The Concept and Framework of Biodiversity e-Science Infrastructure in China

Zhe-Ping Xu*, Jin-Zhong Cui, Hai-Ning Qin, Ke-Ping Ma
Institute of Botany, the Chinese Academy of Sciences, Beijing, China

Abstract: China has rich biodiversity data, such as specimens, images, scientific literatures, field survey data, lab data. However, for a long time, many data are neither available nor accessible. In recent few years, few groups and institutions in China have began working on collection, collation and dissemination of biodiversity information. After the cooperation with Species 2000, EOL (Encyclopedia of Life), BHL (Biodiversity Heritage Library) and other related international projects, biodiversity information in China has been organized and shared in a good way. Also in this time, with the help of biodiversity informatics, the concept of biodiversity e-Science infrastructure has come to appear. In this paper, we suggest a biodiversity e-Science infrastructure based on a Service-Oriented Architect (SOA) and OGC ISO 19119 service standards. Related data, tool, metadata standards, service, annotation, community and application will be integrated into one platform, which may play an important role in the development and application of biodiversity research in the future of China.

I. INTRODUCTION

China has rich biodiversity resources, both living and non-living. The research work on biodiversity is very important for conservation, agriculture, fisheries, industry and forestry. However, the status of the biodiversity in this country is more and more dangerous resulted from population explosion, urbanization, invasive alien species and habitat fragmentation. To prevent this situation, we need more research work in biodiversity subject.

As a data-driven subject, biodiversity research needs massive high-quality data from scattered data providers all over the country(Xu et al. 1999) and the world. Recently, some articles and reports on e-Science infrastructure initiatives recognize the shortage in qualified professionals to manage the increasing stores of scientific data. After the cooperation between Chinese National Committee for Diversitas (CNC Diversitas) and EOL (Encyclopedia of Life, http://www.eol.org) and BHL (Biodiversity Heritage Library in US, http://www.biodiversitylibrary.org/) in 2009, many metadata standards, techniques and tools in biodiversity informatics have been introduced into China. Furthermore, the biodiversity e-Science infrastructure also comes into the talk. The first professional laboratory, Key Laboratory of Biodiversity Informatics, Institute of Botany, the Chinese Academy of Sciences, was launched on February 3rd.

Biodiversity Informatics is a young and rapidly growing field and is the basic support for the construction of biodiversity e-Science infrastructure. There are so many biodiversity database systems, such as GBIF(Global Biodiversity Information Facility), Species 2000, uBio and so on. The International Conference on Biodiversity Informatics held in June 2009 was a milestone for the development of this discipline, which indicated the change from traditional database construction to the construction of Biodiversity e-Science Infrastructure. In recent years, more and more biodiversity e-Science infrastructures appear like LifeWatch (http://www.lifewatch.eu/, Europe), NBII(National Biological Information Infrastructure, http://www.nbii.gov/, USGS), DataONE (Data Observation Network for Earth , http://www.dataone.org/, USGS) and ALA(Atlas of Living Australia, http://www.ala.gov.au, Australia). Their goals are very clear: “e-Science and technology infrastructure for biodiversity data and observatories”(LifeWatch). China should make more effort to construct the biodiversity e-Science infrastructure as the some related database has been finished and the problems are severe.

II. DATABASE BACKGROUND

A. Summary

In recent few years, few groups and institutions in China have began working on collection, collation and dissemination of biodiversity information. You can find almost all information about the biodiversity in China, such as name, images,

* Corresponding author. Tel.: +86-10-62836847
E-mail address: xuzp@ibcas.ac.cn (Zhe-Ping Xu).

specimens, literature, observatory, oceanic data and the community. Several useful websites have been listed in table 1. Some of them are managed by local people, while others are the cooperative achievement with international projects. All these data and the community will be the backbone for the design, preparation and construction of China biodiversity e-Science infrastructure.

TABLE 1 MAIN WEBSITES ABOUT BIODIVERSITY INFORMATION IN CHINA

Subject	The Title of the Website	URL
I. Terrestrial Data		
Names	Species 2000 China Node	http://www.sp2000.cn/
	Catalogue of Life: Higher Plants in China	http://www.cnpc.ac.cn/
	Animal Information Network of China	http://www.animal.net.cn/
	Biological Data Center for Basic Science	http://www.bioinfo.cn/
	Flora of China	http://hua.huh.harvard.edu/china/
Specimen	Chinese Virtual Herbarium for the Plant(CVH)	http://www.cvh.org.cn/
	Specimen from Nature Reserve	http://www.papc.cn/
	National Mineral & Fossil Resource	http://www.nimrf.net.cn/
	National Digital Animal Museum	http://museum.ioz.ac.cn/
	Educational Specimen Resource Center	http://mnh.scu.edu.cn/
	Resource-sharing Platform of Polar Samples	http://birds.chinare.org.cn/
	Specimen from China in the world	http://data.gbif.org/countries/CN
Literature	BHL China Node	http://www.bhl-china.org.cn/
	Chinese Biological Abstracts	http://www.cba.ac.cn/
Image	Chinese Field Herbarium	http://www.cfh.ac.cn/
	Photo Bank for China Plant	http://www.plantphoto.cn/
	Chinese Virtual Botanical Garden	http://www.cvbg.cn/
Community	EOL (Encyclopedia of Life) China Node	http://www.especies.ac.cn/
	Biology Show	http://www.bbioo.com/
	Plant & Animal Forum in China	http://www.planta.cn/
Observatory	Chinese Ecosystem Research Network(CERN)	http://www.cern.ac.cn/
II. Oceanic Data		
Oceanic Data	China Oceanic Information Network	http://www.coi.gov.cn/
	South China Sea Scientific Data Service Platform	http://www.ocdb.csdb.cn/

B. Name (Catalogue)

The name database is always the backbone for biodiversity e-Science infrastructure. Some are locally maintained by single institute, while the Species 2000 China Node is a more wonderful project. For server years, this project has brought a series of databases together, each of them covering a group or groups of organisms in China. The checklist will be freely provided for users by a website supporting the online query to the dynamic checklist and annual checklist, and a CD providing the annual checklist. In 2009 Annual Checklist of Catalogue of Life China, the number of accepted species names is: (1) Plants: Lichen(2571), Fern(2267), Gymnosperm (244) , Angiosperm (29583); (2) Animals: Amphibian(346), Reptile(403), Fish(3233), Bird(1269), Mammal(564), Spider(3300), Encyrtid(405).

C. Specimen

Specimens are very important to do research in biodiversity area (Graham et al. 2004). The national project, “Specimen

Resource Platform", funded by the Ministry of Science and Technology, consists of 6 sub-platforms: animal specimen, plant specimens, educational specimens, the specimens from nature reserve, minerals & fossils and the polar samples. There are in total 1,863 persons from 137 institutions and universities participating in the project. This project digitized 9.10 million kinds of specimens, took photos for 6 million specimens and 1.5 million photos for living field plant..

D. Literature

For centuries, biodiversity surveys and research has resulted in lots of literatures that is scattered in various libraries in China. Digital library would help greatly in building such abstract, full text literature bank and also metadata about these sources. In 2009, Chinese National Committee for Diversitas (CNC Diversitas) signed up a MOU with BHL and formed BHL China working group. Now, this group has bought a copy of ABBYY FineReader SDK and will obtain a set of Scribe equipment from Internet Archive (http://www.archive.org). After 6 months from November 2009, BHL China Node has collected about 1 million page-name (scientific name or Chinese common name) records from about 500 books. These data will give a solid support for CVH, EOL China Node and other research work.

E. Images (Photos & Drawings)

"A picture paints a thousand words". Image data is very important for the identification. Furthermore, as followed the BasisOfRecord element of Darwin Core standard, image data can be integrated into occurrence dataset. If so, it will largely facilitate the management of massive image data, either for living photo or drawings. Just for plant images, we have collected 1.3 million photos, about half of which have been identified to species level. Some can be mapped into Google Map or Google Earth with manual cleaning and GPS handset.

F. Spatial Data

With the spatial data extracted from specimen and image, we can apply these data in more subjects. But data cleaning is a hard work. During 2009, we cleaned about 1.8 million specimens, which have been georeferenced into county-level. In the future, remote sensing and sensor network will introduce more spatial data.

G. Field Observatory Data

The Chinese Ecosystem Research Network (CERN), the member of the International Long Term Ecosystem Research Network (ILTER) and Global Terrestrial Observation System (GTOS), was established in 1988. Now, it has become one of the largest networks in the world, consisting of 40 field stations, 5 sub-centers(Water, Soil, Atmosphere, Biology Aquatic-Eco) and 1 synthesis center (Fu et al., 2010). For tens of years, the biological sub-center has collected massive biodiversity-related data. Recently, the government is introducing remote sensing in monitoring the site and deploying many leading sensor network in some selected stations as pilot projects (Yang et al., 2008). It will make an important role in collecting and managing biodiversity data in the field stations.

H. Community

Although the community is not so developed in current China, all databases above are completed by virtual organization (VO). There are so many resources we can use from the community, including experts, datasets, groups, E-learning materials and etc. In recent years, so many biodiversity communities, like LifeDesks and Scratchpads, are based on Drupal, a smart but powerful Content Management System. Now, the CVH and BHL China Node are also built on this CMS. We can extend the core function of the system as we think.

III. THE CONCEPT AND ARCHITECTURE

Some problems have occurred during the construction of our previous database, such as the linkage among different database, the exchange standard and protocol, the visual analysis and domain modeling etc. Therefore, we need a new design and architecture for biodiversity e-Science infrastructure in national level. Ideally, we can integrate data providers, tools, ontology, data consumers, services and functions in the e-Science infrastructure (Fig. 1.).

There are three layers in the e-Science infrastructure: resource layer, service layer and function layer. Here are the descriptions about these layers:

(1) Resource Layer. The main part of this layer is the dataset (names, specimens, literature, images, even comments from users) from library, herbarium, museum, laboratory, field observatory, sensor network, website, individual users. Besides, it contains other resources like tools, ontology, documents and compute capabilities. Kinds of standards have been used in this

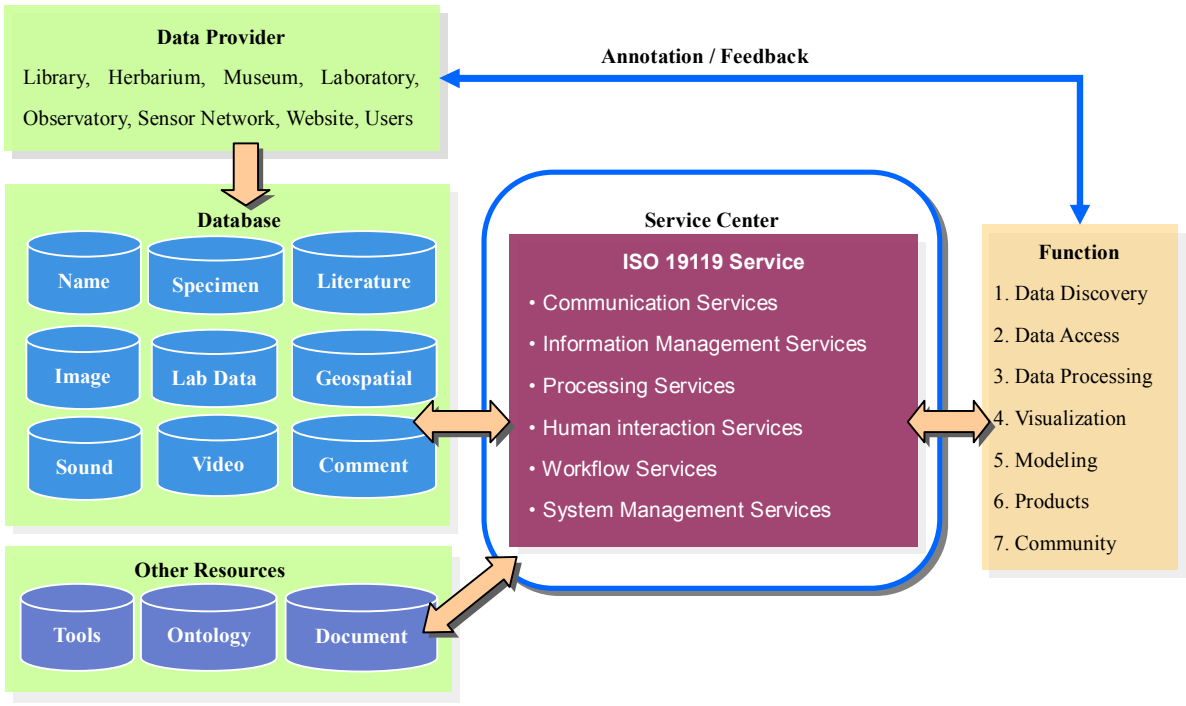

Fig. 1 The Architecture of Biodiversity e-Science Infrastructure in China

layer, such as Darwin Core, EML, Dublin Core, EndNote, IPTC(International Press Telecommunications Council), KML and GML. Some protocols like z39.50 and TAPIR(TDWG Access Protocol for Information Retrieval) are also introduced for the exchange of book catalogues and specimen data.

(2) Service Layer. This layer is based on OGC ISO 19119 service standard, consisting 6 categories: communication services, information management services, processing services, human interaction services, workflow services and system management services.

(3)Function Layer. In this layer, data consumer can get following functions after service layer: data discovery, data access, data processing, visualization, modeling, products and community. More, the users can also make some annotations and feedback to data providers, which make an interaction between the start point and end point of the data.

Here, we take new version of CVH for an example to demonstrate the initial development of our construction for the e-Science infrastructure. In the system, LSID(Life Science Identifiers) has been introduced to integrate all related data for each species in a single page, such as name, images, specimens, literatures, digital botanical garden and related international data source (Fig. 2).You can also know which institutes have specimen data for this species in this page. Darwin Core standard has been introduced to implement the single portal to search for plant fossil, photos and preserved specimen from kinds of user input parameters. Dublin Core will also be referred as the backbone for the implementation of the search engine on this infrastructure. A CMS based on Drupal with hundreds of registered users is aiming to interactively maintain kinds of biodiversity data.

Fig.2 LSID Page in new CVH

IV. SERVICE

As described above, the service of entire China biodiversity e-Science infrastructure is a SOA(Service-Oriented Architect) and based on ISO 19119 service standard. Table 2 below shows mappings between the ISO 19119 service and existing service in current service from related database. The first column describes to the ISO 19119 service types. The second column lists specified service of e-Science infrastructure. The third column is service status in current biodiversity platform in China(○ means no consideration, ◎ means partly finished).

TABLE 2 SERVICE IN THE E-SCIENCE INFRASTRUCTURE

ISO 19119 Service	Service of e-Science Infrastructure	Status
Human Interaction Services	Portrayal Services	◎
	Interaction Services	◎
	Personalization Services	◎
Information Management Services	Data Access Service	◎
	Annotation Services	○
	Identification Services	◎
	Discovery Services	◎
	User Management Services	◎
Workflow Service	Orchestration Services	○
Processing Services:	Spatial Processing Services	○
	Temporal Processing Services	○
	Thematic Processing Services	○
	Metadata Services	○
	Integration Services	◎
	Taxonomic Processing Services	◎
Communication Services	Encoding Services	◎
	Transformation services	○
	Transfer services	○
System Management Services	Monitoring Services	◎
	Quality Evaluation Services	◎
	Security Services	◎

Here are the descriptions of ISO 19119 and the status of the service from current China biodiversity information system.

(1) Human interaction services: the management of user interfaces, graphics, multimedia, and compound documents. Some services can be partly obtained from current information system of China biodiversity, but far from satisfied.

(2) Information management services: the management of the development, manipulation, and storage of metadata, conceptual schemas, and datasets.

(3) Workflow services: the support of specific tasks or work-related activities conducted by humans. The e-Science infrastructure will introduce Kepler tool in the future.

(4) Processing services: perform large-scale computations involving substantial amounts of data. This service can be divided into four sub-types: spatial processing services, thematic processing services, temporal processing services, metadata processing services. This is the most important service in current China biodiversity information system.

(5) Communication services: encode and transfer data across communications networks. It will be based on the construction of the professional community in the future.

(6) System management services: the management of system components, applications, networks, user accounts and user access privileges.

V. CONCLUSION

Although the problems exist and the concept and framework of Biodiversity e-Science Infrastructure in China is completely new, the existing databases indicate that the foundation is not so bad to implement the architecture and service. On the other hand, the concept and the framework Biodiversity e-Science Infrastructure is based on SOA and international ISO 19119 standards, the implementation and experience of which can be applied in other domains.

REFERENCES

[1] Xu, H.G., Gao, Z.N., Xue, D.Y., Wu, X.M.,1999. China National Biodiversity Information Query System. Journal of Environmental Management. 56(1),45-59.

[2] Chen, Z., Wang, H.G.,Wen Z.J., Wang, Y.H.,2007. Life sciences and biotechnology in China. Phil. Trans. R. Soc.362, 947–957.

[3] Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A. T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution. 19(9), 497-503.

[4] Yang, P., Yu, X. B., Zhuang, X. L, Niu. D., 2008. Present status and train of thought of future development of Chinese ecosystem research network(CERN) of CAS. Bulletin of the Chinese Academy of Sciences. 23(6), 555-562 (in Chinese).

[5] Fu, B.J., Li S.G.,Yu, X.B., Yang, P., Yu, G.R., Feng, R.G., Zhuang, X.L.,2010. Chinese ecosystem research network: Progress and perspectives. Ecol. Complex.doi:10.1016/j.ecocom.2010.02.007.