BioResource now !

Cherry blossom in Atami

February 2006

BioResource now! Vol 2.No.2 is here

Download the PDF version of this newsletter at
http://www.shigen.nig.ac.jp/shigen/news/news.jsp

Other information on bioresources is available at

NBRP (http://www.nbrp.jp/index.jsp)
SHIGEN (http://www.shigen.nig.ac.jp/indexja.htm)
WGR (http://shigen.lab.nig.ac.jp/wgr/)
JGR (http://shigen.lab.nig.ac.jp/wgr/jgr/jgrUrlList.jsp)

Information on Resource-related Events

- March 9, 2006
  NBRP symposium at the Tokyo International Forum "Frontline Researches on Bioresources and Life Science"

- March 14 and 15, 2006
  Joint Symposium on Life Science
  at the Tokyo International Exchange Center

- March 19–21,2006
  The Japanese Society of Plant Physiologists Annual Meeting and the NBRP Special Project
  at the University of Tsukuba
  - March 19–21: Panel Exhibition of Plant Resources and Posters
  - March 21: NBRP Symposium
  "Model Plant Resources and Technologies Supporting Plant Researches and Their Application to Crops"

- March 23, 2006:
  Joint Symposium on Life Science at the Senri Life Science Center

- May 6, 2006
  NBRP Symposium at Hotel Granvia Osaka "Post-Genome Research on Yeast and Bioresources"

- May 11–13, 2006
  The Japanese Association for Laboratory Animal Science Panel Exhibition
  at the International Conference Center Kobe
  "The Bioresources of Laboratory Animals: The Present Situation and Outlook" is in the planning stage

Detailed information is available at http://www.nbrp.jp/index.jsp

Introduction to Resource Center No. 6

Global Biodiversity Information Facility (GBIF)

In 2011, GBIF will provide 1 billion data of biodiversity information on 1.8 million species

Professor Hideaki Sugawara, Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics (NIG)

"It is the right initiative with the right goals at the right time"
"In our view, if it did not exist, it would need to be created"
(From an evaluation on GBIF by a 3rd party (2005))

History

GBIF was established based on the proposal of the Organization for Economic Cooperation and Development (OECD). Contrary to its name, the OECD also proposes policies on science and technology to its member countries. At the beginning of the 21st century, the OECD proposed two important policies on science and technology in the field of biotechnology. The first policy was based on the 1999 report by the OECD Megascience Forum Working Group on Biological Informatics and proposed the establishment of the GBIF and the Global Neuroinformatics Capability (GNC) .*

The second policy was based on the 2001 report by the Task Force on Biological Resource Centres (BRC) of the Working Party on Biotechnology and proposed the establishment of a Global BRC Network (GBRCN).**

*) http://www.gbif.org/GBIF_org/facility/OECD_Endorsement
**) http://www.wdcm.org/brc.pdf

The concept of GBIF was the first to be materialized among the proposals made by OECD. The reason for this may be that each nation realized the necessity to obtain comprehensive biodiversity information during the discussion on the Convention on Biological Diversity (CBD) proposed at the RIO Earth Summit in 1992. In fact, even after the proposal of the OECD was published, there was continued active participation by many countries, not just the USA, the country who proposed the establishment of the Working Group to the OECD in 1996, but also other countries such as the Netherlands, Australia, and Mexico. The GBIF was inaugurated in March 2001.

In addition to the Japan Science and Technology Agency and the National Institute for Environmental Studies who were already part of the Japan GBIF node, the National Institute of Genetics (NIG) was designated as the core facility in 2003 and collaborated with the University of Tokyo Graduate School of Arts and Science and the National Science Museum to carry out the undertakings of the Japan GBIF node.

The GBIF websites

http://www.gbif.org/

http://www.gbif.net/

http://bio.tokyo.jst.go.jp/GBIF/gbif/japanese/

Organization

Under the governing board and the executive committee, the GBIF has 3 standing committees which consist of the Science Committee, the Budget Committee, and the Node Managers Committee, and an ad-hoc committee which consists of the Membership Committee (Fig. 1). As a result of an international bid in 2001, the Secretariat as shown on the right side of Fig. 1 is hosted by the Zoological Museum, University of Copenhagen. Dr. James Edwards, who was part of the OECD Working Group, remains with the U.S. National Science Foundation while continuously leading GBIF as the Executive Secretary. There are four subcommittees under the Science Committee which are the DADI, DIGIT, ECAT, and OCB.

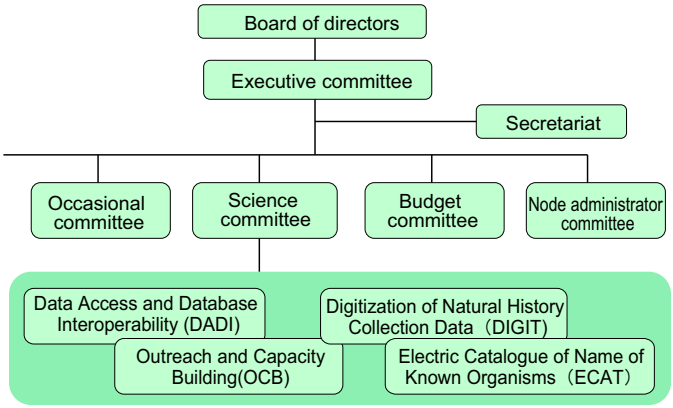

Fig. 1 Organization chart of GBIF

- DADI – Develop standards for data and metadata to answer complex questions involving many disparate types of data from many sources.
- DIGIT – Support the digitization of specimen data and observation data.
- ECAT – Develop a complete electronic listing of scientific names which are the key to all scientific literature and experimental results for species while working collaboratively with each domain.
- OCB – Provide software tools and training to bridge biodiversity information technology gaps and address scientific and technical collaboration in many areas, including repatriation of data and intellectual property rights.

The GBIF Governing Board meeting is held twice a year to make policy-level decisions. Participants of this meeting include 47 countries and economies and 29 non-governmental organizations who have signed the Memorandum of Understanding (MOU). The GBIF and its activities are supported by financial contributions from 26 countries which have the right to vote at the Governing Board meetings. The budget for the fiscal year 2006 is approximately 3 million euros, and a considerable share of this is being borne by Japan along with the USA. Aside from making financial contributions, each member of the Governing Board contributes by sharing biodiversity information and providing a data node. Therefore, it is essential to remember that most of the data are generated by the combined efforts of all the members.

Vision

Since February 2006, the GBIF data portal site (http://www.gbif.net/) allows an integrated search of 86.6 million data records on 98,300 species from 159 websites which are provided by many countries. After the first phase in 2003–2006, GBIF will enter its second phase in 2007 and one of the objectives is to provide one billion data records on the entire known species by the year 2011 (Refer to "Information resources by GBIF" in Fig. 2). In addition, the GBIF aims to establish an information architecture where GBIF information can be integrated with existing information resources such as literature, molecular biology, geographical, ecological, weather, and socioeconomic information (Refer to "Information resources" in Fig. 2). This information architecture will allow anyone from anywhere at anytime to combine information resources according to their needs and search, analyze, preserve, and utilize all kinds of organisms.

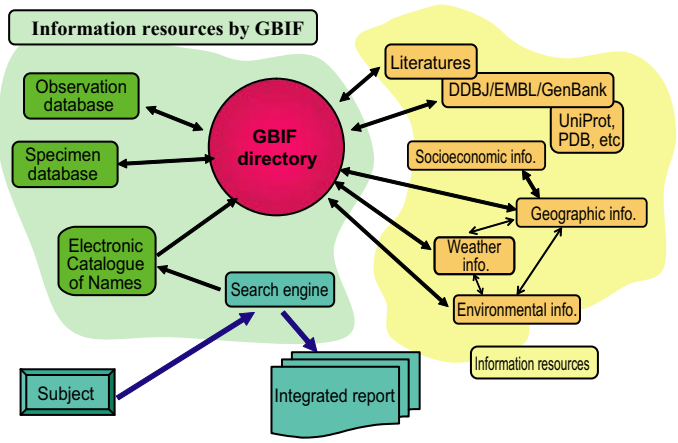

Fig. 2 Integration through web services:
The information architecture to solve biodiversity problems

For example, let us try to predict the movement of a species in a specific area by using a combination of specimen data, base sequence data, socioeconomic information and ecological information which are kept in a natural history museum. At present, you will first have to search for the required information through the Internet. Then you will have to process each data and find a way to combine them. Only after all that can you start analysis.

In contrast, when the solution to biodiversity problems offered by GBIF is fully developed, primary data can be integrated efficiently and easily obtained as shown in Fig. 2. After the subject is entered into the search engine, as shown on the bottom left in Fig. 2, the search engine refers to its scientific name (ECAT), then using the GBIF directory, collects specimen and base sequence data from DDBJ/EMBL/GenBank (http://www.insdc.org/), geographical information, socioeconomic information and ecological information, and returns an integrated report.

Kingdoms covered by www.gbif.net

The pilot project that employs internal and external GBIF information resources is under steady progress. At NIG, experiments to link domestic botanical specimen data to Google Map (Fig. 3) and experiments to integrate all of the GBIF data with DDBJ are being conducted.

Fig. 3 Display of the GBIF data on Google Map
(Part of the botanical specimens preserved in Japan)

Information Technology for a Virtual Database

■ Web services and standardization

The GBIF provides a virtual database in which the real data is scattered over multiple sites on the Internet. Aside from having different database management systems, each site adopted its own approach to information management. Consequently, a technology to enable database interoperability and establish the information architecture indicated in Fig 2 is required. The GBIF has implemented web services (http://www.xml.nig.ac.jp/) as a wrapping technology to enable each site to have a cooperative relationship with GBIF. Each site can use the tools provided by GBIF to provide its own data to the public in compliance with the GBIF standard data structures and the GBIF standard data exchange protocol. Two thirds of the nodes are using DarwinCore and DiGIR (http://digir.sourceforge.net/), and one third of the nodes, mainly the European nations, are using ABCD and BioCASE (http://www.biocase.org/) as the standard data structure and standard data exchange protocol. The Japan GBIF node in NIG uses MySQL to store biodiversity information and utilizes GBIF tools to implement DarwinCore and DiGIR.

Web services are widely implemented by many information resources such as those shown in Fig. 2 and therefore are a favorable technology to integrate theses resources.

■ Globally Unique Identifier (GUID)

The GBIF considers scientific names as a key to link many disparate data sets or other information. As the amount of data which can be mutually referred to by different GBIF sites increases, it becomes clear that it is necessary to specify the specimen or strain of each data source. Moreover, the concept of taxonomic groups for scientific names differs among sites. Therefore, a review for a framework in which a globally unique identifier (GUID) is assigned to every type of biodiversity data record such as physical specimens, strains, institutions or taxon concepts was initiated. If the GUID is defined and employed, it will be possible to rigidly define a physical specimen which is the source of a data, and increase reproducibility and omit redundancy in researches.

Considering that the candidate technologies for GUID are Digital Object Identifier (DOI, http://www.doi.org/) and Life Science Identifier (LSID, http://lsid.sourceforge.net/), a biological GUID seems possible if attention is paid to the characteristics of a biological specimen such as mutation and partial or entire replication or proliferation and if the scope of the GUID application is refined. Moreover, discussions for a biological GUID will surely be made in collaboration with the International Nucleotide Sequence Database Collaboration (http://www.insdc.org/).

Related Projects - Barcode of Life (BoL)

The GUID is a framework in which an identifier is artificially assigned to a biological object. In contrast, the Barcode of Life (BoL) (http://www.barcodinglife.org/) tries to use a gene sequence as an identifier of a species. Although the BoL is not a GBIF project, the GBIF serves as a member of the international consortium of the BoL and shares a close relationship.

DNA barcoding is a technique for characterizing species of organisms using a relatively short DNA sequence. Therefore, a partial sequence that is common within a specific species and differs among species must be chosen. Sequencing and data accumulation has already been initiated in animal specimens using the COL1 gene in the mitochondrial DNA. The results obtained by the BoL will be very useful for all researchers and engineers as well as taxonomists who want to identify biological samples. This technology can be utilized by anyone and is not confined to specialists of biological groups. It can also identify biological samples that cannot be identified by conventional methods. The sequence of the 16sRNA gene in bacteria plays a crucial role in classification and identification, and DNA barcoding at the species level has already been initiated. The BoL is examining an appropriate molecule or molecular group to barcode plants.

According to the BoL website, as of February 2006, 26,398 BoL entries are registered with DDBJ/EMBL/GenBank and 17,834 species have been barcoded.

BEATLES, *"Long and winding road"*

How far will GBIF expands its scale? In the case of insects, it is said that approximately 950,000 species are already known and 8,000,000 species are estimated to inhabit the earth (refer to references). In order to cover all the species, the ECAT in the GBIF, GUID, and BoL will all have to trudge through a long and winding road. However, if we persevere, our goals will surely be reached and that will open doors to a new world.

The Japan GBIF node at NIG will join forces with other domestic and international organizations and research groups and take it one step at a time.

【References】

Watson, R.T., Heywood, V.H., Baste,I., Dias,B., Gamez,R., Janetos, T.Reid, W., and Ruark, G. (edts.)

"Global Biodiversity Assessment-Sumary for Policy-Makers",

Cambridge University Press (1995)

Take a Break..

(1) Can news in the 21st century be written by amateur journalists?

On February 22, the IT media news (www.itmedia.co.jp/news/) announced that Korea's Ohmynews would establish a branch in Japan with support from Softbank Corp. Ohmynews is a new kind of news site where anyone can register and write articles to be published on its website. The standard of publication appears to be low, and the reliability of the articles is essentially judged by the readers themselves. Ohmynews started its service in 2000 and has grown into a powerful news publisher in the Korean society. It is said that Ohmynews greatly influenced the victory of Roh Moo-hyun, the current Korean President, in the 2002 Presidential Election.

In Japan, a civil media and Internet newspaper "JANJAN" seems to follow in Ohmynews footsteps but sadly, it is not as popular. With the initiation of Ohmynews services in Japan, the synergy might popularize this new media in Japan. Personally, I hope that a service that automatically translates the Korean Ohmynews articles to Japanese will also be provided. If this translation service is provided, I would be interested in reading the Korean articles. (N. K.)

Photograph: Original Ohmynews

- ② Science and Google -

Google has penetrated our daily lives to the extent that not a day goes by without us using it. Recently, Google has extended its service and has launched a new Google series in succession. The activities of Google in science were published in Nature: "Google makes data free for all" (**Nature 438, 2005**) and "Mashups mix data into global service" (**Nature 439, 2006**). Nature uses "Google Earth" to map information on areas in the world where the bird influenza broke out. Some of you might also have used "Google Scholar" to search for scientific papers. "Google Map" was mentioned in the GBIF article in this issue. "Google Base" is an interactive database. Anyone can upload a data file to share it with others. Aside from searching and browsing, you can even perform multiple database analysis if you write a program for it. This would allow easy accessibility to information on different fields. Since the Google Base service just started in Nov 2005, it might be a little too early to be expecting a Protein databases (JGI project of DOE) link located next to the Tickets or Wine and food links. On the other hand, I feel that it could be a huge thing if its hidden potential is unleashed. (Y. Y.)

Information Technology Vol. 11

"Bioinformatics in 10 minutes"

"BLAT, BLAST-Like Alignment Tool."

Do you know of BLAT? It is not a misspelling of BLAST. It refers to BLAST-Like Alignment Tool, and as the name suggests, BLAT can produce alignments of base sequences just like BLAST. However, BLAT uses mRNAs as a query and DNA as a target (DNA can also be used as a query and target). In short, BLAT can map mRNA to DNA. Detailed information is available at http://www.genomeblat.com/genomeblat/index.asp.
Let us start using the Windows version of BLAT.

1. Download BLAT from http://www.soe.ucsc.edu/?kent/exe/windows/blatSuite.30.zip.
2. Extract blatSuite.30.zip to an appropriate folder and assign a path to it.
3. Open the command prompt window (Start > Program > Accessories > Command Prompt), type "blat", and press Enter. The usage of BLAT will be displayed. (If you do not know how to assign a path as described in step 2, drag and drop blat.exe into the command prompt window instead of typing blat. Whenever you need to type "blat", use this method instead).
4. Create a new folder on your desktop and name it "test". Place a "target.fa" file (a fasta-format base sequence file for genome) and a "query.fa" file (a fasta-format base sequence file for mRNA or EST) in this folder.
5. Type "cd " in the command prompt window (insert a half size space after cd). Drag and drop the "test" folder you made in step 4 into the command prompt window and press Enter. Type "dir" and press Enter to confirm that "target.fa" and "query.fa" exist.
6. Type "blat target.fa query.fa output.txt" in the command prompt window and press Enter.

A result file, "output.txt" will be created in your "test" folder. When you open "output.txt" with a text editor, you might not recognize the contents. If you want to create a BLAST-like result file, type "blat target.fa query.fa output.txt –out=blast". BLAT has high speed DNA/DNA mapping functions such as fastMap so you might want to consider using BLAT tools in addition to BLAST.

Shingo Sakaniwa

Editor's note: Professor Sugawara kindly contributed the introduction article on GBIF. As suggested by its name, this project aims to achieve something great and it is much anticipated as Japan is making huge financial contributions to its budget. I look forward to using the information on one billion data records in 2011. (Y. Y.)

Contact: Center for Genetic Resources Information, National Institute of Genetics
Yata 1111, Mishima, Shizuoka 411-8540, Japan
Tel: 055-981-6885 (Yamazaki)
E-mail: BRnews@chanko.lab.nig.ac.jp

"translated by ASL translatiaon service and proofread by Sharoh Yip"