Thematic Network: European Network for Biodiversity Information (ENBI)

WP-11

Strategies & Techniques to realize Multi-lingual Access to European Biodiversity Sites through a user-friendly interface on the World Wide Web

Extract from the Final Technical Report ENBI WP-11

Preliminary Confidential Report

B. Ueberschär, Rainer Froese & Sven Mohr, Leipniz-Institute for Marine Science
Kiel, Germany
in cooperation with the MT-Team of the European Commission
Directorate-General for Translation,
and with support of a European Translation Team
Contact: e-mail: bueberschaer@ifm-geomar.de

CONTENT OF THIS DOCUMENT:

A) Introduction:

WP-11 is part of ENBI, an EC supported Thematic Network accommodating 65 European institutes representing 24 countries. ENBI's main objective was to establish a strong network that was identifying biodiversity information priorities to be managed at the European scale. In that context, it was considered to be an important and worthwhile effort to provide the technical framework how to enable access to biodiversity information systems in major languages of the European Community (at present French, Portuguese, Spanish, Dutch, German, Italian and Greek). From surveys it was known that language certainly matters in which biodiversity information are available in the Internet, thus it was supposed that translation into major languages of biodiversity information systems will help user such as decision makers, politicians, manager and the public in general from countries where English is not the native language to access biodiversity information in their own language. The usefulness of a multilingual service was shown for FishBase, the information system on fish which was in the focus of this project as a model system for translation. The request for information in FishBase showed a considerable increase in user hits specifically from developing countries since the first version of machine translation for selected resources (search site, species summary) was implemented (Fig. A-1).

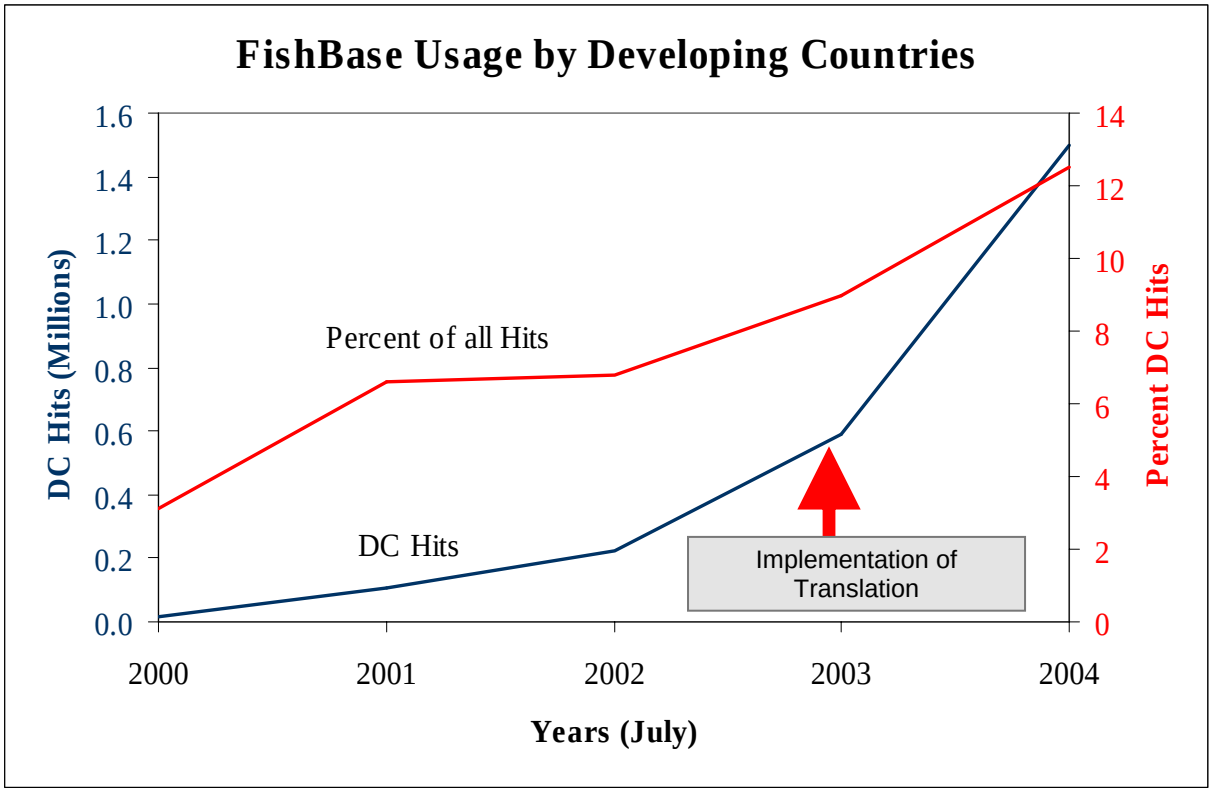


Fig. A-1: Substantial increase in hits to FishBase from developing countries (DC) since a preliminary translation service from English into 7 other major European languages was established in November 2003.

This report is an excerpt of the complete technical report about the tasks and deliverables of WP-11 in the ENBI-project. The purpose here is to deliver a comprehensive and handy document only on the objective how to provide multi-lingual access to biodiversity information in the Internet not including the entire typical framework of final reports of EU funded projects.

Since machine translation is still far from being perfect and not yet a routine application if acceptable results are desired, it is important to be aware about the conditions which are required to produce the desired output.

This document is designed to facilitate the implementation of multilingual services to any biodiversity information system in the Internet. The report starts with an essay on the present status of machine translation in general and specifically for Internet resources in order to set the proper expectations for those potentially interested website owner intending to add machine translation service to their own sites. The title of the essay "*Still a Challenge: Machine translation (MT) in the 21st century*.

Further, the document introduces the framework of general strategies and techniques for translation which should be considered before decisions are made to establish multilingual features to websites.

Finally, the report presents a lists of rules and which are designed to aid in the preparation of English text for machine translation and some specific text-editing examples for FishBase (specifically for English source text translated to German). Some informal notes from the translation team, reflecting their personal experience while preparing the customized dictionaries for FishBase, are attached as well.

This is the first time that machine translation was applied to a non-profit Biodiversity information system in the Internet. In cooperation with the MT-Team of the Directorate-General for Translation of the European Commission the results of this project are considered as a successful approach and the ENBI consortium strongly recommends to use the experience of this project and to consider machine translation for other systems beyond the trial systems of this project (e.g. FishBase, OBIS). Globalisation urges multi-lingual web content and in the next years, the languages of Eastern Europe will be added to the EC-MT system, and work has already begun on Czech and Polish, and with those new EC-member countries, the demand for MT will increase tremendously.

Why to make your website multilingual….

B) Essay

Still a Challenge: Machine translation (MT) in the 21st century

"Now, more than ever, communications and information exchanges are crossing both national and linguistic boundaries. Fortunately, the same computer systems that make such international connections possible can assist in breaking down the language barriers, via machine translation from one language to another. Unfortunately, they are far, far from perfect at doing so. But with careful utilization in appropriate applications, machine translation can open an inexpensive crack in linguistic barriers that would otherwise require costly human translation to scale.
Richard A. Quinnell

In today's modern world, with globalisation of life style, with many people travelling around the world, there is no doubt that English has become the most important "interface" in communication of people with different native languages. While English doesn't have the most speakers (see Table 1 & Fig. B-1), it is the official language of more countries than any other language. However, what are the implications for languages as repositories of culture and identity? The merit of English as a global language is that it enables people of different countries to converse and do business with each other. But languages are not only a medium of communication, which enable nation to speak to nation. They are also repositories of culture and identity. And in many countries the all-engulfing advance of English threatens to damage or destroy much local culture. This is sometimes lamented even in England itself, for though the language that now sweeps the world is called English, the culture carried with it is American.

Rank Total	Language	Primary	Secondary	Total
1	**Chinese***	937,132,000	20,000,000	957,132,000
2	**English**	322,000,000	150,000,000	472,000,000
3	**Spanish**	332,000,000	20,000,000	352,000,000
4	**Russian**	170,000,000	125,000,000	295,000,000
5	**French***	79,572,000	190,000,000	269,572,000
6	**Portuguese**	170,000,000	28,000,000	198,000,000
7	**Arabic***	174,950,000	21,000,000	195,950,000
8	**Bengali**	189,000,000		189,000,000
9	**Hindi/Urdu**	182,000,000		182,000,000
10	**Japanese**	125,000,000	8,000,000	133,000,000
11	**German**	98,000,000	9,000,000	107,000,000

Table 1: Ranking of the most important languages in total numbers of speaker

Thus, there is no doubt that maintenance of local languages, spoken and in literature, is of significance for local culture. Further, the reading of e.g. books, journals newspapers etc. in a foreign language (in English), needs considerable practice which often is not available with many native's whose mother tongue is not English. An example for the need to translate might be the Bible. A total of some 6,500 languages are spoken in the world as a whole, and the complete Bible can now be read in the current number of 405 major languages. Although the number of translated versions of the Bible is still far away from the assumed number of spoken languages, this clearly demonstrates that there is a need for translation, not only "for the book of the books", even with more complex text with uncommon terms and phrases, as e.g. in the science world.

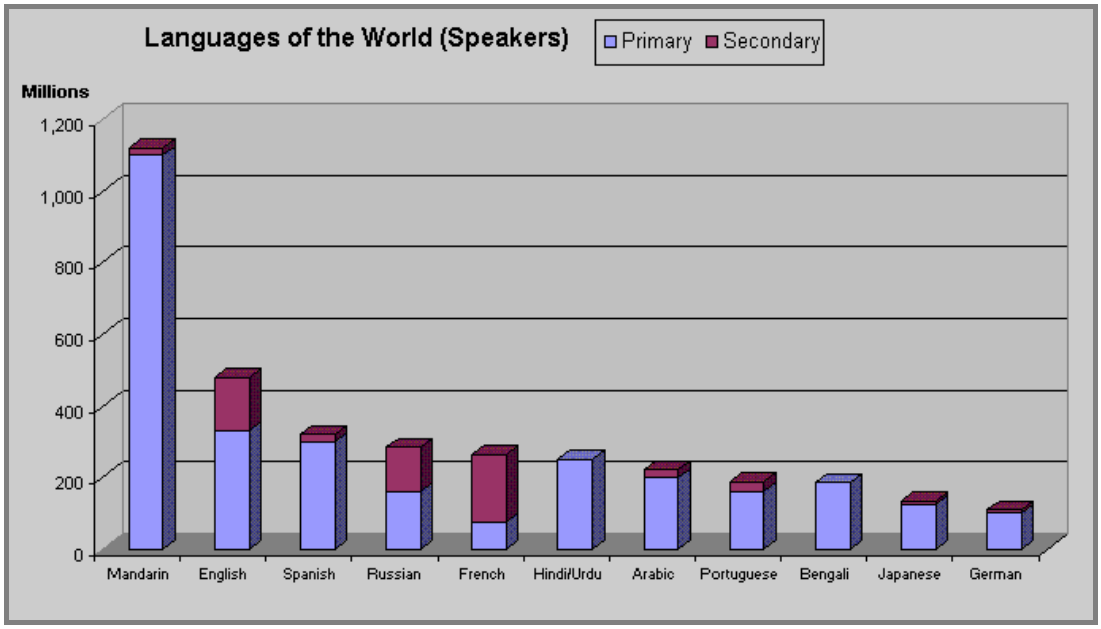


Fig. B-1: Most spoken languages of the World in numbers for primary and secondary speakers.

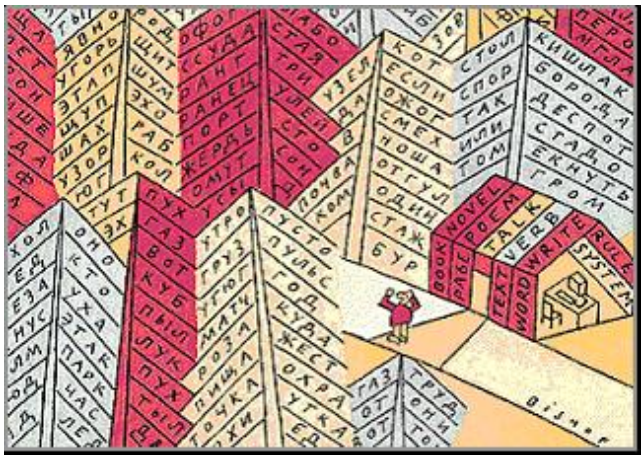
The Internet is also creating new gaps between the rich and the poor. Rich countries with well established educational systems have much greater access to the internet and communications services generally. Although the Internet started off as a communal medium for sharing information, principally among academics, it is increasingly also becoming the tool of trans-national corporations to market their information products around the world. At present, we are moving from an industrial age, in which wealth was created by manufacturing, to an information age in which wealth is created by the development of information goods and services, ranging from media, to education and software. Because it is rich countries generating most of the content on the internet, it becomes a form of cultural imperialism, in which western values dominate and multi-lingual education is considered to be a presupposition to understand Internet content. Since English is the language of the internet, the language barrier is a major reason that poorer countries are often not taking part in this information revolution and are falling further behind.

Why to make Websites multilingual?

Although English is the language of globalisation, it is estimated that by 2050 probably half the world will be more or less proficient, there is no doubt that there is a need for the next decades to present Internet content in other major languages than English. At present, it is estimated that 85% of the Internet's content is in English, but about 45% of Internet users today cannot read English at all (on a global scale).

At present the Internet can be counted in hundreds of millions of pages, and it is growing exponentially at a very high rate. However, it is expected that the non-English speaking web users will soon outnumber the English-speaking users. Thus, it is no longer enough to translate local web sites only to English. In 2006, one expects the Web to reach one billion users and even 70% of them will be non- English speaking. It means that much effort has to be put into localisation of existing web sites and into the creation of new multilingual services, since it is certain that most web users prefer to be addressed in their native language, at least at the top-level pages of services no matter how flawed and error-ridden it may be, rather than to struggle to understand a foreign language text.. Customers, who are addressed in their own language, will stay at a site twice as long. Many owner of Internet sites are aware of this issue and present their content bi- or multilingual (of America's 100 largest firms, 33 had multilingual

websites at the end of 1999, and 57 did a year later). Most of these multilingual presentations are based on manual (static) translation.



Lost in Translation

A further factor will be the growth of access to information sources. Increasingly, the expectation of users is that on-line databases should be multilingual and searchable in their own language, that the information should be translated and summarised into their own language. The European Union pays attention to this demand and is placing considerable emphasis on the development of tools for information access for all members of the community. Translation components are obviously essential components of such tools; they will be developed not as independent stand-alone modules, but fully integrated with the access software for the specific domains of databases. Since Internet content is a very dynamic issue, manual translation is hardly an option, specifically for sites which have naturally a dynamic content with many information being updated in short intervals. Global information systems such as FishBase (www.fishbase.org) are a typical example for those dynamic sites. The wider availability of those kind of databases and information resources in many different languages (particularly on the Internet) has led to the need for multilingual search and access devices with in-built translation modules (e.g. for translating search terms and/or for translating abstracts or summaries).The use of MT in this wider context is clearly due for rapid development in the near future.

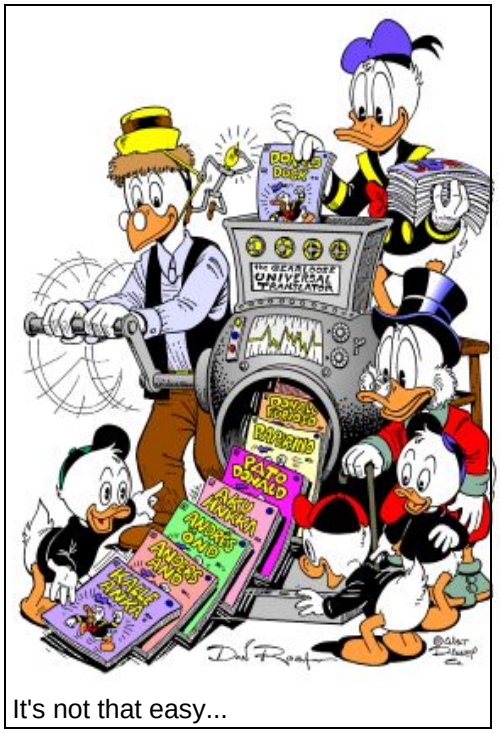
Software companies have already recognised the huge potential market for MT and there are now many systems available for translating Web pages. There is certainly no doubt about the enormous potential for the automatic translation of all kinds of content in the Internet. Only a fully automatic process, capable of handling large volumes with close to real-time turnaround, can provide the translation capacity required, human translation is out of the question. It is now evident that the true niche market for MT is in "cyberspace". While poor quality output is not acceptable to human translators, it is certainly acceptable to most of the rest of the population (Internet user), if they want immediate information, and the on-line "culture" demands rapid access to and processing of information. However, how long poor quality will be acceptable is an open question; inevitably there will be expectations of improvement, and a major challenge for the MT community must be the development of translation systems designed specifically for the needs of the Internet.

Machine translation: Past and Presence

The field of machine translation (MT) was the pioneer research area in computational linguistics during the 1950s and 1960s. When it began, the assumed goal was the automatic translation of all kinds of documents at a quality equalling that of the best human translators. It became apparent very soon that this goal was impossible in the foreseeable future. Human revision of MT output was essential if the results were to be published in any form. At the same time, however, it was found that for many purposes the crude (unedited) MT output could be useful to those who wanted to get a general idea of the content of a text in an unknown language as quickly as possible. For many years, however, this latter use of MT (i.e. as a tool of assimilation, for information gathering and monitoring) was largely ignored. It was assumed that MT should be devoted only to the production of human-quality translations (for

dissemination). Many large organisations have large volumes of technical and administrative documentation that have to be translated into many languages. For many years, MT with human assistance has been a cost-effective option for multinational corporations and other multilingual bodies (e.g. the European Union, US military). MT systems produce rough translations which are then revised (post-edited) by translators. But post-editing to an acceptable quality can be expensive, and many organisations reduce costs and improve MT output by the use of 'controlled' languages, i.e. by reducing (or even eliminating) lexical ambiguity and simplifying complex sentence structures which may itself enhance the comprehensibility of the original texts. In this way, translation processes are closely linked to technical writing and integrated in the whole documentation workflow, making possible further savings in time and costs.

At the same time as organisations have made effective use of MT systems, human translators have been greatly assisted by computer-based translation support tools, e.g. for terminology management, for creating in-house dictionaries and glossaries, for indexing and concordances, for post-editing facilities, and above all (since the end of the 1980s) for storing and searching databases of previously translated texts ("translation memories"). Most commonly these tools are combined in translator workstations – which often incorporate full MT systems as well. Indeed, the converse is now true: MT systems designed for large organisations are including translation memories and other translation tools. As far as systems for dissemination (publishable translations) are concerned the old distinctions between human-assisted MT and computer-aided translation are being blurred, and in the near future may be irrelevant.



It's not that easy...

It is widely agreed that where translation has to be of publishable quality, both human translation and MT have their roles. Machine translation is demonstrably cost-effective for large scale and/or rapid translation of technical documentation and software localization materials. In these and many other situations, the costs of MT plus essential human preparation and revision or the costs of using computerised translation tools (workstations, translation memories, etc.) are significantly less than those of traditional human translation with no computer aids. By contrast, the human translator is (and will remain) unrivalled for non-repetitive linguistically sophisticated texts (e.g. in literature and law), and even for one-off texts in highly specialized technical subjects. However, translation does not have to be always of publishable quality. Speed and accessibility may be more important. What is still often forgotten is that MT is a practical task, a means to an end, and that translation itself (automated or not) has never been and cannot be "perfect"; there are always other possible (often multiple) translations of the same text according to different circumstances and requirements. MT can be no different: there cannot be a "perfect" automatic translation. The use of a MT system is contingent upon its cost effectiveness in practical situations. The principal focus of MT research remains the development of systems for translating written documents of scientific and technical nature; outside the range of possibility are literary and legal texts, indeed any texts where style and presentation are important parts of the "message".

From the beginning of MT, unrevised translations from MT systems have been found useful for low-circulation technical reports, administrative memoranda, intelligence activities, personal correspondence, indeed whenever a document is to be read by just one or two people interested only in the essential message and unconcerned about stylistic quality or even exact terminology. The range of options has expanded significantly since the early 1990s, with the increasing use and rapid development of personal computers and the Internet.

Machine Translation Output Is Not Easily Predictable

MT systems work with natural language: a data set that is infinitely varying, ambiguous, and structurally complex. To translate adequately, an MT system must encode knowledge of hundreds of syntactic patterns, variations, and exceptions, as well as relationships among these patterns. It must include ever-changing vocabulary and specific semantic knowledge about the usage patterns of tens of thousands of words. It must accurately identify the parts of speech and grammatical characteristics of words which may, in different contexts, be nouns, verbs, or adjectives, each having many possible translations. Translation also requires a vast store of knowledge about the world, the intent of the communication, and the subject matter.

A human translator prioritizes and selectively applies linguistic rules based on this knowledge. MT software, unless explicitly coded for each possibility, cannot. Thus, MT will never attain the overall quality of human translation. The primary advantages of MT over human translation are speed, cost, and consistency. An MT system gets much more translation done than is possible manually per time unit, and MT can deliver translations instantly for time-sensitive content. When a term is entered in an MT dictionary, it will translate it the same way every time, unlike human translators who may choose different translations at different times.

Although, machine translation is the only option in an e-business world like today where a large corporation or an organization such as the European Commission may have hundreds of pages on their Web sites with access to databases from which thousands of people may download documents. If all these pages and all these documents are to be translated into a variety of languages, using a human translator would be out of the question, and it would cost millions of dollars and users would have to wait for years to get it done.

The Future: The MT Quality Enhancement Process

SYSTRAN, the system which is in favour for the realisation of the multilingual access in ENBI, has developed the SYSTRAN Review Manager (SRM), which helps the customer to manage the MT quality process by allowing them to change vocabulary and linguistic rules. Users have never before had the power to modify linguistic rules through an intuitive, interactive process. By opening up rule modification, SYSTRAN takes a risk, but one that will almost certainly pay off. Engaging users in the process of improving MT is the surest path to increased acceptance and understanding of the technology. Combined with the SRM, the SYSTRAN Translation Workbench is an interactive XML-based editing tool that incorporates the reviewer's changes as rule modifications. Once it is released, this tool will represent an important advance in MT, both technologically and philosophically. In most MT systems, linguistic rules are not even accessible to the user because they are part of the source code.

Perhaps most importantly, the coming release of the SYSTRAN Translation Workbench represents a shift in the attitude of MT developers toward users. MT systems are extremely complex, and developers have always taken pains to protect the user from making naive changes to the system that could have serious consequences for other contexts. This attitude has been a source of frustration to more sophisticated MT users, who eventually reach a wall on quality improvements after building their dictionaries. Engaging users in the process of improving MT is the surest path to increased acceptance and understanding of the technology.

Overall, making an MT system work for a particular application is a process, not a quick fix. Improving MT is a cyclic process beginning with review of a translation, update of dictionaries and other linguistic resources, and retranslation to validate the effects. In the SYSTRAN system, the SRM acts as a coordinator, managing access to different customization resources and tracking quality (Fig. B-2).

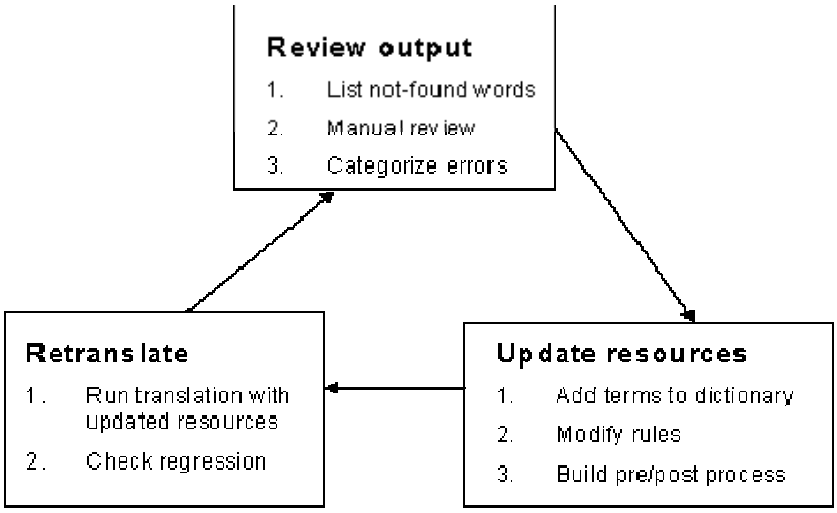


Fig. B-2: MT- Quality Enhancement Process

With potentially thousands of dictionary changes, numerous rule modifications, and changing text, it is a challenge to track customisation activities and measure results. The SRM integrates the three steps into a single-process management program with links to the user dictionary, the source and target texts, benchmark files, and interactive translation testing. In addition, the SRM categorizes errors, assigns levels of severity, and keeps track of statistics on the rates of various error types.

It can be configured as a Web-based application for single or multiple users. In the latter case, reviewers in different locations can access translations, provide feedback, update dictionaries, and even store their own variant translations for a particular word or phrase. For multinational institutions or companies, the SRM allows easy cooperation between sites where different language abilities reside.

After the review process is complete, the dictionaries are saved, the document can be retranslated. Reviewers can also open the dictionary records directly and modify or refine the translations or grammatical tags for an entry. Enhancing the source text is equally important to dictionary building for quality assurance. Translation results tend to be better when the source text is modified to simplify word order and shorten lengthy sentences.

Once the changes to the system are saved, the reviewer can retranslate the text to verify that the new entries are in effect. It is important at this stage to check for regressions. Regressions occur commonly in MT output. They can sometimes originate with an incorrectly coded dictionary entry. For example, a user might supply a translation that is correct in the context of one sentence, but incorrect in another context.

The SRM manages regressions with a color-coding system that shows what portions of the text have changed since the last time it was translated. This feature reduces the amount of time spent on reading and comparing the previous translation with the new version by highlighting the areas for focus.

New MT approaches: Statistical translation

Recently, statistical data analysis has been used to gather MT knowledge automatically from parallel bilingual text. The idea is to let the computer learn automatically by examining large amounts of parallel text: documents which are nearly exact translations of each other. These techniques have not been disseminated to the scientific community in a usable form, however SYSTRAN and other machine translation companies are now applying some of the latest research in natural language processing and new statistical approaches, such as those where scientists are exploring ways of teaching software to translate by feeding it masses of previously translated text.

ENBI and MT: What are the Implications?

The Internet has proven to be a huge stimulus for MT, with hundreds of millions of pages of text and an increasingly global — and linguistically diverse — public. What role will MT play in bridging languages barriers in accessing biodiversity information in the Internet? There is no doubt, that the application of MT is needed to assist in getting information from a database in a foreign language, one that the user does not well understand and ENBI can be part of the stimulus which might help to push forward the accuracy of MT for scientific websites. Since biodiversity information are rather science related results in it´s nature, those resources in the Internet are predestinated for MT, because their presentation can follow the simple rules how to simplify text in order to assist MT (as submitted with this report e.g. for SYSTRAN). The more texts are "standardised", the more they are full of jargon and clichés, the more the text is mundane and uncreative, the more accurate will be the MT output (and eventually the less correction by post-editing is necessary once required). Machine translation works best on standardised input. Creativity is not desired. Unfamiliar word combinations and sentence constructions and telegraphic style with incomplete syntax result in poor MT versions. The more uncreative a text, the better the results. These rules should be carefully considered from website owners who are interested to establish multilingual access to their websites, realised through MT.

The European Union is one of the longest users of MT (apart from the US Air Force), and it is probably the largest user of MT. The EU has developed its own MT-System (EU SYSTRAN) for many language pairs and presently adds real time machine translation of web pages to its services. ENBI has made an agreement with the Translation Department (SdT) to use their system in order establish a Website "on the fly" translation. European biodiversity web sites can avail of this service by showing a "Translate" button on their pages. The cooperation between ENBI and the SdT has good prospects concerning the improvement of MT for Biodiversity information in the Internet. As considered above, the quality of MT considerably depends on the customized activities. Since ENBI has created special Biodiversity dictionaries gradually being integrated into the machine translation service of the European Commission, a good result can be expected after some revisions were applied. What users can finally expect from this approach was expressed by Brian Garr, (former IBM Computer Department):

"Machine translation is a viable technology that can have good value. You just need to set your expectations properly so you get the most out of it."

In the next years, the languages of Eastern Europe will be added, and work has already begun on Czech and Polish, and with those countries being new member of the EU, the demand for MT will increase tremendously.

Compiled from Bernd Ueberschär, (WP 11, Multilingual access to Biodiversity websites; further information at www.enbi.linguaweb.org)

Many thanks to John Hutchins, University of East Anglia, UK, for his comprehensive knowledge on MT.

C) Strategies and Techniques for manual and machine translation, general considerations.

Manual or interactive machine translation? No alternative choice for dynamic website content.

The approach how to deal with translation of Internet resources depends on their nature. Since Internet content is a very dynamic issue, manual translation is an option only for static resources which are barely subject of modification (e.g. labels), such as "Frontpage" content (e.g. "Search site" in FishBase), Homepages (see Fig. C-1), Family lists and Species lists (e.g. common names). It should be kept in mind, that manually translated content needs long-term commitments of translation partner to keep content up-to-date in different languages. Since every change in the web site now needs translation, eventually a list-server-type email approach of sending terms that need translation to all translators needs to be established: the system should automatically notify the translator who can click a link and can enter the translation directly into the translation table. This should be a quick and easy procedure, otherwise, it is too difficult to keep the system content up-to-date in all languages.

Fig. C-1: Homepage of FishBase translated from English into German by means of EC-SYSTRAN® machine translation (no further editing applied). The headline indicates the ECMT-Service.

Specifically for sites which have typically a dynamic content with much information being updated in short intervals, manual translation is not an option (however, with the present quality of MT, in some cases a blend of manual and machine translation for a webpage could be of advantage). Global information systems such as FishBase are a typical example for those dynamic sites. The major content of FishBase is subject to frequent modifications. In general, all dynamic pages, specifically free text as "Species Summary" in FishBase is a typical resource for machine translation (see Fig C-2). Major condition for a reasonable

translation quality is the standardization of source text, and the consideration of rules which are designed e.g. for the SYSTRAN translation engine to aid in the preparation of English text for machine translation (see following paragraphs). A major advantage when applying machine translation is the limited commitment of translation partner. Machine translation may be also applied to glossaries, with careful prepared source text, the result will be acceptable.

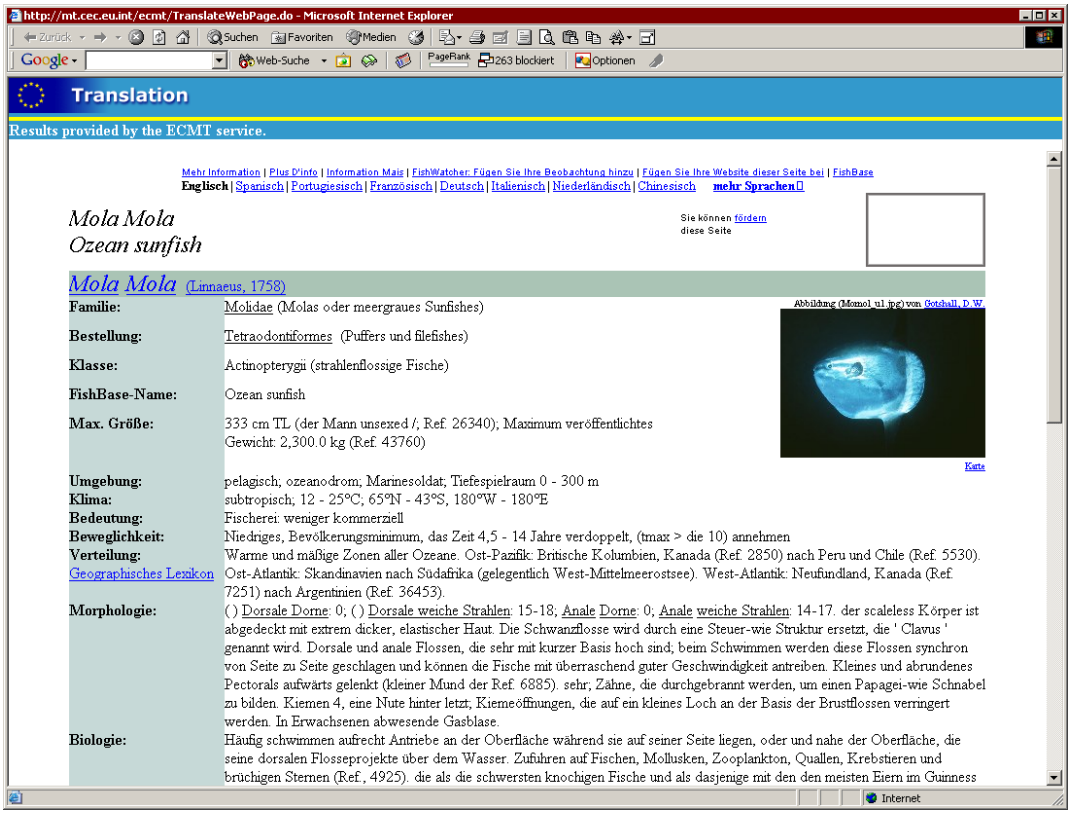


Fig. C-2: Example for a dynamic page: species summary *Mola mola* from FishBase translated from English into German by means of EC-SYSTRAN® machine translation. The headline indicates the ECMT-Service.

History and "State-of–the Art" Translation Engines

The SYSTRAN translation engine (http://www.systransoft.com) is being considered to be the "State-of-the-Art" system at present for machine translation and was in favour for the realisation of the multilingual access in ENBI. However, since the desktop version of SYSTRAN is useful for the basic evaluation of the process of machine translation, it is not appropriate for the translation of website content "on the fly", unless the user has installed and updated it's own version on his desktop (which is out of the question for the majority of user). For those purpose (the user opens a website, wants to see the content in an other language, clicks on the related link, e.g a flag in the desired language which is shown in the page, and a translated page is returned to the user within seconds) the company SYSTRAN offers "SYSTRAN Links", which is a turnkey website translation solution. "SYSTRAN Links" transforms standard websites and content applications into interactive multilingual hubs. "SYSTRAN Links" offers all major European, Asian and Russian languages. However, the fee for this commercial service (from c.a. 25,000 US $) is out of reach for the most biodiversity websites which are in the public domain and maintained by non-profit organisations. There are other companies which offer the same service (e.g. LINGUATEC), but the fee for the service is in the same order of magnitude.

The EC has developed its own MT-System since the 70´s, based on the SYSTRAN engine (EC-SYSTRAN) for many language pairs. Since it is supposed that this service can basically be used free of charge for non-profit projects in the European dimension, the ENBI project requested permission to be connected to those system, and ENBI and the EC-Translation Department came to an agreement which allowed the ENBI project to use their system in order to establish a Website "on the fly" translation for selected biodiversity information systems in the Internet. However, at the date of commencement of ENBI, this service was not yet established and it took more than two years of further technical developments to establish this service (part of the problem was for security reasons). Only some months ago, the MT-Department has finally opened access for WP-11 to their translation service for websites. Although the access is still restricted (using this service requires an individual agreement with the EU-MT Service), the access to real-time translation from the EC-SYSTRAN System it is being considered as a major breakthrough for the goals of WP 11. Facilitated by a very helpful and efficient personal contact between the EC MT-Department and the WP-11 project manager the MT-Department is gradually implementing the special dictionaries which were compiled from WP-11 translation partner into 7 languages on the basis of missing terms in general dictionaries. The missing terms and phrases which were subject to translation were tested against the free-text content of FishBase. Since the results of the machine translation depends to a large extent on the availability of correct translated terms three major dictionaries with specific terms are being hand-coded now from the technical team in the MT-department (noun, proper noun, verb, adjective etc.) and tapped to the EC-SYSTRAN System under consideration of an assigned category (e.g. "Biology", "Fisheries", "Environment" etc.). This three major dictionaries "Biology", "Morphology" and "Distribution" contain more than 3000 specific words which do not appear in general dictionaries; in addition the system is being fed with about 20,000 scientific names which are not to be translated, but advises the translation engine to pay attention to the original name.

Once the EC-Service is implemented into the website of information systems, one single click only on the flag is needed to get the translated webpage (or parts of it) displayed within seconds on the users´ screen. European biodiversity web sites can avail of this service in the future by showing a "Translate" button or flags on their pages (presumed specific dictionaries are delivered). The cooperation between ENBI and the EC-MT-Department has good prospects concerning the improvement of MT for biodiversity information in the Internet. As considered above, the quality of MT considerably depends on the customized activities. Since ENBI has created special biodiversity dictionaries which are going to be integrated in the machine translation service of the European Commission, a good result can be expected after some revisions were applied.

D) Cookery book for the implementation of translation service.

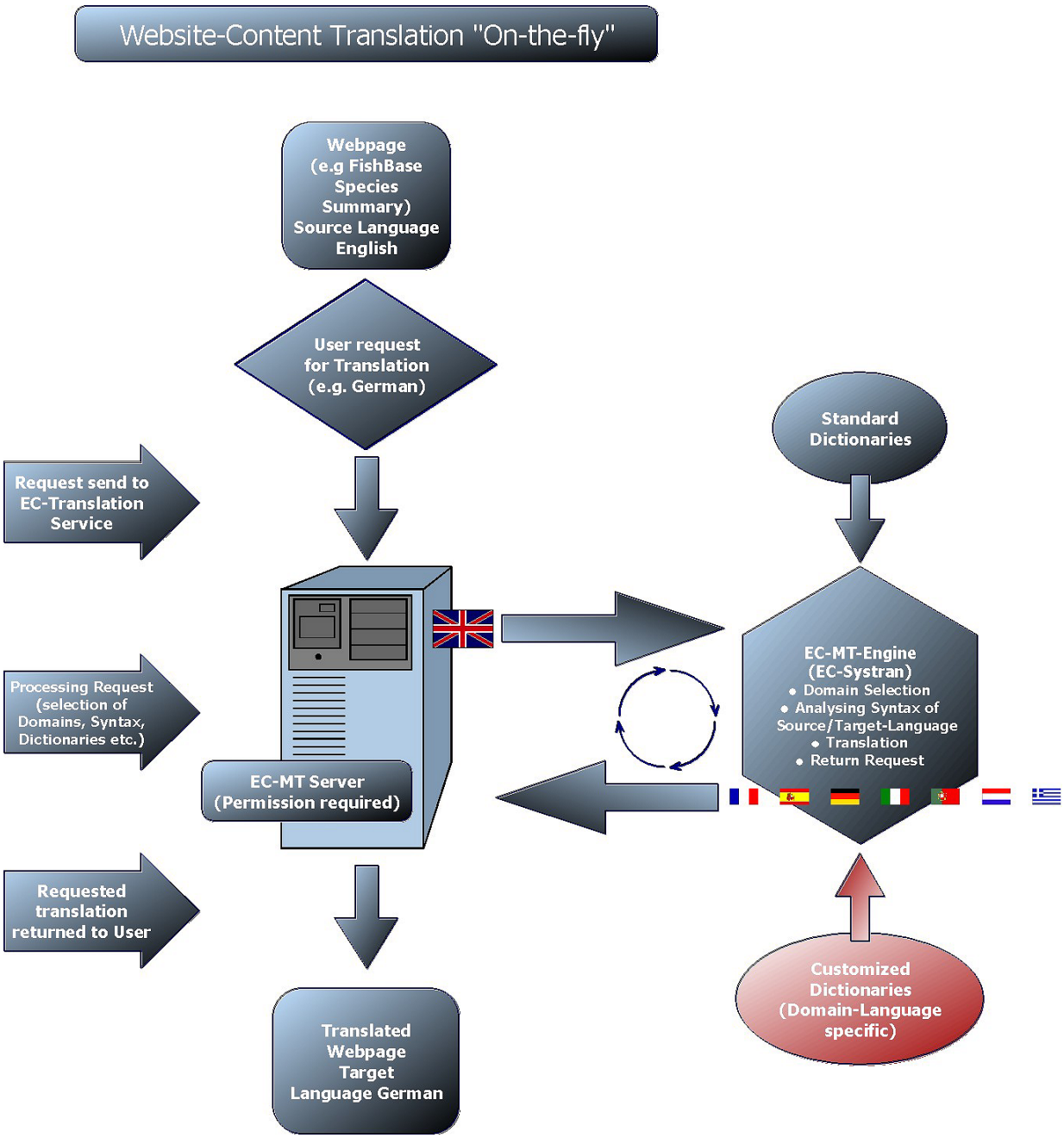
How does it work?
About Machine translation
The principle of machine translation: the translation of a document, from a source language into a target language, is made on the basis of a system of dictionaries and linguistic programs (e.g. SYSTRAN).

Machine translation quality
Machine translation quality mainly depends on the kind of documents (with typing errors, telegraphic style or complex syntax, the result will be poor) language similarities and on specific dictionaries available.

Flowchart Machine Translation:
The plot: A user wants to see an English species summary in FishBase on *Salmo salar* in his own language. She or he is clicking on a link shown in the page and indicating the translation

into the desired language, in this case German (Fig. D-1 & 2). The request is transmitted to the EC-MT server and processed by

(1) analysing the syntax of the source and target language, applying proper translation under consideration of the rules of the target language;
(2) analysing the nature of the source text and identifying assigned domains or categories (a domain or category is specifying the "nature" of the source text, e.g. from Business, Biology, Engineering etc. the source text provider usually sets a proper domain for his resource of information);
(3) consultation of the standard dictionaries and, under consideration of the domains or categories which are identified for this species summary (e.g. Biology, Marine Science..) comparison and correction with terms from the customized dictionaries;
(4) returning the translated page to the users display
(5) and all of this needs only a few seconds!



Salmo salar Linnaeus, 1758

Family:	Salmonidae (Salmonids), subfamily: Salmoninae
Order:	Salmoniformes (salmons)
Class:	Actinopterygii (ray-finned fishes)
FishBase name:	Atlantic salmon
Max. size:	150 cm TL (male/unsexed; Ref. 7251); 120 cm TL (female); max. published weight: 46.8 kg (Ref. 41037); max. reported age: 13 years
Environment:	benthopelagic; anadromous (Ref. 51243); freshwater; brackish; marine; depth range - 10 m
Climate:	temperate; 2 - 9°C; 72°N - 37°N, 77°W - 61°E
Importance:	fisheries: highly commercial; aquaculture: commercial; gamefish: yes
Resilience:	Medium, minimum population doubling time 1.4 - 4.4 years(K=0.29-0.76; tm=3-5; tmax=14; Fec=8,000)
Distribution: Gazetteer	Atlantic Ocean: temperate and arctic zones in northern hemisphere (Ref. 51442). In western Atlantic Ocean distributed in coast drainages from northern Quebec in Canada to Connecticut in USA (Ref. 5723). In eastern Atlantic Ocean distributed in drainages from the Baltic states to Portugal (Ref. 51442). Landlocked stocks are present in Russia, Finland, Sweden and Norway (Ref. 6439) and in North America (Ref. 1998). Appendix III of the Bern Convention (protected fauna; except at sea).
Morphology:	Dorsal spines (total): 3-4; Dorsal soft rays (total): 9-15; Anal spines: 3-4; Anal soft rays: 7-11; Vertebrae: 58-61. Fusiform body (Ref. 51442). Mouth extends only to area below rear of eye and has well developed teeth (Ref. 51442). Vomerine teeth weak (Ref. 7251). Caudal fin with 19 rays (Ref. 2196). Little scales (Ref. 51442). Adults are blue-green colored with a silvery coating and a few spots in salt water; no spots under lateral line (Ref. 37032, Ref. 51442). During reproduction period, in fresh water, it loses the silvery guanin coat and becomes greenish or reddish brown mottled with red or orange, certainly the males (Ref. 37032, Ref. 51442). Few black spots on body, caudal fin usually unspotted and adipose fin not black bordered. Juveniles have 8 to 12 blue-violet spots on the flanks with little red spots in-between (Ref. 51442). Also Ref. 3137.
Biology:	Amphihaline species, spending most of his life in fresh water (Ref. 51442). Young remain in freshwater for 1-6 years, then migrate to the ocean and remain there for 1-4 years before returning to freshwater. It grows up at sea on the continental plate west of Greenland (Ref. 51442). Adults then return to the river of their origin to spawn (Ref. 51442). It returns to sea after spawning, but a lot of adults already die (Ref. 51442). Active during the day. Juveniles feed mainly on aquatic insects, mollusks, crustaceans and fish; adults at sea feed on squids, shrimps, and fish (Ref. 51442). Adults in freshwater which are approaching the reproductive stage do not feed (Ref. 30578, Ref. 51442). Growth in freshwater is slow while in the sea is very rapid. Life history of the salmon can be read from the growth zones in the scales (Ref. 35388). Several lake populations are landlocked. Marketed fresh, dried or salted, smoked, and frozen; eaten steamed, fried, broiled, cooked in microwave, and baked (Ref. 9988). Prefers cool temperature (Ref. 37032).
Red List Status:	, 01-Aug-1996 (Ref. 53964)
Dangerous:	harmless
Coordinator:	
Main Ref:	Page, L.M. and B.M. Burr. 1991. (Ref. 5723)

Fig. D-1:FishBase Summary Page about *Salmo salar* in the standard Language of FishBase (English)

Salmo salar Linnaeus, 1758

Familie:	Salmonidae (Lachse), subfamily: Salmoninae
Ordnung:	Salmoniformes (Lachsfische)
Klasse:	Actinopterygii (Strahlenflosser)
FishBase Name:	Atlantic salmon
Max. Größe:	150 cm TL (Männchen/unbestimmt; Ref. 7251); 120 cm TL (female); max. veröff. Gewicht: 46.8 kg (Ref. 41037); max. veröff. Alter: 13 Jahre
Lebensraum:	benthopelagisch; anadrom (Ref. 51243); süßwasser; brackwasser; seewasser; Tiefenbereich - 10 m
Klimazone:	gemäßigt; 2 - 9°C; 72°N - 37°N, 77°W - 61°E
Bedeutung:	Fischereien: hoch kommerziell; Aquakultur: kommerziell; Sportfisch: ja
Widerstandsfähigkeit:	mittel, Verdopplung der Population dauert 1,4 - 4,4 Jahre.(K=0.29-0.76; tm=3-5; tmax=14; Fec=8,000)
Verbreitung: Gazetteer	Atlantischer Ozean: mäßige und arktische Zonen in nördlicher Hemisphäre (Ref. 51442). In westlichem atlantischem Ozean, der in Küsteentwässerungen aus Nordquebec in Kanada an Connecticut in USA verteilt wird, (Ref. 5723). In östlichem atlantischem Ozean, der in Entwässerungen aus den baltischen Staaten an Portugal verteilt wird, (Ref. 51442). Landumschlossene Bestände liegen in Rußland, Finnland, Schweden und Norwegen vor (Ref. 6439) und in Nordamerika (Ref. 1998). Anhang III der Konvention von Bern (geschützte Fauna; außer auf See).
Morphologie:	Rückenflossenstacheln (insgesamt): 3-4; Rückenflossenweichstrahlen (insgesamt): 9-15; Afterflossenstacheln 3-4; Afterflossenweichstrahlen: 7-11; Wirbelzahl: 58-61. Spindelförmiger Körper (Verweis 51442). Öffnung verlängert nur auf Bereich unterhalb der Rückseite des Auges und hat gut entwickelte Zähne (Verweis 51442). Vomerine Zähne schwach (Verweis 7251). Schwanzflosse mit 19 Strahlen (Verweis 2196). Wenig Skalen (Verweis 51442). Die Erwachsenen sind gefärbt mit einer silbrigen Schicht und einigen Punkten im Salzwasser blaugrünes; keine Punkte unter seitlicher Linie (Verweis 37032, Verweis 51442). Während der Wiedergabeperiode im Süßwasser, verliert sie den silbrigen guanin Mantel und wird grünlich oder gesprinkelt mit Rotem oder Orange, zweifellos die Männer Rötlichbraunes (Verweis 37032, Verweis 51442). Wenige Schwarzpunkte auf Körper, Schwanzflosse unspotted normalerweise und nicht Schwarzes der fetthaltigen Flosse eingefaßt. Jugendliche haben 8 bis 12 blau-violette Punkte auf den Flanken mit kleinen roten Punkten in-between (Verweis 51442). Auch Verweis 3137.
Biologie:	Amphihaline-Arten, die den größten Teil seines Lebens in Süßwasser verbringen, (Ref. 51442). Junge bleiben im Süßwasser für 1-6 Jahre, also wandern zum Ozean aus und bleiben dort für 1-4 Jahre vor dem Zurückkehren zum Süßwasser. Es wächst auf See auf der kontinentalen Platte westlich von Grönland (Ref. 51442). Erwachsene kehren dann zum Fluß ihres Ursprungs hervorzubringen zurück (Ref. 51442). Er kehrt zum Meer nach dem Hervorbringen zurück, aber viele Erwachsene sterben schon (Ref. 51442). Aktiv während des Tages. Jugendliche fressen hauptsächlich Wasser-Insekten, Mollusken, Krebstiere und Fische; auf See Erwachsene fressen Kalmare, Garnelen und Fische (Ref. 51442). Erwachsene im Süßwasser, die sich dem reproduktiven Stadium nähern, speisen nicht (Ref. 30578, Ref. 51442). Das Wachstum in Süßwasser ist langsam, während im Meer sehr schnell ist. Lebensgeschichte des Lachses kann von den Wachstumszonen in den Maßstäben gelesen werden (Ref. 35388). Mehrere Seebevölkerungen sind landumschlossen. Vermarktetes frisch, getrocknet oder gesalzen, geräuchert und eingefroren; gegessen gedämpft, gebraten, gebraten, gekocht in Mikrowelle und gebacken (Ref. 9988). Zieht kühle Temperatur vor (Ref. 37032).
Rote Liste:	, 01-Aug-1996 (Ref. 53964)
gefährlich:	harmlos
Koordinator:	
Hauptref:	Page, L.M. and B.M. Burr. 1991. (Ref. 5723)

Fig. D-2: FishBase Summary Page about *Salmo salar* in the target Language (here German), translation conducted by MT

The following paragraphs depict in greater detail the techniques and strategies which have to be applied for machine translation

(1) At first decide what parts of your web sites should be translated and decide which resources of the system are appropriate for manual or machine translation; e.g., for the news section or free text which is subject of frequent modifications, machine translation is the best option. The database section, e.g. with tables is appropriate for static translation, unless there are resources with much free text which is being modified from time to time (see above); As mentioned above, considering the present quality of MT, in some cases a blend of manual and machine translation for a webpage could be of advantage.

(2) Assemble a team of reliable volunteer translators for the "static" translation into various languages and secure their long-term commitment (see also item (5). As for the number of languages which can be applied in manual translation the only limitations are the financial means for human translator. The translation team should be familiar with the topics of the concerned information system; otherwise they may not be able to translate more sophisticated terms in a reasonable time.

(3) In web pages use placeholders for titles, headers, labels, choices in choice fields, notes. This arrangement helps to make efficient use of the translated terms and webspace and only one lookup-table is needed as a resource for various pages with the same e.g. labels.

(4) Have items in (2) translated by your translation team and stored in a lookup table. With our experience in the ENBI project it turned out, that even knowledgeable translator, who are basically familiar with the topics of the trial systems, had difficulties to translate very specific terms for FishBase and OBIS. There are several resources available in the Internet which could be consulted when special terms are to be translated. Apart from popular search tools as Google (Google helps a lot, when lists of terms are to translate: often, the term needs to be seen in the actual context in the database for correct translation; a search in Google with the respective term mostly return the context from the source of the term, e.g the species summary in FishBase) and the free Encyclopaedia Wikipedia, for more specific translation, the European Union offers Eurodicautom which is the European Commission's "multilingual term bank." When it was first set up in 1973 the development team drew upon the know-how and lexicographic material of two other tools available to Commission translators: Dicautom, a phrasal automatic dictionary launched in 1964, and Euroterm, a translation dictionary developed in 1962-68. The four original languages of Eurodicautom were Dutch, French, German and Italian, to which Danish and English were added in 1973, Greek in 1981, Portuguese and Spanish in 1986, and Finnish and Swedish in 1995. Latin is also present. Although originally developed to meet the needs of in-house translators, Eurodicautom soon became useful to other Commission staff and was later adopted by linguists in other European institutions. Today it is an invaluable tool for translators, interpreters, terminologists and other linguists worldwide over the Internet, where it records a daily average of 120.000 enquiries. Entries are classified into 48 subject fields (ranging from medicine to public administration). A typical entry contains the term itself and its synonyms, together with definitions, explanatory notes, references, etc. At present the term bank contains about five and a half million entries (terms and abbreviations), subdivided into more than 800 collections. Consultation of this and other resources is certainly worthwhile and can facilitate the translator's effort http://europa.eu.int/eurodicautom/Controller.

(5) 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 the respective translation project. Since every change in your web site now needs translation, a list-server-type email approach of sending terms that need translation to all your translators needs to be established; the system should automatically notify the translator who can click a link and can enter the translation directly into the translation table. This should be a quick and easy procedure, otherwise, it is too difficult to keep the system content up-to-date in all languages. For FishBase, such a system is under development; please check the progress with the FishBase manager in case you want to introduce such a system for your information system.

(6) For machine translation, dedicated dictionaries for different types of texts are a key issue for a reasonable translation. Rely on the same translation team which has been hired for manual translation, they are already familiar with the topics of your information system. At present all available translation programs are delivered with a master dictionary which holds only standard vocabulary for each language. This is certainly sufficient if translation is desired for pieces of standard texts and "daily conversation", but for databases and information systems which deal with very special topics, it is absolutely indispensable to compile dedicated dictionaries. Further, many terms have different meaning, for example the word "stocks" are populations in biology, (e.g. fish stocks), but money in business contexts (e.g. stock market). To avoid these errors, domains or categories have to be created which contains the proper translation in the related context. The user of a translation engine has to "advise" the translation engine to search in the proper domain or to assign the correct category for the translation of words and terms (e.g. for FishBase, domains such as Biology, Fisheries, Environment and Science may apply, unless it has his own category "FishBase" established). When using the translation service of the EU, the creation of new domains has to be negotiated with the responsible department.

(7) The ENBI project (WP-11) has developed dedicated dictionaries, based on terms and phrases in FishBase which were not available from standard dictionaries. In accordance with the structure of the free text resources in FishBase, three major dictionaries on Morphology, Distribution and Biology were compiled, each containing more than 1000 terms and phrases which in many cases represents biodiversity terms in a wider sense. These dictionaries were delivered to the EU-MT department and are now gradually implemented from an encoder team in the EU-MT department into the EU-SYSTRAN translation system. Since the financial resources of the ENBI project did not allowed paying for this service, we are dependant on the capacity of the encoder team to deal with this extra work. Hopefully, the EU-SYSTRAN translation engine from the EU may be able to make soon full use of these dictionaries when translation requests are send from FishBase, which will certainly gradually improve the translation quality.
The dictionaries were tailored and compiled for FishBase, however, they contain many terms which are of general significance for information systems which deals with biodiversity. Nevertheless, any other information system with a different focus compared to FishBase which wants to implement the translation service of the EU-MT department has to scan their resources for unknown terms (e.g. free text) which are supposed to be subject of machine translation. The results should be compared with existing translation before new dictionaries are delivered to the EU-MT department in order to avoid unnecessary effort for the encoder team. Please note, that it sometimes makes sense to advise the translation engine NOT-to translate certain terms. This should be also considered when compiling dictionaries.
Some specific technical rules for the compilation need to be obeyed in order to

facilitate the encoding process for the EU-MT system. For detailed advisements please contact the Coordinator of WP-11 in the ENBI project (Bernd Ueberschär).

(8) For many users who want to make use of a biodiversity information system a number of terms are unknown for e.g. decision maker, politicians and the public at large. For those clients, it is certainly useful to offer a glossary of (technical) terms. FishBase has developed for years a glossary with terms which are useful for the whole aquatic world. In order to make this glossary available for other biodiversity sites, the ENBI project has developed a web portal which is suggested as a multilingual portal to existing and new, appropriate glossaries in the Internet. At present it can be considered as a prototype with trial character. For the FishBase glossary, machine translation is offered from English into 7 European languages. The translation is realized through the EU-MT service. Basically, any glossary can be attached to this web portal in the future (assumed permission of glossary owner is granted), the link can be shown in any biodiversity information system in the Internet. See http://filaman.uni-kiel.de/search/

(9) After completion and consideration of all the items (1) to (8), you certainly want to know how to connect your information system to the translation service of the EU-MT department. The technical connection to the EU translation server from FishBase was realized with a PHP script under consideration, other script languages such as PERL may work as well (the choice depends from the general technical environment where your database is running). The script is sending the source text to the translation server of the MT-EU department and formats the reply (the translated text into the desired target language) in accordance with the needs of the webpage which is shown to the user. The script can be provided to the manager of other information systems who are interested to add multilingual website translation "on-the-fly" to their databases. However, please consider that permission from the EU-MT service is required for any other system. Please inquire about detailed conditions and technical prerequisites before planning a multilingual service for your information system.

(10) Following to the implementation of translation facilities, bear in mind, that testing is essential, and certainly improvements at least with the source text and for the special dictionaries will be necessary.

For further information, advisements and in case of any query on multilingual services for data resources in the Internet please contact the manager of WP-11 in ENBI, Bernd Ueberschär e-mail: bueberschaer@ifm-geomar.de

(E) Preparing English source Text for MT

In the following paragraphs, a list of rules is provided which specifically apply when SYSTRAN is used as translation engine, no matter if a PC-Version of the translation program is used or the EU-SYSTRAN MT. Presented in the form of rules, the information herein can help SYSTRAN users to eliminate many of the problems that can occur during corpus analysis. The rules are designed to aid in the preparation of English text for machine translation and it is strongly suggested to obey the recommendations while preparing source text for databases. (Please note: Some of the rules apply only to the PC-Version of SYSRTRAN, e.g. the formatting for not-to-translate. This kind of formatting will be applied properly from the encoder team when submitting your dictionaries to the EU-SYSTRAN). The translation results can be tremendously improved with a proper design of the source text! The following topics are considered:

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

Each list of rules contains explanatory notes on each rule therein, as well as examples wherever possible. Of course, apart from this specific advisements, when preparing text for machine translation it is important in general to write as clearly as possible and to avoid telegraphic style, typing errors, or complex syntax.

Many of these rules address the use of correct grammar and punctuation, as English presents many opportunities for ambiguity and the degree of precision necessary for machine translation is much greater than that needed for human understanding.

Please note: Although the concept of using correct grammar and punctuation to minimize translation ambiguity applies across all languages, the specific rules are different for each language.

Rules for the Use of Articles
The large challenge in machine translation is the resolution of homographs, words that are written the same way but that have different meanings and often are different parts of speech (for example, noun and verb or adjective and verb). In English, the placement of articles often helps to clarify how a word is being used. Articles are also useful for clarity in many other contexts.

The rules listed below apply only to English as source language.

A1. Use articles to reduce the ambiguity caused by homographs.
Instead of: empty file
Use: empty the file
Or: the empty file

A2. Use articles to clarify the function of the present and past participles.
Instead of: moving car
Use: the moving car
Or: moving the car

A3. 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.

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. The rules listed below apply only to the English language.

L1. 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

L2. 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

L3. 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.

Rules for Phrase Structure
Correct and concise English grammar is essential to good machine translation, as is absolute clarity. The rules listed below apply only to the English language.

S1. 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.

S2. 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.

S3. 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.

S4. 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.

S5. 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.

Rules for Punctuation
Punctuation is essential for dividing sentences into their logical components, allowing for their correct interpretation. The rules listed below apply only to the English language.

P1. 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.

P2. 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.

P3. 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.

P4. 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.

***P5. 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.

P6. Limit the use of the slash*.
Instead of: at the beginning/end
Use: at the beginning or at the end

*Unless they have been entered into your user dictionary (UD) prior to translation, two words with a slash between them will most likely be marked as a single Not-Found Word. The untranslated word pair will then be transferred as-is to the translated text.

P7. Hyphenate phrases that modify other words or phrases.
Instead of: man eating shark
Use: man-eating shark

P8. 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.

*In-correct will be translated as the word "in" and the word "correct," instead of as "not correct."

P9. 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.

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.

The rules listed below are useful for most language pairs.

F1. 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.

F2. Use one space after abbreviations, commas, and semicolons.
Instead of: Mr. Smythe
Use: Mr. Smythe

F3. Use the word wrap or the soft return feature of your word processor for all sentences within a paragraph, instead of inserting hard returns.
Instead of: There is no need to press ENTER at the end of each line. It may cause your text to be broken in strange places, and it will convince the translation program that each line is a paragraph. Instead, just keep on typing, and your software will automatically wrap the text around to fit your page, whatever size it may be.

F4. Use a hard return at the end of each paragraph.

F5. Use two hard returns after all titles and headings, unless they end in a punctuation mark.

F6. Use indents, tables, and tabs instead of many spaces.

Other Rules
There are a number of rules that do not fit comfortably into any of the other five categories, and these rules are offered here. Most of these rules are useful for many language pairs.

O1. Use abbreviations consistently.
Instead of: a 3 min. min.
Use: a 3 min. minimum.
Or: a 3 minute min.

O2. 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.

O3. 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.

Imperative Translations
These are sample sentences and their translations for each Imperative Translation Mode available for English into French, German, Italian, Portuguese and Spanish. There are no imperative options available for English Target translations.

English Into French
Sample English sentence using the imperative:
Push the button.

Default Translation:
Poussez le bouton. Formal / Plural Informal

Option:
Pousser le bouton. Infinitive

English Into German
Sample English sentence using the imperative:
Turn the knob.

Default Translation:
Drehen Sie den Hebel. Formal

Options:
Den Hebel drehen.
Drehe den Hebel.
Dreht den Hebel.

Infinitive Singular Informal Plural Informal

English Into Italian
Sample English sentence using the imperative:
Shut the door.

Default Translation:
Chiuda il portello. Subjunctive

Option:
Chiudere il portello. Infinitive

English Into Portuguese
Sample English sentence using the imperative:
Open the door.

Default Translation:
Abra a porta. Subjunctive

Option:
Abrir a porta. Infinitive

English Into Spanish
Sample English sentence using the imperative:
Verify the position.

Default Translation:
Verifica la posición. Imperative

Option:
Verificar la posición. Infinitive

(F) Text-Editing Examples from FishBase

These samples apply specifically for English source text to German, which, however may be also valid for any other scientific-styled text resource.

The way FishBase texts are written (a subject or verb is often missing) 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.

Please note: the text in bold and italics represents the original text in FishBase species summaries.

(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 sub-clauses.

(G) Some additional considerations on machine translation

The following paragraphs depicts some informal notes compiled from the translation team as an outcome of the ENBI-workshops on "Machine Translation of Biodiversity Information Systems".

Since machine translation is still a challenge and results are sometimes not really predictable and often cannot be standardized (because of the many differences between natural target languages), the following notes from brainstorming on translation issues from ENBI-translation workshops reflect practical experience on machine translation and may help to understand some of the difficulties (and possible improvements and solutions) which are connected with machine translation. Potential user of MT-technique may find useful hints for their specific database. For further information contact the project manager of WP-11.

- 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.

- Context sensitivity is another major problem for translation. Terms have different meaning when appearing in another context. That is the reason, why domains or 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, 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 the 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 tests, and it didn't help: the plural form seemed to be overruled by the singular form. So just put the singular form.

- 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 has distributed a corrected file "Distribution" to the translation consortium, which includes almost any country in the world in the correct form. Other systems can make use of this dictionary of countries, regions etc.

- 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.

- 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 into your language. Please mark the added terms in the dictionary to notify the coordinator of the dictionaries!

- 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?
  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 generally associated with a particular noun, best to include the phrase e.g. "amphihaline species" -> "espèce amphihaline" to make sure the correct translation.

(E) Glossary

A useful glossary of terms and phrases used in the world of machine translation, specifically for the SYSTRAN translation facilities.

Acronym: In SYSTRAN terminology, a word, all upper case, formed from the initial letters of other words or parts of a series of words, such as *WAC* for *W*omen's *A*rmy *C*orps.

Adjective: A word used to modify a noun by limiting or qualifying. In English, it can be distinguished by one of several suffixes, such as *-able, -ous, -er,* and *-est,* because it directly precedes a noun or noun phrase, such as *red* in *a red door,* or because it is preceded by a form of *to be,* such as *the door is red*.

Adverb: A word that modifies a verb, an adjective, or another adverb, such as *brightly* in *The sun shines brightly*.

Article: A word used to signal a noun and to specify its application. In English, the definite article is *the,* and the indefinite articles are *a* and *an*.

Auxiliary verb: A verb, such as *have, can,* or *will,* that accompanies the main verb in a clause and helps to make distinctions in aspect, mood, tense, and voice.

Clause: A group of two or more words which contains a verb. One or more clauses make up a complete sentence.

Direct Mode: An interactive mode of translation that allows the translator to enter text to be translated in the SYSTRAN source text window, and to see the translation in the SYSTRAN target text window. See File Mode.

DNT: An acronym for Do Not Translate.

DNT List: A list of words created by the translator that will not be translated by the SYSTRAN Translation System.

Do Not Translate: A translation option that allows the translator to choose to leave certain words untranslated in the Source text.

Document Type: A translation option that allows the translator to specify the general style of the text to be translated.

Embedded clause: A subordinate clause that is embedded in the middle of a main clause. For example, *The camera, which I bought yesterday, is already broken*.

Expression: In SYSTRAN terminology, a noun phrase consisting of more than one noun or any combination of nouns and adjectives, where the Head Word is a noun, such as *lug nut, Library of Congress,* and *red letter day*.

Complete sentences are not valid SYSTRAN expressions; neither are any phrases that contain verbs.

Expression Dictionary: The SYSTRAN dictionary that contains expressions. See Stem Dictionary.

File Mode: A non-interactive mode of translation that allows the translator to send a file to be translated and to choose the translation options for it. See Direct Mode.

Finite verb: A verb form that is limited in tense, person, and number, such as *goes* in *He goes*.

First person: A pronoun that refers to the speaker. For example, *I* in *I see*, or *We* in *We are.* Also, a verb form that refers to the speaker. For example, *am* in *I am*, or *are* in *We are.* Verb inflection rarely indicates person in English, but in other languages, it often does.

Head Word: In SYSTRAN terminology, the noun in a noun phrase which may change in the plural, such as *nut* in *lug nut*, *Library* in *Library of Congress,* and *day* in *red letter day*. The Head Word is not necessarily the same as the Principal Word.

Homograph: In SYSTRAN terminology, one of two or more words that have the same spelling and are different parts of speech (for example, noun and verb, or adjective and verb). For example, *head* as in *head west* or *on the head,* and *light,* as in *the light box* and *light the match*.

Imperative: A verb form that is used to express an order or command. For example, *Eat* in *Eat your vegetables*.

Infinitive: A verb form that is the ordinary dictionary-entry form. In English, it is often used with “to” as in *He wants to eat*. It may also occur without “to”, for example, *get* in *I made them get in line,* or with auxiliary verbs such as “must” as in *We must leave*.

Inflection: An alternation of the form of a word by adding suffixes without changing the basic meaning or part of speech, as in making *rugs* from *rug,* or by changing the form of a base word, as in making *ate* from *eat*.

ISO639: ISO standard codes for the representation of names of languages.

Main clause: A clause that can stand alone as a complete and correct sentence. For example, *It was raining*.

NFW: The acronym for Not-Found Word.

Not-Found Word: A word or string that occurred in the text, but not in the dictionaries used in translation.

Not-Found Word List: An alphabetical list of all words in a text that were not found in the dictionaries used for translation.

Not-Found Word Marker: A mark that the translator can select, which appears in the translation to indicate a Not-Found Word.

Noun: A word that is used to name a person, place, thing, quality, or action, for example, *house*, *flammability*, or *movement*.

Noun phrase: A phrase that functions as a noun, and that has a noun as its head word.

Parse: To break (a sentence) down into its component parts of speech with an explanation of the form, function, and syntactical relationship of each part.

Parser: In SYSTRAN terminology, the module of the computer program that performs syntactic or semantic analysis of the Source text.

Participle: A verb form that can be used with an auxiliary verb. It can also function as an adjective or a noun. See Past Participle and Present Participle.

Past participle: A verb form that indicates past or completed action. It can be used with an auxiliary verb as in *The cake was baked yesterday*, or as an adjective, as in *baked beans*.

Phrase: A sequence of two or more words which express an idea. See also, Noun Phrase, Verb Phrase, and Prepositional Phrase.

POS: The acronym for Part of Speech.

Preposition: A word placed before a noun or noun phrase, indicating the relation of that noun or noun phrase to a verb, an adjective, or another noun or noun phrase, such as *at, by, in, to, from,* and *with*.

Prepositional phrase: A phrase that consists of a preposition plus the noun or noun phrase that it governs, such as *at the park*, *in reference to your letter,* or *from Mars*.

Present participle: A verb form expressing present action, formed in English from the infinitive plus *-ing,* which can be used with auxiliary verbs as in *He is baking a cake*; as an adjective, *the baking rack*; or as a noun, as in, *the act of baking*.

Pronoun: A word that functions as a substitute for a noun or a noun phrase and designates persons or things asked for, previously specified, or understood from the context. For example, *I, it, that,* and *which*.

Proper noun: A noun belonging to the class of words used as names for individuals or places. For example, *Clinton* or *Boston*.

Run-time dictionary: The binary indexed dictionary file used during the translation process.

SDM: The acronym for SYSTRAN Dictionary Manager.

Second person: A pronoun that refers to the listener. For example, *you* in *You see* or *You are.* Also, a verb form that refers to the listener. For example, *are* in *You are.* Verb inflection rarely indicates person in English, but in other languages, it often does.

Semantic: Having to do with meaning.

Sentence: A grammatical unit that is syntactically independent and has a subject that is expressed or understood and a predicate that contains at least one finite verb.

Source: The language of the original text, before translation.

Subjunctive: A verb form that indicates possibility, doubt, or desire rather than fact. For example, *were* in *I wish it were true*, or *start* in *I suggest you start immediately.*

Subordinate clause: A clause that modifies or expands on other clauses. A subordinate clause modifies or expands on another clause. It cannot stand alone, as *that he gave* in *The account that he gave was true*.

Subordinate clause marker: A word which indicates that a clause is a subordinate clause, such as *that, which, or who*.

Subordinate phrase: A phrase that modifies or expands on other phrases.

Syntactic: Having to do with sentence structure.

Target: The language into which the text is translated.

Third person: A pronoun that refers to neither the speaker nor the listener. For example, *He* in *He is*, *She* in *She is, it* in *it is*, or *They* in *They are.* Also, a verb form that refers to neither the speaker nor the listener. For example, *is* in *He is* and *She is*, or *are* in *They are.* Verb inflection rarely indicates person in English, but in other languages, it often does.

Specialized Dictionary: A translation option that allows the translator to choose appropriate subject areas for the document.

UD: The acronym for User Dictionary.

User Dictionary: A dictionary used in translation that has been created by the user for the purpose of tailoring translations to his specific needs. Definitions in a UD override the ones in SYSTRAN's dictionaries.

Verb: A word that expresses existence, action, or occurrence.

Verb phrase: In SYSTRAN terminology, a phrase or other construction used as a verb. Verb phrases are not acceptable entries in a UD.