernd that will collect them to be send to LB team.
Test of ecmt
It seems that it is important that abbreviations are entered with there final dot when necessary, like aff., sp. Ref., because it may confuse the machine, e.g.: "(Ref. 2834). Status of protection …" is translated in "(statut Réf. 2834) de protection …".
The system is very sensitive to the presence/absence articles (a, the) and other liaison words (of, that), and to the correct grammar. The English version should be revised in that sense. Especially, the telegraphic style has to be reviewed, like in the stockdef of the STOCK table.
Also, sentence like cod form schools must be transformed in "The cod forms schools.", or "The cods form schools", because the ecmt does not use cod as a plural, then the translation is totally false (form being translated in that case as shape …).
Other website
After suggestion and/or presentation of 5 websites, the future EurObis one, then the CLEMAM one have been selected first to explore with their authors the possibilities.
Schedule
The current files must be translated for the 31st October deadline, in the following order Biology, Morphology, Distribution.
Conclusion
This workshop was actually needed, because it was difficult to explain by email some difficulties on the translation of the dictionaries. It will boost the achievement of the task.

3. Bettulla Morello, CNR – ISMAR (Ancona)

The workshop was extremely useful (and well-organised) for both the solution of current problems and the identification of future targets and tasks. A succinct list of themes and problems discussed is included below.

Manual translation

All manual translation tasks have been completed to date. The last file received for translation has not yet been implemented into FishBase and thus some labels on the search page are still in English.
Manual translation of static terms (labels ecc…) in the future requires a long-term commitment of the partners, beyond the time-span of this specific project. The Italian partners are willing to grant full cooperation for future static translation tasks.

Machine translation (MT)

Three lists of words have been extracted from FishBase to be implemented into the EU Machine Translation (Systran) according to the following themes: "Biology", "Distribution" and " Morphology". This is the core task for partners. As long as these dictionaries are not compiled the MT cannot be implemented. When implementing new thematic dictionaries into Systran, each term has to be coded according to what it is: noun, verb, adjective ecc… This should greatly improve the quality of the Systran product.
The Italian partners has completed the former two lists and is half way through the third. Translation of each word in the list was done according to the context/s each specific word was set in the FishBase free text. Some problems were encountered and discussed during the workshop:
in Italian, as in French, adjectives vary depending upon whether the noun they refer to is plural, singular, male or female. In many cases an adjective was set in more than one context in the FishBase free text and thus its form would change according to the noun. One translation only was provided for each word in the list. This should not cause problems to Systran, should it be able to adapt the translation to each scenario. The possibility of being able to use the EU version of Systran rather than the personal edition was discussed and implementation of this should enable verification of the above.
Some terms were impossible to translate alone. Examples:

"Snouted": this refers to a "long-snouted" species. The Italian translation will have to be of both words which should thus be kept together in the list of words to translate. The same applies to "raker" in "gill-raker"; "Great" and "Britain" in "Great Britain", "Paul's" in "St. Pauls's Rock" ecc…
This also applies to common names of species. For example in "sailfin molly" the two words should be kept together as in "southern blue whitings".
Some words have different meanings and will also be assigned to different categories according to the context they are set in. For example: "Barbel" can be one of two things (i) a noun referring to the barbel of a fish (ii) an "adjective" when in conjunction with another word as in "barbel zone". Thus the word barbel alone should refer to (i) whilst in "barbel zone" it should be translated as part of the wider context and the two words "barbel" and "zone" should be kept together. Another similar example is "Mid-water" which can be a noun or an adjective depending on the context.
The translation of family names Englicised in "id" and "ids" like ammodytids was discussed and it was decided to, where possible, use the latin name for the family e.g "Ammodytidae". Furthermore, some of these "Englicised" terms do not refer to a family but to a broader group of species e.g. "brachyurids". In this case, though, "Brachyuridae" would not be a valid alternative.
Some english common names refer to groups of species from several families and there may be no corresponding term in other languages: this may render translation impossible. An example is Basslets.
In Italian, as for most languages, categories are a must, but may not be enough: syntax and sentence construction are important. English is possibly too simple a language to use as a template language. For this reason the partners will, during the coming months, use the EU version of Systran (if permission is given) to explore ways in which the source text in English should be compiled in order to gain the best translation results. Trials were carried out during the workshop using the EU version of Systran and the general feeling was that the source text should be as expanded as possible, making use of articles and verbs in every sentence. The main problem here appears to be related to the translation of species diagnoses and distributions where the English is extremely succinct, verbs and articles missing in most cases. Systran cannot cope with this and the source text should possibly be expanded.
Some words have two meanings in English but not in other languages. An example is "to feed" which means both "to offer food to somebody else" and "to eat". This results in very ambiguous translations. Thus "to feed" in the context of a species feeding on another species should be changed to " to eat".
Wherever abbreviations such as "aff." or "ref." are to be translated or implemented into the dictionary it is important to include the full stop after the abbreviation in the translation, otherwise Systran does not manage the translation properly.

The translation of other websites of interest on top of FishBase was discussed. The CLEMAM website, proposed by the Italian partner in the previous workshop, was considered as a possibility as well as OBIS.

4. Nikolaos Lampadariou, Greece

The 2nd Workshop of EMBI was held at the Leibniz-Institute of Marine Research, in Kiel, from 20 to 21 of September. In general the workshop was very well organized and the level of hospitality was very high. The workshop room was comfortable and well equipped, offering among others the possibility for each participant to connect directly to the internet. This was particularly useful, as it gave us the opportunity to exercise online with the under development SYSTRAN Translation Machine (MT).

As I was not initially involved in any of the previous ENBI phases, I had so far only occasionally used Machine Translation and the results were most of the times rather disappointing. Thus so far, my personal opinion was that MT has still a long way to go and that at the moment, it can not be used successfully. However, after a throughout presentation from the coordinator on the state of the art of MT, I realized that MT is a means for finding quick and easy information, even if the translation is not perfect or meaningless sometimes, whereas the alternative is to not find any information at all.

Particularly useful was also the exercise section. There, the simultaneous translation of specific examples into the different languages helped me to better understand how a translation strategy for my own language should be developed. Finally, meeting with the other partners together with an overview on what has been done by each one helped us to set new feasible deadlines and to realize what has to be done within the next final period of the project.

5. Afonso Marques, LMG – Portugal

The 2nd Workshop on "Multi-lingual Access to European Biodiversity Sites" held in Kiel, 20–21. September 2004, was well organised and the Leibniz-Institute hospitality, once more, proved to be of high standards.

Anders Silfergrip and Nikolaos Lampadariou the Swedish and Greek new participants were welcome by Bernd Ueberschär.

This interim meeting was useful since it was possible to each partner, to diagnose, discuss and point out adopted solutions to some of the translation issues.

Manual translation

The Portuguese partners have completed the translation of the given excel format lists. Terms translation was accomplished observing the original context (word file text) and also using Google to easily get the meaning of unknown local terms.

The manual translation files, “Biology”, “Distribution” and “Morphology”, are complete and delivered. Although, suggested recommendations will be implemented in the near future as long as they are applicable in Portuguese.

The use of Anglicised family names like macrourid(s) was discussed and all partners decided to use the Latin name for the family e.g. “Macrouridae”.
English mode adverbs (“lly”) can be directly translated to Portuguese with no major problem, using the Portuguese suffix “mente”.
As to some double common names of species and adjectives referring noun localizations/quantity (e.g. mid-point; half-banded), the two words should be kept together to maintain intelligibility.

Manual translation of static terms in the future requires a long-term commitment of the partners, beyond the time-span of this specific project. The Portuguese partners grant full cooperation for future static translation tasks.

Machine translation

It seems that all partners will face, more or less, the same problems. These are: the context sensibility and the telegraphic style of sentences in the original English text. This is a serious problem to automatic translation. To almost languages, words could have many meanings varying the context. Categorized list of terms is instrumental to solve the first issue and the use of expanded syntax can help the machine to display the correct translated phrase.

English technical phrases can be short and clear. But often is not possible to transpose this syntax to other languages, namely Latin ones. Some investment in the source text should be implemented in order to expand it. Missing verbs and articles are not obvious to the machine and so the final result is often poor. The use, by partners, of an authorized EU version of Systran, to explore the ways in which the source text in English should be compiled, would help to improve the quality of the final result.

ANNEX 4

Date of next Workshop:

A final workshop on ENBI WP-11 will be conducted in January 2005. Since there is the intention to organize a general workshop on ENBI affairs in January in Chania, Crete (Greece), the WP-11 coordinator suggest to conduct the WP-11 workshop at the same place. There is e.g. an **ENBI-Digitisation Workshop** planned, which might be of interest for some of the translation partner and which they might join if desired. The WP-11 workshop is supposed to be scheduled for **Wednesday, the 19th of January**. I strongly suggest to take note of this date and to plan soon in accordance with this schedule. More details on the meeting schedule, hotel accommodations and flights are to be find in the following summary. Please note, that WP-11 will pay for accommodation (up to 3 nights) and flights for translation partner to Chania. However it is requested, that each partner makes his own reservation for accommodation in Chania and flights.

ENBI IN CHANIA JANUARY 2005

Hotel Information

Hotel Porto Veneziano

One of the best hotels in Chania, and one much appreciated by those members of the ENBI Steering Committee who stayed there during last April's meeting, has offered ENBI a superb deal for those participating during the week of the ENBI workshops and meetings. As some of you will know the hotel is on the waterfront in the old port of Chania and adjacent to all the best restaurants and bars of Chania. For details: www.porto-veneziano.gr

The prices per night (including breakfast) for ENBI delegates are as follows:

Single Room:	45 Euro
Double/twin room	65 Euro
Suite:	115 Euro

These prices assume that most ENBI participants use the Porto Veneziano. I cannot obtain significantly lower rates even at lesser hotels. I therefore propose that we use this hotel and provided you agree to this, then I will arrange daily transport between the hotel and MAICh.

Travel to Chania

All flights are via Athens. There are no international direct flights to Chania in winter. There are some flights to Heraklion but the three hour bus journey from Heraklion to Chania more than offsets any advantage of such flights.

The full internal return fare between Athens and Chania is **156 Euro**, but there are always reduced rate offers in January. These reductions are not announced until much nearer the time but in previous years have varied from 15 to 25%. There are two airlines offering such flights, Olympic and Aegean. There are seven or eight flights daily in winter.

I have obtained a selection of prices from some major European cities, as follows:

International Flights with Olympic Airlines. The advantage of using Olympic, is that the fare from Athens to Chania is much lower when part of a combined ticket (typically a reduction of about € 60). I was given the following quotes by Olympic, as examples:

Amsterdam – Chania	€ 351
Brussels – Chania	€ 247
Frankfurt – Chania	€ 300
London – Chania	€ 297
Paris – Chania	€ 355

These examples are for return flights departing Sunday 16th January, returning Sunday 23rd January but there is not much difference for other days or shorter periods provided that a Saturday night is included.

Other International Flights. If you don't want to use Olympic (reliability has been much better lately!) or if Olympic does not fly from an airport convenient to you, then here are some other examples I found on the www. Note that these fares are to Athens.

Copenhagen – Athens	(SAS)	€ 132
London – Athens	(BA)	€ 200
Prague – Athens	(Czech)	€ 300

Two real bargains:

Berlin (Schönefeld) – Athens	(EasyJet)	€ 38 (+ taxes)
London Gatwick – Athens	(EasyJet)	€ 39 (+ taxes)

All these are prices obtainable now. Olympic recommended booking at once. You do not have to pay until one week before and there is no cancellation charge. There may be cheaper flights later but it is a lottery. EasyJet you can book on the web but cannot cancel: their prices usually go up as the date approaches. There must be many other deals available.

Dont ask me to explain the differences – why it cost three times more to fly from (much nearer) Paris than from Copenhagen, for example. Strange are the ways of airlines.

TENTATIVE SCHEDULE IN CHANIA

Sunday 16th January	Excursion
Monday 17th January	ENBI Continuation Workshop

Tuesday 18th January	ENBI Continuation Workshop
Wednesday 19th January	WP6, WP7, WP11, WP13(?) Workshops*
Thursday 20th January	**ENBI Digitisation Workshop**
Friday 21st January	**ENBI Digitisation Workshop**
Saturday 22nd January	**Excursion**

The dates in bold are now fixed.
(* parallel sessions; to be confirmed)

I've suggested two excursions for the benefit of those not staying for the full period. They might be the same excursion repeated. We can decide this later.

Once WP leaders have confirmed (or otherwise their attendance) I will send more information, for example about hotel booking and a more detailed timetable.

Chris Johnson (cjohnson@maich.gr)
Chania, October 4, 2005

ANNEX 5:

ENBI – Work Package 11

2nd Workshop on "Multi-lingual Access to European Biodiversity Sites"

Leibniz-Institute of Marine Research, Kiel, Germany

20. – 21. September 2004

Workshop Agenda:

Monday 20.09.2004:

- **10:00am:** Welcome and General Opening Notes, Coffee, Refreshments, Tea and Cookies available, Introduction of Participants.
- **10:15am**: Presentation **Bernd Ueberschär:** -General Introduction on the progress of tasks in ENBI WP 11 since the last workshop: - "Multi-lingual Access to European Biodiversity Sites-."
- **11:15am:** Translation Partner (10min. each). Reports, Comments, Discussion.
- **12:30pm:** Lunch (Kantine Landeshaus, ~10 min. to walk)
- **13:30pm:** Presentation **Bernd Ueberschär**: Technical Talk: Experience with Manual and Machine Translation. Issues related to Dictionary compilation. Introducing Guidance for Source Text. Introducing Experience and Techniques how to translate. Status of the Cooperation with the EU-Translation Department.
- **14:30pm:** Discussion
- **15:00pm:** Coffee Break
- **15:30pm**: Experience, Exercises and Application of Translation Techniques (please provide your personal experience, questions etc.), all Participants.
- **17:00:** Contract Issues, Processing of Reimbursement Forms, Billing.
- **17:30:** Return to Hotel (Collective Transport, for those who want to change the Dress)
- **19:15:** Departure from Hotel to Dinner Location (Official Workshop Dinner, Restaurant Louf, close to the Place for Lunch) Collective Transport.

Tuesday, 21.09.2004:

- **9:00am –** General Discussions, adjustment of Workshop Schedule in accordance with Partner Travel Schedule. Coffee, Tea and Cookies available.

- **10:00am:**. Presentation **Rainer Froese**: FishBase and Non-European Languages. Translation Issues in the Future (e.g. remote Interface for manual Translation)

- **10:45am:** Discussion

- **11:00am:** To continoue: Experience, Exercises and Application of translation Techniques (please provide your personal experience, place questions etc.), all participants.

- **12:30pm:** Lunch (Kantine Landeshaus, ~10 min. to walk)

- **13:30pm:** Workshop Outcome, Recommendations and Final Notes, Dates and Location for the Final Workshop (recommendation: Chania, Crete, End of January 2004).

- **14: 00pm:** Workshop officially closed.

- **For those who stay until Wednesday morning: If desired, open for further Discussions, visit of FishBase Coordinator Rainer Froese, Library Visit, Various other Issues, joint Dinner.**

ANNEX 6:

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

Nikolaos Lampadariou, 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.geomar.de,

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

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

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

Anders Silfergrip Anders Silfvergrip, Department of Vertebrate Zoology, Swedish Museum of Natural History, POB 50007, SE-104 05 Stockholm, SWEDEN, e-mail andesilf@imap.nrm.se

ANNEX 7:

ENBI Translation Project: Some Text-Editing Examples (specifically for English Source Text to German)

The way FishBase texts are written, a subject or verb is often missing, which causes the program to misinterpret information. For example, in FishBase texts there are examples e.g. of words *feeds, forms*, and *leaps* which are assumed to be nouns, not verbs, because there is no subject; *present* is also taken as a noun rather than an adjective because there is no verb – *is present*. English has a lot of words which can represent several parts of speech, and a telegraphic style denies the computer information (clues) it needs to disambiguate correctly.

Included below are some examples of original text and edited text, with the different translation results. Subject nouns and verbs have been inserted. Even when there is no ambiguity such as that described above, inserting words can still improve the arrangement of a translation, making it clearer.

Writing full sentences. Telegraphic style prevent good results!

(1)

Diagnosis: Head large without deep occipital groove.

→

Diagnose: Ohne tiefe occipital Rinne großer Kopf.

Diagnosis: head is large and without deep occipital groove.

→

Diagnose: der Kopf ist groß und ohne tiefe occipital Rinne.

(Although not in the English text, the article “der” was inserted automatically here.)

(2)

Gill membranes fused to the body and isthmus. Superior trunk and tail ridges continuous, inferior trunk and tail ridges discontinuous, lateral trunk ridge confluent with inferior tail ridge. Brood area of male located under trunk.

→

Kiememembranen, die zum Körper und Isthmus durchgebrannt werden. Überlegenes kontinuierliches, untergeordnetes Kabel der Kabel- und Endstückkanten und unterbrochene, seitliche Kabelkante der Endstückkanten, die mit untergeordneter Endstückkante zusammenfließend sind. Brutgebiet von Mann, das sich unter Kabel befindet.

Gill membranes are fused to body and isthmus. Superior trunk and tail ridges are continuous, inferior trunk and tail ridges are discontinuous, and lateral trunk ridge is confluent with inferior tail ridge. Brood area of male is located under trunk.

→

Kiememembranen werden zu Körper und Isthmus durchgebrannt. Überlegene Kabel- und Endstückkanten sind kontinuierliches, untergeordnetes Kabel, und Endstückkanten sind unterbrochen, und seitliche Kabelkante ist zusammenfließend mit untergeordneter Endstückkante. Das Brutgebiet des Mannes befindet sich unter Kabel.

Note on definite/indefinite articles. In most of the examples, omission of articles in the English does not really pose a problem for understanding the text, so these could be ignored to save time. However, they can also have an impact. In example (2) here, insertion of articles improves understanding:

The gill membranes are fused to the body and isthmus. The superior trunk and tail ridges are continuous, the inferior trunk and tail ridges are discontinuous, and the lateral trunk ridge is confluent with the inferior tail ridge. The brood area of the male is located under the trunk.

→

Die Kiememembranen werden zum Körper und Isthmus durchgebrannt. Die überlegenen Kabel- und Endstückkanten sind kontinuierlich, die untergeordneten Kabel- und Endstückkanten sind unterbrochen, und die seitliche Kabelkante ist zusammenfließend mit der untergeordneten Endstückkante. Das Brutgebiet des Mannes befindet sich unter dem Kabel.

(3)

Distribution: Gazetteer Eastern Atlantic: Norway and Greenland south to Morocco, and from Mauritania to Guinea (Mauritanian Upwelling Region). Seasonally present from Morocco to Mauritania along the edge of the continental shelf.

→

Verteilung: Geographisches Lexikon Ost-Atlantik: Norwegen- und Grönland-Süden nach Marokko und aus Mauretanien nach Guinea (mauritanische Region Upwelling). Saisonal Gegenwart aus Marokko nach Mauretanien den Rand des Kontinentalsockels entlang.

Distribution: Gazetteer Eastern Atlantic: species ranges from Norway and Greenland in north to Morocco in south, and from Mauritania to Guinea (Mauritanian Upwelling Region). It is seasonally present from Morocco to Mauritania along the edge of the continental shelf.

→

Verteilung: Geographisches Lexikon Ost-Atlantik: die Art reicht von Norwegen und Grönland im Norden bis zu Marokko den Süden und von Mauretanien bis zu Guinea (mauritanische Region Upwelling). Sie liegt saisonal von Marokko nach Mauretanien den Rand des Kontinentalsockels vor entlang.

(4)

Feeds on crustaceans, mostly shrimps and shore crabs; fishes, mostly gobies, flatfish, young herring and sand eels.

→

Alimentations sur les crustacés, principalement crevettes et crabes côtiers; poissons, principalement gobies, poissons plats, jeunes harengs et équilles.

It (the fish, the species) feeds on crustaceans, mostly shrimps and shore crabs; fishes, mostly gobies, flatfish, young herring and sand eels.

→

Il nourrit sur les crustacés, principalement crevettes et crabes côtiers; poissons, principalement gobies, poissons plats, jeunes harengs et équilles.

(Verb correctly used instead of noun, but still not quite right – should be *se nourrit de*. For German, a verb is in fact used, but not for French or Spanish, hence a French example!)

(5)

Leaps out of the water when hooked. Utilized fresh and frozen; can be fried, broiled and baked.

→

Sprünge aus dem Wasser wenn eingehackt. Benutztes frisch und eingefroren; können Sie gebraten, gebraten und gebacken werden.

It leaps out of the water when it is hooked. The fish is utilized fresh and frozen; it can be fried, broiled and baked.

→

Es springt aus dem Wasser, wenn es eingehackt wird. Der Fisch ist benutztes frisch und eingefroren; er kann gebraten, gebraten und gebacken werden.

(6)

Biology: Gregarious. Forms schools.

→

Biologie: Gesellig. Die Formschulen.

Biology: gregarious. The species forms schools.

→

Biologie: gesellig. Die Art bildet die Schulen.

Punctuation

Resilience: Medium, minimum population doubling time 1.4 - 4.4 years (K=0.16; tm=3-4; tmax=16; Fec=200,000)

→

Beweglichkeit: Mittleres, Bevölkerungsminimum, das Zeit 1,4 - 4,4 Jahre verdoppelt, (K=0.16; tm=3-4; Tmax=16; Fec=200,000)

The program thinks that *medium* applies to *population*; a colon after *time* would also make it clearer that *minimum population doubling time* is one unit.

Resilience: medium. Minimum population doubling time: 1.4 - 4.4 years (K=0.16; tm=3-4; tmax=16; Fec=200,000).

→

Beweglichkeit: Medium. Bevölkerungsminimum, das Zeit verdoppelt,: 1.4 - 4,4 Jahre (K=0.16; tm=3-4; Tmax=16; Fec=200,000).

The result is not brilliant, now there is a noun instead of an adjective after *Beweglichkeit*, but at least the different “packets” of information are kept together.

For information, the 3 punctuation marks ; : and . all mean “end of translation segment, start analysing a new segment”. This means that for phrases like *Climate: temperate*, the adjective *temperate* is not associated with the noun *Climate*.

Consequently, in the translation, the adjective ending may not agree with the noun gender (the default is masculine ending). This is only a minor nuisance, since the text will still be understandable, and it may not worth trying to edit the text in order to change it.

Summary: Punctuation can help, but writing full sentences is the most useful. Other tips are: use the active rather than passive voice, avoid long sentences or a lot of subclauses.