News Letter

www.enbi.info

Volume 1, No.1 March 2004

Editor: Chris Johnson (cjohnson@maich.gr)

Participants at the first ENBI Meeting, held on 17-19 March 2003 at the Royal National Institute of Natural Sciences, Brussels, hosted by Olivier Retout, Head of International Relations at the Institute, attended by 86 ENBI members

---

In this Issue

News of the ENBI Work Programme

The Background to ENBI's Work Programme

The European Network for Biodiversity Information is organised on a project (workpackage) basis. There are four main clusters of workpackages, allowing that the interaction between closely related workpackages can be managed in an efficient way.

Cluster I Coordinating activities.

1. Network co-ordination & Sustainability and continuity of European activities.
2. ENBI Forums & Inventory of state-of-art.
3. Dissemination.
4. Copyrights & financial issues.

Cluster II Maintenance, enhancement and presentation of biodiversity databases.

5. Cooperation of pan-European checklist and 'Species bank' database projects.
6. Cooperation of pan-European databases on biological collections and specimens.
7. Observational survey data.

Cluster III Data integration, interoperability and analysis.

8. Data management in large scale distributed biodiversity database systems.
9. Interoperability and common access.
10. Generic analysis tools and data mining.

Cluster IV Products and e-services.

11. Multi-lingual access
12. Information services on European biodiversity data.
13. Making non-European biodiversity data in European repositories globally available.

Updates on most of these workpackages appear in the following pages.

WP 2. ENBI forums
Providing ENBI's Communication Platform

From Joaquín Hortal, Real Jardín Botánico (CSIC), Madrid, Spain

Announcing *ENBI Forums'* forthcoming activities: 2nd ENBI *e-conference* and 1st ENBI Workshop

ENBI Forums announces two main events for the first half of 2004, directed to favour the interaction of scientists and biodiversity data stakeholders to the structure, development and implementation of GBIF:

- *The Second ENBI Electronic Conference* will be held from March 9th to 24th. ENBI *e-conferences* are intended to be interactive electronic forums to foster discussion on various topics about the implementation of GBIF, both in Europe and worldwide. You can find up-to-date information about the topics to be discussed and subscribe to the *e-conference* at: http://www.enbi.info/forums/ec2/index.html.
- *The First ENBI Workshop* will take place in May at Pruhonice, close to Prague (Czech Republic), from 25th to 28th. In this workshop, the topics and discussions arising from the two first *e-conferences* will be extensively debated face-to-face in several working groups, including a specific session for European GBIF nodes. You can find updated information and programme, as well as subscribe to the workshop at http://www.enbi.info/forums/ws1/index.html.

ENBI Forums' Web Page

ENBI Forums web pages are now available at http://www.enbi.info/forums/. In these pages, ENBI brings together experts and end-users in specific fields of interest. Here you will find information about general meetings, workshops and e-conferences. These events aim to promote discussion and exchange of opinions, criteria and knowledge on issues of common concern, as well as to provide an insight on the state of activities in Europe. Other information resources, such as links to related projects, expertise databases, GBIF nodes and *ENBI Forums'* partners web pages, are also available. In future updates, an ENBI state-of-the-art will be also included.

ENBI First Electronic Conference: Open Access for Biodiversity Information

The first ENBI *e-conference* took place from 10th to 28th September 2003, with the title "Open access for biodiversity information". Although participation was low, many interesting topics arose, dealing with how information should be stored and accessed, how its reliability and accuracy could be validated, which services should be provided by GBIF, and how. The *e-conference* was divided into three sections, the first two dealing with previously defined specific topics, which were extended to be discussed in the third one.

During the first session, *Sharing biodiversity information*, the amount and reliability of data available through GBIF was discussed. Several contributions pointed out that the data freely

available in GBIF should be only highly reliable information, thus resulting in offering less data. However, other contributors argued that defining which information and data are useful, and which are not, depends of the purpose of the user or of the analyses he/she wants to carry out. A trained user, such as a scientist, would be able to discriminate by his/her own criteria, which information is useful, and/or how to correct errors in data. Since data storage seems not to be a key question, as current hard disks are 'cheap and huge', many contributions committed in favour of no reduction of any information stored in GBIF that might be of potential utility in the future, but to restrict the open-access to the data designated as highly reliable.

The other key issue in the debate was how information should be indexed. Most contributors agreed that a unique number per specimen stored in GBIF should be used. This unique identifier per individual may solve many problems and future changes in the records stored in GBIF. However, there were different opinions about the structure and/or origin of this code. Thus, much debate in this session was was concerned with how it should be developed, centred in two main options: i) constructing the code as pure random numerical codes, in a way similar to GenBank (see http://www.ncbi.nlm.nih.gov/Genbank/), or ii) including comprehensible parts (as well as numbers) in an alphanumerical and partially human-comprehensible code. Here, former initiatives such as DarwinCore or ABCD schema were highlighted as current approaches to the problem.

Under the title *Biodiversity information: What services are needed, and how the data providers should supply them?*, the second session dealt with a particular aspect of GBIF's purpose of making useful biodiversity information available. Although end-users most times look actively for the information they need, thus ensuring certain feedback with GBIF, scientists must also be an active part, both trying to translate society's needs to their work, and generating information accessible and comprehensive (and, of course, reliable enough) for non-specialist users, such as students (not only university, but primary and secondary school students), politicians, non-specialized journalists, etc. A question arising was how to provide end users with useful information. In relation with this issue it was argued (using also several examples of its potential utility) that GBIF, and, more precisely, ENBI, should also focus on developing works and technical reports on concrete problems (e.g. reconstructing socio-historical backgrounds of collections, reconstructing global change effects, economical values of biodiversity, etc.), rather than just providing access for global information. Here, biodiversity database projects must truly involve end-users, not just use them as external advisors.

A hotly debated issue was the adequacy of commercial exploitation of biodiversity data. Most comments supported the free-access concept of GBIF, imposed by the urgent need for both a conservation strategy and the social dissemination of the real value and importance of biodiversity. However, GBIF and data providers need certain budgets to afford the costs of, on the one hand, the development and maintenance of GBIF, and in the other hand, carrying out field surveys, taxonomic identifications, database compilation and maintaining natural history collections. Many contributors viewed pre-pay access to a part of the data as a problem rather than a solution to this issue, and thus argued for an economic compromise by which the EU and the member nations maintain in the future the process of gathering, storing, evaluating and disseminating biodiversity data. Thus, a balance is needed between raising funds and making biodiversity information freely-available worldwide. In relation to this, data providers should also be encouraged to share their biodiversity information through, e.g., by giving them appropriate credits, technical support, analytical tools, rapid data provision, etc.

A new body of experts trained in biodiversity informatics should supply both data providers and end-users with the services and support necessary to make the free-access biodiversity information system a healthy reality, rather than just a well-meaning initiative.

You can access the full contents of this first ENBI *e-conference* through ENBI Forums (WP2) web pages (http://www.enbi.info/forums/ec1/index.html) and in the open-access area of ENBI CIRCA (http://circa.gbif.net/Public/irc/enbi/econ1/library).

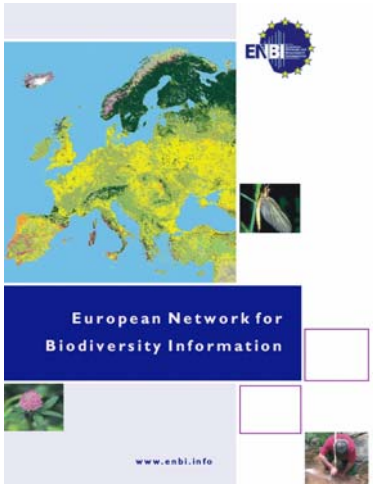
WP 3 Dissemination

Dissemination and Transfer of Expertise developed within ENBI

From Chris Johnson, Department of Natural Products, Mediterranean Agronomic Institute of Chania, Greece

ENBI Brochure

The first product from this Workpackage has been the ENBI Brochure, outlining the structure and function of ENBI, the ways in which interested parties can contribute to or benefit from ENBI, and a list of contacts within ENBI related to the different parts of the programme. To see the ENBI Brochure as a pdf click here



ENBI Newsletter

This ENBI Newsletter is the first of a series of twice-yearly publications that will be an uimportant part of this Workpackage. The editor will be pleased to receive contributions concerning any aspect of biodiversity information. News of events, developments, problems will be very welcome from all ENBI Members for future issues. Please contact Chris Johnson either at cjohnson@maich.g or c.b.johnson@reading.ac.uk. We will try to include one or two longer articles in each issue. If you have an idea for a longer article (like the one on page XX of this issue) please get in touch with the editor for advice first.

ENBI Workshops:

Announcement of Specialised Workshop:
MAKING SPECIES DATABASES INTEROPERABLE

Advanced Workshop
Sponsored by ENBI and GBIF

13 -16 July 2004

Venue: Centre for Biodiversity and Systematics
The University of Reading
Reading, UK

This advanced workshop is intended for those already involved in species databases and will be a participating workshop in the techniques for aggregation and interoperation between such databases.

Particular attention will be given to the technologies, the protocols established by international programmes, wrapper development kits, and ideas for their further development. Participants are likely to be custodians and developers of both species checklist databases and 'Species Bank' databases that hold rich data about species. Developers of interoperable systems and software engineers will be welcome.

Participants from EU countries accepted for the workshop will receive ENBI scholarships covering the full cost of the registration fee (£325), to include all sessions of the workshop and refreshments, but a deposit of £75, returnable on arrival, will be required on acceptance. There will be an accommodation charge from ca. £175 to £125 for three nights, including meals. There will be six sessions (3 nights, four days). The workshop will be limited to a maximum number of 20-25 participants, and will include 'hands-on' computing work.

Programme

Full details of the programme can be viewed at the ENBI Forums page of the ENBI Website (www.enbi.info)

Peliminary application should be made to:
Dr C B Johnson
ENBI Workshop
Department of Natural Products
Mediterranean Agronomic Institute of Chania
73100 Chania
Greece
or by E-mail to: cjohnson@maich.gr

Please give an e-mail address if possible. This will enable application forms and further details to be sent to you by e-mail.

WP 4. IPR, copyrights & financial issues

Kew Workshop on Intellectual Property and Data Repatriation

From Simon Owens, Keeper of the Herbarium, Royal Botanic Gardens, Kew, UK

The workshop held on 29 October, 2003 at the Royal Botanic Gardens, Kew, began by bringing partners up to date with a presentation on Kew's perspective on IP. An important issue to resolve is that of who owns the data and this can have both national and international implications. This presentation and discussion was followed by a practical demonstration of Kew's work on data repatriation.

Recent work on IP and copyright under the auspices of BioCASE and Species 2000 Europa was then presented with focus on copyright and database rights. The session finished with a discussion of questions and issues to be addressed in ENBI. These questions and issues are listed below:-

(1) How to control information flow? Should information flow be controlled and if so by whom?
One risk identified was the impact on third parties by complete access to all information. The group, while agreeing that there was a risk, considered that there were too many countries to control the flow of information and that minimum information could be provided on the web with no perceived problems (Darwin core). Consultation with FISHBASE (www.fishbase.org) was considered to be a step forward to potentially resolving this issue.

(2) Validation of data
Poor quality data, out of date data, and misinformation are among the items which will affect data validity. The group considered the possibility of expert checking or editing of data and considered consulting the International Plant Names Index (www.ipni.org) who face similar problems.

(3) Validity/legality of disclaimers
Many databases now appear with disclaimers. There was no evidence available for their legality or validity and this needs to be investigated.

(4) Data collected during monitoring by multiple providers/observers
Does everyone have to agree to provide these data to databases?

(5) Who owns observational data?
The group could not answer this question and agreed to consult BIOTA (www.biota.org) and groups in Finland and Sweden.

(6) Copyright law: how is it implemented in Europe? Is this causing problems for data providers? How much “breeding” (cultivar) information is free?
The group were not aware of any particular impacts. However, there were examples of information being placed onto the web to prevent patenting e.g. in Peru.
The group noted finally that if information is available, governments can regulate it and whatever comes from it.What happens if information is not available?

WP 5. Cooperation of pan-European checklist and 'Species bank' database projects

First Steps towards Worldwide Collaboration

From Frank Bisby, Centre for Biodiversity and Systematics, University of Reading, Reading, UK

Listing the world's species, the Catalogue of Life, takes a step forward

There is no catalogue of the known organisms on Earth – a fact that surprises many outside the sphere of biodiversity – but a significant step was taken recently towards producing such an index when an international agreement was signed to help develop the Catalogue of Life.

Professor Frank Bisby, of the School of Plant Sciences at the University of Reading, and Dr Michael Ruggiero in Washington, DC, signed agreements with the intergovernmental Global Biodiversity Information Facility (GBIF), based in Copenhagen, Denmark. The partners agreed to use the developing Catalogue of Life, a comprehensive electronic index of all known organisms, as the core species index for GBIF. The Catalogue of Life programme is principally run by Species 2000, based in Reading, and the North American Integrated Taxonomic Information System (ITIS), based at the Smithsonian Institution in Washington.

The GBIF will enable scientists and citizens alike to navigate, extract and analyse the world's vast amounts of biodiversity information. It will enable them to put it to use in generating the economic, environmental, social and scientific benefits from the sustainable use, conservation and study of biodiversity resources (www.gbif.org). In particular, it will make the world's primary biodiversity data on specimens available from the natural history museum collections, botanical gardens, zoos, culture collections, libraries and associated databases all around the world. To do this, it is evolving an interoperable network of the appropriate biodiversity databases and information technology tools.

Professor Bisby, Executive Director of Species 2000, which is a not-for-profit organisation acting as a federation of taxonomic database organisations around the world, said: "The endorsement and partnership that GBIF brings to the programme is expected to make this a major milestone in the already developing global Catalogue of Life programme with partner organisations around the world. These partner organisations specialise in plant, animal, fungal and microbial biodiversity, and work to provide and maintain relevant sectors of the distributed database system. Major partners in the UK include the Natural History Museum, London, the Royal Botanic Gardens, Kew and CABI Bioscience. It is anticipated that the synonymic species checklist pioneered by the Catalogue of Life partnership and made available through this new agreement, will play a key role in the name-service and indexing functions of the GBIF portal."

WP 6. Co-operation of pan-European databases on biological collections and specimens

Practical steps to improve an emergent computerised network of biodiversity information.

From Malcolm Scoble, Department of Entomology, The Natural History Museum, London,UK

Ever more biological collection- and specimen-level databases are being created across Europe as the culture of providing integrated computerised access to data spreads across the region. In Workpackage 6, we aim to realise the broader *ENBI* objective of database co-ordination by taking practical steps to improve materially an emergent computerised network of biodiversity information.

- The *BioCASE* (Biological Collections Access Service for Europe) network architecture (the *ENHSIN* - European Natural History Specimen Information Network - successor) is now established as a suitable network design and is ready to connect 100 data providers around Europe. A crude test portal is operational and demonstrates the networking of several providers of ABCD-standard data. The wrapper has been tested by installing it on different operating systems and database products, as well as with different models of content presentation. Until now the wrapper has been installed for 13 different databases and is being tested by at least 5 other institutions. More than 50 pages of documentation have been created to support data providers in the process of installing and configuring the wrapper. A *BioCASE* provider training workshop was held at the international *TDWG* (Taxonomic Database Working Group) meeting in October 2003, where 30 attendants were instructed in the working of the network architecture and the installation and configuration of the wrapper. Progress provides not just a material contribution to *ENBI* Workpackage 6, but also helps builds the pan-European Network through content development, training and communication. The involvement of this work with other European and wider international efforts encourages ENBI outreach. The project is being carried out by Professor Walter Berendsohn and his team at the Botanischer Garten & Botanisches Museum, Berlin-Dahlem.

- *LepIndex* is a database, constructed by Dr George Beccaloni of the Entomology Department at the Natural History Museum, London. It provides access to species-level information in a unique archive of index cards to a species rich Order of insects (the Lepidoptera – butterflies and moths). Each index card can be treated as the equivalent of a specimen, and *LepIndex* provides access to digital images of these cards. The nomenclatural data contained in the archive provides a connection with Workpackage 5 of *ENBI* and with *Species 2000 Europa*, thus broadening the networking component of the *ENBI* demonstrator. The *ENBI* project has added images of 127,049 cards to the database in a searchable format at the following website: http://www.nhm.ac.uk/entomology/lepindex/)

- Digital imaging standards for type specimens. This component of the Workpackage will commence formally in the second year of ENBI with a workshop to be held at the Staatliches Museum für Naturkunde in Stuttgart, Germany, in March 2003. To inform the first workshop, a demonstrator database of a selected group of freshwater fishes is being constructed by Dr Darrell Siebert and his team at the Zoology Department of the Natural History Museum, London, with the incorporation of radiographs and images of whole specimens.

WP8 - Data management in large distributed biodiversity database systems

Analysis, identification, and characterization of the distributed information management requirements for ENBI

From Ozgul Unal & Hamideh Afsarmanesh, University of Amsterdam – Informatics Institute, Amsterdam, The Netherlands

This workpackage primarily aims at the analysis and design of a cooperative federated information management system inter-linking and integrating the varied biodiversity information types (e.g. on taxonomy, collections, observation, etc.) that are distributed among different databases. This goal will be achieved in close collaboration and based on the common interoperability approach of WP9. As such, the WP8 will mainly analyze and introduce an integrated / unified data description (schema) and a federated database management architecture, in order to integrate the *wide variety of information*, whilst WP9 addresses the *diversity of data sources* through an approach for adaptation and harmonization using a common interoperability approach.

The first step towards the achievement of WP8 objectives will be a full analysis, identification, and characterization of the distributed information management requirements for the ENBI biodiversity application domain. The identified requirements will include both the modeling constructs (entities and concepts) and the functional (access and manipulation) requirements for the target federated information management architecture. As a part of this feasibility analysis stage, a study of related development approaches (e.g. Java, CORBA, XML, WSDL), and data description models in the biodiversity domain will be carried out and reported.
Based on the identified distributed information management requirements, the individual components of the federated database architecture will be specifically designed and tailored to the ENBI application domain.

The first deliverable of WP8, titled *D8.1a-'Draft report on Distributed Information Management Requirement Analysis'*, has been developed by University of Amsterdam – Informatics Institute. This is a restricted deliverable that has been distributed since October 2003 within the group of experts from different biodiversity projects related to ENBI and is now available in the library of Cluster III Interest Group in ENBI-CIRCA.
The main purpose of the deliverable is to identify the applicable information management requirements within the scope of ENBI, in order to form the basis for succeeding phases in WP8. As such, the following subjects are included in the deliverable:

1- *General description of information management in biodiversity domain.*
2- *Preliminary analysis of the environment, related research, standards, technologies and paradigms.*

However, *D8.1a* corresponds to only the "first part" of the Requirements Analysis of ENBI in WP8. This report will be followed by a more detailed, final requirement analysis report, titled *Deliverable D8.1b-'Final report on Distributed Information Management Requirement Analysis, including characterization of ENBI database schemas'* that will be released through the ENBI-CIRCA early in 2004.

WP 10. Generic analysis tools and data mining

Analysis tools for biodiversity databases

From Michael Malicky, Oberösterreichisches Landesmuseum, Linz, Austria

We are just preparing our first deliverable, which will be ready by the end of December and include a list of analysis tools for biodiversity databases. A special focus will be on web enabled tools.

From 21st to 28th September 2003 the 18th SIEEC Symposium was held in Linz/Austria. In my speech I talked about the future of our biodiversity database ZOBODAT. Our major goals are to link the database to GBIF, this link should be espablished soon, as we already registered to GBIF. Furthermore I promoted ENBI to the audience. The majority of the people there were entomologists from central Europe. (Austria, Germany, Switzerland, Czech Republic, Romania, Ukraine).

In November 2002 the European Invertebarte Survey held a small meeting. ENBI, BioCASE and GBIF where mentioned there and promoted. In 2003 a prototype mapper, which statically linked 15 Invertebrate species to a European Map was generated and is now accessible for a first example for mapping data from different databases at: http://www.zobodat.at/biowww/zobo/eis.php.

WP11. Multi-lingual Access to European Biodiversity Sites.

Still a Challenge: Machine translation 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. 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.

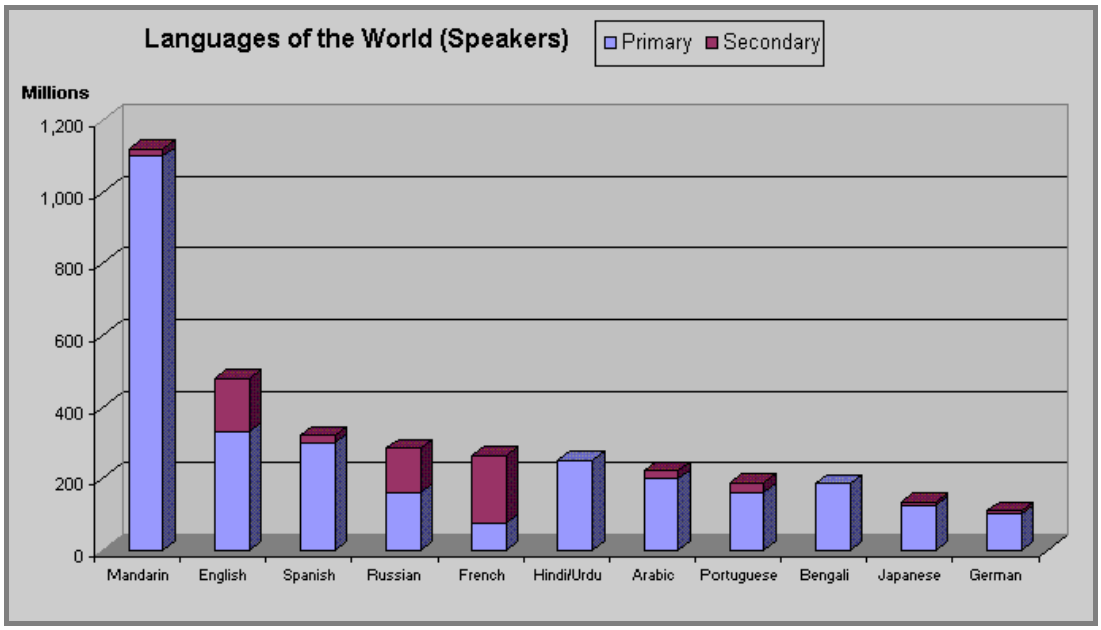


Fig. 1: Most spoken languages of the World in numbers for primary and secondary speakers.

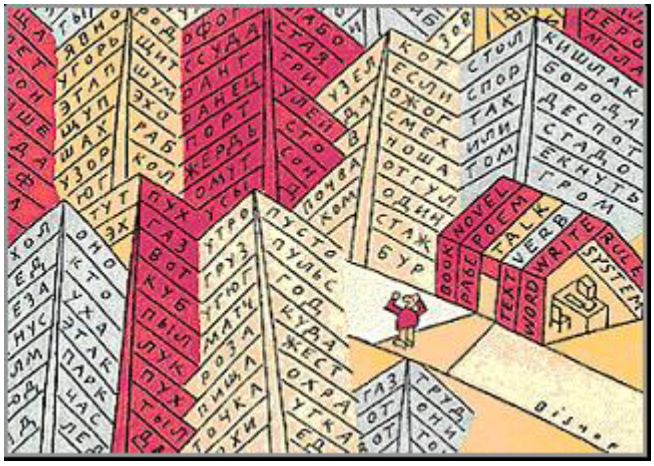
The Internet is also creating new gaps between the rich and the poor. Rich countries with well established educational systems have much greater access to the internet and communications services generally. Although the Internet started off as a communal medium for sharing information, principally among 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 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 2005, 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 kinds 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 machine translation (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 Present

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). 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 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.

Its not that easy...

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 beginnings 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. 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.

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. 2)

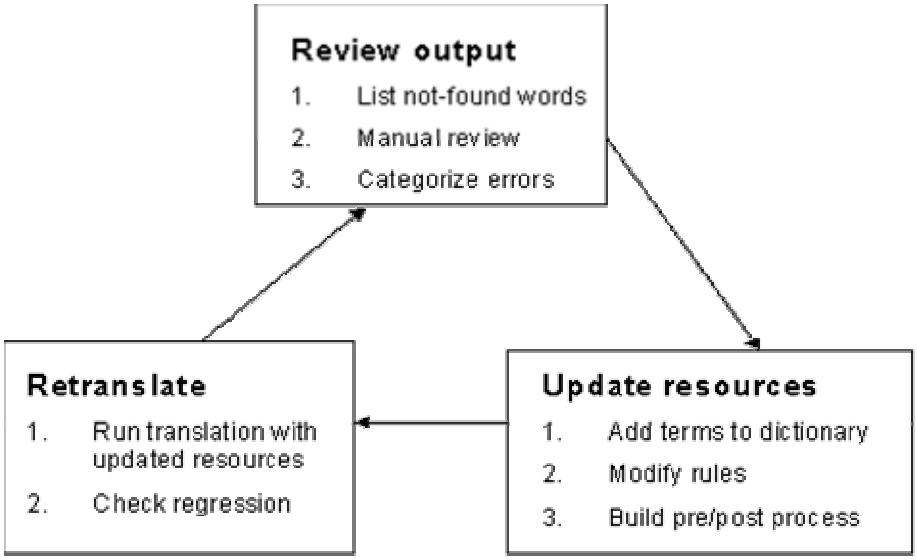


Fig. 2: MT- Quality Enhancement Process

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.

Update Resources and Enhance Source Text

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.

Retranslate and Validate

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 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 e.g. by 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 result in poor MT versions. The more uncreative a text, the better the results. These rules should be 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 will create special biodiversity dictionaries to be integrated in 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, (IBM Pervasive Computing):

"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 becoming a member of the EU, the demand for MT will increase tremendously.

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.

A brief, but not so serious summary on Machine Translation's Past and Future from 1629 to 2264

1629 René Descartes proposes a universal language, with equivalent ideas in different tongues sharing one symbol.

1933 Russian Petr Smirnov- Troyanskii patents a device for transforming word-root sequences into their other-language equivalents.

1939 Bell Labs demonstrates the first electronic speech-synthesizing device at the New York World's Fair.

1949 Warren Weaver, director of the Rockefeller Foundation's natural sciences division, drafts a memorandum for peer review outlining the prospects of machine translation (MT).

1952 Yehoshua Bar-Hillel, MIT's first full-time MT researcher, organizes the maiden MT conference.

1954 First public demo of computer translation at Georgetown University: 49 Russian sentences are translated into English using a 250-word vocabulary and 6 grammar rules.

1960 Bar-Hillel publishes his report arguing that fully automatic and accurate translation systems are, in principle, impossible.

1964 The National Academy of Sciences creates the Automatic Language Processing Advisory Committee (Alpac) to study MT's feasibility.

1966 Alpac publishes a report on MT concluding that years of research haven't produced useful results. The outcome is a halt in federal funding for machine translation R&D.

1967 L. E. Baum and colleagues at the Institute for Defense Analyses (IDA) in Princeton, New Jersey, develop hidden Markov models, the mathematical backbone of continuous-speech recognition.

1968 Peter Toma, a former Georgetown University linguist, starts one of the first MT companies, Language Automated Translation System and Electronic Communications (Latsec).

1969 In Middletown, New York, Charles Byrne and Bernard Scott found Logos to develop MT systems.

1978 Arpa's Network Speech Compression (NSC) project transmits the first spoken words over the Internet.

1982 Janet and Jim Baker found Newton, Massachusetts-based Dragon Systems.

1983 The Automated Language Processing System (ALPS) is the first MT software for a microcomputer.

1985 Darpa launches its speech recognition program.

1986 Japan launches the ATR Interpreting Telecommunication Research Laboratories (ATR-ITL) to study multilingual speech translation.

1987 In Belgium, Jo Lernout and Pol Hauspie found Lernout & Hauspie.

1988 Researchers at IBM's Thomas J. Watson Research Center revive statistical MT methods that equate parallel texts, then calculate the probabilities that words in one version will correspond to words in another.

1990 Dragon Systems releases its 30,000-word-strong DragonDictate, the first retailed speech-to-text system for general-purpose dictation on PCs.

1991 The first translator-dedicated workstations appear, including STAR's Transit, IBM's TranslationManager, Canadian Translation Services' PTT, and Eurolang's Optimizer.

1992 ATR-ITL founds the Consortium for Speech Translation Advanced Research (C-STAR), which gives the first public demo of phone translation between English, German, and Japanese.
1993 The German-funded Verbmobil project gets under way. Researchers focus on portable systems for face-to-face English-language business negotiations in German and Japanese.

BBN Technologies demonstrates the first off-the-shelf MT workstation for real-time, large-vocabulary (20,000 words), speaker-independent, continuous-speech-recognition software.

1994 Free Systran machine translation is available in select CompuServe chat forums.

1997 AltaVista's Babel Fish offers real-time Systran translation on the Web.

Dragon Systems' NaturallySpeaking and IBM's ViaVoice are the first large-vocabulary continuous-speech-recognition products for PCs.

Parlance Corporation, a BBN Technologies spin-off, releases Name Connector, the first large-vocabulary internal switchboard that routes phone calls by hearing a spoken name.

1999 A televised newscast is automatically transcribed with 85 percent accuracy.

Logos releases e.Sense Enterprise Translation, the first Web-enabled multiple translator operating from a single server.

IBM releases ViaVoice for the Macintosh, the first continuous-speech-recognition Mac software.

Kevin Knight, of the University of Southern California's Information Sciences Institute (ISI), leads a multi-university team that develops Egypt, a software toolkit for building statistical MT systems. Egypt examines bilingual texts for statistical relationships, analyzes those patterns, and applies what it has "learned" to its translation functions.

2000 At MIT's Lincoln Laboratory, Young-Suk Lee and Clifford Weinstein demonstrate an advanced Korean-English speech-to-speech translation-system prototype.

USC's ISI performs backward machine-transliterations of proper nouns, which are replaced with phonetic approximations.Southern California translates to "Janoub Kalyfornya" in Arabic.

2001 Carnegie Mellon University's Language Technologies Institute (LTI), led by Jaime Carbonell, constructs speech-to-speech translation for "small" languages like Croatian or Mapudungun, spoken by Mapuches in Chile.

USC biomedical engineers Theodore Berger and Jim-Shih Liaw create a new Berger-Liaw Neural Network Speech Recognition System (SRS) that understands spoken languages better than humans do. Ford says the technology will be incorporated into its cars to facilitate communication at fast-food drive-thrus.

2002 NowHear offers an agent-based newsreader device that translates articles from thousands of publications worldwide, delivering them as MP3 audio files.

2003 Text of Joyce'sUlysses is run through Cliff's Notemaker, a new omnidirectional literary interpreter and summarizer. Program: "Your professor didn't read it either. Don't worry about what your essay says, just include the words Dublin, pub, and fuck."

2004 Dragon Systems' NaturallyCursing software is added to wristwatches to ease communication at multilingual construction sites.

2005 Employee at Allstate Insurance files suit against the company, citing emotional distress from the collective chatter of coworkers using speech recognition input devices.

GeoCities pulls down 350,000 homepages for failing to use GeoCities Controlled English, a 1,000-word edictionary designed to interface with its language translation software.

2006 It's that .001 percent part that got us," moans NASA director Rafu Sanjali, after the fourth disastrous attempt to land a robot-controlled vehicle on Mars was foiled by the use of "99.999 percent accurate" MT technology.

2007 Microsoft pulls its "What do you want to think today?" campaign after reviewers unanimously trounce the company's much-anticipated Thought Recognition Interface (TRI).

2008 L&H's Travel Sunglasses offer real-time translation of road signs, marquees, and menus into a wearer's native language.

2009 CorconText introduces FinalCopy, a Japanese-to-English documentation translation program that uses AI-based semantic networks to reduce the need for human editing of output.

2012 Saruzuno embeds its Lexical Disambiguation System (LDS) into smartcards equipped with membrane microphones so travelers can converse with store clerks in dozens of languages.

2017 The Russian-made Durok II language tutor is used to train customs-and-immigrations bots (DNA-based servant-devices) employed at US points of entry.

2020 Teaching a child reading and writing is a waste of time," declares Yeo Kiah Wei, Singapore's minister of education, who cancels the subjects in schools. "Children needn't be burdened with such an onerous task as deciphering tiny markings on a page or screen. Leave it to the machines."

2021 PigLatin Furby reveals parents' plans for divorce. Dozens of toddlers are traumatized.

2043 Tower of Babel is completed in Iraq (formerly Babylonia) after a 4,000-year delay, thanks to NEC Technologies' Neutral Language.

2045 Telepathy system developed by Europeans. Users wear adhesive patches containing thought recognition and MT technology, plus a high-speed wireless transceiver.

2058 The Reformed Rifkin Institute (RRI) is awarded a patent for its invention of a symbio-parasite that feeds on the electrical impulses in the speech center of the human brain, then excretes a translated signal that can be understood by anyone who inserts the creature in their ear. The estate of Douglas Adams files suit, claiming prior art.

2108 Procter & Gamble researchers use their newly developed Distributed Tachyon Swarm System (DTSS) to learn that diphtheria bacteria band together as a hive mind capable of communication.

2264 "Humans are dumber than bags of hair," declares Entity 296. "Only the most naive scientist would try to develop a technology to understand those smelly lumps of protoplasm," it states. "The noises they emit from the holes in their heads are ultimately less enlightening than cosmic static."

Compiled by Kristin Demos (kvdemos@yahoo.com)
and Mark Frauenfelder (mark@well.com)

WP 12: Information services on European biodiversity data

Feasibility Studies, State of the Art Inventory, Glossary

From Christian Köppel, Verlag für Interaktive Medien GbR, Germany

1. Feasibility studies and design (pilots) The objective of work package 12 is to provide information about needs of European users on biodiversity data. In this frame also a few feasibility studies and design (pilots) will be supported in 2004.The following proposals had been submitted and were reviewed by an advisory board:

- Illustrated Digital Keys to North Sea Biota. - Rob van Soest (University of Amsterdam, NL)
- A pan-European key to the aquatic stages of mayfly. - Kearon McNicol (FreshwaterLife, UK). A model for the use of information on biodiversity in the context of urban ecology. With special reference to socio-economical aspects. - Manfred Ade + Barbara Di Giovanni (College for management and design of sustainable development, Berlin, DE + ENEA, Rome, IT) Database creation that will provide karyological summaries for the vertebrates (excluding fish), distributed in Europe. - George Mitsainas (University of Patras, GR)

During a workshop at Kew Gardens (28.10.2003) the positive and negative aspects of the evaluation of each proposal were discussed. It was pointed out that the pilots should set an example for future developments. In this view we are not interested in the product of a specific pilot study, but we rather want to learn how the process and mechanism can be maximised. Interesting pilot projects are those that will involve many databases from different institutes, together with IT tools from other institutes or networks).

2. Inventory of the State of the Art in Europe. The following two services are being established by V.I.M. (programming, design and content) together with WP 12.

2.1. Glossary

Content: app. 2,000 acronyms and abbreviations:

- technical terms, e.g. XML, html, http, LDAP
- terms in the area of Biodiversity informatics, e.g. GIS
- organisations, e.g. GBIF, ENBI, GTI, Species2000
- museums and institutions, e.g. SMNS, MNHN, ETI
- working groups and committees, e.g. TDWG, CODATA, ABCD

Features are:

- short descriptions, explanations and comments
- links with automatic and periodical URL-verification
- online editing by registered userscontinuous updating

The Glossary is available at http://www.enbi.info/forums/.

2.2. Biodiversity databases and database projects

The aim of this service is to give a fast overview on “Biodiversity databases and database projects” incl. contacts (e.g. responsible officer) with focus on Europe. It gives also the opportunity to present the own databases in the ENBI network.

Service:

- for finding new partners e.g. in EU projects (framework VI)

- for general information: who is who
- for content search: where to find observational data, conservation data, images

Methodology: Online data input and data editing by registered users. Member of the ENBI network (> 130 members, who represent at least one institution or organisation) organising the data input and data editing of their own databases and projects in this online database.
It is planed to circulate the invitation for online editing first in the ENBI network, then in other projects like Species 2000 Europa, Fauna Europaea, Euro+Med PlantBase, ERMS, etc.
The advantages of the online data input and data editing are that the database owners/holders have the most extensive knowledge about their databases (e.g. content, address, url), and online data entry is the best and fastest way to minimise errors and to correct them quickly.
There is no overlap with the BioCASE database (A Biological Collection Access Service for Europe)? The database is very simple, much less detailed than the BioCASE database, which has a taxonomic backbone for querying information of collections in museums and institutions.
The database about "Biodiversity databases and database projects" is available at ENBI Forums website.

3. Next steps

- Identification of priorities of specific user groups
- Tool (server based software + survey) to assess user-friendliness of biodiversity information services

---

NEXT NEWSLETTER.....

For the next newsletter, to be published in October 2004, we would like to receive as many items as possible from any ENBI members. Please send any news items you have to:

Dr C B Johnson, Editor ENBI Newsletter,
Department of Natural Products, MAICh, 73100 Chania, Greece
Or by E-mail to cjohnson@maich.gr

We will also include reviews on subjects related to ENBI's activities. Please consult the editor before submitting review articles.

ENBI Newsletter is published as part of ENBI WP3 at the Mediterranean Agronomic Institute of Chania, Greece.
The European Network for Biodiversity Information is funded by the fifth framework programme of the European Community.