



GBIF WORK PROGRAMME

2003

Approved by the GBIF Governing Board at GB5
October 2002, San José, Costa Rica

TABLE OF CONTENTS

	Page
EXECUTIVE SUMMARY.....	1
INTRODUCTION.....	1
HISTORY AND VISION OF GBIF.....	1
GBIF WORK PROGRAMME FOR 2003.....	2
INTEGRATION OF WORK PROGRAMME ELEMENTS	3
IMPLEMENTING MECHANISMS AND PARTNERSHIPS.....	5
MEANING AND INTERPRETATION OF PERFORMANCE INDICATORS.....	6
CROSS-CUTTING ISSUES.....	6
DESCRIPTION OF WORK PROGRAMME AREAS.....	6
ICT STRATEGY.....	6
DADI WORK PROGRAMME.....	10
ECAT WORK PRORAMME.....	13
DIGIT WORK PROGRAMME.....	16
OCB WORK PROGRAMME.....	18
PARTICIPANT NODES.....	21
FUTURE DIRECTIONS.....	22

EXECUTIVE SUMMARY

The GBIF Work Programme for 2003 concentrates on six programme areas:

1. Establishing the GBIF information system
2. Developing standards for interoperation of biodiversity databases (DADI)
3. Helping to complete the Electronic Catalogue of Names of Known Organisms (ECAT)
4. Promoting the digitising of natural history collection data (DIGIT)
5. Preparing the foundation for a comprehensive plan for outreach and capacity building (OCB)
6. Providing tools and recommendations for the development of GBIF Participant Nodes and for databases that wish to affiliate with GBIF. (This component is contained in the other five work programme areas and does not have a separate budget of its own.)

The budget for the Work Programme elements (in US Dollars) is as follows:

	2002 Expenditures	2003 Expenditures	TOTAL
ICT System	\$90,000	\$160,000	\$250,000
DADI	\$30,000	\$270,000	\$300,000
ECAT	\$0	\$620,000	\$620,000
DIGIT	\$65,000	\$820,000	\$885,000
OCB	\$35,000	\$605,000	\$640,000
TOTAL	\$220,000	\$2,475,000	\$2,695,000

INTRODUCTION

History and Vision of GBIF

The Global Biodiversity Information Facility (GBIF) is a co-ordinated international scientific effort to enable users throughout the world to discover and put to use vast quantities of global biodiversity data, thereby advancing scientific research in many disciplines, promoting technological and sustainable development, facilitating the equitable sharing of the benefits of biodiversity, and enhancing the quality of life of members of society.

GBIF had its genesis in a working group of the Megascience Forum of the Organisation for Economic Cooperation and Development (OECD), which concluded that:

- The biodiversity information domain is vast and complex, but critically important to society.
- At present, existing biodiversity and ecosystems information is neither readily accessible nor fully useful.
- Recent technological and political developments present opportunities for OECD countries to show leadership in the area of biodiversity informatics.

The delegates to the Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level in Paris on 22–23 June 1999 endorsed the recommendation from the OECD Megascience Forum that a Global Biodiversity Information Facility be established, as a free-standing organisation with no direct ties to the OECD and with open-ended participation.

GBIF is based on a Memorandum of Understanding (MOU), which is signed by each Participant. By signing the MOU, a Participant agrees to seek to:

- participate actively in the formulation and implementation of the GBIF Work Programme;
- promote the sharing of biodiversity data in GBIF under a common set of standards;
- form a node or nodes, accessible via GBIF, that will provide access to biodiversity data;
- as appropriate, make other investments in biodiversity information infrastructure in support of GBIF; and
- contribute to training and capacity development for promoting global access to biodiversity data.

GBIF formally came into existence on 1 March 2001. For the first several months, it concentrated on forming a Governing Board, selecting a site to host the GBIF Secretariat, and hiring an Executive Secretary and other Secretariat staff. Now that these activities have been accomplished, GBIF is putting into place its first Work Programme, which will be considered by the Governing Board at its fifth meeting, in San Jose, Costa Rica, in October 2002.

GBIF Work Programme for 2003

GBIF is developing an interoperable network of biodiversity databases and related information technology tools. Near-term GBIF developments will focus on species- and specimen-level data. Mid-term developments will concentrate on expansion of content, additional improvements of search engines and tools to combine data from different sources. In the long term, GBIF will provide a portal that enables simultaneous queries against biodiversity, molecular, genetic, ecological and ecosystem level databases, which will facilitate and enable "data mining" of unprecedented utility and scientific merit.

This is the first GBIF Work Programme. As such, it is structured to put the foundations into place for the implementation of the GBIF Business Plan, and to position GBIF (in cooperation with its strategic partners) to become the premier global access point for species- and specimen-level information about the world's biological diversity.

The Work Programme for 2003 focuses on six key developmental areas:

- Two areas focus on the ICT infrastructure needed for GBIF:
 1. **ICT STRATEGY:** Establishing the GBIF information and communications technology strategy and services.
 2. **DADI WORK PROGRAMME:** Developing the standards for interoperation of biodiversity databases.
- Two areas focus on developing content for the GBIF System:
 3. **ECAT WORK PROGRAMME:** Working with GBIF's partners to complete the Electronic Catalogue of Names of Known Organisms, which will serve as an authority file for all of biology.
 4. **DIGIT WORK PROGRAMME:** Promoting the digitising of natural history collection data.
- One area focuses on allowing the international community to fully participate, contribute and benefit from GBIF's mission:
 5. **OCB WORK PROGRAMME;** Putting the foundation in place for a comprehensive plan for outreach and capacity building.
- One area focuses on facilitating the development of the Participant Nodes:
 6. **PARTICIPANT NODES:** All of the work programmes contain components that provide tools, best-practice handbooks, and other materials to help in establishing GBIF Participant Nodes and to aid databases that wish to affiliate with GBIF.

Each of these six areas is discussed in greater detail below.

Integration of Work Programme Elements

Clearly, all parts of the Work Programme are highly interactive. In several cases, progress in one work programme area depends on completion of a study or handbook being developed in another area. Table I provides a timeline for various programmatic activities and indicates their integration. A project which demonstrates elements of the work undertaken in the above key development areas should be identified so that its implementation shows the progress made towards these goals and contributes strongly to the component of the OCB work programme developing proof of concept products.

TABLE I - COORDINATION OF WORK PROGRAMME ELEMENTS															
2002				2003											
	Oct.	Nov.	Dec.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
IT		PTK 0.2	Central Portal V.1	PTK 1.1	Data Repository Tool, Nodes Workshop		XML Schema Repository for Standardization	Data Validation Tool	Archiving and filing system						Central Portal V 2.0
ECAT						White Paper on requirements for taxonomy / names database (Linneaus Core) v. 0.1	Standards for interoperability		Agreements with CoL RFP for Geographically based Indexes Assessment of State of Universe White Paper v. 1	Name Service V. 0.1					Name Service V. 1.0
DIGIT	Digitization Criteria Workshop TDWG Meeting Brazil	Initiate Review of a) current digitization software b) of current digitization efforts		Complete Review of a) software b) of current digitization efforts ----- Release RFP's for new initiatives ----- Initiate Review of DIGIT Universe		Complete Review of DIGIT Universe ----- Draft Handbook V. 0.1 Available									Content from some RFP's Available
DADI	NODES mtg: Present vision/ request nodes' requirements; Contract: survey of reusable elements		Collate Use Cases; White paper on collection standards;	White paper: GBIF&TDWG; Architecture v. 0.1; White paper on reusable elements		2003 collectn. standard; Prioritized Use Cases; Contract: Geog. Svcs.; External arch. review	Architecture v. 1.0 approval; Open Source Toolkit Projects; Collection Service interface published		White paper: Geog Svcs.; GBIF.net on-line; General Resource Service interface published						Year 1 Use Cases operational
OCB			Policy Paper on Repatriation	Pilots a) digitization b) repatriation	GBIF Recruitment strategy ----- Best Practices Workshop	SBSTTA 8 Outreach Strategy	Nodes Workshop ?					World Conference on Protected areas (Case Study)	Nodes Workshop ?	SBSTTA 9	Workshop on repatriation
MEETINGS	GB5, NODES, TDWG			Open Forum on Metadata Registries, Santa Fe 20-24/1		SBSTTA 8	GB 6					World Conference on Protected areas		SBSTTA 9	Workshop on repatriation

Implementing Mechanisms and Partnerships

In 2003, the majority of activities proposed fall into the following implementation categories:

- *Gathering information*, using such mechanisms as holding workshops to gather best practices for digitisation of data or determine the needs of Participant Nodes, and preparing white papers on such topics as a review of current software packages for biodiversity interoperability.
- *Preparing manuals and tools*, such as a handbook on digitisation methodologies or a Standards Validation Tool for data providers to use.
- *Working with GBIF partners* to, for example, help speed up the preparation of the Catalogue of Life.
- *Providing seed money*, via Requests for Proposals from the community, in specific targeted areas, such as development of open-source components for the toolkits for Participant Nodes. In most cases, we expect that seed-money will pay for up to 20% of the total cost of a project and be for amounts up to USD 50,000. Decisions regarding funding of particular projects will be made by the Secretariat staff, with advice from special panels of experts that will be convened to review seed-money proposals. Criteria to be used in making decisions will be published in a “Requests for Proposals” document, which is expected to be released early in 2003.
- *Capacity building* activities such as training of actors involved in DIGIT, DADI, ECAT, and participant nodes.
- *Positioning GBIF for the future*, through such activities as developing strategies for increasing awareness and visibility of GBIF, and preparing a plan for increasing GBIF membership. (Although it is not a specific component of the Work Programme, a key element in the enlargement of GBIF will be the operationalising of the Supplementary Budget, which can be used to bring individuals from developing countries to Governing Board meetings, and for other purposes specified by the donors to the Supplementary Budget.)
- *Outreach to the related scientific community*. GBIF staff, in concert with the relevant science subcommittee members and other interested partners, will identify and promote liaisons with other related information domains, such as genomic, ecological and climate data, and with the computer science community.
- *Developing partnerships*. As is recognized in the GBIF Business Plan and the Memorandum of Understanding, it is imperative that GBIF be an active partner with its Participants and with other international organisations that focus on biodiversity informatics. Such partnerships can avoid duplication of effort and thereby dramatically speed up the provision of biodiversity data. Several strategic partnerships have already been developed, and many more are needed. We also anticipate that there will be fruitful possibilities for partnering with many of the Participant Nodes.
- *Promoting coherent national investments in relation to the GBIF work programme*. GBIF will never have enough funding by itself to achieve its overarching goals. It is estimated that GBIF Participants, through their national and organisational funding activities, are spending, and will continue to spend, at least 10 to 20 times as much as GBIF on GBIF-relevant programs. Therefore, it was recognized from the beginning that one of the Secretariat’s major roles is to help to synchronize the biodiversity informatics spending programs of the GBIF Participants. This will be especially feasible with the seed money awards that are

outlined in the 2003 Work Programme, where GBIF funding will be combined with money from the Participants' national and organisational funding programs and from private sources. The Secretariat will be an active intermediary and facilitator in helping to steer, coordinate and promote the use of funding for biodiversity informatics purposes. The Secretariat will develop a strategy to encourage the relevant authorities in GBIF participants to increase the budgets for collections and associated information and communication technology.

Meaning and Interpretation of Performance Indicators

Each Work Programme element contains quantitative and/or qualitative performance indicators. These are not meant to be rigid benchmarks, but instead are guidelines that the Secretariat hopes to achieve. Especially in the beginning stages of formulating the GBIF Work Programme, we will have to use adaptive management to fine-tune the indicators as the year progresses. However, progress toward meeting the indicators will be a very useful metric to be considered by the outside panel which will undertake the third-year review of GBIF.

Cross-Cutting Issues

A range of issues cut across all of the work programme elements, and will be developed in a collaborative fashion throughout the year. These include:

- *Developing a strategy for recruiting new Participants to GBIF.* Although this is specifically mentioned in the OCB work programme, all parts of the Secretariat will participate in this activity, and a document will be presented to the Governing Board regarding the proposed strategy.
- *Intellectual Property Rights (IPR).* A wide range of IPR issues will be addressed throughout the year, in workshops and by the development of white papers. In particular, IPR related to data ownership will be clarified.
- *Repatriation of data.* A workshop and white paper on this important topic will be developed, and a best-practices handbook will be prepared.
- *Data Sensitivity.* Methodologies will be researched for the best ways for data providers to block access to sensitive data, such as locality data for endangered species.

DESCRIPTION OF WORK PROGRAMME AREAS

The work programme is presented in the following sections. It should, however, be noted that although the work programme areas are described as separate elements, in reality all of the program areas work very closely, often interpenetrably, with each other.

1. **ICT STRATEGY:** Establishing the GBIF information and communications technology strategy and services.

GBIF's mission of making biodiversity data freely and openly available over the Internet would not be possible without developing a comprehensive ICT strategy and procuring the necessary hardware and software to put that strategy into effect.

The GBIF ICT system plan for 2003 focuses on two major goals: (1) preparing a simple but powerful group collaboration environment for the Governing Board, Committees, and Secretariat staff to use, and (2) putting in place the interoperable system for sharing biodiversity information. The first version of the full GBIF information system, to be called gbif.net, will be in place for testing by June 2003, and by December of 2003 users from the general public should be able to begin to access data through the system.

The key actions of the GBIF ICT strategy for 2003 include:

- *Develop the GBIF central portal.* This will be the gateway to all data and information in gbif.net.
- *Provide tools for the participant nodes.* In order to facilitate the quick start of the participant nodes and their interoperability with GBIF central services, a portal toolkit will be made available.
- *Develop tools for data nodes.* Guidelines and tools will be offered to help data nodes make their data available in standard formats, validate their data and register those data.
- *Develop exchange standards.* XML-based data exchange standards for names, taxa, specimens, collections, etc. will be developed. The appropriate committee structures and collaboration services for the standardization process will be erected.
- *Develop a robust ICT architecture.* A coherent architecture of GBIF services across the programmes will be created and maintained.
- *Provide a collaboration environment.* GBIF participants will be provided an effective collaborative work environment on the Internet.

Elements of the GBIF information system

A coherent ICT strategy that describes the architecture, standards, services, and policies of the GBIF information system will be created.

The components in the architecture are identified in Figure 1. They include the following:

- *GBIF central portal.* This will be the hub of communication and data exchange in the GBIF network, providing information on news, events, links, standards and availability of data and services throughout the network, as well as related search functions.
- *Participant portals.* These will duplicate, at the participant level, most of the functions of the central portal, plus tools for the participants to coordinate their part of gbif.net, including quality assurance over the data nodes.
- *Online databases and data repositories.* The data of gbif.net will reside on data nodes which will advertise the existence of the data through the GBIF central registry. The data nodes can either use their own database tools to achieve interoperability or export the shared data in standardised document format into a locally-owned data repository.
- *Tools for data nodes.* A standards validation tool, a digitisation tool, a repository development tool, and other tools are required so that the nodes can effectively provide the data.

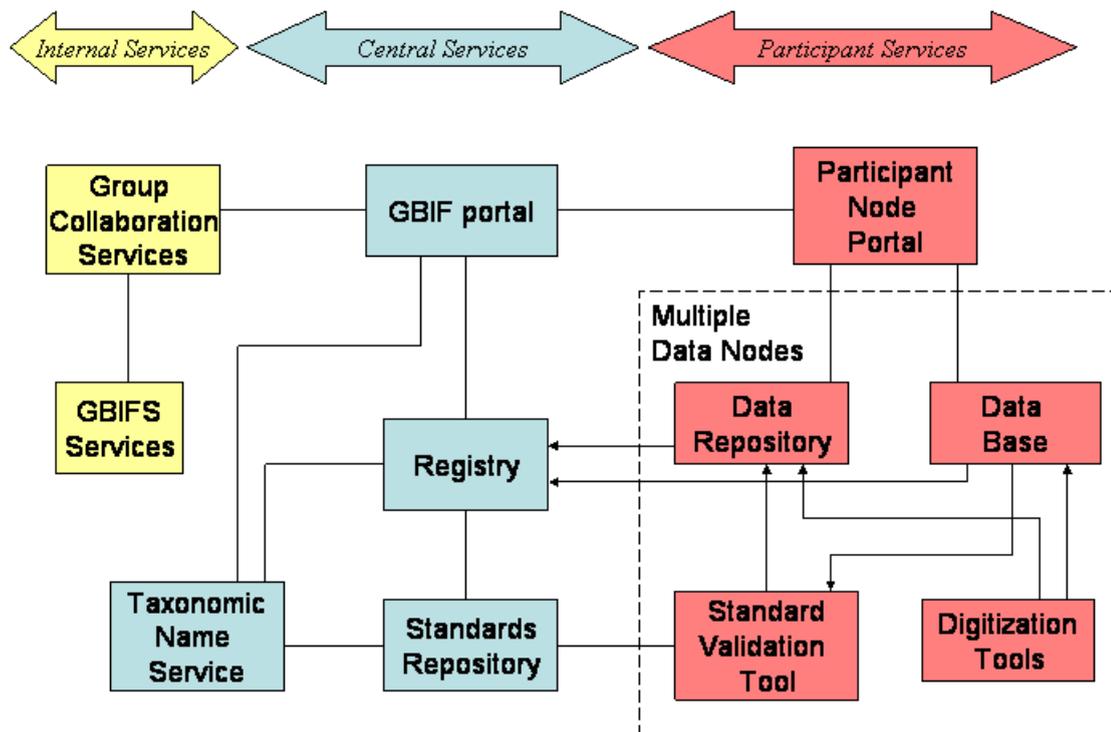


Figure 1. Overall architecture of the GBIF network.

- *Registry.* The registry is the central phonebook and switchboard of gbif.net. It keeps track of the services that the nodes provide and what data they actually advertise. The contents list and search services of the central portal are largely derived from the registry.
- *Taxonomic name service.* Derived from the Electronic Catalogue of Names of Known Organisms, the taxonomic name service provides the key index for the registry and is also a linking mechanism to a wealth of other data outside of GBIF.
- *Standards repository.* The standards repository provides a platform for archiving and disseminating GBIF's data interchange standards, as well as information on their status and a discussion platform for refining them.
- *Group collaboration services.* Such services include document sharing, mailing lists, discussion areas, directory services, and project management support, which will all be provided by the CIRCA groupware system. CIRCA has been chosen because it is supported by the European Commission for research projects, is open source and has open interfaces, and has potential for local installations at interested nodes.
- *Internal GBIF Secretariat services.* These services include the physical network infrastructure, security, and a document archiving and filing system.

In developing this strategy, several possible alternatives were considered to facilitate interoperability of the distributed data. The preferred choice for sharing data is to use locally-held data repositories. Each repository holds the data in a standardized format that will make it much easier to formulate cross-platform queries. This approach avoids the risks and overheads having

each data node put its complex production database on line, provides more robustness for the network, and makes archiving and caching services feasible.

Key operating principles and communication structures in implementing the GBIF information system are the following:

- *Close communication with nodes.* The Participant Node Managers Committee (NODES) is instrumental in the operational and coordinating aspects of gbif.net. Working practices for NODES will be developed. User needs will be reviewed. A major focus of NODES' activities will be to introduce the GBIF approach and services to the data nodes.
- *Standards committees.* GBIF, in particular the DADI STAG, will work closely with standards-setting groups such as the Taxonomic Databases Working Group (TDWG) to ensure timely development and wide adoption of a new generation of XML-based data exchange standards.
- *An independent review of the proposed ICT architecture* will be performed to ensure its wide and fast adoption.
- *Open source tools* will be used by the Secretariat whenever possible. For instance, the Zope application server is used as the basis for the portal toolkit. This will ensure the possibility to disseminate GBIF tools as widely as possible, and will allow commercial entities to write value-added applications, since the source code of the tools will be publicly available. This will help to open a marketplace for biodiversity informatics support and services, just as the open-source nature of the large sequence databases has created such a marketplace for genomics.
- *Keep the data at the data providers.* In order to be in line with the IPR policy that GBIF has adopted and to comply with national regulations, data will be kept at data bases and repositories that are locally owned.
- *Support and training* will be provided for the participants, data providers and gbif.net users. A dedicated helpdesk@gbif.net will be available.
- *A coherent set of ICT policies* to support and secure the operations will be defined.
- *Close collaboration with the host institutions* on ICT infrastructure issues and internal services of the Secretariat will be important.

Deliverables include:

September 2002	CIRCA (group collaboration environment) in place, for use by Committees, Secretariat and Governing Board members
December 2002	Central GBIF portal in operation
February 2003	Data repository tool in place, so that Participants can start advertising data through the GBIF system
April 2003	GBIF data standards repository in place; test users can begin to access data through gbif.net
June 2003	Central data registry is in place
June 2003	Archiving and filing system for the Secretariat is in place
December 2003	General-public users can begin to access data through gbif.net

Programme Goals and Budget

Goal	2002	2003	Total
Prepare a simple but powerful group collaboration environment	\$30,000	\$30,000	\$60,000
Put into place the interoperable system for sharing biodiversity information	\$60,000	\$130,000	\$190,000
TOTAL	\$90,000	\$160,000	\$250,000

2. **DADI WORK PROGRAMME:** Developing the standards for interoperation of biodiversity databases.

The Data Access and Database Interoperability (DADI) Work Programme has a very large scope including all data interfaces within the GBIF Network (although some of these will be managed through the ECAT and DIGIT programmes). Interfaces which are wholly internal to the development of the ECAT facility or to digitising collection data are here treated as belonging entirely to their respective Work Programmes.

The Scientific and Technical Advisory Group (STAG) for DADI met in San Diego, USA on 27 and 28 June, 2002. These sessions covered many aspects of implementing large-scale bioinformatics data systems and produced a timetable for a set of prioritised actions. These actions have been incorporated into this document within the short-term and medium-term goals.

Some requirements for the GBIF system are well established:

- The GBIF Network will be a distributed collection of databases offering biodiversity data and made accessible through the Internet as a set of Web Services.
- Data ownership (the right to remove or change previously published data) will remain with the publishing nodes.
- Access to data should be free to all users of the system.
- GBIF should support the use of any appropriate technologies rather than mandating specific tools.
- GBIF should seek to bring together any data relating to biodiversity science (allow scientific researchers to get intelligent answers to intelligent questions). The short-term emphasis is on the core foundations (species-level and specimen-level data).

These requirements could be satisfied in a number of ways and a variety of technologies have been used by different biodiversity informatics projects. In particular some existing initiatives are using CORBA or the z39.50 network protocol to provide the application infrastructure. However the currently preferred model within the IT industry for combining disparate data sources to create complex systems is through the use of Web Services technologies, particularly based on XML document exchange. This is the approach which has been adopted for the GBIF Network because of the following key advantages:

- A Web Services model allows participants to use existing tools and databases with a minimal additional layer of software wrappers.
- XML-based Web Services operate naturally across standard HTTP connections and do not require special access through firewalls.
- The type of interoperability problems which hindered the development of CORBA solutions are removed because the interfaces are described independently of specific implementations.

- The suite of XML technologies includes open standards to support the registration of Web Services and the description of their interfaces, as well as to exchange structured data and present it to users.
- XML tools and libraries are available for all popular development languages, including Java, Perl and PHP.
- Validation of XML documents can be largely automated using standard features of the technology.
- Once components are provided as Web Services, they will be ready for reuse within any future applications requiring access to the same data.

A complete Web Services solution has the following characteristics:

- A well-defined Web Service interface. This is a “contract” between any system which wishes to provide data of a certain type and any system wishing to use that data.
Example: A Find_Specimens_By_Species service may require the requestor to provide a valid species name and the provider to supply summary data for any specimens of the species held by the institution in a standard format. The actual implementation (programming language, hardware, etc.) is not defined. The interface may be implemented in many different ways, all of which will be examples of the same Web Service.
- A mechanism for registering Web Services which implement the interface and for requestors to locate systems providing the service.
Example: GBIF may define an interface for all providers of the Find_Specimens_By_Species service to register the existence of their node implementing the service. Another interface would allow requestors to locate institutions able to handle such requests.
- Actual registered implementations of the Web Service.

This model applies very naturally to GBIF’s requirements. Nodes within the GBIF Network should implement one or more of a set of defined Web Service interfaces to give access to their data holdings. GBIF should provide the mechanisms for registering these service implementations and the infrastructure for allowing users to search and browse their contents.

A suite of Web Services will need to be defined to support all of the functions which GBIF is required to provide. For the present Work Programmes the following services will need to be defined:

- *Collection Data Service*
Includes the definition of all interfaces required to locate and retrieve data from nodes holding data on collection specimens. Also includes the interfaces required by GBIF to index holdings. This area is already being addressed by the TDWG/CODATA Access to Biological Collections Data subgroup (both for data exchange standards and for the DiGIR exchange protocol).
- *Taxonomic Name Service*
Includes the definition of all interfaces required to resolve and validate possible taxonomic names and to support navigation through taxonomic hierarchies. This is the area covered by the Electronic Catalogue of Names of Known Organisms work programme. The interfaces will require standardization similar to that required for the Collection Data Service (and again the DiGIR protocol is likely to be a key part of the solution).

- *General Resource Service*

Includes the definition of an interface allowing providers to register miscellaneous URLs holding data relevant to different species.

- *Geographic Information Services*

Include access to all facilities required by GBIF to process location data (gazetteer services, coordinate mapping, etc.). This may be managed by identifying existing accessible web services.

In each case, an evolutionary approach is to be expected. The initial requirement is to establish a basic but robust foundation for the GBIF Network. GBIF will start by identifying the core minimal subsets of data and function which *must* be provided and by helping service providers to meet these standards before proceeding to other optional data items and functions. Future versions of these Web Services will therefore extend the interfaces to provide additional non-core elements once the central foundations are secure.

This network will require the definition of several types of standards:

- Standards for the form of the data managed by the system (e.g. specimen records, taxonomic objects)
- Standards for the Web Service interfaces to be used to access the data (e.g. *Find_Specimens_By_Species*, *Get_Specimen_Detail*, *Find_Species_By_Family*)
- Standards for the Metadata describing the data managed by each node
- Standards for the key data fields to be stored within the central index to allow resources to be located
- Standards for the process and interfaces used to register Metadata with GBIF

The early stages of the development of the GBIF Network will focus on defining appropriate standards in each of these areas, using XML technologies (including XML Schema, SOAP, WSDL and UDDI). These designs will be part of an overall architecture for the GBIF Network.

The basic user requirements for the GBIF Network are fairly clearly understood but it is important to establish a clear statement of functional requirements before development starts. Therefore one of the earliest goals of this Work Programme is to collate a set of Use Cases describing how the GBIF Network will be used.

The GBIF Network must produce some working infrastructure as quickly as possible. To make this possible during 2003 some elements are likely to be developed only in a minimal form. The following restrictions are likely:

- Collection data exchange standards will be based on standards available early in 2003 (probably a version of the Darwin Core), with a later move to fuller standards (e.g., CODATA/TDWG standards).
- In 2003 GBIF will use a single server installation. In the future GBIF will need several instances of its server to be maintained (and shadowed) in different locations, to ensure that the services are always available.

Programme Goals and Budget

Goal	2002	2003	Total
Establish the data standards and interchange mechanisms required to integrate species-level and specimen-level data within the GBIF Network	\$0	\$90,000	\$90,000
Identify and develop the foundational components of the network as early as possible	\$30,000	\$40,000	\$70,000
Provide tool kits to assist node managers to bring their databases online as rapidly as possible	\$0	\$140,000	\$140,000
Promote a community (and open-source) development model for GBIF components	\$0	\$0	\$0
TOTAL	\$30,000	\$270,000	\$300,000

3. **ECAT WORK PROGRAMME:** Working with GBIF's partners to complete the Electronic Catalogue of Names of Known Organisms, which will serve as an authority file for all of biology.

For the global enhancement of biological research and resource management, a central file of the names applied to the organisms of the Earth is needed. This file or data store will have to be freely accessible to everybody at any time and should provide the user with reliable data, organized in a structured manner.

A reliable file of names, synonymies and classification can serve as a resource to the world's biologists and with time, the index may seek endorsement as an authority file for taxonomy. Moreover, in order to make the integration of data in the GBIF Network possible, a computerized index of names is essential.

Through the last decades, a number of organizations have been concerned with creating structured lists of names and the process is already much advanced. GBIF can enhance this process and at the same time meet its needs for an Electronic Catalogue of Names of Known Organisms (ECAT) by promoting and supporting these initiatives and linking their datasets into the GBIF network structure.

ECAT is conceived as a knowledge set – an electronic catalogue of names of all known species of organisms, including viruses, micro-organisms, fungi, plants and animals. It is important to work towards *breadth* (rapid coverage of all known species – currently estimated at 1.7 million species), and *depth* (including responsible taxonomic opinion as to a workable set of accepted species, with associated synonymy and links to alternative treatments). Because of the way binomial nomenclature works for all organisms except viruses, a minimal requirement is that this Catalogue should contain names at the species rank, clearly placed in genera. In practice these names of genera and species will be thought of as part of one or many taxonomic hierarchies or phylogenies, and there may be a requirement for at least one hierarchy for organizational and other purposes.

No single organization has sufficient in-house expertise to provide or maintain the whole ECAT. To meet the goal, ECAT will require contributions from the entire taxonomic community. It is thus

important that the programmes developed to assemble ECAT are open and participative, and integrated with the workings of the taxonomic profession as a whole. While this is a long-term social problem that is beyond the scope of ECAT, efforts must be made to work with the Global Taxonomy Initiative (GTI), the GBIF Outreach and Capacity (GBIF-OCB) building group, and others to assist taxonomists in understanding why they need to participate and to provide better tools for them to do so.

ECAT needs to be accessible to users when and where needed even if distributed in an array of participating systems. We may think of a variety of access mechanisms that may be portals or distribution media as well as Web Services, enabling programmers to build computer programs capable of automatically connecting to, and downloading from, the Catalogue through the Internet.

The programme will need to have ongoing changes with development of new access and infrastructure tools. Definition of appropriate standards as well as implementing ECAT as a core function in the GBIF Network are to take place in close cooperation with the GBIF Data Access and Database Interoperability (DADI) work programme.

The ECAT mission is to create a comprehensive catalogue of names of organisms to serve as

- a reference for the taxonomic community and common users at large,
- an authority file for taxonomic names, and
- a core dictionary file for the GBIF Network – GBIF Name Service

ECAT is foreseen as consisting of two major phases. In the first phase, the work programme will focus on bringing existing name resources of various scope and ownership together in a unified data store available to everyone through the GBIF portal. In many cases, these will be regional or local checklists. Additionally, the indexing of names of organisms not yet treated by indexing initiatives will gain high priority. These names have two main origins: they may be names of newly described organisms or they may be names found in organism-groups where no indexing or revisionary work has been captured by electronic media.

In the second phase, ECAT will focus integrating the accumulated data into Global Species Databases (GSDs). To do this, ECAT will need to facilitate the formation and ongoing activities of groups of experts who will work to scrutinise and harmonise the regional and local datasets.

The rationale behind initially focussing on already existing check-lists and names datasets as the primary object for inclusion in the ECAT is the need for fast results and a catalogue “backbone” for the GBIF network. Obviously, the ultimate goal of ECAT is to have all known names of all known organisms on file and it will thus serve as a Global Species Database for all taxa, but integrating as many disparate data-sources as possible – including regional check-lists and otherwise specialised lists – will be the fastest way of getting to a very large dataset. An ongoing task will be to harmonise the systematic hierarchies of the datasets in order to form full GSDs incorporating data from various sources. The system should be able to accommodate multiple, incongruent hierarchies and the authority of the reviser should be noted.

For the purpose of dissemination of the names data, policies and mechanisms for Intellectual Property Rights as well as Source Recognition will have to be worked out. This will take place as a joint GBIF Secretariat task. It shall be emphasized that data providers shall remain custodians as well as owners of data provided to ECAT.

ECAT shall be developed as a web-service, integrated in the GBIF Network. This Name Service will require open standards both for the integration of the various data sources into the core dataset and for dissemination of the data. It also involves the formation of a set of standards for names storage; number and names of fields, indexing methods, parental linking methods, synonymy, taxonomic scrutiny and credit, common names, source recognition etc. A survey of the formats of existing databases will be needed – GBIF partner organizations could contribute already existing knowledge. Provisions for source recognition as well as a method for giving ECAT a finite “reality” – versioning – will have to be thoroughly worked out. Integration with the DADI Work Programme will be essential for formation of these resources.

Another core activity for the ECAT programme will be to contact potential data providers. To facilitate this, a thorough survey of the state of indexing of the world shall be undertaken, to help in clearly identifying existing indexes and highlighting organismal groups where little or no indexing has taken place.

- Initially workshops will be held, bringing representatives of data-owners together for discussion of format and terms for integration of individual data sets in ECAT. Data sources could be regional check lists for particular organism groups as well as large regional indexing projects and Global Species Databases.
- Indexing resources that have a quality or level of completeness indicating that a limited amount of work or funding would bring them to a state where they could be integrated into ECAT should be given the option of applying for grants that would ease their completion.
- In areas – geographical as well as systematic – where indexing is still in its beginning, GBIF might play an active role in the forming of new organisations to carry out indexing activities. Two main paths could be envisioned. GBIF could work with local decision makers to encourage the building of entirely new geographically oriented indexing organisations – here, co-funding might be a possible mechanism. Likewise, GBIF could contact the taxonomic communities for organism groups whose indexing state is poor, encouraging the formation of indexing bodies or organizations. GBIF would primarily contribute with know-how and infrastructure.
- Finally, GBIF could support development of on-line (or other electronic) tools for databasing of species names and hierarchies.

Programme Goals and Budget

Goal	2002	2003	Total
Assess the status of global and regional species checklists	\$0	\$10,000	\$10,000
Initiate collaborations with contributors	\$0	\$100,000	\$100,000
Produce a white paper on standards and on the functioning of the Name Service	\$0	\$10,000	\$10,000
Promote initiatives for names-gathering organisations and speed up already ongoing projects	\$0	\$500,000	\$500,000
TOTAL	\$0	\$620,000	\$620,000

4. **DIGIT WORK PROGRAMME:** Promoting the digitising of natural history collection data.

It is estimated that there are approximately 3 billion specimens in the world's natural history collections. If the data associated with these specimens were digitised to allow dynamic access, it would provide a treasure trove of information about the Earth's biota. Improved access to this digital specimen data will not only aid traditional systematic research but will also result in the utilization of this data by a wide range of additional scientific disciplines, facilitate assembling the "tree of life", support environmental decision makers who have not traditionally had easy access to this wealth of information and facilitate the repatriation of data from the developed to the developing world.

The DIGIT program in its initial phase is tasked with facilitating the digitising of the estimated 3 billion specimens found in the world's natural history collections, including those in natural history museums, herbaria, living organism collections, etc. and exploring technologies to make the resulting digitised data easily available so it can be analysed and integrated in new and innovative ways. In the initial phase the DIGIT program will concentrate on the capture and geo-referencing of basic label data associated with museum specimens but it is expected that as this task progresses the work program's future phases will expand to include other types of information including digital images, sonograms, field notes and eventually observational records. It is recognized, however, that while it is important to move existing non-digital biodiversity data into a digital format, it is also important to insure that the vast amounts of new data being recorded each year are being captured and documented using modern information management technologies. While in its initial phase the DIGIT work program will concentrate on retrospective data capture, the DIGIT "Best Practices Handbook" and the DIGIT training modules will include information and recommendations on the latest approaches to modern biodiversity data capture and biodiversity information management. It is hoped that this will help insure that the digitisation task does not have to be repeated for the new data that is being recorded yearly.

Currently there is no comprehensive estimate of the size or distribution of our natural history collection resources. Some regional reviews have been completed for specific taxonomic groups but there is no accurate global assessment of this vast information store. In addition, globally, there are no comprehensive reviews of the current digitisation efforts or the tools and techniques that are being utilized. If we hope to measure progress and develop benchmarks for success, it is necessary in the initial phase of the DIGIT program to undertake a preliminary global assessment of our natural history collection resources and a review of the current digitisation efforts, tools and approaches that are being used. Once this assessment has been completed, it will be possible to document successful approaches and cost-effective solutions. From this information a "Best Practices Handbook" will be developed to communicate to the digitisation community at large the lessons learned from other projects.

If the limited DIGIT budget is to have maximum impact on the global digitisation effort, it will be necessary to target key areas for investment and to build on the experiences and efforts of the many existing successful digitisation efforts. While there are hundreds of thousands of specimen records currently accessible on the internet, there are also numerous already existing databases that are not currently available due to such problems as a lack of quality assessment and quality control, inadequate georeferencing and/or the need to migrate them into modern interoperable database

formats. Rapid progress in increasing the number of specimen records accessible on the internet can therefore be made by bringing these existing data stores on-line.

In the long term, DIGIT investments will have an even greater impact by supporting new projects that are characterised by one or more of the following attributes: will develop innovative digitisation technologies and methodologies (including novel approaches to quick and accurate geo-referencing of specimen data); have the potential to produce rapid results; have a unique scientific impact; or have a high potential for capacity building.

Therefore, in its first year, DIGIT will fund not only new digitisation efforts but also identify and support the completion of the work needed to make a number of existing databases internet accessible.

It must be recognised that due to the sheer size of the digitisation effort, the DIGIT work program funding will only be able to supply short-term support for start-up projects, facilitate communication of best practices within the digitisation community, support short-term proof of concept pilot projects or limited tool development to fill critical technological gaps, and support a limited number of projects to make selected existing databases internet accessible.

It will be necessary for GBIF Participants, through their national and organisational funding activities, to be responsible for the vast majority of the funds necessary to accomplish the digitisation effort. It is envisioned that the DIGIT work program will achieve its goals through partnerships with institutions and national, regional or thematic networks that are either content providers, developers of technological solutions or suppliers of services and information of benefit to the larger digitisation community.

The goals for the 2003 Work Programme include:

- Develop preliminary baseline estimates of the current status of the global digitisation effort
- Develop a best practices handbook for digitising specimen data
- Develop priorities for tackling the digitisation effort.
- Support efforts to make selected existing specimen databases internet accessible
- Initiate new digitisation efforts, via seed-money.

Benchmarks for DIGIT include:

January 2003	Report evaluating existing digitisation software is completed
January 2003	Report reviewing existing collection-based digitisation efforts is completed
January 2003	Request for proposals for initiating new digitisation efforts or making existing specimen databases internet accessible
April 2003	Version 0.1 of handbook summarizing best practices in digitisation of specimens is completed
October 2003	Preliminary quantitative estimates of the numbers of specimens in natural history collections are completed
December 2003	Digitised data from funded initiatives begins to become available

Programme Goals and Budget

Goal	2002	2003	Total
Develop preliminary baseline estimates of the current status of the global digitisation effort	\$0	\$40,000	\$40,000
Develop a best practices handbook for digitising specimen data	\$40,000	\$30,000	\$70,000
Develop priorities for tackling the digitisation effort	\$25,000	\$0	\$25,000
Initiate digitisation efforts, via seed-money	\$0	\$750,000	\$750,000
TOTAL	\$65,000	\$820,000	\$885,000

5. **OCB WORK PROGRAMME;** Putting the foundation in place for a comprehensive plan for outreach and capacity building.

The Outreach and Capacity Building (OCB) Work Programme focuses on allowing the international community to fully participate, contribute and benefit from GBIF's mission by providing or facilitating adequate institutional and human capacity and by promoting GBIF widely within the international community, particularly policy- and decision-makers. OCB cuts across and supports the activities of the other work areas of GBIF.

The Scientific and Technical Advisory Group (STAG) for OCB met in Pretoria, South Africa on 14-15 July 2002. The STAG recommendations were vast and valuable and many of them have been incorporated in the OCB and other GBIF work areas. Some of the suggestions made by the STAG (e.g., that GBIF agree to take over orphaned databases) were impossible for GBIF to implement at this time.

In general the actions considered under the OCB work programme present a suite of options from the development of strategies and concept papers to concrete pilots, case studies and training workshops. The main idea is to explore new options for GBIF's growth, benefit from existing experience and expertise and learn from some specific pilot projects and case studies.

As recommended by the OCB STAG we have given an important priority to support training in digitisation of collections. GBIF considers that it is crucial at this point and time to generate content for the databases that become affiliate with GBIF, and to foster in-country expertise on how best to digitise and utilise these data. The DIGIT and OCB work programmes will therefore work closely together to gather information about digitisation needs and technologies, to produce a best-practices manual for digitisation, and to hold training workshops on digitisation technology. The workshops will also focus on how to use that data (i.e. for decision-making purposes).

The enclosed summary work programme also includes some cross-cutting issues, including consideration of intellectual property right (IPR) issues and repatriation of data and information. Many biodiversity-rich countries have expressed their interest in repatriating information to the countries of origin, and this view has been supported by many GBIF participants. Among others, GBIF participants see that GBIF can and must play a key leading international role in the efforts dealing with repatriation of data and making biodiversity information available through the internet. As this very important role has not been fulfilled by any other international organization thus far,

GBIF will derive not only leadership but also international recognition and appreciation from such activities. Furthermore, GBIF members have expressed that this would be a very concrete contribution of GBIF towards the implementation of the Convention on Biological Diversity particularly in regards to article 17.2 which deals with exchange of information. In practical ways we envisage dealing with repatriation of information within a context of training and scientific and technical cooperation. Via an interconnected and mutually supportive modular approach and activities we will (a) analyse existing experiences and approaches in the repatriation of data and information pursued thus far, (b) facilitate scientific exchanges between scientists and staff from developing countries and large collection-based facilities (mostly in the developed world), (c) provide hand son experiences on how to digitise collections and how to use the recently acquired data and information, (d) facilitate the development of agreements/MoUs with institutions holding large international collections, (e) benefit the scientific and international community at large by making these data readily available through the internet, and (f) using the experience acquired, convene an experts' meeting in early December 2003 to address the best ways to make progress in this important area.

Another cross-cutting issue identified at the STAG is IPR related to biodiversity data. It was agreed that as a first step GBIF will hold a workshop on IPR and commission a white paper on this issue.

Specific Activities

In the Outreach area, the OCB Work Programme for 2003 stresses the need to promote and raise GBIF's visibility and focuses on how best to increase GBIF's membership. The Outreach area includes the following goals:

- Increase awareness and visibility of GBIF and its potential projects, and add value to existing initiatives.
- Increase GBIF membership.
- Develop "proof-of-concept" product(s) illustrating ways to use the GBIF system, and use these products to reach out to GBIF providers and users

Main activities include:

- Development of an outreach strategy for policy and decision makers which identifies key target audiences and main biodiversity negotiations and fora (i.e. Convention on Biological Diversity) that we want to influence and promote specific areas of collaboration and synergies.
- Develop a GBIF recruitment strategy, identifying priorities, deadlines and ways and means (including collaboration from supporters or "GBIF champions").
- Identification and promotion of suitable GBIF products (in coordination with the other GBIF areas)

Benchmarks for the Outreach component include:

February 2003	Strategy for recruiting new GBIF Participants is developed
March 2003	Outreach strategy developed
December 2003	GBIF participates in and influences at least at 3 major international initiatives/negotiations (i.e. CBD-SBSTTA, Protected Areas Congress, etc.)
December 2003	GBIF increases its current membership by at least 10%
December 2003	Proof-of-concept products have been developed and made available to all Participants

In Capacity Building, the OCB Work Programme for 2003 includes the following goals:

- Assist ECAT to produce regional checklists, especially in developing regions
- Provide assistance to countries in setting up their national nodes
- In collaboration with DIGIT develop a training program in digitisation of specimen data and data capture and on how best to use the acquired data.
- Encourage and facilitate scientific collaboration, hands-on training and repatriation of data and information from large collection-based institutions to countries of origin, including the facilitation of agreements with those institutions.
- Hold workshops on how to use GBIF ICT tools

Main activities include:

- Disseminate GBIF recommended standards and formats and assist partners in their use.
- In coordination with NODES develop training courses for node development covering (a) software packages available, (b) standards and protocols, (c) information management and networking. Facilitate a mentor program that leads to tangible cooperation among countries.
- Support ECAT in the wide dissemination of a functional requirements document and production of regional checklists, including the facilitation of agreements with relevant regional and sub-regional initiatives.
- Conduct a training needs assessment in digitisation (2002)
- Facilitate 2 pilot projects and organize 2 training modules on digitisation and on how best to use the acquired data.
- Organize hands-on training modules and experts workshop on repatriation of data
- Facilitate the production of white papers on IPR issues and repatriation of data.
-

Benchmarks for the Capacity Building area include:

January 2003	Consultant delivers study on digitisation training needs, gaps and survey of existing training resources and opportunities (2002 budget).
February 2003	White paper produced on repatriation of data and information models, approaches and ways and means (2002 budget)
March 2003	Consultant produces white/concept paper on IPR (particularly regarding data ownership)
December 2003	2 pilots and 2 training modules on digitisation and use of data particularly for developing countries are developed (in coordination with GBIF members)
December 2003	3-4 pilot projects are in place regarding repatriation of data and information
December 2003	Experts' workshop on repatriation of data and information models, approaches and ways and means. Promote support from international initiatives and aid/cooperation agencies in repatriation of data efforts particularly for developing countries.

Regarding institutional capacity building, GBIF should not be a major provider of hardware and software. Rather, GBIF can act as a facilitator in the acquisition of hardware and commercial software via such activities as partnering with corporations, funding institutions and aid/cooperation agencies who may be willing to assist in this endeavour.

Summary proposed budget for OCB:

Area/Activity	2002	2003	Total
Outreach + cross cutting issues (in 2002 will include the preparation of white papers on IPR and repatriation of data)	\$25,000	\$50,000	\$75,000
Capacity Building	\$10,000	\$555,000	\$565,000
Total	\$35,000	\$605,000	\$640,000

6. PARTICIPANT NODES

The Participant Nodes are at the heart of the GBIF system, as it is through these nodes that the majority of data will be made available to gbif.net. Tools and services to help the nodes are found in all of the work program elements, including the following. Funding for these Node activities, totalling at least \$380,000, is included in the budgets for the various Work Programme areas.

October 2002	Request to Participant Node Managers for requirements for Toolkit for Participant Nodes
January 2003	Portal Toolkit available for use by Participant Nodes
March 2003	Draft design proposal for a Collection Node Toolkit
March 2003	Draft design proposal for a Participant Node Toolkit
April 2003	Establish open source project to produce Collection Node Toolkit
April 2003	Establish open source project to produce Participant Node Toolkit
April 2003	Hold workshop of Participant Node managers to identify institutional and human capacity-building gaps
April 2003	Prepare report summarizing status of Participant Nodes, as well as needs and recommendations on how to overcome shortcomings
October 2003	Hold training course for Participant Node Managers
December 2003	Best practices manual for how to develop a node is ready for distribution

Area/Activity	Total
Toolkit for Participant Nodes (DADI)	\$40,000
Toolkit for Collection Nodes (DADI)	\$100,000
Identifying institutional and human capacity gaps for nodes (OCB)	\$10,000
Developing best practices and models handbooks for nodes (OCB)	\$20,000
Training courses for node managers (OCB)	\$130,000
Developing portal toolkit and making it available to node managers (ICT)	\$30,000
Data validation tool (ICT)	\$20,000
Data repository tool (ICT)	\$30,000
Total	\$380,000

FUTURE DIRECTIONS

In addition to building on the activities laid out for 2003, it is planned to organize e-mail discussion groups and brainstorming meetings to emphasize the following projects for their possible inclusion in the future GBIF work programme:

- Fleshing out the two remaining work programme areas: designing and implementing SpeciesBank and developing a digital library of biodiversity data and literature resources;
- Developing interoperability with additional kinds of data (e.g., molecular data, physiological data, behavioural data, ecological data, phylogenetic data, etc.);
- Working with authors, publishers and other data providers to set up on-line data-entry services (not just concentrating on the legacy data already present in natural history collections);
- Exploring the possibility for GBIF to host on-line community monograph projects (such as those advocated by Charles Godfray in his paper in the 2 May 2002 *Nature*);
- Working with scientific societies to get their support for on-line publication of new taxonomic descriptions;
- Discussing with the various nomenclatural committees the possibility for GBIF to work with other organisations to act as a registrar of new taxonomic names and to make available on-line all new species descriptions;
- Keeping abreast of new advances in the ICT world, such as massively parallel computing and the GRID.
- Continuing an assessment identifying the variable needs of the end users of biodiversity data in the context of the GBIF work programme.