Initial/Basic Standards and Protocols for IABIN

Prepared by Boris Ramirez, IABIN Secretariat

1 PURPOSE

This document makes recommendations on a basic set of data standards and communications protocols that would enable the envisaged connectivity and interoperability of IABIN. These standards are an initial set and IABIN´s protocols and standards will always be subject to evolution and constant review by IABIN Technical Working Group.

2 BACKGROUND

IABIN has adopted 11 guiding principles for interoperability formats, standards and protocols:

1. Seamless access to all types of IABIN data and information regardless of where it resides and interoperable with both CBD-CHM, GBIF and other networks.
2. Open, widely supported, non-proprietary standards;
3. Compatibility with emerging standards of key regional, global and national biological information networks;
4. Minimization of technology restrictions imposed by the network architecture;
5. Phased, incremental development;
6. Scalability, so that standards will be usable and applicable at different network scales: global, regional and nacional.
7. Inclusion (e.g. facilitate local-language queries) in the design of applications;
8. Expertise and capabilities are shared throughout the network;
9. Respect for Intellectual Property Rights, Traditional Knowledge Rights and rules for access and benefit sharing of Genetic Resources in accordance with the CBD principles and guidelines, and national legislations.
10. Future extensibility and backward compatibility.
11. Minimization of cost while ensuring reliable user services

3 JUSTIFICATION AND PROPOSALS

System architecture adopted by IABIN will base on flexible, widely support software standards in web-based software development, and have an inherent capability to support the requirements for a distributed system. IABIN System architecture will be applied to the IABIN Gateway and to the design of Project Network accessed by the Gateway. The intent is to minimize the number of technology restrictions that are imposed on data providers while establishing a limited number of standards that ensure interoperability.

The following nine areas have been suggested in IABIN documents which areas where IABIN have to adopt a standard and now need to be formally adopted.

1. System Architectures
2. Data Transport
3. Presentation Language
4. Data Encoding
5. System Access Protocols
6. System Interface Descriptions
7. Registry Services
8. Metadata Formats
9. Geospatial Interoperability
10. Document Formats
11. Graphic Formats

3.1 System Architectures

IABIN is envisioned as a distributed system in which the IABIN partners play a key role in the development and maintenance of the information that constitutes the network, while the IABIN Secretariat plays the role of facilitator. In most cases, data providers will store and maintain source data, and be responsible for releasing only data that they wish to. Architectures supported by IABIN should be based on flexible, open software standards in web-based software development, and have an inherent capability to support the requirements for a distributed system. Moreover, the IABIN endorsed system architectures should also be designed to support component-based software development methodologies that allow different groups to develop system components independently. The following are the architectural alternatives:

z39.50

z39.50 is a mature information retrieval standard which has been particularly popular within the library community. It has particular relevance to biodiversity informatics since it has been the foundation for the development of The Species Analyst project.

The complexity and relative obscurity of the protocol limits its suitability as a foundation for use by IABIN. The Species Analyst network is also moving away from it in the near future. Continuing activities to fuse z39.50 with XML (see page 3, Data Encoding, for explanation on XML) may however lead to its more natural inclusion within a web services architecture.

Web Services

A Web Services model is currently the architecture of choice for development of truly global networks of heterogeneous data providers. It offers the greatest degree of technological separation between different providers because all

communications between systems are based on XML document exchanges. This provides a highly flexible model with good support for multiple languages. This model is being rapidly adopted as the standard both within the general e-Business and e-Commerce communities, but also within the biological informatics community. As an example, GBIF has adopted this architecture to support its global network.

Proposal

Given the system architectures options available, IABIN adopts the Web Services network architecture, but IABIN will also provide support to z39.50 based network architectures.

3.2 Data Transport

HTTP over TCP-IP

HyperText Transfer Protocol (HTTP) is the standard protocol that enables users with Web browsers to access HTML documents and external media. Transmission Control Protocol/ Internet Protocol (TCP/IP) is the ISO standardized suite of network protocols that enables information systems to link to other information systems on the Internet, regardless of their computer platform. TCP and IP are two software communication standards used to allow multiple computers to talk to each other in an error-free fashion.

As the foundational technologies of the Internet, TCP/IP and HTTP are the only logical choice for a globally accessible data transport.

Proposal

IABIN adopts the HTTP over TCP/IP communication standard for data transport.

3.3 Presentation Language

HyperText Markup Language (HTML)

HyperText Markup Language (HTML), is a markup language used to create documents for World Wide Web applications. HTML has evolved to emphasize design and appearance rather than the representation of document structure and data elements. HTML 4 is the presentation language most widely used to develop web pages.

Extensible HyperText Markup Language (XHMTL)

Extensible HyperText Markup Language (XHTML), is a extension of HTML 4 using rules of XML (see next section). It is a hybrid between HTML and

XML. This language allows describing data in XML format. Not all browser support XML so XHTML provides and intermediary solution and can be interpreted by XML and HTML browsers.

Proposal

IABIN adopts HTML 4 and XHTML as its presentation language**.** The use of metatags within the HTML document is recommended to support with search engines visibility and display.

3.4 Data Encoding

eXtensible Markup Language (XML)

The eXtensible Markup Language (XML) provides a clear foundation for improved interoperability and data transfer within the Web Services architecture. XML is a platform independent language for exchanging and validating data between heterogeneous systems. It provides good support for multilingual data exchange and is well-supported by freely-available cross-platform tools in a wide variety of programming languages. Direct support for XML is appearing steadily in database management software and other key tools. GBIF has also adopted XML as its standard for data encoding.

Proposal

IABIN adopts XML as its standard for data encoding.

3.5 System Access Protocols

Simple Object Access Protocol (SOAP)

System access protocols are used to develop interfaces between systems that need to exchange data, instructions, requests, or responses. The Simple Object Access Protocol (SOAP) is an open standard with wide acceptance in the software development community. SOAP is an XML-based lightweight protocol designed for exchange of information in a decentralized, distributed environment and is ideal for exchanging messages between different computer systems. SOAP can be used to implement cross-platform messages between different systems. Such messages typically form a request for the target system to perform some task. SOAP manages the definition and exchange of parameters as part of the request. The protocol also handles the return of response data back to the requestor. SOAP can potentially be used in combination with a variety of other protocols; however, it is typically used in combination with HTTP and the HTTP Extension Framework.

Distributed Generic Information (DiGIR)

DiGIR (Distributed Generic Information) is an access protocol initiative adopted by the TDWG/CODATA Biological Collections Data Subgroup, and is managed as an open source project (http://digir.sourceforge.net/). DiGIR seeks to use XML documents to define and handle federated search requests based on any chosen data exchange schema. It is in use today by projects such as MaNIS (Mammal Network Information System) to exchange specimen and observation data in the Darwin Core metadata format (see below for an explanation on this under Metadata formats). Although its roots are in the biological informatics, the DiGIR protocol can be used with other data formats to develop networks for other applications.

SOAP is expected to continue achieving broad acceptance within the general software development community. DiGIR is expected to develop a significant following within the biological informatics community and may provide advantages for biological applications. GBIF is supporting both protocols for the development of its network. GBIF has also adopted DiGIR as its standard for its Network and GBIF country nodes.

Proposal

IABIN adopts both the SOAP and DiGIR system access protocols for its distributed networks.

3.6 System Interface Descriptions

Web Services Description Language (WSDL)

Before an external system can utilize a web service, it requires information on how to access and communicate with that service. The Web Services Description Language (WSDL) addresses this need by defining an XML grammar for describing network services as collections of communication endpoints capable of exchanging messages. WSDL allows Web Services interface descriptions to be stored as XML documents for distributed systems and serve as a recipe for automating the details involved in applications communication. Tools or systems can access the XML documents and subsequently understand how to access and utilize the service. GBIF supports WSDL for web service interface descriptions.

Proposal

IABIN adopts WSDL as its system interface standard.

3.7 Registry Services

Universal Description, Discovery and Integration (UDDI)

Registry Services provide a central point to allow users to locate web service providers. Universal Description, Discovery and Integration (UDDI) is one of the more widely accepted Registry Services and is supported by a broad array of software development tools. UDDI creates a standard interoperable platform that enables users and applications to quickly, easily, and dynamically find and use Web services over the Internet. UDDI also allows operational registries to be maintained for different purposes in different contexts. UDDI is a cross-industry effort driven by major platform and software providers, as well as marketplace operators and e-Business leaders within the OASIS standards consortium. GBIF and others networks have adopted UDDI.

Proposal

IABIN adopts UDDI as its Registry Services standard. IABIN will maintain and create its own registry services and each data provider could be registry into IABIN service. This standard will be using for IABIN for exchange information about how to access the data in each node or data supplier.

3.8 Metadata Formats

Metadata means, literally, "data about data." Metadata includes data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation. The initial metadata formats for IABIN are the following:

Metadata format for Bibliographic Data

Dublin Core – Dublin Core is a standard which defines a basic set of metadata elements which may be used to describe digital resources. GBIF also uses Dublin Core. This standard will be using for create bibliographic metadata: publications, images.

Metadata format for Specimen Collections and Observations

Darwin Core – The Darwin Core (DwC) is a metadata profile describing the minimum set of standards for search and retrieval of natural history collections and observation databases. It includes only the core data elements that are likely to be available for the vast majority of specimen and observation records. This standard is utilized within both the Species Analyst and REMIB networks, among others. DwC is also a GBIF approved data standard.

ABCD Schema – The Access to Biological Collections Data (ABCD) Schema is the product of a joint TDWG and CODATA initiative to develop a standard

for distributed data retrieval from specimen collection databases. The schema supports data exchange for all kingdoms, and for both specimen and observation records. The ABCD Schema is a GBIF approved data standard.

Metadata format for Spatial Data

CSDGM (ISO 19115) – The Content Standard for Digital Geospatial Metadata (CSDGM) was developed by the Federal Geographic Data Committee (FDGC) to provide a common set of terminology and definitions for the documentation of digital geospatial data. The standard was developed by the Federal Geographic Data Committee, an agency in the United States representing a 19-member interagency committee composed of representatives from the Executive Office of the President, Cabinet-level and independent agencies. The FGDC is developing the National Spatial Data Infrastructure (NSDI) in cooperation with organizations from state, local and tribal governments, the academic community, and the private sector. The NSDI encompasses policies, standards, and procedures for organizations to cooperatively produce and share geographic data. The FDGC standard is utilized extensively throughout the Western Hemisphere.

Metadata format for Biological Spatial Data

CSDGM with Bio Profile (NBII) - The purpose of this standard is to provide a user-defined or theme-specific profile of the FGDC Content Standard for Digital Geospatial Metadata to increase its utility for documenting biological resources data and information. This standard supports increased access to and use of biological data among users on a national (and international) basis. It also helps to broaden the understanding and implementation of the FGDC metadata content standard within the biological resources community. This standard also serves as the metadata content standard for the National Biological Information Infrastructure (NBII) and the IABIN catalogue services. More information on this metadata standard is available at: metadata.nbii.gov

Other Data Themes

It is recognized that specific metadata standards will be required for other biological themes supported by IABIN (e.g., species, protected areas, neotropical flora, etc.). In many of these themes, predominant or emerging standards do not exist. As one of its primary functions IABIN will facilitate the development of new standards or adoption of existing standards through consensus building processes that involve the major players within the theme of interest. The first step in this process is to have the IABIN stakeholders identify the priority themes of interest and pertinent players.

Proposal

IABIN adopts the metadata formats above as its standards

3.9 Geospatial Interoperability

The Open GIS Consortium (OGC) is an international industry consortium of 258 companies, government agencies and universities participating in a consensus process to develop publicly available geospatial interoperability specifications. Open interfaces and protocols defined by OpenGIS® Specifications support interoperable solutions that "geo-enable" the Web, wireless and location-based services, and mainstream Information Technology, and empower technology developers to make complex spatial information and services accessible and useful with all kinds of applications.

The most mature and prevalent geospatial interoperability standards are the OGC Web Map Services (WMS), Web Feature Services (WFS), Web Coverage Services (WCS) and the Catalog Web Services (CWS) standards.

Proposal

IABIN adopts the WMS, WFS, WCS, and CWS standards and commits to evaluating the other emerging standards as developed by the OGC.

3.10 Document Formats

The recommended electronic formats for document exchange within the IABIN network are: HTML, PDF, and ASCII (plain text).

3.11 Graphic Formats

The recommended electronic formats for graphic exchange within the IABIN network are the formats recommended with use in a web browser: PNG, JPEG, GIF, SVG and WebCGM. The imagines will be protected by watermarks technologic.

Summary of IABIN Basic Standards and Protocols

Part of IABIN Data Architecture	Standard or Protocol Adopted
System Architectures	Web Services with support for Z39.50
Data Transport	HTTP over TCP-IP
Presentation Language	HTML 4 and XHTML
Data Encoding	XML
System Access Protocols	SOAP, DiGIR
System Interface Descriptions	WSDL
Registry Services	UDDI
Metadata:	
o For Bibliographic Data	Dublin Core
o For Specimen Collections and Observations	Darwin Core and ABCD Schema
o For Spatial Data	CSDGM (ISO 19115)
o For Biological Spatial Data	CSDGM with Bio Profile (NBII)
Geospatial Interoperability	WMS, WFS, WCS and CWS and emerging standards develop by Open GIS Consortium (OGC)
Document formats	HTML, PDF, and ASCII
Graphic Formats	PNG, JPEG, GIF, SVG, WebCGM