Report of the

GBIF Metadata Implementation Framework Task Group

(MIFTG)

September 15, 2009

Task Group Members

Matthew B. Jones (Co-chair)
National Center for Ecological Analysis and Synthesis (NCEAS), USA

Nic Bertrand (Co-Chair)
Centre for Ecology and Hydrology, U.K.

Jörg Holetschek
Botanic Garden & Botanical Museum, Germany

Vivian Hutchison
National Biological Information Infrastructure (NBII), USA

Burke Chih-Jen Ko
Taiwan Biodiversity Information Facility, Academia Sinica, Taiwan

Ángela Suárez-Mayorga
Colombia GBIF Node, Colombia

Melanie Meaux
NASA/GCMD Ocean and Polar Sciences Coordinator, USA

William Ulate
GBIF Science Committee, Chair IDA; The Nature Conservancy (TNC), Costa Rica

David Watts
Australian Antarctic Division, Australia

GBIF Liaisons

Tim Robertson
Global Biodiversity Information Facility Secretariat, Denmark

Éamonn Ó Tuama
Global Biodiversity Information Facility Secretariat, Denmark

Executive Summary

The Global Biodiversity Information Facility (GBIF) aspires to expand beyond their historically successful focus on species point occurrence data and become a major provider of discovery and access services for a wide variety of biodiversity data types. A distributed metadata catalog system that describes and makes accessible general information on datasets of primary biodiversity data is recognised as an essential component of GBIF to achieve this objective. In 2008, GBIF convened a working group which reviewed the existing GBIF informatics architecture in regard to metadata and delivered a set of general recommendations on a strategy for incorporating metadata as a core component of that architecture [GBIF08].

In this report, the GBIF Metadata Implementation Framework Task Group (MIFTG) recommends best practices for deployment of metadata systems to support the "technical, social and policy framework" needed for publication of primary biodiversity data.

The principal usage scenarios for which this catalog should be designed are data discovery, human interpretation, and analytical reuse of "primary biodiversity data", defined as a collection of measured values that pertain to an organism. These data will likely cover diverse scientific areas such as species distribution and abundance, measurements of characteristics of organisms, physiology, ecological processes, behavior, experimental data, and others.

The recommendations in this document span the gamut of implementation issues that GBIF will need to address when establishing a metadata network. The most critical recommendations, however, surround choices of metadata specifications, the architecture of the metadata system, and the interaction of GBIF with existing metadata catalog initiatives.

Metadata Specifications. The MIFTG recommends that GBIF should accept, store, index, and search metadata in multiple formats that are in common use in the ecological and biodiversity communities. These formats include Ecological Metadata Language, the FGDC Biological Data Profile, and the ISO 19115 geospatial metadata specification, among others. In addition, we recognize that crosswalks among metadata specifications are typically lossy and therefore the GBIF metadata catalog must be able to return metadata in the original format in which it was contributed to GBIF. This approach differs from many other networks that use internal representations to store metadata and cannot return the original documents.

Metadata content. GBIF minimum requirements for metadata provision should be trivial in order to promote participation and adoption of the GBIF system. The minimal acceptable metadata record might only include the Identifier, Title, Creator, Contact, Metadata Publisher, and Abstract for a data set. Despite these modest requirements, GBIF should still highly recommend that metadata additionally include geographic coverage, temporal coverage, taxonomic concepts, methods, data quality (linked to domain specific controlled vocabularies), provenance, thematic keywords, structured entity and attribute descriptions, measurement units using a controlled vocabulary, physical format of the data, distribution information, access control, and intellectual rights. In addition, GBIF should recommend that a full, detailed, and high quality metadata record is in the best interest of scientific advances. In order to ameliorate language incompatibilities among GBIF members, we also recommend that required metadata must be provided in English, with an optional additional translation to one or more other languages. Finally, each metadata record and data object should possess a location-independent, globally unique identifier which can be used to retrieve the metadata object and serves to differentiate each version of the object (i.e., the ID is idempotent).

System architecture. Because of network latency and accessibility issues at continental scales, we recommend that GBIF should build a distributed system of regional nodes, each containing a replica of all metadata. These regional nodes will provide rapid and reliable access to the metadata

system from all country nodes, and will enable GBIF to improve fault tolerance and load balancing. This architecture differs from the current architecture of the specimen data that is centralized at the GBIF Secretariat. To achieve this architecture, each regional node must replicate metadata to other regional nodes when record changes occur, rather than waiting for periodic harvests of the whole collection of a node. Because each version of a metadata record will have a unique identifier, the replication process can be more efficient and more timely than current harvesting approaches. In addition, the use of a replicated set of regional nodes will allow GBIF to develop a 'virtual portal' that provides the appearance of a centralized search facility but is actually implemented to provide services from the best regional node based on load-balancing and failover considerations. Finally, the GBIF metadata catalog system should expose one or more standard query APIs for programmatic access so that many third party tools and systems can be used to both access and contribute metadata to the system.

Community alignment. GBIF is undertaking this initiative in a global community that already has an abundance of metadata cataloguing initiatives and data sharing efforts that are well established. Groups such as the National Biological Information Infrastructure, the Knowledge Network for Biocomplexity, the World Data Centers, and DataONE have existing systems and significant expertise that would benefit GBIF. GBIF should strive to collaborate with these existing groups in order to not reinvent systems that already exist. In addition, GBIF should adopt, or adapt, existing technology where it meets most of the needs of the catalog project, and work to contribute system improvements back to the broader informatics communities through participation in open source projects. In addition, GBIF should develop and pursue an implementation plan for the catalog system that builds infrastructure in an incremental fashion. Recognizing that the software engineering team for GBIF is small, it will be crucial that an incremental development strategy is adopted that produces working systems with initially limited features but that then evolve and improve over time. Finally, in engaging with the community, it is critical that GBIF provide both attribution and branding for original metadata providers in a way that avoids the feeling that existing initiatives have been subsumed by the GBIF brand. This will encourage participation in the network and help ensure the utility of the metadata catalog system for the broader science community.

The remainder of this report provides a detailed set of recommendations that complement these general principles and that will enable GBIF to develop a metadata catalog system that is broadly useful to the global biodiversity community.

Table of Contents

1. Introduction

GBIF aspires to expand beyond their historically successful focus on species point occurrence data and become a major provider of discovery and access services for a wide variety of biodiversity data types. These data types could include evidence of species distribution (images, sounds, tissues), ecological information (e.g., abundance, population cycles, behavior), habitats (including their geospatial representations), and species characteristics (i.e. natural history attributes, genes). A distributed metadata catalog system that describes and makes accessible general information on sets of primary biodiversity data is recognised as an essential component of GBIF.

In 2008, GBIF convened a working group which reviewed the existing GBIF informatics architecture in regard to metadata and delivered a set of general recommendations on a strategy for incorporating metadata as a core component of that architecture. Continuing that work, this document, produced by the GBIF Metadata Implementation Framework Task Group (MIFTG), will focus on the actual implementation issues associated with building a global metadata catalog for GBIF. Our goal is that the GBIF network follows best practices in deployment of metadata systems and that the metadata requirements are in place to support the "technical, social and policy framework" for publication of primary biodiversity data that is being addressed by the GBIF Data Publishing Framework Task Group.

In order to identify the implementation of a system like the one described above, it is important to consider the following issues:

1. Current metadata handling by GBIF is limited by available means for capturing and describing the context of the data (both tools and metadata specifications), but these limitations have been identified already and both the Secretariat and the members of the community are contributing solutions that may be useful depending on the scale.

2. These recommendations are intended to support a long-term strategy for metadata management in the GBIF network. Given this, the recommendations must be coherent with the expected data contents that GBIF is going to manage, and the future developments in data and metadata management that may be envisioned henceforth.

3. In order to gain acceptance and be deployed, the recommendations provided herein must be compatible with the conceptual and technological infrastructure already developed by GBIF members.

1.1. Intended uses of the GBIF Metadata Catalog

In general, it is desired that metadata should allow a prospective end user of data to discover data of interest, learn how to acquire those data, and understand their fitness-for-use through reading natural language descriptions of the data (see Michener et al. 1997, Jones et al. 2001, Jones et al. 2006). Further data processing such as integrating data sets, interpreting data, and drawing conclusions are semantic capabilities that are desirable features that nonetheless may be considered beyond the current scope of the GBIF Catalog [GBIF-EML08].

The main function that the Catalog should support, in its global scope, is a global data discovery service that can present a unified view of the distributed collections that are present in GBIF member nodes. Such a discovery service requires a centralized metadata search portal that integrates the regional, national, and thematic metadata catalogs that are already in use. We also recognize that the degree of completeness of metadata (how detailed metadata is) will determine how well the GBIF Catalog will support effective discovery of relevant primary biodiversity data.

The GBIF Catalog System should also support human interpretation of data by providing natural language descriptions of the data and the methods used to acquire those data. It is also desirable

that the metadata provide support for analytical reuse of the data by leveraging structured metadata to facilitate semi-automated machine processing of the data, and potentially machine interpretation through the use of ontologies.

1.2. Definition and scope of GBIF biodiversity data to be catalogued

For the purposes of the GBIF Catalog implementation, "primary biodiversity data" is defined as any measured value or set of values that pertain to an organism. This definition is more expansive than the species point occurrence data that the GBIF Network has been working with until now. It is necessary to include other types of biodiversity information available to become an effective mechanism that allows for broad data discovery.

The data comprised within this definition would be available in many different formats and representations. The data could be categorized according to several criteria. Scientifically, data that conforms to the above definition would include both species occurrence information and a variety of types of observational and experimental data. Some examples of data that should be included in the GBIF metadata catalog include species distributions and abundance scientifically determined through surveys or experiments like tracking data for a population (birth and mortality rates, migration data), data on characteristics of a species, including measures on individuals of such species (for example: weight, fat content, genetic information), phylogenetic data, derived data from gene sequences, and data on ecological processes (such as plant transpiration rates, photosynthetic efficiency, and behavioural observations). Each of these types of data can be collected in various temporal and spatial contexts. Geospatially, the types of data could consist of a single point defining a coordinate in a certain projection where a measure was taken, a line representing a path used for data acquisition, a polygon to indicate an area from where data was acquired, and a grid or coverage to symbolize a map of assigned values, among others. The GBIF metadata catalog needs to accommodate these diverse data types and sampling contexts in order to be successful and relevant to the biodiversity science community.

1.3. Contents of this report

The remainder of this report provides an overview of implementation issues that GBIF will need to address when building a metadata catalog system. Each section presents these issues as a problem statement, a series of recommendations, and a general discussion of the issue. In section 2, we discuss the important issue of aligning the GBIF initiative with existing national and global metadata initiatives that have been under way or are currently arising. In section 3, we review existing metadata specifications and address which specifications should be supported by GBIF. In section 4, we review issues related to building a network of metadata servers and the software that might support such a network. In section 5, we address metadata editing and provision, and in section 6 we provide an overview of issues concerning controlled vocabularies that GBIF should consider. Finally, we have two detailed appendices, one providing a comparison of metadata server systems, and one providing a comparison of metadata editing software.

2. Alignment with Related Metadata Initiatives

2.1. Problem statement

Given that the usage of primary biodiversity data extends to various domains, and its significance is revealed only when the data are put together with data from other knowledge realms, data discovery across diverse domains is as important as that within biodiversity. Along with the establishment of GBIF and its success in providing access to some 177 million species occurrence records, other leading initiatives have also made huge progress in data sharing, standards refinement, and

technology innovation. GBIF can benefit from the experiences and results of these initiatives, especially concerning the wide scope of biodiversity data types defined in the first section. It is necessary to foster interactions between GBIF and all of those initiatives, not only for the development of the metadata catalog system, but for GBIF to achieve its goal of bringing the benefits of biodiversity research to other fields.

GBIF can benefit from collaborating with related metadata initiatives. Many organizations have already been through every stage of implementation that GBIF will undertake to build its catalog system today. By engaging with these organizations, GBIF will understand the specific backgrounds of each domain which will help in assessing needs of its own diverse user groups. In addition to user needs, the technologies in use in these fields are also great models that can be referred to by GBIF. In this way, GBIF will not need to reinvent every piece of software to build its own system from the ground up. Ideally, it may only need to modify existing solutions to achieve its own needs.

2.2. Recommendations

R1. GBIF should adopt, or adapt, existing technology where it meets most of the needs of the catalog project.

The scope of the GBIF metadata framework initiative is large and overlaps significantly with other technology development efforts throughout the world. GBIF will be more likely to succeed in creating a metadata framework if it utilizes existing software, either wholesale or by making changes and additions to existing packages to meet GBIF's needs. By contributing to open source initiatives, GBIF will also help advance the quality and effectiveness of these existing frameworks for the broader community.

R2. GBIF should seek to collaborate on any new development in order to maximize impact of its development resources.

New developments usually begin after identifying needs that cannot be satisfied by existing solutions. It could be a rearrangement of a workflow using existing software, enhancements of software, or a series of new coding efforts. Once requirements are identified, GBIF should coordinate with relevant initiatives to review their scope, and to decide on a future roadmap involving collaboration on development. In this way, GBIF can gain maximum return on resources invested in software development while achieving a solution that works for itself and others.

R3. Any software developed should be made available as open source

The concept of open source has been proven as a working model for software development. It allows for more creative ways of problem solving in biodiversity informatics and encourages cooperation and community building. This is especially important as developments in biodiversity informatics mostly rely on public funds.

R4. GBIF should develop a comprehensive list of metadata specifications pertinent to various communities and make sure the metadata catalog supports these, and keep it updated

To keep its metadata catalog system interoperable with others, GBIF should closely follow refinements of the metadata standards designed by major initiatives. A comprehensive list would help users identify the corresponding metadata elements across standards as well as help developers update crosswalks between standards.

R5. GBIF should promote metadata best practices (e.g., through training activities, web, etc.)

In conjunction with the development of its metadata catalog system and its own metadata specifications, GBIF should design training courses for its participant nodes in a similar manner to that provided the for recently released Integrated Publishing Toolkit (IPT). GBIF should develop

online documentation, including writer's guides and "How-to" guides that describe details of metadata provision by GBIF participants.

2.3. Discussion

There are four roles associated with metadata initiatives: "data provider", "data aggregator", "technology developer" and "standard developer". An organization may play single or multiple roles, and with each role, may provide or use single or multiple products. For example, NCEAS developed EML and Metacat, and provides data, so it is a standard developer, technology developer, and data provider. NBII developed BDP and hosts a metadata clearinghouse, so it is both standard developer and data aggregator. By clarifying roles with which an initiative is associated, GBIF will identify its unique niche in the ecosystem of metadata initiatives and develop strategies to align with them. When the same interests are pursued, GBIF can seek collaboration as recommended [R2]. Table 1 and Table 2 list the features of different metadata initiatives.

We have recommended that GBIF should work with most relevant metadata initiatives. We highlight several of these initiatives that have particular relevance, including the International Long Term Ecological Research (ILTER) Network, DataOne, SONet and Dublin Core Metadata Initiative (DCMI), and we list additional initiatives in Table 2.

ILTER

The International Long Term Ecological Research (ILTER) Network consists of worldwide members that support data gathering and coordinate at local, regional and global scales. In order to create an ILTER-wide data catalog, EML has been adopted as its metadata standard and a core set of elements including title, keywords, abstract, creator, and spatial and temporal coverages will be generated. Also, participants agree to document the core elements of EML in English and the native language thus cross-language data discovery will be maintained at a least satisfactory level (Vanderbilt *et al.*, 2008). Recently, the Virtual Data Center (VDC) (an NSF Interop Project) was launched to provide a “cyberinfrastructure that enables open, stable, persistent, robust, and secure access to well-described and logically organized data”. In this project, GBIF participates with collaborators from Oak Ridge National Laboratory, USGS National Biological Information Infrastructure, National Evolutionary Synthesis Center, National Center for Ecological Analysis and Synthesis. (http://www.lternet.edu/news/Article224.html)

ILTER and GBIF are similar in their distributed member constitution. While ILTER is strengthening its metadata discovery mechanism across countries, lessons learned from issues tackled in ILTER would be valuable to GBIF. Collaborating with ILTER on technology development for the Virtual Data Center would also benefit GBIF.

DataOne

DataOne is a project with the aim of establishing distributed information technology architecture for long-term environmental data access and archiving at global scales. Data and metadata in DataONE will be broadly replicated to ensure accessibility and allow understanding of the biodiversity and environmental patterns and processes that are fostered by ecological, environmental, and earth science studies. DataOne will consist of geographically distributed Member Nodes that contribute data and metadata to a series of replicated Coordinating Nodes that handle services like distributed authentication, fault tolerance, and geographic, taxonomic, and temporal search. A major focus of DataOne is to be financially and technically self-sustaining after ten years.

While GBIF is implementing its distributed metadata catalog system, features and goals of DataOne make it an important project to work with, especially if the software infrastructure for GBIF can

interoperate with DataONE. It should be noted that GBIF has already signed a letter of collaboration promising to work with DataONE in building a global data access network.

SONet

The Scientific Observations Network (SONet) has been formed to initiate "a multi-disciplinary, community-driven effort to define and develop the necessary specifications and technologies to facilitate semantic interpretation and integration of observational data." In the working group, a semantic, unified, and extensible core data model will be defined for diverse scientific observation and measurement data types to represent and exchange observational data, thus enabling interoperability across data repositories and systems. This core model will be developed for use in annotating and searching for datasets and for building data integration services. SONet is addressing the needs of different users, including informatics tool developers, information managers, data providers and data consumers that need to handle extensive heterogeneity in observational data.

As GBIF will be extending its realm beyond species-occurrence data, we expect heterogeneity issues to become of utmost importance in determining the utility of the GBIF catalog. In order to improve its catalog design in cross-disciplinary data discovery, we suggest that GBIF work with SONet to develop appropriate solutions to the semantic representation of data.

DCMI

The Dublin Core Metadata Initiative is an independent and international organization engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. In order to provide simple standards to facilitate the discovery, sharing and management of information, DCMI develops and maintains a core set of metadata terms as well as guidelines and procedures to help implementers define and describe their usage of Dublin Core metadata in the form of Application Profiles. Discussion and cooperation platforms are also set up for specific communities like education, government information, corporate knowledge management.

DCMI standards have broad usage scenarios beyond biodiversity. Its experience in engaging with diverse communities to promote the usage of the standards would be valuable to GBIF.

Table 1: Use* of common metadata specifications by representative organizations

Project	EML	BDP	CSDGM	ISO19115	Dublin Core	Darwin Core†	DIF	Dryad Application Profile	CF
AKN						✓			
ALA						✓			
DataONE	✓	✓	✓	✓	✓	✓	✓	✓	
Dryad					✓			✓	
EMODNET				✓					
EuroGEOSS				✓					
FAO				✓					
GCMD			✓	✓			✓		
ILTER	✓			✓					
JaLTER	✓								
KNB	✓	✓	✓		✓			✓	

NBII	✓	✓	✓	✓	✓	✓			
NCEAS	✓	✓							
NEON	✓								
OBIS						✓	✓		
OOS			✓						✓
PISCO	✓								
TERN	✓								

* By "Use", we mean the primary metadata standards that are promoted by the initiative for their data collection, not necessarily all specifications that they might exchange with other networks

†DarwinCore is used for documenting attribute information associated with species occurrence, like natural history collections or species observations, on a per-record basis. It does not focus on the background information of a particular dataset.

Table 2: Key technology initiatives

Name	Metadata standards developed	Cyberinfrastructure developed
DataONE		Interoperability API, federated catalog, registry
Dryad	Data Citation format	DSpace-based metadata catalog
EEA	GEMET	
GCDML	Genomics metadata	
GCMD	DIF, GCMD Vocabulary	DIF Authoring Tool, Metadata catalog
GeoNetwork		GeoNetwork
GEOSS		Registry
Humbolt Institute		Cassia
NBII	BDP, Biocomplexity Thesaurus	
NCEAS	EML	Metacat, Morpho, Metacat Registry, EarthGrid
NOAA		Mermaid
OGC		WMS/WFS/WCS
OPeNDAP		OPeNDAP
ORNL		Mercury
SONet	Observation ontology, vocabularies	
TDWG	NCD, DarwinCore, TCS	
U North Carolina		iRODS
USFS		Metavist

In the best scenario, GBIF would work with these major initiatives to implement an interoperability mechanism across distributed metadata catalog systems such that metadata submitted to one repository would be automatically replicated and synchronized to all of the catalog systems within the network. All metadata would be stored in its original format and mapped to a common model

for searching.

3. Metadata specifications

3.1. Problem statement

In order to achieve its mission, GBIF must seek data and metadata contributions from international partners. Much of this work will already have been completed and documented by these organizations in metadata records, thus presenting GBIF with the challenge of smoothly incorporating this information into its infrastructure. Organizations contributing data and metadata to GBIF will likely already be established in their practices of developing and storing metadata. These metadata records will be developed in different standards and formats, creating a challenge of capturing this documentation in its most robust and useful form. Full documentation in the form of complete and detailed metadata is necessary for data to be correctly understood and used. Metadata crosswalks are effective to a point; however some content will be lost, ultimately, if a system is solely dependent on them. A GBIF metadata system will need to support several metadata standards in order to capture metadata from various global sources effectively.

3.2. Recommendations

R6. Metadata should be able to describe multiple types of primary biodiversity data.

Data should include specimen occurrence, species distribution data, quantitative surveys of species abundance, ecological data on species characteristics, experimental ecological data on organisms, ecological process data, genetics and physiology of organisms, organism behavior, and species response to abiotic factors and phylogenetic studies. Metadata records contained in GBIF should aim to describe this data as completely as possible.

R7. Metadata should support data discovery, interpretation, and analytical reuse

Metadata records are essential to the discovery and understanding of complex scientific data. However, most metadata-driven search systems are notoriously bad at the recall/precision trade-off. More semantic information is thus needed to increase precision without loss of recall. At a basic level, metadata records need to be robust enough to be discovered and interpreted, and the GBIF system should support this activity. To support analytical reuse, metadata should describe a dataset in enough detail to be able to use it for analysis and to reuse it for purposes different than the original intent. In order to accomplish such a task, metadata records submitted to GBIF should be encouraged to have data documentation that contains such detail as information about the entities and attributes, in order that more advanced uses of metadata can occur.

R8. Metadata should support search/browse by space, time, taxa, and theme

Searching and browsing metadata records is a requirement for the GBIF system to be efficient. Records should contain information that allows them to be searched in many ways, including geographically, and by date, taxa and theme.

R9. Metadata should support search/browse by name of provider/name of organization

Metadata records in the GBIF system should identify the name of the data and metadata provider and the name of the organization associated with the data and metadata.

R10. Metadata should support search by related publications

Metadata records can be used as a citation source for datasets. For example, new datasets created by

using multiple data sources reference the data in citation form in the metadata record, thus creating a system of reference. Metadata records can serve as a system of citation in which data creators can be credited with references to their datasets. GBIF should recognize this activity and other uses by supporting a search by related publications.

R11. GBIF should accept metadata in multiple formats that are in common use.

There are multiple metadata standards in use. Current widely used formats include: Ecological Metadata Language (EML Versions: 2.0.1, 2.1.0), International Organization for Standardization (ISO - 19115 and various profiles), Content Standard for Digital Geospatial Metadata (CSDGM), Biological Data Profile (BDP), Geography Markup Language (GML), Darwin Core (DwC), Dublin Core, and Directory Interchange Format (DIF). The GBIF system should be designed in such a way that it can accept metadata in any of these standards to protect the integrity of the records.

R12. GBIF should provide crosswalks to enable retrieval of metadata in multiple standards commonly in use in order to aid interoperability (e.g., provide conversion among all of EML, BDP, ISO 19115).(See related recommendations R27 - support of multiple metadata models; R28 – ability to return original, contributed format)

Crosswalks between these widely used standards exist; however, since the results are often lossy, GBIF should accept and store metadata records in their original standard form. Crosswalks should be implemented in the GBIF system between the standards in order to enhance interoperability and user experience in retrieval and assessment of records.

R13. When approached for a recommendation about which metadata standard to use, GBIF should recommend a standard most appropriate for the data being described.

Different metadata standards are particularly useful for certain types of information. For example, EML has a focus on tabular datasets, and the CSDGM likewise has an emphasis on geospatial data. The same can be said of the other major standards. Metadata contributors to GBIF should be encouraged to use one of the established standards that is most appropriate for the type of data being described.

R14. GBIF minimum requirements for metadata provision should be trivial, but GBIF should accept very detailed metadata in any of the standard formats. The minimal acceptable metadata record might only include the Identifier, Title, Creator, Contact, Metadata Publisher, and Abstract.

Recognizing that GBIF should collect as much data as global partners are willing to offer, it is known that some data will be offered without detailed metadata documentation. In such cases, GBIF should accept the data with minimal metadata associated with it, although GBIF should highly encourage more detailed records be prepared as a best practice. Minimal metadata might only include Identifier, Title, Creator, Contact, Metadata Publisher, and Abstract.

R15. GBIF should highly recommend that metadata additionally include geographic coverage, temporal coverage, taxonomic concepts, methods, data quality (linked to domain specific controlled vocabularies), provenance, thematic keywords, structured entity and attribute descriptions, measurement units using a controlled vocabulary, physical format of the data, distribution information, access control, and intellectual rights.

An additional layer of metadata fields should be highly recommended from GBIF, so that the search/retrieval capabilities of the system are fully utilized, and metadata can be more efficiently used for analysis.

R16. GBIF should recommend that a full, detailed, and high quality metadata record is in the best interest of scientific advance, and that providers should provide more complete metadata than the minimum requirements and recommended fields .

Detailed metadata records afford the most return from a scientific analysis viewpoint, as records themselves can be used as sources for data analysis. GBIF should recommend that detailed metadata records be submitted to the system for such purposes. Further, data can be more effectively understood and assessed for reuse if the record is robust.

*R17. Required metadata *must* be provided in English, with an optional additional translation to one or more other languages. The optional translation should be provided in a format determined by the standard being used. Recommended metadata fields *should* be provided in English, but *may* be provided in any language as determined by the contributor.*

GBIF represents a global organization. With such an orientation, GBIF should require that minimal metadata fields required by GBIF should be represented in English; however, they may also be provided in an additional language. If a provider creates both English and additional language representations, GBIF should maintain the record in all languages provided.

R18. GBIF should develop conventions or solutions to indicate that one metadata record represents an alternate-language translation of another, and/or that two or more fields in a document represent multiple translations. GBIF should do this in conjunction with other initiatives working on metadata standards.

In an attempt to minimize the effect of duplicate records, GBIF should devise a system to recognize versions of a metadata record representing different languages. Similarly, GBIF should recognize that the same field in a record can be represented in multiple languages, while still referring to the same dataset. As other major metadata initiatives are developing conventions in this area, GBIF should develop this system in conjunction with these initiatives. Additionally, GBIF should work with standards organizations that maintain metadata specifications to support mixed-language documents.

R19. Each metadata record and data object should possess a location-independent, globally unique identifier which can be used to retrieve the metadata/data object and serves to differentiate each version of the object (i.e., the ID is idempotent).

Globally unique identifiers are an increasingly important aspect of metadata management, particularly when an organization such as GBIF will incorporate the metadata records of many partner organizations. Unique identifiers serve the important function of preventing duplication in records, particularly as the number of records contained in the GBIF system continues to grow. Additionally, identifiers provide a streamlined system for replication and metadata updates to occur. Such identifiers will also enable GBIF to interact with other data sharing initiatives. GBIF should support any of the common mechanisms for representing identifiers, as long as retrieval of the content associated with the identifier always produces the same byte stream. This allows processors to reliably know when metadata content has changed, allows metadata to be replicated unambiguously to multiple locations, and tremendously simplifies the synchronization of metadata records across multiple systems.

3.3. Discussion

The GBIF approach to metadata records contributed from international partners should be flexible enough to accept records created in a variety of standards, and in multiple languages (with English being the mandatory language for core fields). Only in accepting metadata records created in a variety of standards can GBIF expect to include records already in existence. Asking a provider to make available metadata records in a different standard from one already in use for that organization is asking too much. Therefore, GBIF should encourage the use of metadata specifications that suit the data being described, thus promoting a suite of standards. Additionally, with a well designed metadata system, GBIF will be able to offer the most detailed type of record – that which is most complete as a result of being left in its original form. GBIF should encourage its

providers to submit as detailed records as they can, but should recognize the importance of a dataset that is only described with certain core fields as also being valid. Robust metadata records are the most desired, however, due to the increasingly complex data analysis that can be performed by using detailed metadata records. Finally, global unique identifiers will provide a metadata tracking system that will allow users to recognize versions of a record, and will allow GBIF to track potentially thousands of records with relative ease, and allow for efficient replication of records in a metadata catalog system and network.

4. Metadata catalog system and network

4.1. Problem statement

As part of its mission to organize the world's primary biodiversity data, GBIF has a need to collate metadata on the wide variety of biodiversity data collected throughout the world. Thus, GBIF needs to create and maintain a global, distributed, and replicated metadata management system for collating, searching, browsing, and distributing metadata. The system must be global in order to accommodate the federation of data at continental and global scales. The system must be distributed in order to accommodate the local needs of GBIF participants and to address issues in variable internet connectivity across continents. The system must be replicated to support fast local access to the metadata, to support failover in case of regional node outages, and to guarantee the long-term preservation of the metadata.

4.2. Network Architecture Recommendations

R20. GBIF should build a distributed system of regional nodes, each containing a replica of all metadata.

By distributing full replicas of all metadata holdings to regional nodes on each continent, GBIF will ensure fast access to the data and be able to provide a reliable, fault-tolerant and efficient virtualized search portal. The number of regional nodes and their location should be tuned over time to allow for global coverage and accessibility while still minimizing cost. To the extent possible, these regional nodes could be operated by existing metadata systems in order to reduce maintenance costs.

R21. Each regional node must replicate metadata to other regional nodes when record changes occur using a GBIF-prescribed replication protocol.

The architecture for this system must require that each regional node must accept metadata records in any of the accepted specifications, catalog it, and replicate the original metadata file to each of the other regional nodes using the GBIF-prescribed replication protocols.

R22. Each regional node should also provide a harvesting interface that exposes metadata via their unique identifiers.

Harvesting protocols such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) are commonly used by indexing systems (e.g., by GCMD to harvest metadata among partners) and should be supported by the virtual portal and regional nodes. However, because harvest is typically done far less frequently than replication, all regional nodes must provide the replication services described in the previous recommendation.

R23. GBIF should choose one or more regional nodes with adequate technical infrastructure on each continent to serve as a metadata replica in that region

Regional nodes will contain a full replica of all metadata GBIF collates across all country nodes.

Consequently, their location, scale, bandwidth, and other characteristics should be carefully chosen so that each continent has reliable and efficient access to at least one regional node on its continent. More than one regional node may be required for adequate performance on some continents. Using existing country nodes and other initiatives that provide similar services (e.g., DataONE) should be evaluated before deploying entirely new nodes.

R24. GBIF should develop a 'virtual portal' that uses the regional nodes for failover (in the event of network or node outage) and load balancing across the regional nodes

Although the overall network is distributed, users should only need to know a single address to access the network. This address would provide a virtual portal that could be used to both submit and discover metadata from the system. The virtual portal should evolve over time to provide increasing levels of fault tolerance and geographic load balancing as the system grows and matures (see discussion below).

R25. GBIF needs a registry to maintain list of regional nodes and their relevant service endpoints

The distributed system will need to rely upon a registry of service endpoints for both the regional nodes and for metadata providers. The envisaged GBIF GBRDS registry system could encompass this function, or another registry such as the EarthGrid registry could be modified to meet the needs of the system.

R26. GBIF should develop and pursue an implementation plan that builds this infrastructure in an incremental fashion.

GBIF should recognize that developing the specifications and infrastructure for such a system will require several years to design, develop, and deploy. In order to make adequate short-term progress, a staged implementation plan should be designed and then utilized to deploy the system in stages. For example, one trajectory might be to build a single, centralized metadata catalog node first at the Secretariat, and then add in replicated regional nodes as the catalog technology matures, and then add virtual load-balancing to the search portal. Regardless of the exact details of the implementation plan, it should be incremental with staged deliverables and should be realistic about the amount of new software development that can be done with GBIF's small engineering staff.

4.3. Metadata catalog system recommendations

GBIF must build this overall metadata framework by establishing metadata catalog systems at each of the regional nodes. Developing such a system would be difficult, and instead GBIF should adopt an existing open-source system that it can adapt to its needs. By contributing to the development of existing, open-source systems, GBIF will reduce the scope of the development work it needs to undertake and simultaneously contribute to the improvement of catalog systems that can be used by other like-minded organizations. There are several candidate systems that could be considered for the basis of a metadata network. These systems should be evaluated using the following criteria to determine a suitable system to adapt for GBIF's needs.

R27. The metadata catalog system must support multiple metadata models natively

Because conversion among metadata specifications is almost always lossy, the system must be able to support multiple metadata specifications that are common use in the community (e.g., EML, BDP, ISO19115; see list in section 3). Although some existing systems such as Metacat allow new metadata specifications to be used without any code changes, this is not typically the case. GBIF should use a metadata system that can accommodate new metadata schemas and versions of those schemas without code changes.

R28. The metadata catalog system must be able to return the original contributed metadata object

Because the originally contributed metadata is likely the richest, the GBIF metadata catalog should be able to return an exact copy of the original metadata record in its original metadata format.

R29. The metadata catalog system must support unique versioning of metadata and data objects using globally unique identifiers to differentiate revisions

The GBIF metadata system must be able to use globally unique identifiers to store and retrieve metadata objects in order to efficiently know which metadata and data objects are present in each of the regional nodes. Strict adherence to the use of global identifiers will allow GBIF to build an efficient system in which moving metadata through the system is simple and error-free.

R30. The metadata catalog system must support replication and harvesting of metadata (and data) from providers

While many metadata catalog systems only support harvesting of metadata records, it is critical from an efficiency perspective to primarily support replication that is initiated at the provider node. Metadata providers are aware when records change and can initiate timely replication events, allowing the whole network to remain closely synchronized. In addition, harvesting protocols such as the Open Archives Initiative Protocol for Metadata Harvest (OAI-PMH) should be supported to accommodate systems that use this common protocol.

R31. The metadata catalog system must support search and discovery

The most important use case for the GBIF metadata framework is supporting the search and discovery of data holdings via the metadata catalog. The system should support free text queries as well as structured queries, particularly using Keywords and Spatial, Taxonomic, and Temporal coverage metadata. In addition, the search engine should support arbitrary logical queries against the native metadata models in which records are provided, even if these are not as efficient as the more optimized space/time/taxonomic search options. Finally, it would be useful if the discovery system allowed users to find data sets associated with particular journal publications and associated with particular scientists. This type of cross-indexing between data and contributors and publications is not available in existing systems but would be extremely useful for researchers.

R32. The metadata catalog system must support metadata in XML serializations

All commonly used metadata standards for biodiversity data can be represented in an XML syntax and validated by either an XML Document Type Declaration or an XML Schema. Thus, this is the natural serialization that must be broadly supported. Systems may also support serializations in alternate syntaxes such as JSON (JavaScript Object Notation) and RDF (Resource Description Framework), but at this time the community has not yet established metadata content schemas that use these alternate serializations commonly.

R33. The metadata catalog system must support input from multiple metadata editors

The metadata catalog system should not require use of a particular metadata editor. It should be simple for users to choose a metadata editor, save a valid metadata document from that editor, upload that document to a GBIF country node or regional node, and have the document be accepted by the system. This will allow a wide variety of editors tuned to particular user communities to flourish, and will increase overall participation in the network. GBIF should make it extremely easy to upload metadata to the GBIF regional nodes by developing extensions to some common,

open source metadata editors that allows them to upload metadata directly into the GBIF network. Morpho and Metavist are two commonly used editors.

R34. The metadata catalog system must support international language documents and queries

Current metadata on biodiversity data are expressed in a wide variety of languages. Although GBIF should require at least minimal metadata to be in English (see R17), the metadata catalog system needs to be able to accommodate records that are wholly expressed in the world's languages, including both one byte and two byte character languages. Thus, the system should support character encodings such as UTF8 that allow multibyte characters. In addition, the system would be more globally useful if it supported search and result sets to be returned in multiple different languages based on user preferences for their session when that language exists for a record.

R35. The metadata catalog system should support conversion from one metadata model to another and ability to return these alternate formats on request

Each metadata specification can be translated to others, often with loss of information. These converted metadata documents should be accessible from the search portal for people that need to access them using software that might require one particular metadata format. However, the global identifiers for these converted documents should be adjusted to reflect the differing content between the different versions of the record.

R36. The metadata catalog system should be redistributable under an acceptable open source license

GBIF should both take advantage of, and contribute to, the open-source movement in order to amplify its development of resources by building on top of existing systems.

R37. The metadata catalog system should support sorting of search results

Result sets should be sortable, a feature present in most systems.

R38. The metadata catalog system should support logical queries and filters on individual metadata fields from multiple standards

Users should be able to construct logical queries that combine multiple search conditions in novel ways. The search conditions that should be accessible should include the commonly indexed fields that span standards (e.g., spatial and temporal coverage), but should also include the fields that might be specific to one particular metadata specification (likely with a reduction in performance due to non-optimized queries). This will allow uses to build custom queries that exploit the content of particular metadata specifications.

R39. The metadata catalog system should collect access log statistics on all operations that create, read, update, or delete records

The utility of the metadata system can only be demonstrated by its use; the metadata system should keep detailed log statistics on all system operations.

R40. The metadata catalog system should maintain a summary of holdings

The system should be able to report on the aggregated holdings of particular institutions, countries, and other logical organizational levels.

R41. The metadata catalog system should enforce access control restrictions on non-public metadata for read and write by metadata editors

Although GBIF focuses on publicly-accessible biodiversity data, various contributor networks manage records that have restricted accessibility. GBIF should support these groups by providing

an access control system that allows users to specify which individuals and groups can read and change records. This is particularly important for determining who can update a record (by providing a new version that obsoletes the original), as the system should be supporting multiple metadata ingestion routes. Such an access control system implies access to a common user directory across data providers for authentication. For simplicity, GBIF could use a distributed LDAP system such as the one used in the Knowledge Network for Biocomplexity, or they could use emerging standards such as Shibboleth and the InCommon federation. GBIF should consult with other groups that are trying to build a global network of scientists, such as DataONE, in order to potentially find ways to operate synergistically.

R42. The metadata catalog system should register with one or more node registries to advertise services available.

Each regional node should be listed in a node registry so that its capabilities and services can be accessed by clients. This may include the planned GBIF GBRDS registry as well as emerging global service registries such as the one maintained by GEOSS.

R43. The metadata catalog system should expose one or more standard query APIs for programmatic access by client applications

Metadata and data management applications (e.g., Morpho, Metavist), data analysis applications (e.g., Matlab), and scientific workflow systems (e.g., Kepler) will all benefit from a common programming interface for accessing the GBIF system. Existing interfaces for querying diverse metadata standards such as EarthGrid, XQuery, SRU/SRW, and OGC geoservices should be supported by the catalog system as appropriate.

R44. The metadata catalog system should provide attribution and branding for original metadata providers

Original metadata providers have incentive to create and maintain records when they are given credit for their data and metadata contributions. Building a search portal that emphasizes the institutional brands and names of data providers is critical to widespread adoption.

R45. The metadata catalog system may expose metadata records to other search engines (e.g., provide site index for Google, Yahoo)

This global metadata network would be most useful if it were also accessible through common search portals such as Google. The catalog system would gain utility if it were to expose metadata records in a way that is machine indexable by crawlers, and provide appropriate site indexes to major portals (e.g., Google's site index file).

R46. The metadata catalog system may provide bookmarkable queries

Users may benefit from searches that are bookmarkable so they can return to rerun the search.

R47. The metadata catalog system may provide subscription services to new metadata records (e.g., RSS feed on query)

Users may benefit from subscription search services that notify users when new data matching a particular search become available.

R48. The metadata catalog system may provide thesaurus services for searching and access by other editors/clients

Users may benefit from thesaurus services that help improve the recall of searches by exploiting known relationships in controlled vocabularies.

4.4. Discussion

The current GBIF infrastructure for specimen data creates a centralized index in one system in the Secretariat. The task group recommends that GBIF create a more distributed metadata network by replicating copies of the full metadata holdings to regional nodes on each continent (Figure 1). This architecture will allow countries close to each regional node to easily access and contribute to the regional node. Metadata contributed to each regional node would be replicated (pushed) to each of the other regional nodes in a timely manner whenever a new record is created or an existing record is updated (typically within minutes). This rapid replication of metadata records is enabled via use of globally unique identifiers that unambiguously flag when a record has changed, and therefore when it must be replicated. This approach differs significantly from existing GBIF approaches, such as the repeated, wholesale re-harvesting of specimen records from country nodes even when records have not changed. The use of a replication architecture such as this will require contributing nodes to uniquely identify records and to conform to a standard replication protocol, a minimal requirement providing great gains for the global data network that will be created.

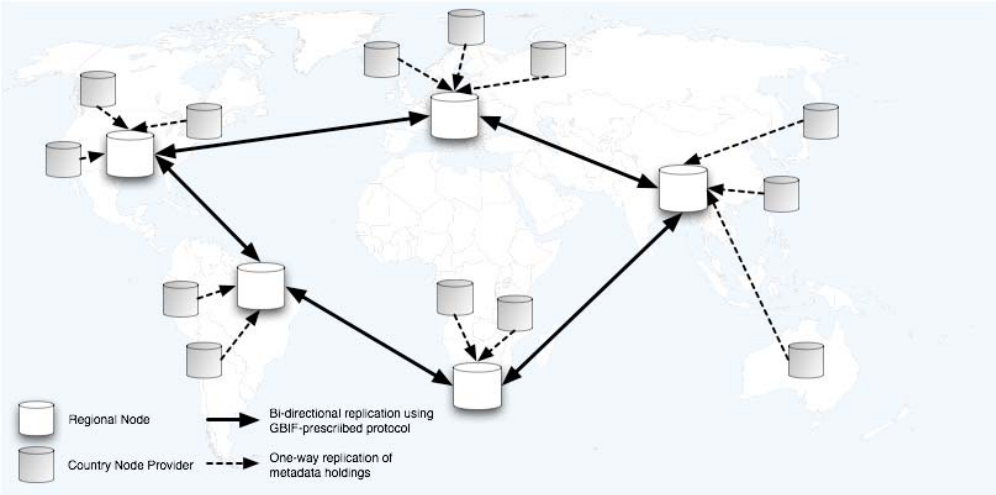

Figure 1: An hypothesized, distributed metadata catalog for GBIF. White cylinders represent regional nodes, while grey cylinders represent country nodes. Each regional node collates metadata from associated country nodes, and replicates changes to those metadata to the other regional nodes. Thus each regional node has a complete copy of all metadata in the system. Replication is used rather than harvesting to improve the currency of metadata records. A virtual portal would be established and run at each of the regional nodes, allowing rapid access to the whole metadata store from each region, as well as effective load-balancing and failover capabilities.

In addition, GBIF should provide the illusion of centralized access to metadata via the creation of a virtualized portal that is, in fact, distributed over the regional nodes. Each regional node would be able to provide all of the services of the metadata system. GBIF could evolve the system through three phases of the virtual portal. In the first phase, one of the regional nodes would act as the master node, and other regional nodes would only replace its services when the master node was unavailable (e.g., during network outages, system upgrades, etc.). In the second phase, all of the regional nodes could be used in round-robin load balancing to improve system efficiency and scalability. Any of the regional nodes could be removed from the round-robin rotation during outages or maintenance periods. In the third phase, a more sophisticated load-balancing solution could be employed that would direct clients to geographically-close nodes in order to limit

bandwidth problems over slow connections across continents, while still maintaining the capability for failover as needed.

Several metadata systems could be potentially used as the basis for the GBIF network. See the Appendix 2 in Section 10 for a comparison of systems.

5. Metadata editors

5.1. Problem statement

There are many metadata editors currently in use. GBIF recognises that each domain has invested significant resources in developing its own network and tools and does not wish to impose additional costs and impositions in order to acquire metadata and data by asking networks to change tools.

5.2. Recommendations

R49. GBIF should create a web-based editor for the GBIF portal for individuals to register their datasets. This should collect the mandatory and recommended list of fields

There will be communities that do not have ready access to established metadata tools. The GBIF metadata entry tool will allow any individual to submit metadata and reduce the barrier for data submission. This tool must be targeted and tested against relatively inexperienced users. In order to conserve resources, GBIF should evaluate adopting or extending an existing web-based editor, such as the Metacat Web Registry editor or the Mercury web-based metadata editor, as an alternative to developing a metadata editor from scratch.

Automated generation of parts of the metadata record from the associated data should be done whenever possible. Possible data elements include geographic, taxonomic and/or temporal coverages. This will be possible if the metadata and data are tightly bound and the tool can effectively trawl the data. If the dataset is updated with new or revised records then changes should be reflected in the metadata. Of course if the data is not available then manual edits of these fields will still be required.

R50. GBIF should support editors that have the following criteria

The following are criteria that should be used to evaluate the suitability of any editors. It is not expected that any given editor satisfies all criteria:

1. XML input from other editors/sources that are already in place.
2. Ability to edit entity-attribute information.
3. Support auto-capture of metadata elements from the data.
4. Support multiple schemas. Provide a validation service to those schemas.
5. Capacity to validate as you edit.
6. Require partially edited records to be saved and kept for later edits.
7. Ensure fields such as creator, contact, etc. can be easily replicated to reduce effort.
8. Copy from existing records to reduce editing effort. Create author/node-specific profiles including validation rules such as spatial extents.
9. Interface must be well designed for the audience it targets.
10. Control access to records including the ability for the editor to specify access of individuals, groups or public.
11. Metadata editors should support internationalization of the user interface and underlying software components.

12. Metadata editors should support input of metadata in multiple natural languages.
13. Off-line editing should be possible. There should be support for mobile devices for in-field editing and creation of records.
14. Consider open source versus commercial for encouraging the deployment of tools.
15. Shippable desktop application or a service-based tool (e.g., web site)
16. XML input from other editors/sources that are already in place.

It is recognised that most tools will not support all the criteria but this list may give guidelines to continued development and improvement of metadata editors.

R51. GBIF should support any metadata editor that outputs metadata that are valid according to the previous accepted list.

GBIF should allow the use of any metadata editor that is useful to the community as long as it produces metadata in one of the accepted formats for ingestion by the GBIF network.

R52. GBIF maintains a list of recommended tools against the feature set

GBIF should evaluate multiple metadata editors to highlight their strengths and weaknesses for various applications or domains. Appendix [A] contains a list of known metadata editors as a starting point for the continued evaluations and ongoing recommendations by GBIF and associated partners.

6. Controlled vocabularies

6.1. Problem statement

In any structured document, there are certain data elements that require a degree of commonality and community-accepted definition that then allows for discovery of similar or related information. Controlled vocabularies provide this mechanism and allow users some confidence that data discovery via such keywords will return a complete set of results.

We recognize the important developments being made in use of ontologies and RDF to represent metadata. However, comprehensive ontologies have not yet been accepted by the community and there are complexities that have not been addressed by existing tools for deploying ontologies to science audiences.

Controlled vocabularies for measurement parameters/characteristics/attributes/variables would be extremely useful, but there are no accepted vocabularies for these yet, and groups such as SONet/SWEET/GCMD/ will be producing them over the next few years.

6.2. Recommendations

R53. Providers should use controlled vocabularies in any metadata field for which an appropriate vocabulary exists, and should use a multi-lingual thesaurus when appropriate

To aid discovery of similar or related metadata, it is important that common metadata elements are described in a controlled manner. It is recognised that many metadata systems do not have mechanisms to ensure use of controlled vocabularies. This process should encourage such developments.

Using multi-lingual vocabularies (e.g., GEMET, NBII Biocomplexity Thesaurus) will aid in understanding and interpretation of data in different languages. If there are two competing vocabularies, then the multi-lingual version is the preferred.

R54. The GBIF vocabularies registry is a valuable service, but should be extended to include a canonical identifier for each vocabulary, and should work to be consistent with other vocabulary registries (e.g., oasis, info, srw)

As more vocabularies are developed and used, it is imperative to trace the origin of the elements that make up the vocabulary and have a shared understanding of the meaning and definition behind them. The identifier of the vocabulary should use existing identifiers from other registries where possible. If one does not exist, then GBIF should construct and publish the identifier. GBIF should be prepared to create synonymies of identifiers and a capacity to resolve synonyms if needed.

R55. Providers should reference the canonical identifier for a vocabulary when listing it in a metadata document (e.g., in the keywordThesausus field in EML)

At present there is no well-defined and consistent means of referencing an identifier of a vocabulary or a vocabulary term. The proposed GBIF registry should provide an unambiguous citation method for each vocabulary and the terms they contain.

R56. GBIF should create an applicability statement identifying which vocabularies are most appropriate for particular fields in particular metadata standards (e.g., use ISO country code in 'country' field)

Some vocabularies will be global in use but some will be domain specific. To ensure compatibility across all metadata records, it is important that users use the appropriate and community agreed vocabularies. An applicability statement will provide confidence that metadata records are using the most appropriate vocabulary.

The selected vocabulary should be sufficiently modest in size to encourage acceptance by data providers. Conversely, large and/or complex vocabularies defeat the purpose of data discovery if they are only implemented by part of the metadata network.

R57. The GBIF vocabulary registry should support registration of new and existing vocabularies by third parties

Apart from some very general vocabularies, existing vocabularies are currently relatively difficult to find and understanding their current status (in development, ratified, etc.) is also an issue. If GBIF maintains a registry of acceptable vocabularies then the biodiversity community can have improved confidence in choosing the correct vocabulary for a particular metadata field. It will also allow the community to identify potential gaps and encourage development of new vocabularies.

Table 3: List of example vocabularies

Name	Description	Purpose and scope
GEMET	Thematic and multi-lingual	Very high level of all types of subjects. Appears restricted to common terms.
ISO Country Codes	2 and 3 letter country codes	ISO 3166 is the accepted International Standard
GCMD Science Keywords	Five-level broad classification on earth science data	All sciences. Limited to broad biological classification terms.
NBII Biocomplexity Thesaurus	Thematic and multi-lingual	All biological sciences.

7. Conclusion

Following the conclusions of an earlier working group that provided a set of general recommendations on a strategy for incorporating metadata as a core component of the GBIF architecture [GBIF08], the GBIF Metadata Implementation Framework Task Group was convened

to advise on the practical design of the metadata catalog system. Having adopted an expansive definition of “primary biodiversity data” as any measured value or set of values that pertain to an organism, the scope of the GBIF metadata catalog was set to cover a wide range of biodiversity data. Based on this requirement, the task group provided recommendation on metadata specifications, metadata catalog and network systems, metadata editors, controlled vocabularies and the alignment of GBIF’s efforts with other major initiatives involved in metadata projects and activities. This document reflects the consensus reached by the task group members and can serve as the basis for further comments from the wider GBIF community.

8. Bibliography

DataONE. 2009. Data Observation Network for Earth. http://dataone.org/

GBIF08. Metadata Requirements for Datasets delivered via the Global Biodiversity Information Facility (GBIF) Network. http://www2.gbif.org/GBIF-metadata-strategy_v.06.pdf

GBIF-EML08. Developing a Metadata Profile for GBIF based on Ecological Metadata Language Document v. 01, 27 June 2008. http://wiki.gbif.org/dadiwiki/wikka.php?wakka=FilesUpload/files.xml&action=download&file=metadata-profile-development.pdf

GCMD. Global Change Master Directory http://gcmd.nasa.gov

GEMET. GEneral Multilingual Environmental Thesaurus http://www.eionet.europa.eu/gemet/

ISO. Country codes - http://www.iso.org/iso/country_codes.htm

Jones, M B., C. Berkley, J. Bojilova, M. Schildhauer. 2001. Managing Scientific Metadata. IEEE Internet Computing 5 (5): 59-68.

Jones M B, Schildhauer M, Reichman O J, and Bowers S. 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics. 2006. 37:519–544.

Knowledge Network for Biocomplexity (KNB). http://knb.ecoinformatics.org/

Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG. 1997. Non-geospatial metadata for the ecological sciences. Ecol. Appl. 7:330—42.

Vanderbilt, K.L., Blankman, D., Guo, X., He, H., Li, J., Lin, C., Lu, S. L., Ko, B., Ogawa, A., Ó Tuama, É., Schentz, H., Wen, S., and van der Werf, B. 2008. Building an information management system for global data sharing: a strategy for the International Long Term Ecological Research (ILTER) Network. Pages 156--165 in Gries C. and Jones M.B. 2008 (editors). Proceedings of Environmental Information Management Conference 2008.

9. Appendix 1: Metadata editor comparison matrix

	GCMD DIF Author	Morpho	Mercury Editor	GeoNetwork MEST	Metavist	MERMAid	SMMS Intergraph	TkME	EU Portal INSPIRE Editor	Arc Catalog	Metacat Registry	IPT Metadata
Tool version	2.4.0	1.7.0	4.7.5	2.4.0	2005	1.2	5.1.13	2.9.9	1.07 build719	9.3	1.9.1	1.0rc1
XML input from other editors/sources	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Ability to edit entity-attribute information	No	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	No	No
Support auto-capture of metadata elements from the data	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	Yes
(1)Support multiple schemas.	No (DIF), but external	No (EML), but external	Yes (FGDC, Dublin Core,	Yes (FGDC, Dublin Core,	No (FGDC-BDP)	Yes (FGDC: includes Biologic	No (FGDC-BDP)	No (FGDC)	Yes (ISO 19115, ISO	Yes (FGDC, ISO 19115,	Yes EML-based, but can	Yes (Darwin Core; EML,

Comment [MBJ1]: Really supports all of these, or is it only a subset of the data these specs provide? Is this native support, or support after conversion to an internal model?

Comment [MBJ2]: Really? Does it support all the specs fully, or just subsets? Native support for each, or after conversion to native model?

	conversion via XSLT	conversion via XSLT	Darwin Core, Z39.50, ISO 19115, EML)	ISO 19139)		al, Shoreline, Remote Sensing; EML)			19119)	ISO 19139)	convert with XSLT to BDP	TAPIR)
(2)Provide validation service to those schemas	Yes	When uploading to Metacat		Yes	Yes	Yes		No	Yes	Yes	Yes	No
Capacity to validate as you edit	Yes	Yes	No	No	Yes (when opening)	Yes	No	No	Yes	No	Yes	No
Allow partially edited records to be kept for later edits	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No, but automatic metadata creation can be turned off	Yes	not applicable
Fields such as creator, contact etc can be easily replicated to reduce	Yes (via contact lookup)	Yes	Yes	No	No	No	Yes	No	No	No	Yes	No

effort												
Copy from existing records to reduce editing effort.	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes	No
Create author/node-specific templates including validation rules such as spatial extents	Yes	No	No	No	No	Yes	Yes	No	No	No	Yes	No
Interfaces must be well designed for the audiences it targets	web app	desktop app	framed web app	web app	desktop app	framed web app	desktop app	desktop app	web app	desktop app	web app	web app
Ability to have control access to records include	Yes (Public vs. Private)	Yes (Full role-based access control	No	Yes	No	Yes	No	No	No	Yes	Yes	Yes

the ability for the editor to specify access to individuals, groups, or public		across institutions)										
Metadata editors should support internationalization of the user interface and underlying software components	Yes	Yes (English and Chinese versions (v1.6.1).	No	Yes	No		No	Yes (Spanish, Indonesian, and French versions)	Yes	No	Yes, possible to be translated into other languages	Yes
Metadata editors should support input of metadata in multiple natural	Yes	Yes	Yes	Yes	Yes		No	No	Yes	Not complete	Yes	Yes

Comment [MBJ3]: Melanie says yes – check to find out which languages currently exist

Comment [MBJ4]: Find out which languages currently exist

Comment [MBJ5]: Which languages now?

languages												
Off-line editing	No	Yes	No	No	Yes	No	Yes	Yes	No	Yes	No	No
License	neither	GNU GPL		GNU GPL			commercial	open source		proprietary commercial	GNU GPL	Apache License 2.0
Shippable or a service base tool	service based	shippable	service based	both	shippable	service based	shippable	shippable	service based	shippable	shippable	both

10. Appendix 2: Metadata catalog software comparison matrix

	Metacat 1.9.1	Mercury	GeoNetwork	GCMD MD
Version examined	1.9.1		2.4.1	9.8.1
Implementation language	Java	Java	Java	
Most recent publicly downloadable release	1.9.1	None found	2.4.1	None found
Is redistributable under an OSI-certified open source license	Yes (GPL)	No	Yes (GPL)	No
Supports replicating metadata to other nodes when record changes occur.	Yes	No	No	Yes
Provides a harvesting interface that exposes metadata via their unique identifiers.	Yes	Yes	Yes	Yes
Provides a 'virtual portal' that uses the regional nodes for failover and load-balancing	No	No	No	Partial (local load balancing and failure recovery)

Comment [MBJ6]: Need to verify that it only supports harvesting

Comment [MBJ7]: Need to clarify if this means they support replication in addition to harvesting. What replication protocol do they use?

Provides a registry of other nodes and their relevant replication endpoints	Yes	Yes (for harvest list)	Yes (for harvest list)	
Supports multiple metadata models natively without code changes.	Yes (Can store, retrieve, and perform structured search on all fields from any metadata specification)	Partial (Doesn't store full metadata record in original format; extraction code must be updated for each metadata standard)	Yes (ISO19139, FGDC and Dublin core)	Partial (Stores metadata file from any specification for retrieval; extraction using XSLT for searching)
Can return the original contributed metadata object	Yes	No	Yes	Yes
Requires unique versioning of metadata and data objects using globally unique identifiers to differentiate revisions	Yes (Metacat Identifier, LSID)	Not required (but supports DOIs)		Yes
Supports replication and harvesting of metadata (and data) from providers	Yes/Yes (Metacat, OAI-PMH)	No/Yes	No/Yes (GeoNetwork, WebDAV, OAI-PMH)	Yes/Yes
Supports search and discovery	Yes	Yes	Yes	Yes
Supports metadata in XML serializations	Yes	Yes	Yes	Yes
Has programming API for 3rd party metadata editors to use to insert and update records	Yes	No	No	No
Supports international language documents and queries	Yes	No?	Yes	Yes
Supports conversion from one metadata model to another and ability to return these alternate formats on request	Yes			Yes
Supports sorting of search results	Yes	Yes	Yes	Yes
Supports logical queries and filters on individual metadata fields from	Yes	Yes	Yes	Only on DIF

Comment [MBJ8]: Need to verify

Comment [MBJ9]: Need to verify

Comment [MBJ10]: Need to clarify which GUIDs are required

Comment [MBJ11]: Need to clarify capabilities

Comment [MBJ12]: Need to verify -- other APIs exist, but I couldn't find a metadata insert/update xml service, although there is an MEF import API call

Comment [MBJ13]: Need to verify. Couldn't find references to this API in the documentation.

multiple standards				
Collects access log statistics on all CRUD operations for reporting	Yes	Yes		Yes
Maintains summary of holdings	Yes	Yes		Yes
Enforces access control restrictions on non-public metadata for read and write by metadata editors	Yes (Full role-based access control; LDAP support)	No	Yes (Full role-based access control; LDAP support; Shibboleth support)	Yes (Public/Private)
Registers with GBRDS node registry, and possibly other registries (e.g., GEOSS)	No (but yes for EarthGrid)	No	No	No (but yes for GEOSS)
Exposes one or more standard query APIs for programmatic access (e.g., OGC WMS, EarthGrid query protocol, SRU/SRW, XQuery/XPath)	Yes (EarthGrid,, XPath, OGC WMS)		Yes (custom XML web service API)	Yes (OGC WMS)
Provides bookmarkable queries	No	Yes		No
Provides subscription services to new metadata records (e.g., RSS feed on query)	No	Yes		Yes
Provides thesaurus services for searching and access by other editors/clients	Partial (ontology search in prototype)		Yes	Searching using GCMD thesaurus interface
Exposes metadata records to other search engines (e.g., provide site index for Google, Yahoo)	Yes (Google)	Yes		Yes
Provides attribution and branding for original metadata providers	Yes	Yes		Yes

Comment [MBJ14]: This feature is vague – unclear what it really means

Comment [MBJ15]: The extent of branding varies significantly among the catalogs. What are we really looking for here?