Technical report No 34

The Global Biodiversity Information Facility: Architectural and implementation issues

"What would Linnaeus have invented,
if he had lived in the 21st century and grown up with the Internet?"

Prepared by:
Hannu Saarenmaa
European Environment Agency

August 1999

European Environment Agency

Cover design: Rolf Kuchling, EEA
Layout: Pia Schmidt, EEA

Legal notice

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int)

Printed in Copenhagen

Printed on recycled and chlorine-free bleached paper

European Environment Agency
Kongens Nytorv 6
DK-1050 Copenhagen K
Denmark
Tel: +45 33 36 71 00
Fax: +45 33 36 71 99
E-mail: eea@eea.eu.int

Content

Abstract

This paper supports the OECD Megascience Forum for Biological Informatics on the technical aspects of its plan for the Global Biological Information Facility. The paper identifies the capacity to produce homepages for all species of organisms as the main goal of GBIF, but in such a way that the homepages are dynamically derived from online databases. GBIF should lead to an accelerated rate of describing new species and to new information markets on biodiversity, and complement the Clearing House Mechanism of the Convention of Biological Diversity. It is seen as the critical success factor that an infrastructure is erected for biodiversity similar to what exists for molecular biology. Its cornerstones are regional centres that provide longevity and co-ordination, a distributed object-oriented database architecture based on co-operating agents, data interchange with XML, and seamless use of both existing and new databases. At the heart of the infrastructure, a new Biological Addressing System is suggested that maps the volatile but commonly used scientific names to stable Biodiversity Identifiers that are derived from IPv6. A separate treatment for the name and taxon concepts is deemed essential in this architecture. Finally, issues for research and education are discussed.

—

Please also see acknowledgements and release notes at the end of document.

1. Introduction

The Megascience Forum[1] of the OECD is a committee that brings together science policy makers from OECD Member governments to discuss ways of strengthening international co-operation on very large scientific projects and programmes. In 1995 the Megascience Forum established a series of Working Groups to look at the major challenges in some critical areas. Biological informatics was one of these, and its work has proceeded in two subgroups, which are neuroinformatics and biodiversity informatics.

State of information management in biodiversity is currently very unsatisfactory. There are many valuable national services, such as Costa Rica's INBio[2], Mexico's CONABIO[3], and USDA ITIS[4], and also some global initiatives such as DIVERSITAS[5], Species 2000[6], and the Tree of Life[7]. However, these actions are not coordinated with each other, and subsequently, there is no common information architecture that would easily allow contributions of data, information, and knowledge, so that there would be an easily accessible shared global information repository on biodiversity. Developments around the Clearing-House Mechanism (CHM) under the Convention on Biological Diversity[8] (CBD) don't seem to aim at one, either.

There are many reasons for the current situation, but probably the leading one is the strong tradition in taxonomic research that does not see taxonomy as an informatics science. Subsequently, the Codes of Nomenclature [1, 2] that in principle should facilitate interoperability, have not supported modern approaches to information management. However, in the 4th edition of the ICZN the door is now opened for electronic publishing of new names.

Therefore, additional layers and supporting facilities must be sought for. The work of the Subgroup on Biodiversity Informatics of the OECD Megascience Forum's Working Group on Biological Informatics is probably the most ambitious of these. It aims at Global Biodiversity Information Facility (GBIF), which can be simply characterized as a new global service that should bring information management within biodiversity to a level, which is comparable to the services available today for molecular genetics.

The work of the Subgroup on Biodiversity Informatics is now completed and their report [3] was presented together with the reports of the other Working Groups [4] to the meeting of OECD science ministers in 1999. The report was adopted.

Such reports are aimed at political decision makers, and cannot be very specific on technical solutions or about their exact and desired outcomes. Therefore, there is need for supplementary analysis, inventorying of options, and planning that will support the implementation work when it starts. The current work presents such findings and visions. They have been born during the preparation of the above reports and some of them have been taken up in the main reports. However, most of the material below should be seen as technological options and possible avenues to pursue after the work begins. These findings are not repeated in the above-mentioned reports, and conversely, this work dies not repeat the content in the Working Group reports, either.

[1] http://www.oecd.org/dsti/sti/s_t/ms/index.htm
[2] http://www.inbio.ac.cr/
[3] http://www.conabio.gob.mx/
[4] http://plants.usda.gov/itis/
[5] http://www.lmcp.jussieu.fr/icsu/DIVERSITAS/
[6] http://www.species2000.org/
[7] http://phylogeny.arizona.edu/tree/phylogeny.html
[8] http://www.biodiv.org/

The first section establishes the broad vision for the technical implementation of the GBIF. We first examine the **goals and desired outcomes** in some detail, trying to identify them as accurately as feasible at this stage of planning. Then we examine what kind of **strategy** might deliver these outcomes. Finally we look at the **conditions and elements** that would need to be in place for the strategies to work. We start from the big picture, and then consider the goals and means against several different timespans.

The discussion in the current paper covers species-level biodiversity informatics only. Notwithstanding the importance of information at the more aggregated levels: habitats, sites (including protected sites), and ecosystems, they are left for a later scrutiny. However, this is only done for the purpose of concentrating on the most critical issue solvable now using the currently emerging technologies, the large user community of ecologists and taxonomists, and the current political processes such ad the CBD. It is also assumed that if the infrastructure for the biodiversity informatics at species-level is erected now, it will form the basis for solutions at the more aggregated levels later. Moreover, information on habitats and sites already have been tackled by major projects, such as those of the WCMC and European Union's CORINE and Natura 2000 projects. Also the technological solutions for managing habitats and sites through GIS and databases are already available, and the recent breakthrough of Internet may not have created a similar opportunity in these areas, as it has done with species information that is much more distributed. This opinion is formed also knowing that at the gene-level, working solutions have been found through the GenBank in USA, European Molecular Biology laboratory's DNA Data Library, and DNA Database of Japan. There are lessons to be learned from that positive experience.

2. Desired outcomes

2.1. Target situation

2.1.1. Homepages of all species on Internet

Homepages for all the species of organisms on planet is the overwhelmingly most important target for GBIF. It is a concrete thing that links most pieces of information together. Without it, other goals would be much more difficult to achieve. This means that GBIF shall concentrate on real data, not just metainformation or information locators. However, the main focus would not be on observational records, but on their summaries.

Here, species homepages must not be taken as mere static HTML pages. Instead, species homepages must be understood as dynamically created *information pages*, visualizations of some core objects living in and being linked to a global biological information infrastructure. This infrastructure for species information is the main outcome of GBIF, the capacity to produce homepages is just one of its results.

In computer terms, we are talking here about distributed objects that are permanently stored in on-line databases. These objects have attributes such as scientific name, synonyms, publication where described, diagnostic characters, geographic distribution, and models for life cycle. These objects are also linked to other similar objects like their host and parasitic species objects, records and images of type specimens, superior and subordinate taxa, habitat objects, and even field observation records. It must be possible and it is even desirable to have parallel instances of these species objects maintained by different research groups. These parallel instances can then be linked together by queries to generate a big picture of the species as it is known on planet.

The word *species* above could be replaced by *taxon or clade*. However, it must be made clear that the above goal is not dependent on any particular interpretation of systematics or even an organizing methodology thereof. Traditional scientific names are not either used as keys or as a linking mechanism. They are mere associated data. In other words, a taxon object for which the homepages are required, shall be seen as a separate concept from a name object. The former is almost synonymous to a biological population, and can have individuals or their cohorts as instances and hence is open-ended. A taxon object can even be temporarily unnamed. The latter is classical data; it is a closed data set defined by the Codes and human publishing activity.

A new addressing system for the taxon objects will have to be created, simultaneously maintaining the Linnean one as data. Such an addressing system can be handled by technical interoperability protocols, such as Z39.50, LDAP, or CORBA, available on Internet from the industry and object standard bodies. The possibilities for data interchange have recently become much easier and practical with XML.

To visualize the goals further, a taxon object looks, when fully printed from one perspective, much like (the original) description of that taxon in a scientific journal. From another perspective, it could look like a colourful description of the animal in a biological field guide. Several of these perspectives will be possible to derive from the same basic objects using XML representation. While one homepage of a species might be interesting, but of limited use, having them all and linking them together will be a different matter altogether.

Prototypes of species homepages or taxonomic databases can already be found in many Internet sites. Some of the below sources provide excellent ideas on how the GBIF may sometimes look to users. However, none of these have been designed as an infrastructure that would scale up to a global level.

- Tree of Life project[9]
- Costa Rican InBio. This service also contains world's first descriptions of new species *Tetranema gamboanum* Grayum & Hammel 1996 and *T.floribundum* Grayum & Hammel 1996 that appeared as preprint on Internet[10]
- *Cosmotrice lobulina* (Denis & Schiffermüller 1775), a sample Lepidoptera homepage in XML[11]

2.1.2. Description rate of new species multiplied

A current problem is that only about 18000 new species are described per year. Despite a general increase in scientific funding and productivity elsewhere, this rate has not changed during the past 30 years. With the current rate of extinction, which happens to be about the same magnitude, millions of species will disappear even before they become known to mankind. It is critical that the current slow pace of research could be increased. This, however, shall not be pursued for only to satisfy scientific curiosity, but to provide better basis for protective measures and natural resource utilization.

There are lots of reasons for the sluggishness of taxonomic research, but the technical problems in comparing specimens in collections world-wide is probably a leading cause. Another one is the mixing of keys and data: In the current international codes of biological nomenclatures, time lags of years have *purposefully* been introduced in the process of accepting new names in fear of instabilizing the name system (i.e., keys that link information). If taxonomic information was available on Internet as outlined above, these time lags could be removed. The time lag from a field observation to the description of a new species and the information being available to other researchers could drop from years to days. This is what happened in molecular biology when GenBank was introduced in 1987, which greatly has contributed to the recent success of biotechnology.

2.1.3. Multiplier effects and new information markets

Now what would be the wider benefits of an on-line infrastructure for species information? It provides a linking mechanism to lots of things. Observation data could be better linked to habitat data, which knowledge is needed for preservation programmes. Probably nowhere on Earth the only known site of some species is deliberately destroyed if conservation agencies only held the information and compensation mechanisms be available. If this knowledge were available on Internet, new kinds of better-targeted international funding programmes could be invented to assist poorer nations and landowners in protecting the nature where it most matters.

Under pressure from human population, biodiversity will only be saved if it is valued. Tropical countries with rich nature would gain most if biodiversity's value increased. Information of biodiversity as stored in the above-described network may not necessarily be free for other than academic research and those who contribute to this information. It might be possible to attach a micropayment to each query, because it naturally would cost for the service provider to produce this information. Who would pay? Companies doing bioprospecting are one clear customer. Countries' own biological surveys could provide the information and bioprospecting could be done on Internet. Other customers could be

[9] Also see footnotes on page 4.
[10] http://www.inbio.ac.cr/papers/Tetranema/tetpage.html
[11] http://jaguar.eea.eu.int/life/animalia/invertebrata/insecta/lepidoptera/bombycidae/cosmotriche/lobulina.xml

ecotourism businesses who need to attract customers with information on their biological resources.

These possibilities are partly imaginary and cannot be fully foreseen. But it is probably safe to conclude that if information on biodiversity was available on Internet, it would be likely to increase its real value and generate new kinds of opportunities and markets. It would fully take care of the need to repatriate information to the originating countries, which has been agreed under the CBD. There will be a wide range of incremental uses for this information, such as education and life-long learning. In any case it will create a new marketplace for global funding schemes for development programmes.

2.2. Immediate desired outcomes (one year)

In this phase, the only expected outcomes are international recognition of GBIF, adoption of its vision, its work programme, plans for the next five years, and establishment of the first regional centres. These include the following:

- GBIF shall be found a place in the framework of the CBD. It complements the CHM, which otherwise focuses on technology transfer and metainformation. CHM provides the real data layer into which CHM can link to.
- As it is proposed that OECD adopt GBIF concept, it is instrumental that OECD provide GBIF with its core funding to get started.
- GBIF should find its way in national and international research and development programmes. In fact, EU's 5th framework programme for research and development is going to recognize care of the ecosystems, including biodiversity, as one key area.
- Endorsement of GBIF by the major information providers in museums shall be obtained. This will probably be a slow process, as GBIF necessitates some rethinking in codes of nomenclature and general acceptance of electronic publishing – almost heretic thoughts to many. With a proper infrastructure and funding schemes to support these actions, acceptance should be possible to obtain.
- General public shall be made well informed about GBIF. Its educational aspects shall be well presented.
- GBIF's major organizational architecture shall be installed in the beginning of project. A steering body must be established and first centres nominated and their planning work started. Each of the centres should operate a regional network, whose major players shall be identified and pulled into round tables.

2.3. Near-term desired outcomes (less than 5 years)

By the end of the first five years covered by GBIF's first multiannual work programme, it shall be possible to attain the following results:

- It is possible to locate and produce homepages of 100,000 taxon objects on Internet.
- Publication of new species on Internet becomes viable as the supporting infrastructure gains credibility.
- Issuing of unique globally accepted identifiers (keys) for all used names has been started. This will enhance interoperability between systems.
- Programme for making information of type specimens available on Internet has been started. This will make possible the remote identification of specimens.
- Software architecture of GBIF created. A master plan for interoperability has been created and is being implemented. This master plan likely is two-pronged: use existing databases, but with unique global keys, and
- start development of new category of interoperable servers for biodiversity informatics (GBIF Server). This software is based on distributed objects and will allow an increased degree of interoperability. It shall incorporate a new addressing (naming) system for taxonomic objects. It shall also produce a turn-key package of software

and organizational aid for capacity building programmes and technology transfer especially made for the biodiversity-rich 3rd world.
- Basics of infrastructure, i.e., all regional centres and most national networks started.

2.4. Mid-term desired outcomes (5 to 15 years)

Typically, it takes about fifteen years between the invention of a new technology until its impact really becomes obvious in the society. In case of GBIF, these expected impacts after 15 years of operation are:

- It will be possible to locate home pages of 1,600,000 species on Internet. That approaches about 90% of known species. This mere fact will convince users to look at GBIF first for any biodiversity information.
- Make publication of new species on Internet the preferred practice. Up until now it has been an additional task to make a preprint or electronic version of species descriptions available. However, about 15 years from the start of GBIF, electronic descriptions are the ones used. Printed copies and the traditional publishing process remain, but are increasingly recognized necessary only for archive and safety reasons.
- The rate of new species descriptions increased from 18,000/y to 50,000/y, tripling from what it has been. This increase is due to faster access to species descriptions and to digitized type specimens. Especially, research is accelerated in 3rd world countries who now have access to a first class virtual library at low cost.
- It will be possible to make global views of distribution maps, synthesis of information by querying the distributed species objects from across hundreds of different servers across the world.
- The backlog of issuing unique globally accepted keys for all used names has been cleared. These unique keys are now used by all taxonomic databases.
- The backlog of making information on type specimens available on Internet has been cleared. Major efforts have been made by institutions that own these materials to digitize them.
- Use of GBIF Servers widespread, development efforts synchronized and a user community evolves. Standalone GBIF Servers spread especially in less advantageous countries.
- A scheme for remote identification with multimedia has been established. As the basic character data is available in species objects and type specimens have been digitized, new types of knowledge-based applications will be devised to automate diagnoses. Moreover, remote experts can be consulted with videoconference and unknown specimens can be digitized and identified by remote experts.
- Description of new species from finding one to publishing drops to an average of 3 months.

2.5. Long-term effects (15-50 years)

- It is possible to locate home pages of 4,000,000 (all by then known) species on Internet.
- However, it remains as fact that a full inventory of all species on planet will only be accomplished in 100-200 years' time. A dampening the rate of new descriptions per year, which in its peak may reach 100,000 will become now apparent.
- Make publication of new species on Internet the only acceptable practice. Descriptions in printed media alone are no longer recognized valid by codes of nomenclature.
- The time lag in description of new species from finding to publishing is reduced to 1 month. Type specimens are digitized by default.
- Transfer of all known old species descriptions to Internet has been completed.
- It will become possible to do new kind of integrative ecology by linking species objects to each other, generating models of ecosystems and foodchains. The species

objects will contain more and more dynamic models of their life cycle, habitat requirements, responses to environmental pressure. This allows decision support and what-if scenarios being run against global biodiversity data.

- Increase in appreciation of taxonomy and systematics as modern subjects will be attained as a result of all these endeavours.

3. Strategies

This section makes an attempt to outline the key actions necessary to achieve the above goals.

3.1. Critical success factors (overall)

An infrastructure for biodiversity informatics must be erected similar to that already in place for molecular biology. This means establishing regional centres for each major biographical region that is the keeper of the original information originating from that area. To complement that, national centres and networks shall exist. In addition, thematic networks across regions may exist.

Standards on biodiversity objects must be agreed upon the same way standards have been created for other information types on Internet, such as addresses, certificates, encryption and encoding. However even with an enthusiastic user community, these standards alone are not enough. Infrastructure is needed to provide longevity, archiving and co-ordination, i.e., credibility to the system so that it can be trusted information that shall last thousands of years.

A two-pronged strategy for connectivity and technological development shall be adopted: existing databases, systems and networks shall be used, supported, linked to and enhanced in a medium term. Meanwhile, development of new types of technological solutions that make full use of the newly emerged Internet technologies will be initiated and gradually rolled in.

User acceptance shall be obtained with an open dialogue, participation, continuous demonstration, and putting together existing "building blocks" in innovative ways. This way understanding and acceptance of the wider concept is built and useful products are delivered in the interim.

Data harmonization and conversion have to be addresses systematically. Especially, synonymy will become a major problem with a much higher rate of species descriptions as envisaged here. This can only be tackled by the proposed new taxonomic addressing system of taxon objects, which also maintains and supports the Linnean one as data.

3.2. Immediate (one year)

Immediate actions that should be taken within the first year of GBIF's existence are the following:

- A vision paper for GBIF shall be written and with a high visibility be published in each of the regions. A conference shall be held to establish a momentum behind GBIF and review the needs of the user community, crystallize the vision, and review the plan.
- A user needs assessment should be done to drive the priorities for development and help in identifying resources. The concept of species homepages, especially when implemented the way they are proposed here will be unfamiliar to many. Demonstrators and pilot applications will be built and the ideas refined with users.
- Review of existing information services that could be built on and co-operate with GBIF will be done.
- Decisions by appropriate authorities shall be made to establish GBIF's legitimacy. Following endorsement by OECD, those of CBD's 5th COP, EU and USA must be sought. Task forces shall be appointed to propose these legal bases. Following their

adoption, a permanent task force shall be appointed to act as the management board of GBIF.
- Strategic and financial planning will be key activities during this period. Core funding should be secured. Certain amounts should be allocated to GBIF as a whole to be used following a global evaluation of priorities.
- The cores of the first regional centres shall be established. These centres work under the authority of the management board but shall also respond to regional priorities and conditions. Their funding from local sources shall be possible.

3.3. Near-term (less than 5 years)

In five years time, several of the above activities will continue, but also the following ones shall take place:

- The remaining regional centres and their co-operative mechanisms shall be established.
- Establish national centres. These are typically existing major institutions already engaged in taxonomic work. Reorientation to providing more electronic content in addition to curating collections is necessary in almost every country.
- Establish user community. Following the initial conference, regular communication must take place in a similar forum. Discussion within the user community on GBIF's vision will have to evolve.
- Establish a quality assurance programme for the information being provided. This is likely to involve an accreditation mechanism of taxonomy experts who can review information that can be found in networks. For instance, a specimen is observed and digitized somewhere in the tropics by a local expert. This specimen can be examined by a foreign licensed specialist who can attach his/her electronic signature as a validation to the identification.
- Establish research programme to biodiversity informatics. This will be done by the appropriate regional authorities such as EU and USA.
- Establish XML Data Type Definitions (DTD) for taxon, name, and related objects.
- Establish software architecture that allows dynamic linking of parallel homepages. A proposal must be made for the Internet Task Force about the standard content and interfaces of a taxon object.
- Start development of a new category of software for biodiversity information management. The current relational databases are not ideal for maintaining this object-oriented data. New software is also needed for free distribution to less advantageous countries. Development shall be orchestrated in a process similar to the Linux and GNU initiatives on Internet.
- Establish a standard for electronic publishing and securing of species descriptions.
- Adopt a standard for electronic description of characters.
- Harmonize existing systems and especially establish a standard for unique global keys for names. This is likely to be an extension to the codes of nomenclature. Only a clearing-house like the GBIF can issue these keys. These keys act as the linkage between the Linnean naming system and the taxon object system.
- Support activities for converting existing older data sets to electronic form and digitizing collections.

3.4. Mid-term (5 to 15 years)

- Fully develop national, thematic, and special interest networks. These networks ultimately will provide the main content to GBIF such as observation data. However, for them to be fully empowered, the quality control mechanism have to be put in place and the software solutions have to be made available. With these networks, the amount of data in GBIF will surge.

- Make efforts to empower 3rd world countries with biodiversity information. Megabiodiversity countries have to adopt GBIF and start reaping benefits from information exchange with their customers. This will mean that a powerful capacity building programme has to be in place to support taxonomic research in the 3rd world. Funding from all international organizations has to be channelled to this.
- Commercialize services around software. The GBIF Server software shall be free, but consulting on its use and customization become needed services.
- Establish an interoperability measure to make syntheses of geographic distribution, observation data, and simulation of food-chains. Methods providing this functionality, which goes beyond simple information queries shall exist on all taxon objects.

3.5. Long-term (15-50 years)

- Establish a new addressing system for taxa. The Linnean naming system will remain but only as one user interface to a more robust distributed addressing system managed on the Internet. This new addressing system will grow naturally from the infrastructure erected, and should not be seen as a goal by itself – and definitively not as a threat.

4. Components of GBIF and their functions

4.1. GBIF infrastructure

This section lists bodies that could be part of the GBIF directly. These include its centres, national, and thematic components plus a global steering and resourcing element.

4.1.1. Centres in each biogeographic region

The GBIF will have to adopt a maximally distributed infrastructure, but nevertheless it needs one or several bases for support actions. These actions include:

- metadatabase management, especially issuing of globally accepted keys
- archiving and safekeeping of material, which essentially means two things:
  - ◊ mirroring all the region's GBIF Servers and creating a permanent storage of their material
  - ◊ publishing otherwise fully electronic material to comply with the codes of nomenclature
- co-ordination of capacity building actions; this is a a very large decentralized activity but needs a support base
- co-ordination of software development; this again shall be decentralized, but needs a support base
- forum for negotiations: for instance talks on what information shall be made public in the interest of developing a common infrastructure, and what will be retained as national or owners' sellable property
- information dissemination and promotion
- telematic network management and second-level helpdesk to national networks

It would be unrealistic to assume that one centre could cover all of the planet – priorities and issues vary across the major biogeographic regions so much. Except for the largest countries, it is also questionable whether just national centres will do: for many countries, the job of setting up a GBIF centre will be a very demanding task. It can also be assumed that more plentiful support could be channelled into GBIF from regional research and development programmes to their own centres. It is also safer to establish some, but not too much, redundancy. In network access, distance is becoming less and less an issue, but it still may occasionally happen that a regional centre is more accessible than a global one would be. In any case, these regional centres will have to be in close co-operation and mirror each others' information.

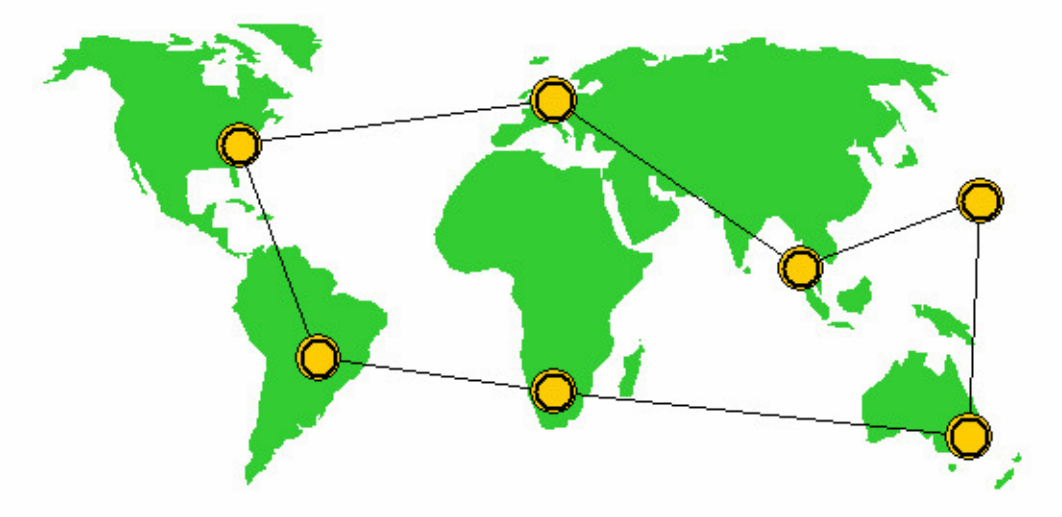

Figure 1. The possible regional centres in each biogeographical region, including one for marine biodiversity.

4.1.2. National centres and networks

The CBD makes it clear that countries are sovereign owners of their biological resources. This has already been interpreted in some cases so that also access to information on biodiversity is restricted. Because of these facts, it is clear that the national component in GBIF is essential.

In each country there shall be a national centre for GBIF. This would typically be a department in a central natural history institution. Its main tasks include:

- co-ordination of the national network of information providers
- together with pertinent administrations that govern natural resources and biological surveys, co-ordination of national interests in bioprospecting
- co-ordination of relationships toward larger GBIF
- first-level helpdesk to users
- promotion of electronic information production
- further dissemination of GBIF's material
- ensuring that GBIF's services become available in local languages as needed

In each country, a national network shall be organized. Typically this entails co-operation with a large number of research groups, individual scientists, scientific societies and their membership. It is foreseen that these all will be offered GBIF's capacity building packages such as software, helpdesk from the national centre, and training on quality control issues.

4.1.3. Thematic networks

It will not be enough to channel all activities in a hierarchical manner through national centres and networks. Direct contacts between special interest groups of particular taxonomies will naturally work the same way they do today across borders and continents, but enhanced with GBIF's support and technology.

4.1.4. Research and development networks

There are several important non-taxonomic networks for common research and development issues that shall be maintained under GBIF:

- capacity building programme, whose function is to help developing nations to establish biodiversity surveying, identification, and informatics functioning within 3rd world. This programme shall cooperage closely with the CBD CHM, which has similar goals, although not directly aimed at taxonomic work
- taxonomic support programme, whose function is to identify gaps in taxonomic knowledge and steer resources in poorly known subjects
- standards committee
- technological research and development programme which aims at advancing the functionality of GBIF's underlying information architecture, network solutions and software.

4.1.5. GBIF secretariat

In addition to the above geographically or thematically limited bodies and functions, it has already been decided that a common secretariat oversees implementation. It is GBIF's secretariat's task to support geographical and thematic networks in a balanced way.

GBIF secretariat takes care of the legal foundation of GBIF and its political support. It shall closely cooperate with the CBD and other international bodies to co-ordinate actions. It shall support a committee that is the highest decisive body within GBIF.

The secretariat shall not endeavour into tasks that can be handled by some of GBIF's geographical or thematic components. However, GBIF secretariat may, together with geographical components support directly priority areas in thematic networks. Especially, it shall fund directly GBIF's research and development networks.

4.2. User community

This section briefly reviews those elements of GBIF that are not directly administered by it, but cooperate with it, provide or use its information, and form its customer base. The meaning of GBIF to their activities is discussed.

- Museums and collections: These are the traditional keepers of biological material and hence, information. However, in the past curating the information has not been well organized. Only in rare cases do collections yet computerize their accessions, let alone make material browsable through networks. With GBIF, curating information shall become a much more important task. Over a long timespan, each collection shall be computerized. Key materials, such as type specimens shall be digitized. Emerging of GBIF should mean changes in activities of all museums and collections. This, of course, should not mean even further drain of resources from traditional work, but should be motivated otherwise.
- Scientists working in taxonomy and systematics will see a major change in their working practices. Armed with tools that can access information world-wide and with a taxonomy workstation to contribute to it, there will be an increase of productivity and appreciation of this profession. Taxonomy will be recognized more as an information science – not as basic biology as today.
- National biological, forest and other surveys will have an important role to contribute to the data base of GBIF. Their scope will be increasingly wide going beyond immediately usable natural resources. They will also be better equipped to monitor much wider groups of organisms.
- Other observers, such as bird watchers and amateur collectors will have a platform to which contribute their observations. This will motivate them even further and data collection will accelerate from the past. Typically they will submit all their observations into GBIF's national servers or even operate their own servers. This requires that the quality control mechanism is fully developed.
- Scientific societies will probably have to tackle the question of formally accrediting their members that are willing to validate data in GBIF.

- Standard bodies such as International Commission on Zoological Nomenclature, International Organization on Plant Information that today control the information infrastructure will find their work supported and very much changed with GBIF.
- Companies using biodiversity resources will, technically, have much easier access to information. However, they will have to be ready to pay for certain data.
- Capacity building companies will have a whole new business area in helping governments and citizens to contribute to the GBIF.
- Schools and citizens will have a cornucopia of new information available for study and learning.
- Ecotourism businesses use GBIF as a prime source for promoting their services.

4.3. IT architecture and technological solution

GBIF should have a well-defined common information architecture. This architecture can be created in many ways, but probably the most modern approach would be to build on the on the concepts of distributed objects and intelligent agents. It shall be described in a published document, and contain models for data, processes, agents, and other relevant definitions that comprise an enterprise model.

The purpose of the current document is not to discuss the data models for GBIF. However, Figure 2 has been produced to make clarify the concepts and make clear the distinction between a biodiversity and a taxonomic information system. A purely taxonomic database could probably work without an addressing system that spans over multiple distributed information sources, whereas a distributed biodiversity information system might not work without one.

It shall be decided early on, which level of detail in models will be shared and what is left for individual projects to solve. The main choice is whether to develop a shared ontology only, which can be later refined to detailed data models in individual projects, or whether to continue to the design phase and promote shared models at that level.

Regardless of what technological approach will be adopted, it is obvious that XML (Extended Markup Language[12]) will have a major role in data interchange. It is a priority to create the necessary DTDs for biodiversity objects very soon.

After the ground for work defining the information architecture has been done by GBIF's engineering standard bodies, the results will be made public and global co-operation is invited to implement and refine it. Only core elements will be designed by GBIF itself. GBIF's role is merely to guard the integrity of the GBIF software so that it does not become overly fragmented. This eventually requires a small but extremely competent staff. In other words, the development model known from the GNU and Linux projects will be adopted.

The GBIF will build on the existing taxonomic databases and co-operates, for instance, with the Species 2000 initiative that aims at interoperability between them. It must be fully understood that even in the best case that developing new technological solutions went well, it will take years before existing systems can be converted. Some of them never will. Building on existing systems is a mandatory intermediate solution and also a backup strategy in event of resource shortage or technological development problems. However, in the long run, GBIF will develop its own data architecture from bottom up. This architecture is outlined in below sections.

[12] http://www.w3.org/XML/

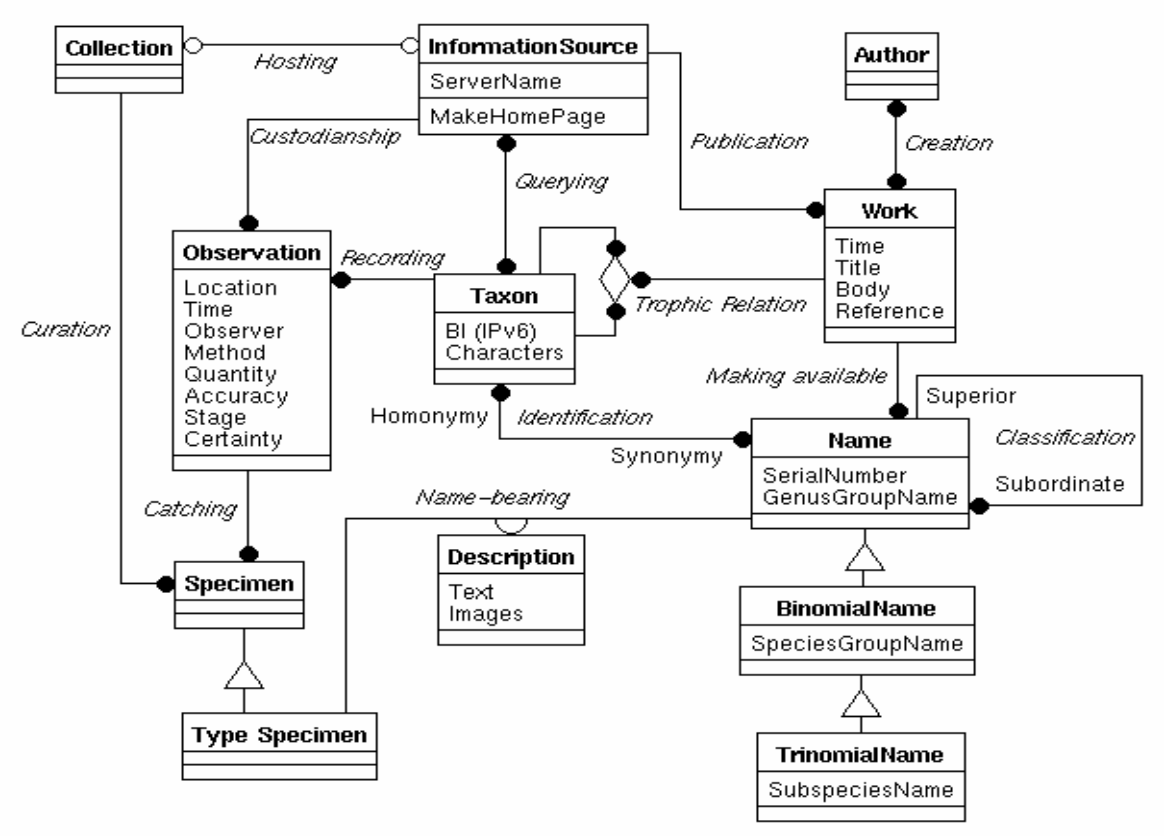

Figure 2. Conceptual model for GBIF core data, which spans from taxonomy on the right to inventories on the right. Only selected most important attributes for the objects are shown. In the Taxon object, many other attributes needed to produce the home pages will have to be added, such as those on life cycle, distribution, etc., see some of the examples in section 2.1.1. It is essential to notice the separatedness of the name and taxon concepts. It is not yet clear whether the classification loop that can be seen in Name is also needed for the Taxon object. Notation after [5].

Internet is the only carrier medium for GBIF. Without such a general-purpose infrastructure, GBIF were not possible. Network access in most biodiversity-rich countries has been a problem, but such restrictions are increasingly going away. Already now email is the most reliable communication medium to some parts of the world, like parts of the former USSR.

4.3.1. Biodiversity Addressing System

At the heart of GBIF is a new linkage between keys of names and addressing taxa. The Linnean naming system was never created as an indexing system that could locate information beyond traditional publications. Any attempt to stretch it to that direction would be dangerous.

Therefore, it will be necessary to erect a new Biodiversity Addressing System (BAS), a new service for GBIF that maps all used scientific names and taxa to stable numbers. The BAS will resemble in many ways the Domain Name System (DNS)[13] of Internet [9], which maps commonly used names of computers to unique IP numbers of four bytes. For instance, this computer is known as cat.eea.dk, which is synonymous to cat.eea.eu.int and www.bioshare.org, but they all map to an IP number 194.182.237.193, which is unique

[13] http://ds.internic.net/rfc/rfc1034.txt

on Internet. The uniqueness is guaranteed, because a hierarchical, Internet Assigned Number Authority[14], issues them.

There is no reason why a similar arrangement would not work for names and taxa. All available names, valid or invalid, would simply have to be translated to a unique stable number that will be called Biodiversity Identifier (BI) below. Like IP numbers, probably four bytes (4 billion choices) will be sufficient for BIs, but it might give more flexibility to use one or two more bytes. The upcoming IPv6 standard[15] [6] has enough address space (16 bytes) to devote a part of it for biodiversity.

The question of homonymy, the same name used for two different taxa, cannot be solved automatically using DNS-like services. When a query using a homonym is issued, it should be escalated back to the user for clarification. Information pages of the related taxa could be provided in order to assist in this conflict resolution. However, the number of homonyms is much lower than the number of synonyms, and this problem should not be a major roadblock.

Hence, in summary, two kinds of identifiers are needed:

1. Simple serial numbers for all names ever used. (These are the "available names" as defined in [2].) All taxonomic databases currently use some keys, but these have not been standardised, and every database uses its own keys. Like the names, the numbers will be anchored to the digitized records of type specimens and their descriptions. These will be part of the "Catalogue of Known Life" (see [3]).
2. As a taxon is a different concept from a name, and could be even unnamed (see [7, 8]), also taxa need addresses, the BI. As it will be necessary to search for taxa on Internet, an addressing system compatible with that of Internet could be designed. It is not clear yet whether there should be some form of logical mapping between the serial numbers of names and addresses of taxa, or could they even be the same numbers. In any case these will be the heart of the "SpeciesBank" (see GBIF main proposal).

[14] http://www.wia.org/pub/iana.html
[15] http://www.cs-ipv6.lancs.ac.uk/ipv6/

4.3.2. Possible technical approaches

What are then the possible technical approaches to create this addressing system? At least the following ones can be thought of, and each of them has some merits:

1. Build on the existing Domain Name System (DNS), and especially its newer implementations on IPv6.
2. Build on directory services, especially the Lightweight Directory Access Protocol (LDAP).
3. Build on Z39.50.
4. Build on CORBA and iiop interoperability architectures and intelligent agents.

DNS is a mature service on Internet and can provide a lot of building blocks (see [9]). IPv6 has enough address space to support DNS also in future and in new areas. It comes with a built-in multicasting functionality that will allow distributing queries on taxa very efficiently. Using the existing DNS implementations (bind) will require a lot of expertise from users, and could even lead to physical network problems. Bind implementations will not be able to deal with homonymy without supporting services. So, it can probably be concluded that DNS does not go all the way to provide a solution, but it will be used together with some other services. However, creating indexes and address space for biodiversity that are IPv6 compatible, will be important.

LDAP, the Internet-version of X.500 directory services, is quickly gaining ground. It is ideal for managing metainformation of people, organizations and data sources in an hierarchical way [10]. It is not suited for transactions and merging of different like data other databases. LDAP is very flexible, and an addressing system of distinguished names (dn) to represent the entire tree of life there would certainly be possible. LDAP is based on objects that can have the above BIs and synonyms as attributes. Reasonable implementations that also allow replication of data between servers are available. However, the essential replication mechanisms are non-standard.

The Z39.50 protocol has been used widely for interoperability of metadatabases. For each application area, a customized profile that defines the attributes must be created. These profiles have been defined, for instance for US Government directories (GILS) and environmental data (GELOS). The complexity of creating and using these profiles has meant that Z39.50 has only had limited success in some well standardised areas such as libraries.

The final alternative listed above means abandoning existing systems and building a new one from scratch. The current state of software industry and Internet standard give good possibilities for that. Interoperability is finally truly possible with distributed objects and CORBA. Given the long term perspective of GBIF and the limitations of any of the above alternatives listed above, this approach is the most appealing, and will be discussed in more detail below. Also IPv6 and LDAP will have roles in this architecture.

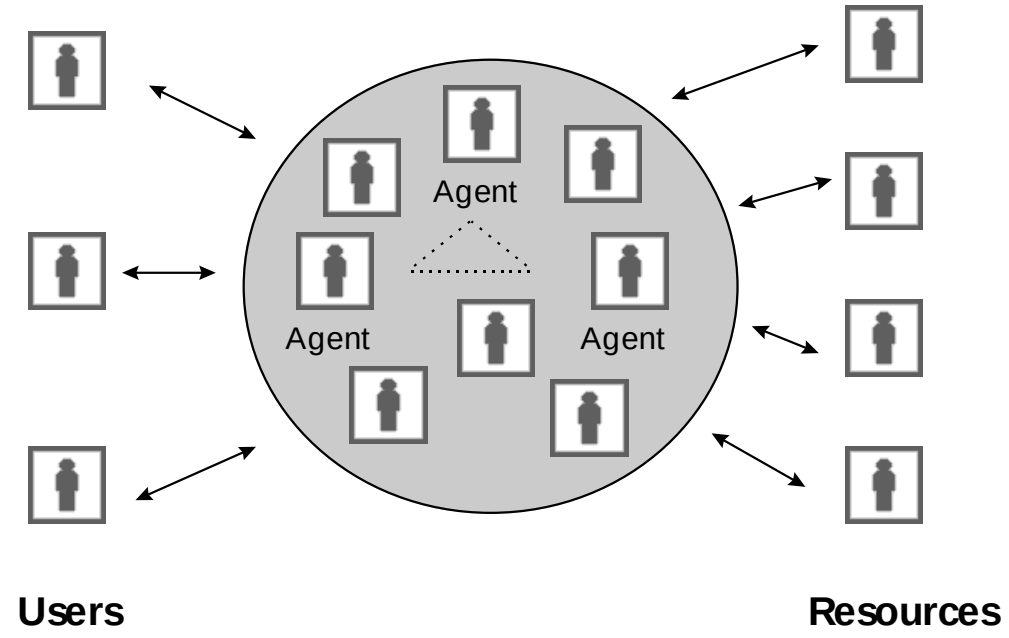

Figure 3. Overall architecture of the community of interacting and communication agents. Modified from [11].

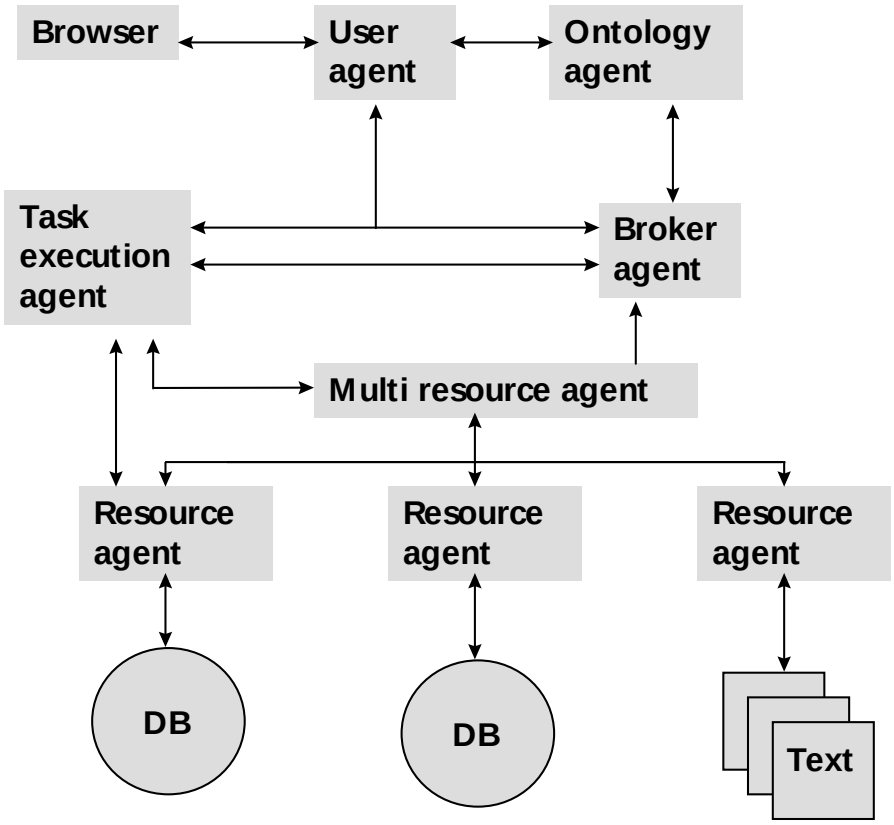

Figure 4. The different kinds of agents useful in GBIF architecture. Modified from [12].

4.3.3. General concepts of distributed objects and intelligent agents

CORBA[16] (Common Object Request Broker Architecture) will provide the needed interfaces for interoperability across Internet. CORBA is a standard for distributed objects maintained by the OMG[17] (Object Management Group). Many OMG member companies are now providing commercial products that support these standards and/or are developing software that use this standard. CORBA provides the mechanisms by which objects transparently make requests and receive responses. CORBA is an application framework that provides interoperability between objects, built in (possibly) different languages, running on (possibly) different machines in heterogeneous distributed environments. CORBA does not use http but another related Internet protocol, iiop.

However, CORBA alone is not sufficient for GBIF – it is an interface of a too low level for that. On top of CORBA, and agent framework will be created that allows flexible construction of GBIF services. The general ideas of the InfoSleuth[18] agent architecture (see [11, 12, 13]) are probably appropriate for GBIF. However, other similar architectures do exist.

InfoSleuth has been created by MCC to allow data access across Internet. InfoSleuth is similar to popular Internet web search engines in that it maintains a central database of information sources, which can be queried. Unlike the current internet search engines, InfoSleuth locates information in remote databases, not static HTML pages on remote web servers.

Figure 3 shows how this architecture works. Users communicate with a "cloud" of cooperating agents who retrieve the information from multiple, usually heterogeneous databases. This is possible as the agents have knowledge of the data and share a common ontology (structured vocabulary) with the user. There are several specialized agents, such as ontology agent that holds the semantic knowledge, broker agent who knows where the databases are, and resource agents who map the data models of the databases to the common ontology. This is illustrated in Figure 4.

4.3.4. Agent framework of GBIF

The general principles described above will have to be implemented into a framework that is specific to the problems of biological information and taxonomy. Especially, the ontology agent requires special treatment, as it will have to deal with the peculiarities of taxonomic information. There probably will have to be two different ontology agents: one in the sense it is generally used on InfoSleuth, and another specialized on taxonomic name-address resolving. The broker agent and multi-resource agents are probably no different regular InfoSleuth counterparts. The resource agents, however, will again have to be able to cope with the taxonomic name resolving and also be able to deal with images of type specimens, etc. Finally, there is a need for an archiving "spider" agent. This architecture is illustrated in Figure 5 and each of these functions is described below.

[16] http://www.acl.lanl.gov/CORBA/
[17] http://www.omg.org/
[18] http://www.mcc.com/projects/infosleuth/

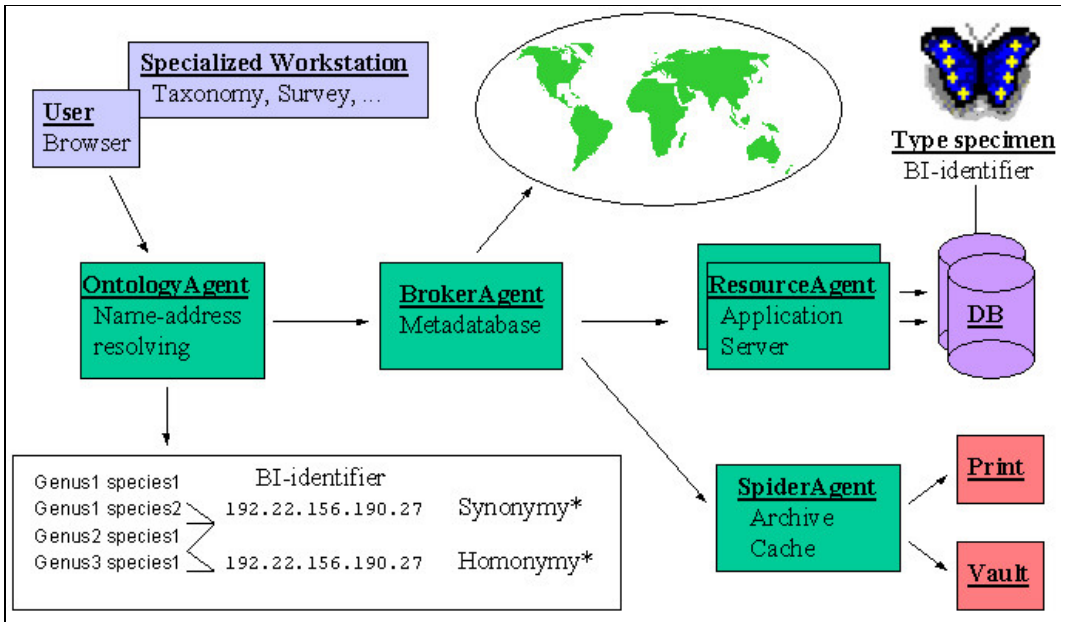

Figure 5. Application of InfoSleuth architecture to GBIF.

The BIs would be created by the ontology agent of GBIF and passed on through other agents to the actual databases. There, the resource agent would map the BIs back to the names used in the database, if they are different. However, this problem should diminish in time, as BIs could also be used as universal keys in any taxonomic database, which may not be directly linked to GBIF. The GBIF shall issue BIs. Probably this task can be subdelegated to thematic networks of expertise on particular taxonomies.

4.3.5. Metadata management, broker and spider agents

Somewhere in GBIF and an instrumental part of BAS, there must be a decentralized data registry, a cloud of broker agents who knows where all the other servers are. These registries reside in the regional and perhaps also on national centres. It is an open question how much information these registries need to have about the other servers beyond knowing that they exist.

When this metadata registry is available, it will be possible to automatically query all data in GBIF for archiving purposes. This is the task of a specialized spider agent. Such archives will naturally reside in regional centres, and they will in time grow into order of magnitude of tera- and petabytes. Not all data may be archievable, though, as the owners may want to keep some confidential or sellable data so tight that it cannot be released. This question shall be answered only through practice.

Data validation and a quality control is a big issue in GBIF. Therefore, all species and specimen objects will have to be taggable by accredited experts with their approval or rejection on certain data. This means that a directory service of taxonomic expertise, which also can contain information on user certification must be erected.

4.3.6. Data management and resource agents

There are two cases for data management: 1) the native GBIF server, and 2) an existing relational database that does not natively support GBIF objects.

Ideally, the GBIF will develop its own standard database server for biodiversity information. This shall be optimized to the task of linking persistent taxonomic and other objects in disparate servers in the same way, say GIS systems are optimized for spatial data management. This is a new category of application software that does not exist

today. Most likely a pure object-oriented database will be needed for the task, but details shall be found out later through detailed scoping studies.

The GBIF server supports direct access to species objects. In essence, the species object has attributes such as names, type specimen data, diagnostic characters, life cycle, host and parasitic organisms, etc. Initially when GBIF gets started, the values of these attributes are just strings of text, and the difference to relational databases is not big. When GBIF gets filled with data, these strings will be replaced gradually with pointers to other similar objects, which can reside on any GBIF server.

Species objects also boast methods (encapsulated program code) that make use of the data in attributes. There will probably not be a major separate application programme for GBIF. Instead, these species objects' method collections will gradually become more comprehensive. The methods are invoked directly with CORBA.

In addition to species objects, the database can also store specimen objects and field observation objects ("catch" objects). These should make it usable for managing data on collections and surveys.

In order to facilitate access to taxonomic information held in existing relational databases or other databases that do not directly support the species objects, *resource agents* have to be employed. These map the ontological information from queries and BAS data to the data model of the database.

4.3.7. Data interchange

A GBIF application server could also be a native web server that can talk http to its human user, or it can be make use of a standard web server. This is needed when the user wants to have a homepage of a species generated. A method in the species object or the resource agent can generate the XML code and the extensible style sheet and send those to the user with *http*. However, to each other the GBIF servers talk *iiop*.

The desired format for data interchange in GBIF is obviously XML. Data Type Definitions (DTD) for taxon, name, and other objects shown in Figure 2 must be created for that purpose and registered at W3 Consortium (and at GBIF). Creating these DTDs is a different task from standardising data models between data bases. A DTD only defines the rules what data elements exist for interchange purposes does not mandate any data model directly. This is a major advantage. It is the task of the resource agent to manage the XML documents with the DTDs in and out from the databases.

Moreover, extensible style sheets (XSL[19]) should be created for visualising the taxon objects at user browsers. It is the job of the user agents to manage these style sheets.

A GBIF database should be able to query other databases dynamically with the assistance of resource agents. For example, producing a distribution map of a species would require querying hundreds of servers. GBIF server should be designed such that when a species object in it receives such a query, it can contact other servers for additional information.

4.3.8. Knowledge management

GBIF will in the long run have a strong knowledge-based component (artificial intelligence) that allows users to identify specimens they have by just entering their characters. This requires that somewhere there are expert agents that map these characters to data that is available in the networks. In a distant future, this could in some cases be automated with pattern matching on images.

[19] http://www.w3.org/Style/XSL/

4.3.9. User interface

The sole interface to GBIF is the Java-enabled web browser.

A user of GBIF will have to logon, if they are to enter any data. This is required for quality control. This means that the GBIF data registry will have to have a directory server of users, which also contains their credentials as experts of certain taxonomies. There must be a certification authority that issues digital certificates that makes it difficult to bypass the quality control. It is probably the responsibility of national authorities or scientific societies to rank their experts.

In the plain GBIF interface that comes up when the home page of any GBIF server is loaded, there should be a predefined user agent that can connect to the data registry.

In more advanced forms, there can be specialized agents on the user's workstation that enable him or her to perform the daily routines on taxonomic work in an efficient way. We can, for instance, define a *Taxonomy Agent* that has functions such as scanning of specimens, comparison of names, identification of specimens, building-up of queries with species characters, and so on. A *Surveyor Agent* would be optimized for field data input and could also generate maps with geographical patters of species distribution. A *Curator Agent* would boast applications for specimen data management.

4.3.10. Turn key solution

Finally all the above components shall be packaged into a user-friendly package that can be deployed next to any biological collection or research lab. There can be competing products made by different companies. It is essential that through GBIF's capacity building programme, this GBIF server is available at low or no cost at all to any potential contributor or user of biodiversity information.

4.4. Needs for research and education in biodiversity informatics

The solutions to GBIF which have been lined out above should make it obvious that in future, there will be equally strong linkages from taxonomy to information systems science as there currently are to systematics, the broader biological science. Until now, taxonomy and systematics have been inseparable, taxonomy's role usually being just to help systematics to put labels on the more generic findings on evolution and speciation. In future, taxonomists will have to study also the input, storage, retrieval, and synthesis of biological information as a whole. This will be a paradigmatic shift, which will lead to a redefinition of taxonomy, or to birth of a new discipline, which could be called biological informatics or biodiversity informatics. This comes close to bioinformatics, which is an established science dealing with the computational aspects of molecular biology and genetics. Of course it is a possibility that the concept of bioinformatics be widened to cover also other aspects of biology: species, habitat, and ecosystem information.

Regardless of which name that this new approach will finally adopt (biological informatics, biodiversity informatics, environmental informatics, widened bioinformatics, widened taxonomy), it is necessary to review here the foundations of this science. In the following we use the word *biodiversity informatics*.

4.4.1. Theoretical foundation and limitations

The domain of biodiversity informatics is the discovery, diagnosis, organization, acquisition, storage, retrieval, quality assurance, and synthesis of information that is concerned with life. These aspects are described below:

Discovery

The processes of finding new objects of study such as specimens, taxons, populations, or sites.

Identification
The process of diagnosing, describing, and naming the findings.

Classification
The process of arranging the objects into a theoretical framework.

Organization
The process of designing conceptual models for the objects of study, linkages between them, and the computational models and information systems used to manage them.

Acquisition
The process of collecting the objects and entering information of them in systems. Possibilities of automating the data entry in advance or on-demand.

Storage
The mechanisms that can keep the objects and information related to them safely stored for indefinite periods of time.

Retrieval
The mechanisms that allow universal, fast, and controlled access to the objects and information on them.

Validation
The mechanism for quality assurance, i.e. the process of advertising findings and data to review, followed by peer feedback that approves or disapproves it. This is a particularly important step, as it could allow even a less credible data to become available, become quality controlled, and the making the credibility of data a known factor.

Accreditation
The process of assigning trust levels to validators.

Synthesis
The process of analysing information on the objects in relation to other objects at same or different hierarchy levels, over time and across geographic locations.

The above definitions retain the management of the actual objects together with the management of information. Museums do have established procedures for the former, and it is a matter practical convenience whether they be included in these definitions. However, we must keep in mind that a huge chore of digitizing physical objects awaits, and possibilities of automating that process by any means remain very important (acquisition). Therefore, the physical specimens have been mentioned here. It should also be noticed that the first three topics above constitute traditional taxonomy.

4.4.2. Research priorities for biodiversity informatics

Following the above descriptions of needs, technological solutions, and cornerstones of the discipline, we should now be able to identify some priorities for research on biodiversity informatics. These are areas, where our current knowledge is insufficient for implementation of GBIF. This list is by no means exhaustive.

- Design of the new taxonomic address system. There are no real technical obstacles for implementation of this system, as described in section 43. However, as it has a major role in GBIF, it is important that it be designed with extreme care and with high quality. Multiple, competing implementations could also be encouraged.

- Biodiversity objects. What are their attributes and implementations of operations.
- Distributed queries and metainformation. How to design efficiently a distributed database that knows about all the other taxonomic databases.
- Ontology and its mapping to resource agents. Thanks to TDWG, taxonomic databases in plant sciences have converged very well during the past ten years. Many of them now share common features in data models. However, we are still a long way from similar development in zoological databases. To create a common higher level data model, an ontology, of all these remains a challenge. Only after one has been created will it be possible to issue truly distributed queries to GBIF. How to map this ontology into the data model of each particular database will remain a subject of study for a long time.
- Biodiversity server technology. How to create an agent framework and interfaces that allows co-operative efforts of many developers. How to package the above and other functionalities into an efficient software solution.
- Automated digitization of types. Digitization can of course proceed manually, but it should be possible also to put some collections under robot control that can automatically or on-demand place specimens under video input devices. None of that has ever been tried.
- Assistance in identification based on digital images. As the digital libraries grow, there will be a possibility to automate or at least limit the choices in identification of new specimens by automatically matching new data to the old. This will probably entail use of neural networks an other AI techniques.
- Character description language. Instead of using just digital images, often a more accurate description can be done by human interpreters of data. The current free form descriptions used in species descriptions should be supplemented with a formal language that allows exact pattern matching.
- Quality assurance and accreditation mechanisms. Until now, taxonomic information has literally always been 100% correct, and less than that has not been acceptable. This assumption can no longer be assumed true, but how to deal with other levels of certainty has to be defined. Examples to approaches can be found in medical diagnosis, where a long tradition exists in dealing with uncertainty. Moreover, how to define the levels of trust to experts and how to test them remains an open question.

4.4.3. University curricula in biodiversity informatics

There are very few individuals in the world that have accumulated sufficient knowledge both in information systems science and biodiversity in order to lead development efforts of major projects. At the moment, there are no universities in the world that have recognized biodiversity informatics or biological informatics at a level of professorships or educational programs. However, some pioneering institutions have combined environmental and information sciences into one curriculum. Examples:

- Charles Sturt University, School of Environmental and Information Science[20]
- EU Canada Curriculum on Environmental Informatics[21]
- Washington University in St. Louis, Environmental Informatics and Systems Analysis[22]
- University of California at Davis, Information Center for Environment[23]
- University of Guelph, Computing Research Laboratory for the Environment[24]
- University of Sunderland, Centre for Environmental Informatics[25]

The lack of university curricula in this area is rather surprising, given the multitude of governmental and international initiatives in this area and proliferation of biodiversity

[20] http://life.csu.edu.au/seis/
[21] http://eccei.crle.uoguelph.ca/
[22] http://capita.wustl.edu/ME567_Informatics/
[23] http://ice.ucdavis.edu/
[24] http://cfc.crle.uoguelph.ca/html_docs/crle.html
[25] http://cei.sunderland.ac.uk/core.htm

information services in the Internet in general. It could be argued that if a biology student takes advanced courses in information systems science, he or she may have accumulated enough knowledge to become a practitioner of biological informatics. However, such double training is time-consuming and does not directly address issues of information management that are special to biodiversity.

A possible curriculum in biodiversity informatics should be built on basic courses from biology, especially taxonomy, environmental sciences, natural resource management, management science especially on quality assurance, political science for international aspects, and computer science. After such a broad basic training, an integrated and in-depth coverage should be provided on each of the issues listed above under "Theoretical foundations".

5. Conclusion

GBIF as outlined here aims at a singular goal: species information and its supporting infrastructure. However it is assumed that when this has been firmly established, other more aggregate themes will grow around it. Solving all the problems at the same time may not be feasible, and addressing ecosystem level informatics before the ground work has been done would certainly be premature. How to link species and habitat information will become a key question soon after GBIF becomes operational, and a number of different issues come to play when a level in hierarchy changes.

GBIF should be closely linked to the Clearing-House Mechanism of the Convention on Biological Diversity. However, GBIF is different from the current plans of CHM in that it concentrates on real data management. GBIF complements CHM.

GBIF is extremely distributed. It will ultimately consist of thousands of servers. (A good indication of the volume can be seen from a book such as [14] "The insect and spider collections of the world".) One key position will be held by the museums that are keepers of the type specimens. Therefore, it might be advisable to establish regional and national centres in these institutions.

One question brought up by several reviewers of this paper was on what will happen if several sources provide homepages of the same taxon? Will it create confusion? In fact, in a biodiversity information system, it will be desirable to stimulate such parallelism for most of the content. Otherwise, distribution maps, identification services, etc., could not be created. However, this does not mean that parallelism was desirable for names, which will probably be best managed in a centralised manner. One should clearly understand the difference between a taxonomic database of names and a biodiversity information system of taxa (species, populations, specimens).

It is essential that GBIF does not become a bureaucratic centralized activity. It is really only needed to establish the infrastructure and a loose framework of co-operation. It will need to have some central support actions, but they are merely to safeguard the process. The current plan may have a weakness such that it makes a top-down approach to the problems by introducing a radical vision in a somewhat technology-driven manner. However, this has been done purposefully believing that a user acceptance can be achieved when the vision is simple enough, and that there is a general understanding that the breakthrough of Internet must mean some fundamental changes in the way the biological community works. A unique opportunity has been created to address problems in biological information management that have plagued the world for a long time and are getting out of hands with the current dramatic loss of biodiversity. If these big ideas are generally accepted, the key role of species information understood, smaller issues as on how actually to build on existing systems, involve their users, and making gradual progress are easier to sort out. Not all the details of that work have been addressed in this plan.

This plan makes only limited reference to the situation in biodiversity-rich tropical countries, where taxonomic expertise is scarce and the challenges are huge. In order to GBIF to succeed, these countries must be closely involved. GBIF will facilitate remote access to published information and collections, i.e., repatriation, and hence address two important issues. However the capacity building programme mentioned here should be more elaborated, perhaps in another document.

Discussion on funding is left to another forum, but it is likely that a small staff of 20 for each of the regional centres and the secretariat will be enough. Majority of funding should come from national and supranational research programmes. The requirement at

the national level is huge, and would require capacity building in many countries. Where would the resources come from? The only sustainable answer is "natural selection" – with GBIF, taxonomy and related activities will gradually be considered valuable ways of spending funds, and they will fare better in competition on resources with other sciences and activities.

The main risk inherent in above plan is probably the possibility for lack of adoption from the user community. Nothing like this has been tried before in the 250 years of existence of taxonomy as science. It is clear that scepticism will prevail a long time even if the implementation went smoothly. Raising expectations too high is also a danger, as benefits start to cumulate exponentially only after a certain amount of information is already there. This will take at least five years. If GBIF is initially given lots of resources, this may cause irritation in the resource-strapped taxonomic community, endangering the process. Therefore, it would be advisable that funding of basic taxonomic research be also increased at the same time GBIF gets started (also see [15, 16, 17]). In fact, new young taxonomists who now do not have positions available for them would be the ones most likely to adopt GBIF first.

6. References

1. Greuter, W., Barrie, F., Burdet, H.M., Chaloner, W.G., Demoulin,V., Hawksworth, D.L., Jorgensen, P.M., Nicolson, D.H., Silva, P.C., Trehane, P. &; McNeill, J. (Eds) 1994. International Code of Botanical Nomenclature (Tokyo Code). Regnum Vegetabile No. 131. Koeltz Scientific Books, Konigstein.
2. Ride, W.D.L., Sabrosky, C.W., Bernardi, G. & Melville, R.V. 1985. International code of zoological nomenclature. 3rd Edition. 338 p. International Trust for Zoological Nomenclature, in association with British Museum (Natural History), London.
3. Report of the Subgroup on Biodiversity Informatics. OECD Megascience Forum, Working Group on Biological Informatics. 7 October 1998. Manuscript 32 p.
4. Summary report. OECD Megascience Forum, Working Group on Biological Informatics. June 1998. Manuscript 25 p.
5. Rumbaugh, J., Blaha, M, Premerlani, W., Eddy, F. & Lorensen, W. 1991. Object Oriented Modelling and Design. 500 p. Prentice Hall, Englewood Cliffs, New Jersey.
6. Huitema, C. 1996. IPv6: The new Internet protocol. Prentice Hall, Upper Saddle River, New Jersey.
7. Berendsohn, W.G. 1995. The concept of "potential taxa" in databases. Taxon 44: 207-212.
8. Berendsohn, W.G. 1997. A taxonomic information model for botanical databases: the IOPI model. Taxon 46: 283-309.
9. Albitz, P. & Liu, C. 1992. DNS and BIND. 381 p. O'Reilly & Associates, Sebastopol, California.
10. Netscape Directory Server version 3.0, deployment guide. 1997. 156 p. Netscape Communications Corporation. Mountain View, California.
11. Jacobs, N. & Shea, R. The Role of Java in InfoSleuth: Agent-based Exploitation of Heterogeneous Information Resources. MCC Technical Report MCC-INSL-018-96, March, 1996. Presented at the IntraNet96 Java Developers Conference.
12. Bayardo, R., Bohrer, W., Brice, R., Cichocki, A., Fowler, G., Helal, A., Kashyap, V., Ksiezyk, T., Martin, G., Nodine, M., Rashid, M., Rusinkiewicz, M., Shea, R., Unnikrishnan, C., Unruh, A. & Woelk, D. 1996. Semantic Integration of Information in Open and Dynamic Environments. MCC Technical Report MCC-INSL-088-96.
13. Pitts, G. & Fowler, J. 1998. Collaboration and knowledge sharing of environmental information: the EDEN project. 6 p. IEEE International Symposium on Electronics and Environment, Chicago, Illionois, May 4-6, 1998. (In press).
14. Arnett, R.H.Jr., Samuelson, G.H. & Nishida, G.M.. 1993. The insect and spider collections of the world. 310 p. Sandhill Crane Press, Inc., Gainesville, Florida.
15. Anon. 1998. 101 uses for a dead bird. Nature 394: 105.
16. Anon. 1998. Museum research comes off list of endangered species. Nature 394: 115-117.
17. Arnold, N. 1998. 101 uses for a natural history museum. Nature 394: 517.

Acknowledgements

This paper summarizes ideas arising from discussions with several people, most notably John Busby from the World Conservation Monitoring Centre, Steve Blackmore and David Vaughan of British Museum of Natural History, Walter Berendsohn of University of Berlin, Doug Brutlag of Stanford University, Marek Rusinkiewicz of MCC, and Ulla Pinborg and Søren Roug of the European Environment Agency. Their contributions are gratefully acknowledged, but none of them should be held responsible of the interpretations and possible biases of this draft. The paper also derives from the author's experience in participating in the BIN21 network, designing the European part of the Clearing-House Mechanism for Convention of Biological Diversity, and construction of EIONET, the European Environmental Information and Observation Network. Images 3 and 4 have been adopted from MCC originals.

Release Notes

This is a live document[26]. Further revisions are foreseen, and suggestions for improvement appreciated by the author[27].

This work is a scoping study by its nature. It has not been officially endorsed by GBIF, OECD, EEA, or other institutions.

An earlier version #10 of this paper was presented and distributed at OECD Megascience Forum for Biological Informatics[28], held at National Science Foundation, Arlington, Virginia, March 29-31, 1998.

For version #12, which was the first one to go to the web (May 1998), the title was made more specific and the abstract was added. Section 4.1.1. got an illustration and more justification on why regional approach might work. Sections 4.3.1. through 4.3.6. were expanded and revised substantially. Section 4.4. and its subsections were added.

For version #13 (August 1998), the section 4.3. generic part and 4.3.1-2. were expanded with a discussion of the alternative technological approaches. Figure 2 was added and discussed in 4.3. and 4.3.1.

For version #14 (October 1998), the section of references was added, as well as a short description of the conclusions of the OECD Megascience Forum's Subgroup on Biodiversity Informatics.

This is version #15, prepared for printing in the Fall of 1999. It comes a new numbering of sections and contains new material on the role of XML.

[26] http://www.eionet.eu.int/gbif/gbif-implementation-latest.html
[27] hannu.saarenmaa@eea.eu.int
[28] http://www.oecd.org/ehs/icgb/biodiv8.htm