GBIF Data Access and Database Interoperability

An Introduction to GBIF Biodiversity Informatics

Version:1.0 (Final)

Donald Hobern
DADI Programme Officer, GBIF Secretariat

September 2003

Contents

History

Date	Author	Version	Notes
13 Aug 03	Donald Hobern	0.1	Created
20 Aug 03	Donald Hobern	0.2	Comments from Giorgos Ksouris (figure in section 3); comments from Larry Speers (section 1)
23 Sep 03	Donald Hobern	0.3	Updated images to match Powerpoint version
4 Nov 03	Donald Hobern	1.0	Minor corrections

Introduction

The Global Biodiversity Information Facility (GBIF) exists to make the world's biodiversity data freely and universally available by developing an international network of databases offering web access to primary information on the world's organisms.

This network must be flexible enough to handle the needs both of data providers and also of users who wish to access biodiversity data. The GBIF Secretariat has developed an architecture to achieve this goal. It is described in detail in the following documents:

GBIF Biodiversity Data Architecture
http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/gbifbiodiversitydataarch_1/

GBIF Metadata Standards
http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/gbifmetadatastandards_1/

GBIF Strategy for Exchange of Specimen and Observation Data
http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/gbifspecimenobservationd_1/

These documents (in particular the *GBIF Biodiversity Data Architecture*) are already large and will continue to grow as further detail is included to describe the software being developed. The present document serves as an overview of the key concepts from these other documents. It should be seen as an introduction to Biodiversity Informatics as addressed by the GBIF Network.

The sections which follow describe the operation of the GBIF Network by considering a series of key topics. Each of these topics provides an important insight into what the GBIF Network represents. Together they show how the network will support the global exchange of biodiversity data.

This document is an independent overview, but includes references to related sections within the other documents listed above. These references are indicated at the start of each section.

As far as possible, this document has omitted all unnecessary technical detail (although some important computing terms and technologies have been mentioned in footnotes).

The GBIF Secretariat gladly acknowledges that the concepts presented here are not new innovations. Much of what is described has already been implemented on a smaller scale within other biodiversity data networks. These other efforts have led the way for GBIF to follow. The attempt here is to provide a secure foundation for expanding such efforts to a global level and to cover all forms of life.

1 The GBIF Network handles primary biodiversity data

⇒ *GBIF Biodiversity Data Architecture* - ***3.6 Biodiversity Data Model***

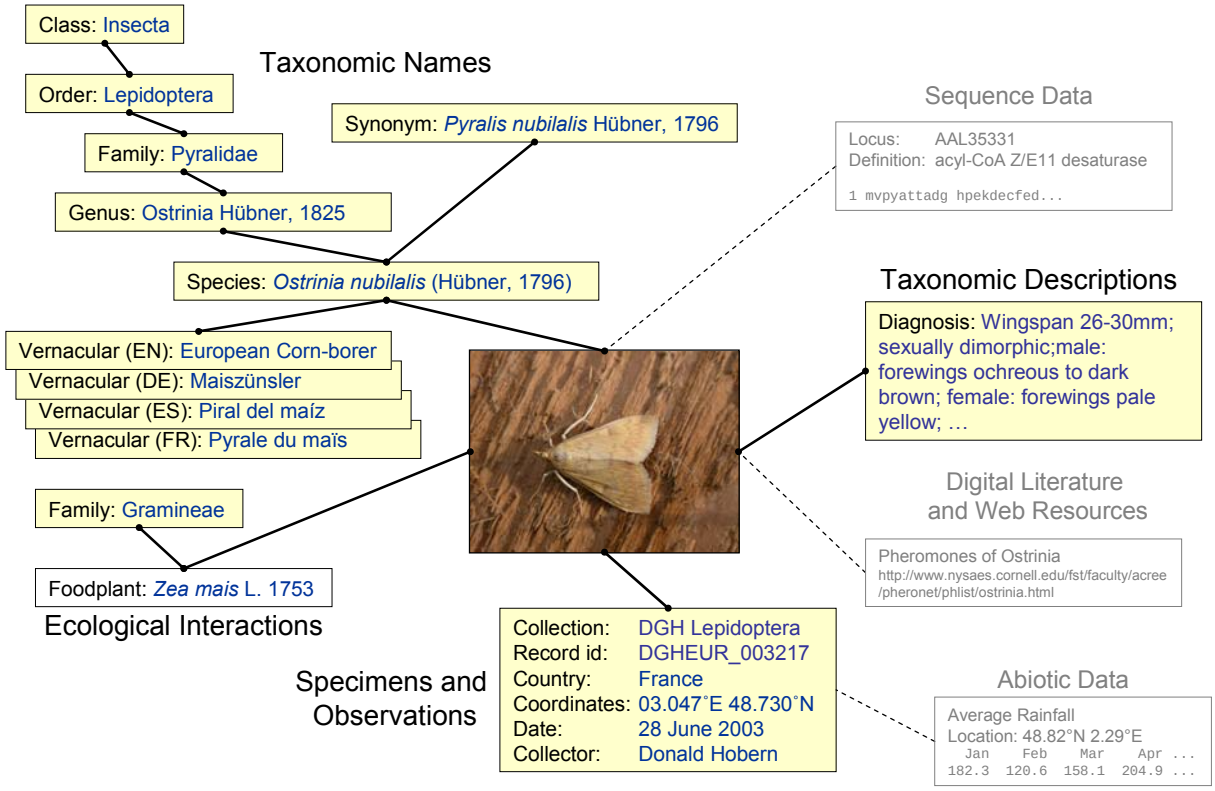

Figure 1 - Some of the primary data elements to be handled by the GBIF Network. The linkages shown in this image are a tiny part of the huge web of data relationships which need to be modelled. The grey boxes represent data which are external to the GBIF Network.

Biodiversity informatics is a branch of computer science dealing with information about living organisms. GBIF has been established to develop models and tools for handling these complex data.

GBIF will not itself be responsible for all of the data which may be of interest to biodiversity researchers. Many areas are already addressed by other organisations and networks. In particular there are major international collaborations to handle the exchange of ecological and genetic data.

GBIF's particular focus is on data about species and about individual specimens representing those species, although appropriate links will also be made to relevant ecological and genetic data. Even within the more restricted domain of species and specimen data, the range of information is enormous and the data is currently held in hundreds of differing formats.

The following is a simple classification of the biodiversity data for which GBIF is responsible:

1. Taxonomic data, including:
    a. Taxon names (the responsibility of the ECAT Work Programme)
        i. Scientific names, including data on synonymy
        ii. Vernacular names
    b. Taxonomic descriptions, including diagnostic keys
2. Taxon occurrence information (primarily species-level, but including data for taxa at different ranks where appropriate):
    a. Specimen records (from natural history collections)
    b. Observation records
3. Links to other taxon-level information, including:
    a. Information on taxon biology and life history
    b. Ecological interactions
    c. Genetic data
    d. Sound and image resources

Within this classification, GBIF's immediate (2003) goal is to build the network required to support exchange of Taxon names (*1a*, above) and Taxon occurrence information (*2*, above). The other classes of information will be addressed in subsequent years.

By establishing a global network to exchange these data, GBIF will help to bring about two significant outcomes. The first of these will be to liberate the wealth of information currently found only in the paper literature and the records of natural history collections, and to make it accessible for computer-based analysis. The second outcome will be to create an environment in which future biodiversity science can capture and publish its findings electronically from the very first, rather than as an expensive subsequent step. Researchers in such an environment will be able to ask and answer questions which today it would be impractical to address.

2 Users need data structured according to standards

⇒ *GBIF Biodiversity Data Architecture* - ***3.1 Basic Principles***
⇒ *Strategy for Exchange of Specimen and Observation Data* - ***throughout***

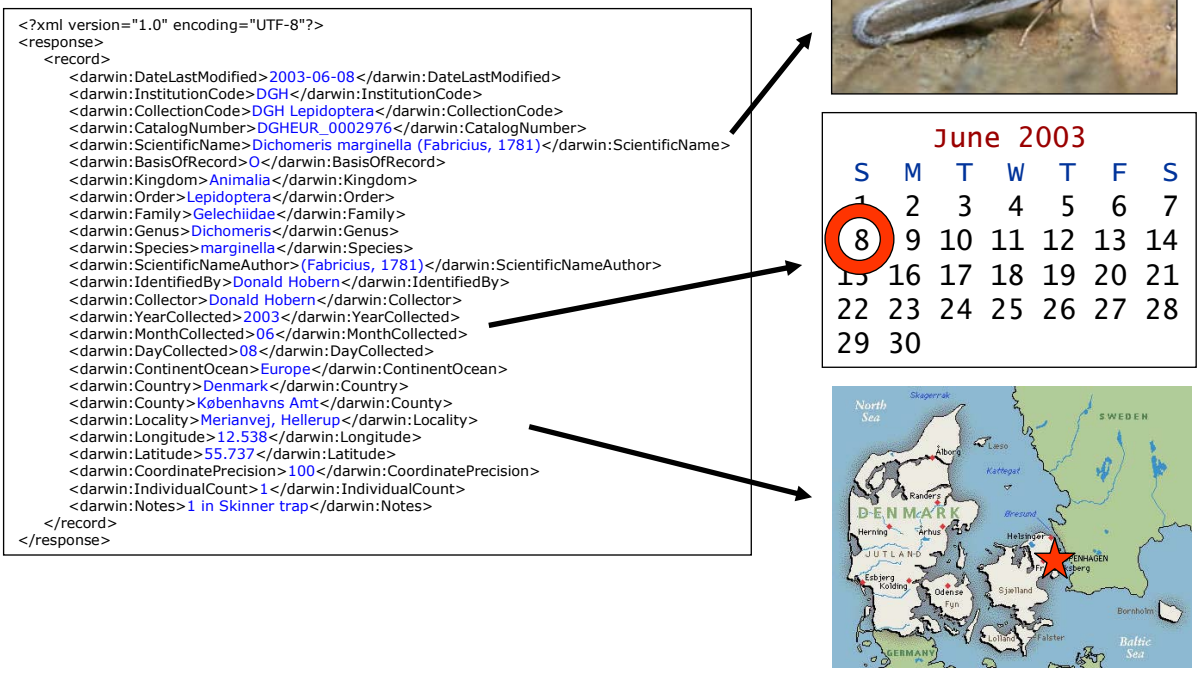

Figure 2 – An observation record for a moth presented in a structured format, in this case using the Darwin Core elements. The structured data elements allow a user or application to determine with confidence what taxon, date and location are referenced by the record.

The human race possesses a vast range of information on the world's plants, animals and micro-organisms. This knowledge exists in many different forms, including an enormous literature (primarily paper-based) and large collections of specimens in museums and herbaria around the world. In recent years a start has been made on converting this information into electronic formats. These electronic materials are again in many different forms, from plain electronic versions of human-readable documents to highly-organised databases designed to support complex user queries.

GBIF aims not only to identify where this information is held, but also to help users to retrieve and process it as simply as possible. In this context "users" will include both human users, who may use a web browser to access the data, and also software applications, which will be able to use the GBIF Network to

select and access data for further processing. The diverse needs of all of these users will only be met if the information is easy to process. This means that the GBIF Network must ensure that all of its biodiversity data is made available in a highly structured and standardised fashion.

To achieve this goal, GBIF does not need to impose a common database format on every institution holding relevant biodiversity data. Such an approach could never work. There are many different software packages in use around the world, all with different strengths and tailored to meet the needs of different communities. These tools store their information in varying levels of detail and often in radically different database formats. The network must support diversity in the choice of tools and technologies.

GBIF therefore addresses the need for standardisation by focusing on how the data are shared and transferred from the providers to the users. This involves considering two questions:

1. Which are the important data which users need to access?
2. In what form should these data be exchanged?

These questions greatly simplify the process of integrating the world's biodiversity data.

Consider the different institutions which hold data on specimens and observations of different taxa. The data which a herbarium holds on a plant specimen may be very different from those which a culture collection holds on one of its bacteria cultures or which a zoological museum holds on a fish specimen. Users require a network which will allow them to access the data held by any of these institutions, including the information which is specific to just one class of institution. The GBIF network must therefore support the exchange of as much detailed information as possible.

This complexity can however be managed. Despite the need to support all of the specific detail, there is a common subset of information offered by all of these institutions and by many others (including for example networks of field observers). In practice this common subset contains the elements which are most frequently of help to users when searching for biodiversity data. It includes the following essential details:

1. Name of the taxon to which the organism has been assigned
2. Location where the specimen was collected or the observation made
3. Date on which the specimen was collected or the observation made
4. Where the specimen is held and how to access more information

These four elements will support most of the basic queries which users need to issue to find relevant data on any given taxon[1]. Many of these users will then wish to retrieve more detail on the specimens and observations that they have found, but efficient handling of these basic details can provide a solid foundation for web-based biodiversity research.

This helps us to answer the first question above. GBIF needs all data providers to provide at least these four items. If they can do this much, their data will offer at least some value to researchers. Additional data may also be provided, but this should be regarded as the basic minimum.

It is also possible for us to answer the second question. All data should be exchanged in a way which makes it as simple as possible to compare and merge information from different sources. This means that all providers should use a common data format with a fixed structure which clearly defines how the information is to be shared. In the case of specimen data, GBIF has identified two different standards as suitable formats for this data exchange, the ***Darwin Core*** elements and the ***ABCD Schema***. Both of these standards allow data on individual specimens to be shared as electronic documents which can be transmitted across the Internet. Data which are exchanged using these standards are said to be "structured" – each standard piece of information appears in a defined position and a defined format. The Darwin Core is a relatively simple model which corresponds to the four core elements described above. The ABCD Schema is a much richer model which will allow data providers to share most of the information which they may hold on any specimen. It includes a large number of optional elements which may not apply or for which data may not be available for many specimens.

These standards cover specimen and observation data. GBIF is also working with the Taxonomic Database Working Group (TDWG) and others to develop and establish appropriate standards for exchange of other classes of biodiversity data, particularly taxonomic name data and taxonomic descriptive data.

The GBIF model for data exchange is therefore that all data providers should use a common, structured data model to share the core information from their databases. These data models will also support the exchange of much more detailed data when such data are available.

[1] More complicated queries may use these elements to perform an initial filter of the available data to identify potentially relevant records. The full query can then be executed against the detailed data for each of these records.

3 Web services support the exchange of structured data

⇒ *GBIF Biodiversity Data Architecture* - ***3.3 Web Services Architecture***
⇒ *GBIF Biodiversity Data Architecture* - ***5.1 GBIF Portal Components***

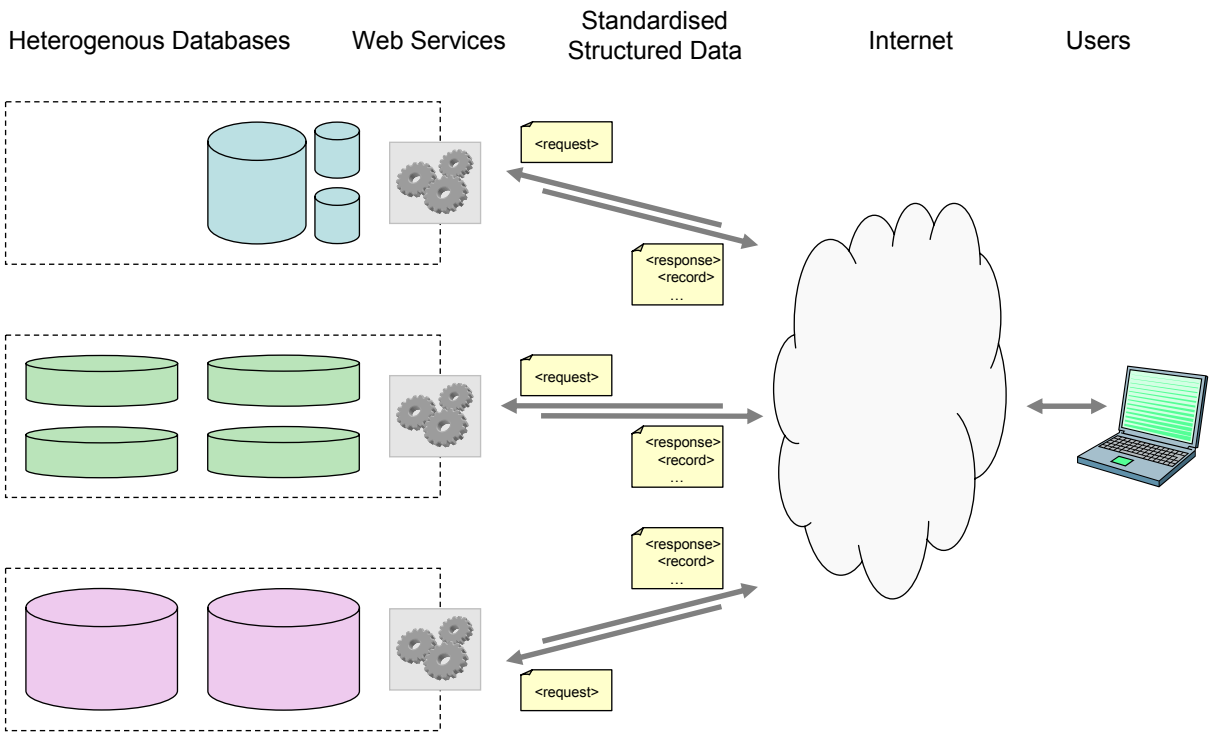

Figure 3 – This diagram shows how web services can allow tools to access data from a wide variety of different data resources using the same interface and data standards. Institutions acting as data providers offer one or more electronic services, each associated with a 'binding' (the web address at which the service can be accessed) and a predefined 'technical model' which identifies the data standards used by the service. Users of these services see them all as similar regardless of how the underlying systems are implemented.

So far we have discussed *what* data should be shared by biodiversity data providers. The next question is *how* they can share these data. As has already been mentioned, different institutions hold their data in different formats and include different elements, according to their individual needs and the software selected for managing their records. GBIF therefore needs a simple model which will allow institutions to share their data using the structured formats described in the previous section, regardless of what formats they use in their own databases.

Fortunately recently developed technologies on the Internet provide an ideal solution for this problem. GBIF will use ***web services*** to make all of the connections between different providers. A web service is a program which can handle remote requests from across a network. Users and applications

submit these requests using the same mechanisms used by a web browser when it downloads pages from a web site[2]. The web service processes the user's request (e.g. "Return data on all specimens of *Pathera leo* from after 1900") and responds by sending back an electronic document, normally some kind of structured data[3].

Each web service will implement a defined and documented interface. One such interface could be for requesting specimen and observation records. Another may be to find synonyms for a given taxon name. An interface can be defined for every operation required within web-based biodiversity informatics. Programs can then be written to process these requests and return the matching data from the provider's own databases. All that is required to add a web service to an existing database is a (relatively small) program which can receive a service request and convert it into a database query for the particular database model concerned. These programs are called service ***wrappers***.

A suitable web service interface has already been defined for exchange of specimen and observation data. This is the ***DiGIR*** (*Distributed Generic Information Retrieval*) protocol, which is already being used by several other networks. It allows users and applications to supply search criteria to find specimen and observation records. These criteria may include the name of a species or a rectangle defining an area on the earth's surface. Wrappers have already been developed for collection owners to map their database models to Darwin Core format and to publish their data on the Internet using DiGIR. These wrappers are reusable and can be configured for use with many different database formats.

GBIF is defining a range of other web services. These will include:

- A Taxonomic Name Service supporting requests for taxonomic name data. This will provide access to the information addressed by GBIF's work programme to develop an Electronic Catalogue of Names of Known Organisms (ECAT).
- Alternative versions of the Specimen/Observation Service which use DiGIR to exchange data structured using the ABCD Schema as well as Darwin Core
- A General Resource Service for data providers to share the addresses of web pages and other resources which hold information on different taxa
- A Name List Service for data providers to publish lists of taxa for various purposes (e.g. lists of protected species, lists of notifiable pests, checklists for a given region or national park)

[2] These exchanges use HTTP (*HyperText Transfer Protocol*), a standard process which allows applications and users to retrieve any kind of electronic document across the Internet.
[3] The requests and responses are normally formatted in XML (*eXtensible Markup Language*), a standard method for defining the content and structure of an electronic document.

Web services provide GBIF with a simple pattern from which to build a data network. This network has the following characteristics:

- Each data provider may offer one or more web services.
- Each web service is of a predefined class, offering a particular type of data in a particular format.
- Every web service offers access to a number of data records.
- Every data record relates to a named taxon, which allows it to be related to other data offered by other web services.

This model can be used for specimen and observation data, for taxonomic name data, for taxonomic descriptions and for any other miscellaneous web resources.

4 GBIF Data Nodes provide biodiversity web services

⇒ *GBIF Biodiversity Data Architecture* - ***3.4 GBIF Nodes***
⇒ *GBIF Biodiversity Data Architecture* - ***3.10 Globally Unique Identifiers***
⇒ *GBIF Biodiversity Data Architecture* - ***5.3 Data Node Components***

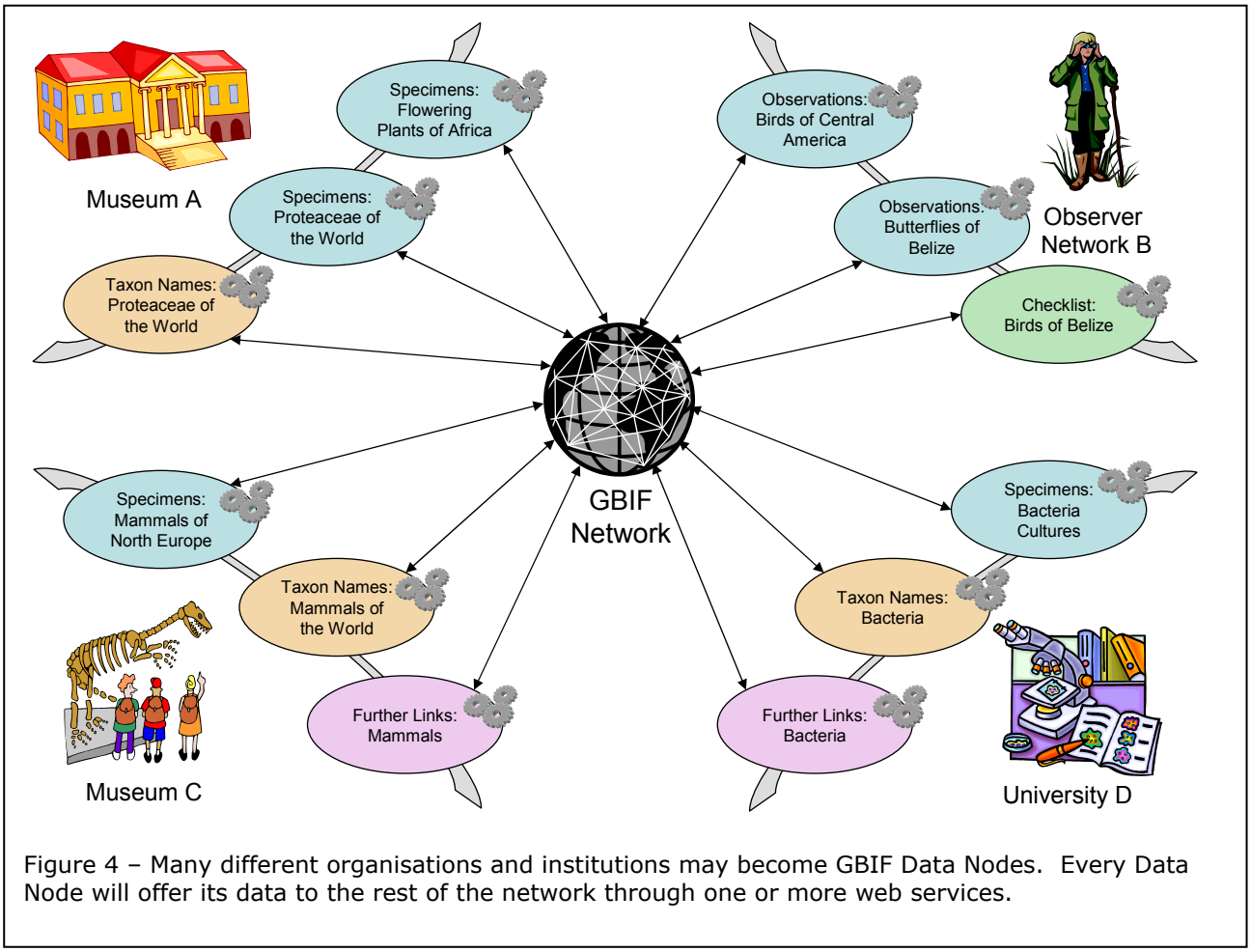

Figure 4 – Many different organisations and institutions may become GBIF Data Nodes. Every Data Node will offer its data to the rest of the network through one or more web services.

The primary responsibility of data providers within the GBIF Network will be to develop a Data Node which implements one or more of the standard GBIF web services. A Data Node may include be an existing web site or one newly developed as part of the GBIF Network. It may provide a range of information other than that handled by the standard GBIF web services. However its linkage with the rest of the GBIF Network will be through the web services it implements.

This means for example that an institution which holds data on specimens or observations of organisms should install a DiGIR wrapper which can process the standard requests for such data. It may also make these same data accessible to users in other ways. For instance there may already be an

existing web site offering web pages on each specimen. Such alternative interfaces may provide access to quantities of information which are not exposed through the DiGIR interface. The GBIF structured data models will include pointers to this additional information. The ABCD Schema already includes mechanisms for including these links and GBIF is investigating adding appropriate elements to the Darwin Core.

Data Nodes have one other important responsibility in relation to the data that they share. Users need to be able to relocate data items at a later time. Therefore every record made accessible through any of the GBIF service interfaces should include an identifier which allows it to be distinguished from all other records within the same service, even if some of the data within the record is subsequently modified. These identifiers may take any appropriate form. They may be numbers taken from a log book or card index, or they may simply be database keys. The important point is that it should be possible for a user to revisit the service and request the same identified record again. These identifiers do not need to be unique across different service implementations. The presence of these identifiers has several other benefits:

- If a data provider deletes or modifies a record, it is simple to notify the network of the change.
- If a user wishes to submit comments on a particular record, it is easy for the network to identify the correct record to the provider when it transmits the comments.
- In some cases the same data may be made accessible to GBIF via multiple paths (e.g. if a Data Node is also a member of another network that is sharing data with GBIF). In this case the identifiers may help GBIF to recognise the presence of duplicate records.

5 The GBIF Network maintains a central registry of Data Nodes

⇒ *GBIF Biodiversity Data Architecture* - ***3.8 Services Registry***
⇒ *GBIF Biodiversity Data Architecture* - ***5.1 GBIF Portal Components***
⇒ *GBIF Biodiversity Data Architecture* - ***6.1 Services Registry (UDDI)***
⇒ *GBIF Biodiversity Data Architecture* - ***6. 2 Services Metadata Index***
⇒ *GBIF Metadata Standards* - ***throughout***

Data Node	Type of data	Taxon	Region	Records
Museum A 	Specimen/Observation	Flowering Plants	Africa	327000
	Specimen/Observation	Proteaceae	World	23000
	Taxonomic Names	Proteaceae	World	1500
Observer Network B 	Specimen/Observation	Birds	Central America	68500
	Specimen/Observation	Butterflies	Belize	4200
	Name List	Birds	Belize	587
Museum C 	Specimen/Observation	Mammals	North Europe	1800
	Taxonomic Names	Mammals	World	8000
	General Resources	Mammals	World	600
University D 	Specimen/Observation	Bacteria	World	1200
	Taxonomic Names	Bacteria	World	5000
	General Resources	Bacteria	World	400

Figure 5 – The GBIF Portal will incude an electronic registry of all web services in the network. This will allow users to locate relevant services based on the class of service, the taxa covered, the geographic scope and other factors.

The GBIF Data Nodes described in the previous section all offer compatible data to any user or application which accesses their web services. However before users can access these data, they first need to be able to locate the Data Nodes and their services. The GBIF Network will therefore include a central registry of Data Nodes. This will be maintained as part of the GBIF Portal operated by the GBIF Secretariat (although multiple copies will ultimately be shadowed to different locations around the world).

The central registry will contain two kinds of information[4]:

[4] This information is known as *metadata*, a term which refers to data which provide information about other data (in this case data which describe the data on offer through the web service).

1. Technical information about the web service and how to access it (including the web address and the data standards supported for data exchange)
2. Descriptive information about the particular data which can be retrieved through the service (including the taxonomic, geographic and temporal scope of the data and descriptions of how the data were collected).

The first set of information is required so that the actual data can be found and accessed correctly. The second set allows users to choose particular web services based on the descriptions of what the services have to offer. When combined together, the two sets of information can give a user full control over which data they use from the GBIF Network.

The information is added to the central registry by the data providers themselves. Tools are being developed to automate this process and to incorporate it within the task of setting up a web service. After the information is added to the registry, a check is performed to make sure that it does indeed identify a genuine provider of biodiversity data. The web service only becomes an active part of the GBIF Network after this check has occurred. This is to prevent the network from being abused, e.g. for advertising purposes.

6 GBIF maintains an index to biodiversity data

⇒ *GBIF Biodiversity Data Architecture* - ***3.6 Biodiversity Data Model***
⇒ *GBIF Biodiversity Data Architecture* - ***5.1 GBIF Portal Components***
⇒ *GBIF Biodiversity Data Architecture* - ***6.3 Biodiversity Data Index***

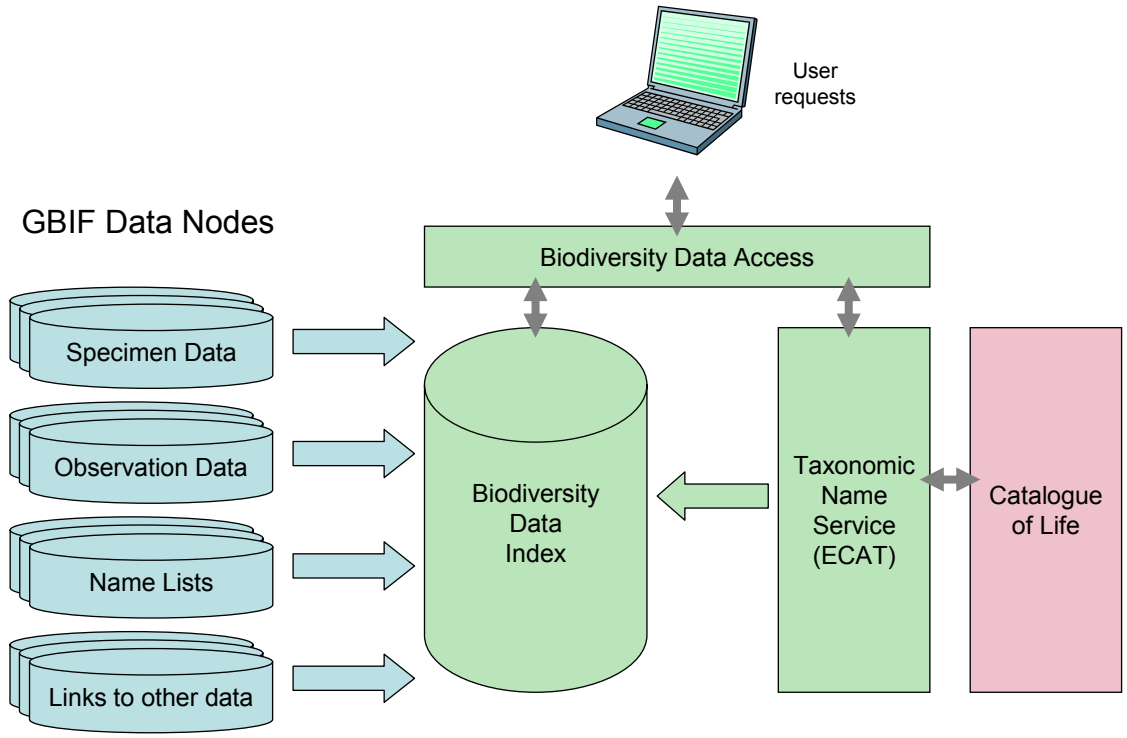

Figure 6 – The GBIF Portal will contain an electronic index of all data records in the network. These records will be indexed according to the taxon name each contains. This index is a central component of the network and ensures that relevant data records can be found quickly without having to visit every Data Node.

The central registry of web services gives access to data providers and allows users to select services based on their descriptions. However this will not in itself make it easy for users to issue search requests with a global scope. Consider a user who wishes to find information about specimens of *Plodia interpunctella* (Lepidoptera, a moth that occurs globally as a pest of stored foods). A simple approach would be to send a request to every GBIF web service which has a description indicating that it may hold data on moth specimens[5]. As the number of data providers increases, this filtering of web services by their descriptions could still leave hundreds of relevant services to

[5] Note that this will include institutions whose descriptions simply identify that they hold data on insects or animals in general.

check. It will therefore become more and more costly for the network to process even a simple request.

To solve this problem the GBIF Network will maintain an index to the biodiversity data distributed throughout the network. The GBIF Portal will send requests to every web service in the network and use the results to construct an index containing a few key elements from every record in the network. These key elements are just those which will most help in processing user requests. For specimen and observation records these elements are the same ones identified earlier (in *2 Users need data structured*), i.e. the taxon name associated with the record and the location and date of the record, along with the identifiers required to retrieve the full record from the provider.

With this index in place, our user can submit a request to the GBIF Portal for all specimen records relating to *Plodia interpunctella*. The GBIF Portal can now process this request directly. It can immediately return a set of Darwin Core or ABCD Schema records each holding the core elements for one of the matching items in the network. Each of these records will include a link to the appropriate data provider for the user to retrieve the complete record. This core information returned by the GBIF Portal will be fully adequate for displaying an overview of the results to a user with a web browser.

In other cases the user or application issuing a request may have no wish to see these partial records. This will particularly be the case when the request is issued by an application which processes detailed information. In such a case the application may indicate within the request that the response should include full detail for all matching records. The GBIF Portal will then itself issue the necessary requests to each of the matching providers and combine the detailed results into a single response. Note that in this case the response may take much longer to generate and there is a high probability that some of the data providers may be temporarily unavailable, in which case the results will still not be complete.

A central index offers another benefit. The species *Plodia interpunctella* was originally described by Hübner in 1813 as *Tinea interpunctella*. The GBIF Portal can use the Taxonomic Name Service (ECAT) to identify such synonyms. It is then easy to locate all records indexed under any known synonym and for these to be included in the results returned to the user. It would be very difficult to provide this same function without holding this core index data in a central location.

7 The GBIF Portal offers a gateway to data

⇒ *GBIF Biodiversity Data Architecture* - ***3.9 Data Aggregation***
⇒ *GBIF Biodiversity Data Architecture* - ***4 Web Services***
⇒ *GBIF Biodiversity Data Architecture* - ***5.1 GBIF Portal Components***

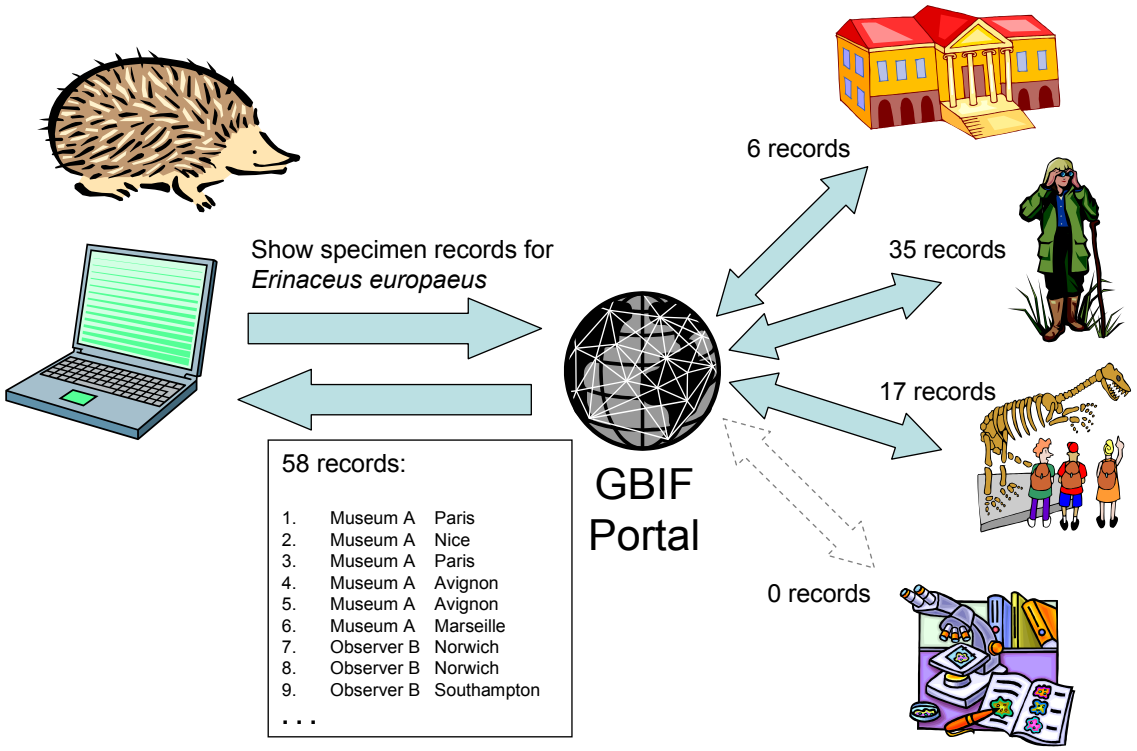

Figure 7 – The GBIF Portal uses the information in the biodiversity data index, along with the data in all of the Data Nodes, to provide an integrated view of the whole network. It can therefore process the same type of requests as an individual Data Node, but its responses can include information about records from throughout the network.

The previous sections have already explained the services registry and biodiversity data index which will be major components of the GBIF Portal to be operated by the GBIF Secretariat. These facilities will be accessible to users visiting the GBIF Portal site, but they will also be made accessible for use by other web sites and applications. To do this the GBIF Portal will itself implement a range of web services. These will operate in exactly the same way as the web services implemented by Data Nodes, and users will be able to use the same standards and tools to access them.

Some of these services are similar to those offered by the individual Data Nodes. The GBIF Portal makes use of the central index to make itself appear

somewhat like a single Data Node giving simultaneous access to all of the data in the whole network. Some of the other services will be more specialised.

The key services offered by the GBIF Portal will include:

- A Metadata Service supporting searches through the services registry. This will allow web applications to find appropriate web services.
- A Specimen/Observation Service offering a consolidated view of all specimen/observation data in the whole network. This will provide access to the index data described in the previous section, optionally filled out with detail from the individual providers.
- A Taxonomic Name Service offering data from the Electronic Catalogue of Names of Known Organisms which GBIF is developing to provide data on taxon names, including synonyms and vernacular names and relationships between taxa at different ranks.
- A Name List Service which will handle access to other lists of taxon names such as regional checklists and red lists.
- A General Resource Service offering a consolidated view of all web URLs which have been registered as holding information on different taxa.
- A User Feedback Service, which allows users to send comments relating to any service or record and ensures that these are transmitted to the appropriate contact individuals.

Other web sites may use any of these services to access data for inclusion in their own displays. Software developers may also develop web applications which manipulate and process data accessed from these services (along with data retrieved directly from the Data Nodes).

8 GBIF Participant Nodes can offer tailored information

⇒ *GBIF Biodiversity Data Architecture* - ***3.4 GBIF Nodes***
⇒ *GBIF Biodiversity Data Architecture* - ***5.2 Participant Node Components***

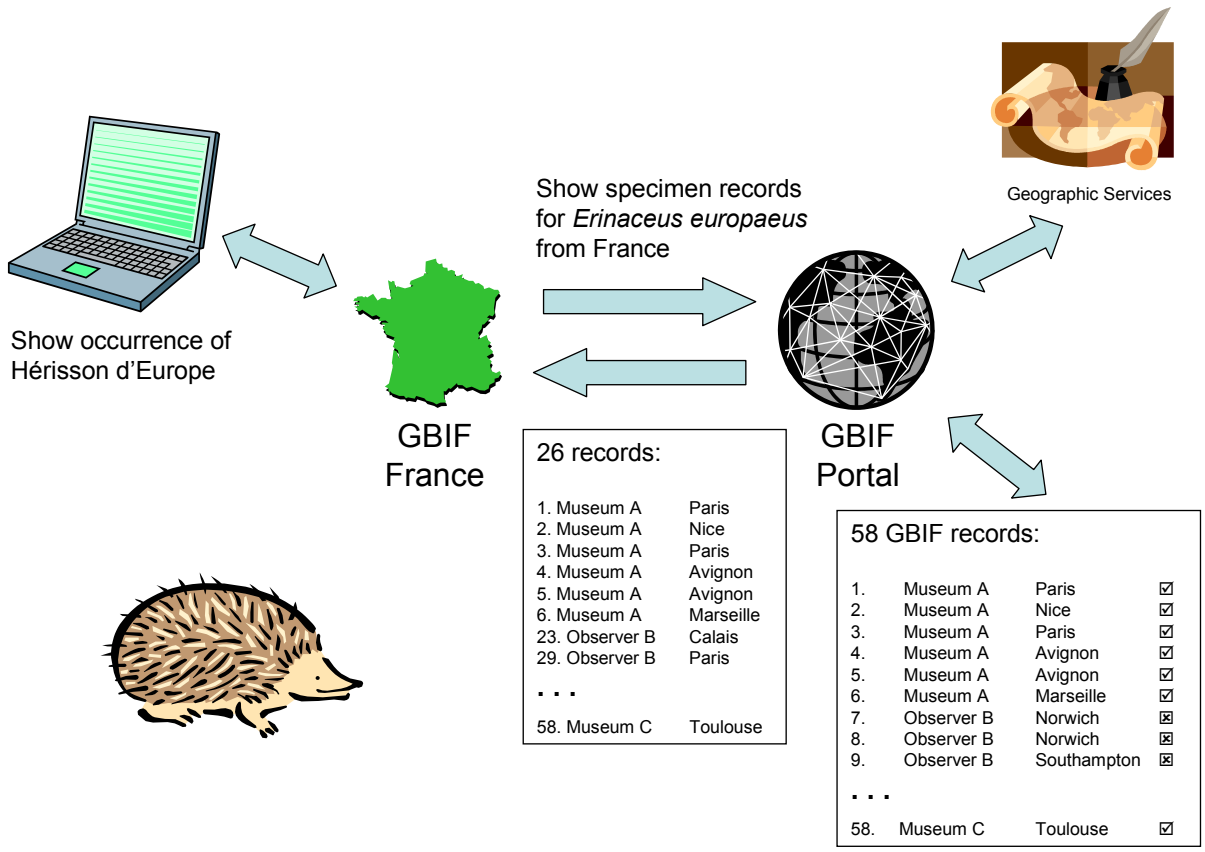

Figure 8 – Just as the GBIF Portal makes use of the Data Nodes' web services to provide an integrated view of the whole network, Participant Nodes can make use of the GBIF Portal's web services to create their own tailored views of the data. In this case a user visits the French node and requests records relating to the European Hedgehog (*Erinaceus europaeus*). The French node can easily satisfy this request by forwarding it to the GBIF Portal with an extra parameter restricting the results to specimens collected in France. This allows the French node to show relevant data even if the specimens are held in institutions outside France.

The Data Nodes and the GBIF Portal both provide essential services to the GBIF Network, but the GBIF Participant Nodes also play a key role. These are web sites developed by countries and organisations which have become GBIF Participants. These sites will normally provide a tailored view of biodiversity data of relevance to the Participant's region or area of special interest.

Some of the data accessible through these nodes may be unique to the node concerned. For example a Participant Node may include data on subjects not currently covered by the GBIF Network. However these nodes may also

construct their own data services based on the data accessible through the GBIF Portal and Data Nodes.

An example will show the benefits to Participant Nodes from being able to access the centralised services of the GBIF Portal. A Country Participant Node could develop its own portal giving access to all of the Data Nodes within that country. This would probably give a user access to most of the relevant data. However many specimens are preserved in collections outside the countries from which they were collected. Data on these specimens would be invisible to a Participant Node which simply checked its own national institutions for data. The Node may therefore instead send a request to the central Specimen/Observation Service hosted by the GBIF Portal. The request might for example be equivalent to "Return details of all specimens of armadillos (Dasypodidae) from Mexico". The GBIF Portal could then process this request for the Participant Node as follows:

1. Use the Taxonomic Name Service to resolve the taxon name Dasypodidae into a set of species names (including synonyms).

2. Search the biodiversity data index for all specimen records from Mexico which have been indexed under any of these names.

3. Return a record for each matching specimen in Darwin Core format.

The Participant Node can then easily display these results in a web browser with forward links to all of the detailed records, whether these are held in Mexico or elsewhere.

Participant Nodes are also encouraged to develop other web services of their own and to offer these for use by other nodes in the network. One example could be a web service which can accept a DiGIR request, forward it to the GBIF Portal to get the matching specimen and observation records and then generate a map showing the position of each specimen. The document returned from this service would be the map generated by this process.

9 The GBIF Network supports flexible applications

⇒ *GBIF Biodiversity Data Architecture - **5.1 Basic Principles***

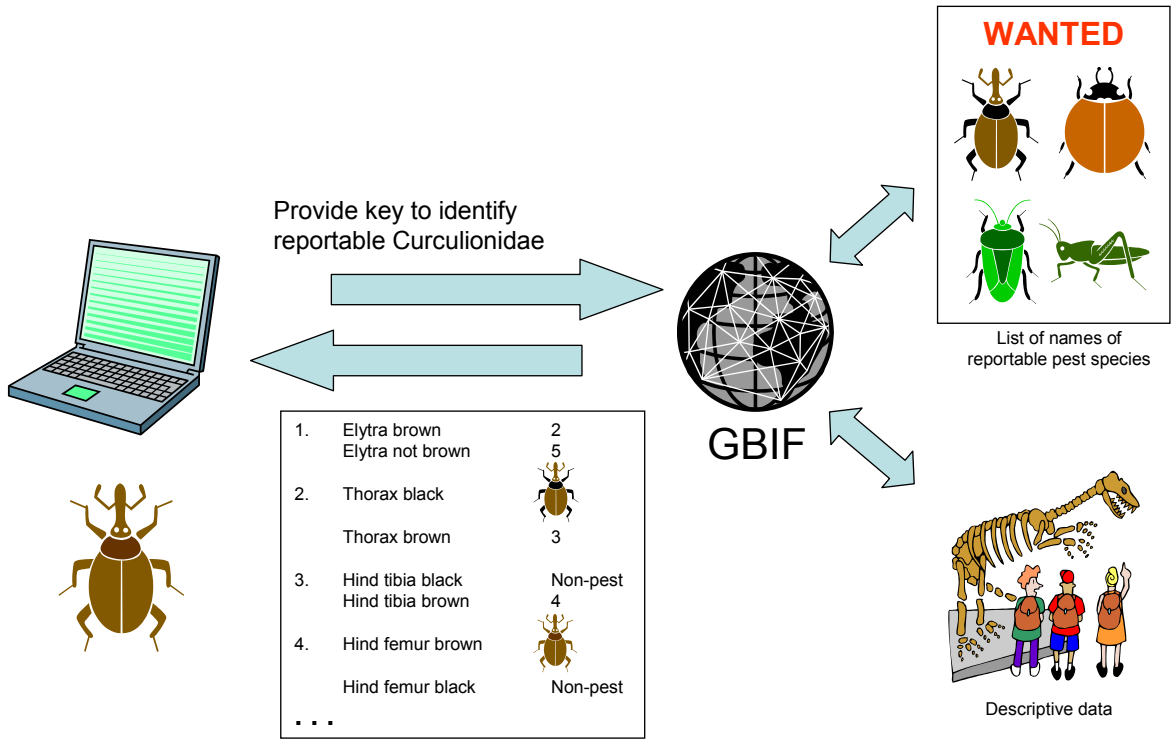

Figure 9 – An example of the kind of application which GBIF may ultimately support. A customs official requires assistance in identifying a weevil which may be a pest species. The application retrieves the list of taxa which are of interest (pest species) and data including the diagnostic characteristics for the relevant species. It can then construct a key allowing the official to identify the species of concern.

The GBIF Network is being constructed as far as possible to use common data standards and access mechanisms at every point. In particular the same web service interfaces will be implemented by Data Nodes, the GBIF Portal and Participant Nodes, all offering compatible data at different levels of granularity and coverage.

Over the next few years this consistency should make it possible to develop new web-based biodiversity applications which can use any of these different nodes as their data source without modification. A researcher with an interest in a single collection or in the holdings of a single institution may use such a tool to open a connection directly to the Data Node concerned. Those with a global interest in a taxon may choose to direct all of their requests via the GBIF Portal. Those with a geographical interest may make use of the tailored

services of a Participant Node. In all cases it should be possible for the same processes and tools to work unchanged.

The use of common standards for the exchange of taxon names should have a similar effect. An application which makes use of a taxonomic hierarchy could connect to the GBIF Portal's Taxonomic Name Service and allow a user to move around the complete tree of known organisms. It could alternatively connect directly to a Global Species Database providing data on a particular taxonomic group, or even to a regional checklist or red list in place of a taxonomic authority file.

The use of structured data model will make it relatively simple to build flexible and reusable biodiversity data applications.

10 The GBIF Portal monitors data usage

⇒ *GBIF Biodiversity Data Architecture* – ***4.6 User Feedback Service***
⇒ *GBIF Biodiversity Data Architecture* – ***6.5 User Logs***

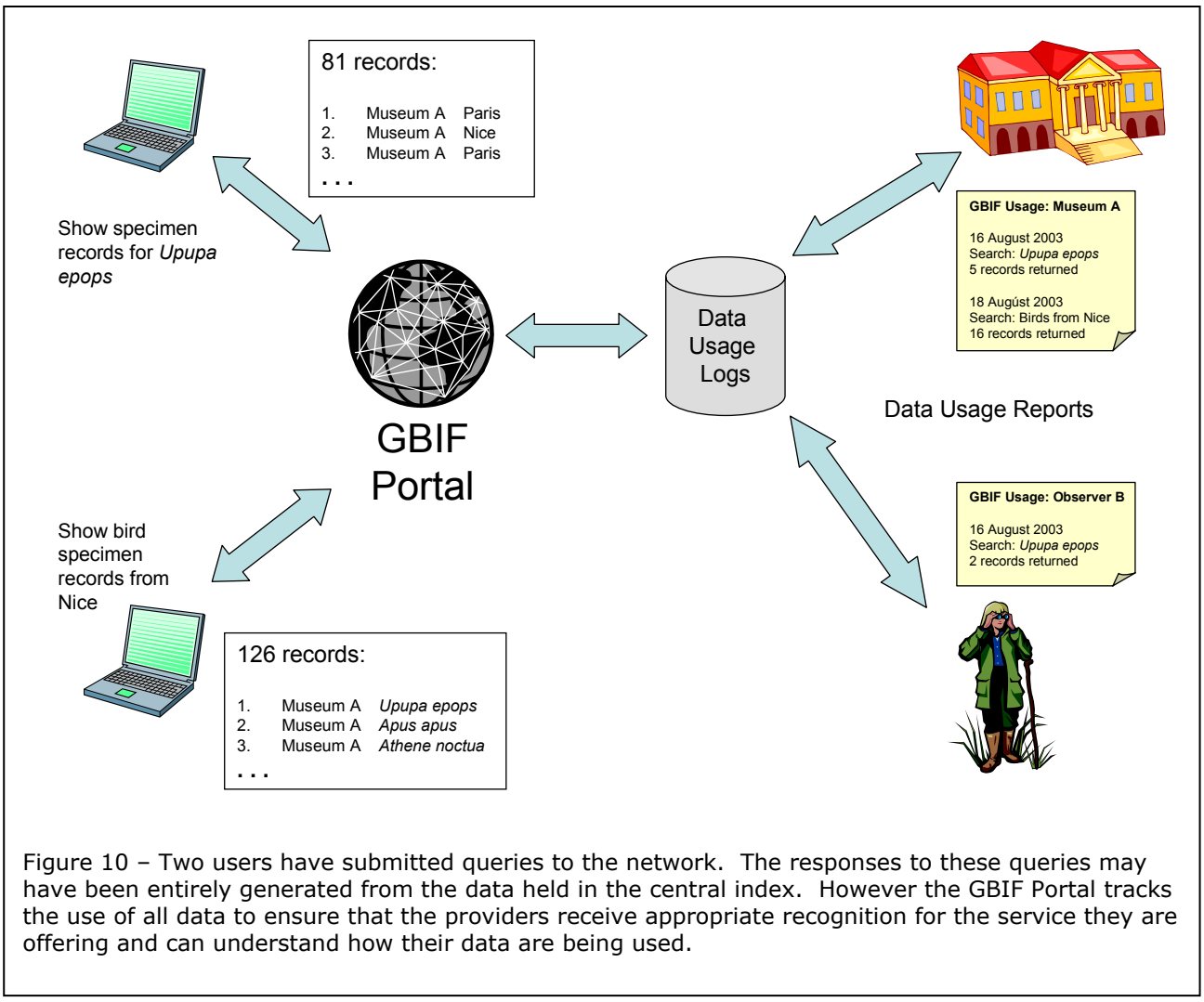

Figure 10 – Two users have submitted queries to the network. The responses to these queries may have been entirely generated from the data held in the central index. However the GBIF Portal tracks the use of all data to ensure that the providers receive appropriate recognition for the service they are offering and can understand how their data are being used.

The GBIF Portal is also responsible for ensuring that the Data Nodes are notified of data usage. All user requests will be logged along with information about which records were returned as part of the reply. The logs will monitor the occasions when the GBIF Portal simply returns core data directly from the index as well as those when the GBIF Portal first retrieves additional detailed data from the Data Nodes. This information will ensure that the Data Nodes can understand how their data are being used. It is hoped too that this will be a means by which data providers can receive some credit for what they share.

In addition to these summaries of data usage, users will have the opportunity to send comments to the provider of any data. These comments could be technical matters relating to the presentation of the data, or queries and

suggestions in relation to the data themselves. In all cases the GBIF Portal will identify the relevant web service and data records concerned and then forward the comments to the manager of the appropriate Data Node.

These feedback mechanisms will make the GBIF Network a more interactive medium from which both users and providers receive benefit.