Chinese Human Genetic Resources Sharing Infrastructure

—Services and Technical Architecture

MA Li-Guang[1, 2], CAO Yan-Rong[1], MA Xu [2,*]

[1]State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and National Resources Research, CAS, Beijing, China

[2]National Research Institute for Family Planning, NRIFP, Beijing, China

*Corresponding author malg@lreis.ac.cn, caoyr@lries.ac.cn, genetic@263.net.cn

Abstract: The objective of this article is to achieve the Chinese human genetic resources' information sharing services. Firstly, the data collection, integration, classification and standardization as well as the data taxonomy and the data processing toolkit were discussed in details, the specification for human genetic resources data description was established and the data processing tools was disseminated at first. Secondly, based on the service-oriented system architecture method, the human genetic resources sharing infrastructure was strategical designed, the services capabilities of the framework was interpreted in details, the intelligence, networking, virtualization and security co-ordination platform were designed comprehensively. Finally, the human genetic resources information-sharing system (or platform) and resource databases were constructed and the information shared services hierarchically were realized between sharing alliance members and public. The results demonstrated that the technological capabilities of the platform will lead new research opportunities promoted the human health research work.

Keywords: human genetic resources;sharing infrastructure; reference model ; data standard; data mining

1. Introduction

The National Infrastructure Project of Chinese Human Genetic Resources (NIPCGR) is a comprehensive web-based management information system for Chinese human genetic resources sharing and services [1]. It is the important component of National Science and Technology Infrastructure Program.

NIPCGR will construct an environment where many different initiatives interoperate to provide rich mechanisms for human genetic resources information access, storage, knowledge exploration, analysis, and application services with international situation [2-5].

The NIPCGR infrastructure major capabilities are: 1) An interoperable platform to support the exchange and use of services for discovery, data access and data analysis about human genetic resources. 2) The provision of information sharing environments, which dedicated to use the services for resources management [6, 7] and thematic services; 3) Provision of the appropriate use of resources through security mechanisms and IPR and attribution policies [8].

This article is organized as follows: Section 2 provides a compilation of relevant standardization about human genetic resources and description formats are then described grouped by themes. Section 3 lists the current resources providers and their contribution for NIPCGR. Section 4 gives an overview of relevant services or function provided by the platform, Issues such as technical architecture and services organization as well as the intellectual property and the resources security are discussed. Finally, section 5 gives a conclusion about this article and the research work.

2. Standards and Taxonomic Data

Standards are at the heart of any undertaking that strives for interoperability or sharing of data, services, and tools [8]. The interoperability of NIPCGR infrastructure should consider the following contents such as standardized data formats, data transmission protocols or interface specifications and the specific identifiers of the resources.

2.1. Standards

Chinese human populations are distributed in a vast territory with complex topography and a varied natural environment. The human genetic resources are based on the Chinese multi-ethnic state (56 ethnic groups) comprising 1.3 billion people and aims to represent their patterns of genetic diversity. Chinese human genetic diversity is characterized by abundant minority groups with varied sizes, lifestyles, histories and disease and marriage patterns. Such data would be an invaluable asset for both the study of Chinese populations and for research on human evolution worldwide.

In China, many institutions have collected and stored human genetic samples such as blood, cell lines and tissue samples from all over the whole country in the past few years. The main tasks of the project are firstly, to integrate the existing resources and format the useful information into the standard metadata description which used for information storage, and secondly, to stimulate the collection of the further human genetic resources that are most needed and strengthen the development of the management environment and tools [9].

The critical thing here is to establish data specification for human genetic resources. From the information viewpoint and considered the intrinsically characteristics of human genetic resources' objects, the hierarchy system of human genetic resources standardization was constructed for the first time, show in Fig.1.

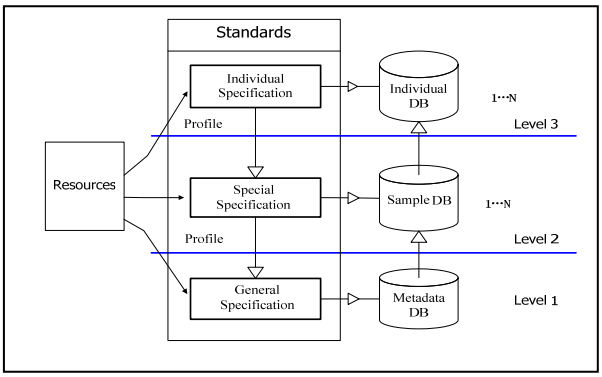

Figure 1.NIPCGR Standards

General specification of the resources is the core elements for both the metadata database and web services. The metadata structure is beneficial for users when they access the database and discovery relevant information and resources. The standardized terms of the metadata are made sensibility and feasibility for specific retrieval. Special specification is used to describe the sample intrinsically information and the individual specification is used for the potential information in the sample for research and analysis.

2.2. Taxonomy and Identifier

Human genetic resources research will need to critically revise the taxonomy of the organisms involved in the specific study. NIPCGR should come to an agreement with the different sources of taxonomic backbone information to be able to offer their usage in the technical framework. Taxonomic methods have been changed over time due to better understanding of the relationships between resources and different providers. Multiple taxonomies can exist concurrently due to differing scientific opinions about the relationships. This means that the scientific classification about the genetic resources can be ambiguous.

A taxonomy concept and the unique identifier are brought up to avoid the unambiguous identification of a specimen, both items are including in the general specification contents, information such as the donor's age, sex, and diagnosis coded according to the tenth Revision of the International Classification of Diseases (ICD-10). A Taxonomy Concept is a unique identification of a specimen, restricting the identity of a specimen to a single identifier. Any resources and information that are generated by NIPCGR or that are retrieved into the infrastructure and subsequently processed should be identifiable by the globally unique identifier as a foundation for provenance and citation as well as the attribution policies. This also makes information discovery more computationally intensive.

2.3. Toolkit for Data Collection

NIPCGR uses the data standards and specifications develop a toolkit for human genetic resources collection and integration in accordance with the uniform templet and requirement [10]. The toolkit is installed on the data provider's web server. The core part of the provider software is the wrapper software. It is an XML/CGI database interface written in object-oriented program language. This toolkit therefore transfers the data into a standard XML format which can access to the selected structured databases directly.

For different resources providers, they can build a standard template for their resources by this toolkit and then upload the final data files to NIPCGR. After the successful certification, the data will be imported to the permanently database. This way it is easy to build new DB modules for different resources providers. And this service will aid the mass data standardization and management. The toolkit function and the main interface show in **Fig.2.**

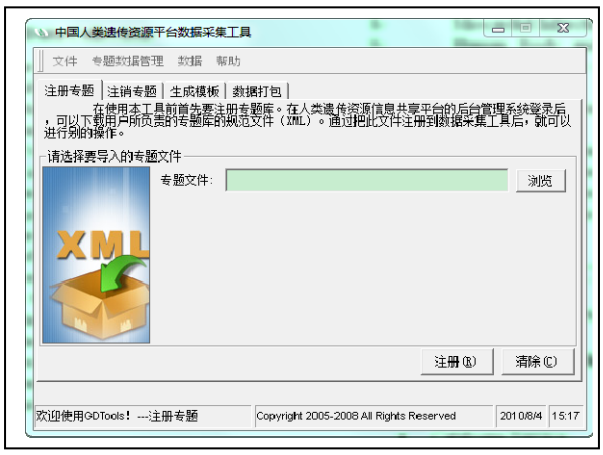

Figure 2.NIPCGR Data Collection Toolkit

978-1-935068-43-3 © 2010 SciRes.

3. Catalogue of Database

So far, more than 30 cooperative partners as data providers participating into the NIPCGR. The unstructured, semi-structured, or structured data and/or a provider of functions in terms of services have been stored in the NIPCGR database. These resources are with very heterogeneous nature and contain information in a variety of types and formats [11-13].

The database of the NIPCGR which holds the resource repository is the focus issue of this paper. It includes disease subclass data, minority population genetic resources and special population-genetic resource data, all of which are unique in China population. It is invaluable for the study of the origin and evolution of the Chinese populations as well as for improving medical treatment.

Until now, we have built different thematic database for variation purpose. Whole database have a collection about 300,000 records and classified into 27 thematic database lists in table1.

Table 1.Thematic Database and Records in NIPCGR

No.	Thematic Database	Records
1	Minority Ethnic Population	19072
2	Qinghai-Tibet Plateau Population	7511
3	Genealogy	2109
4	Tumor	12236
5	Hepatitis B	3000
6	Insulin Resistance	10650
7	Hypertension	20000
8	Miocardial Infarction	29910
9	Human Body and Organization Tissue	49150
10	AIDS	100
11	Isolated Population	2565
12	Obesity Population	6507
13	Gynecology Disease	14351
14	Birth Defects	19407
15	Space Mutagenesis Cell	5119
16	Stroke	6720
17	paternity identification	600
18	letterman	389
19	kidneys	8199
20	embryo	615
21	prospective cohort between rural and urban	38712
22	reproduction	6183
23	mental disorder	4146
24	human skeleton	4056
25	Liver	2096
26	Forensic Science	30000
27	Longevity population	2080

4. Services-Oriented Sharing Environment

4.1. Service-oriented Architecture

OASIS-SOA-RM defines Service Oriented Architecture (SOA) as a “paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains”. The person or organization offering a capability is called a Service Provider and the entity having a need to be solved by one or more capabilities is called a Service Consumer [14].

The following Fig.3 illustrates a basic service-oriented architecture. It shows a service consumer at the right sending a service request message to a service provider at the left. The service provider returns a response message to the service consumer.

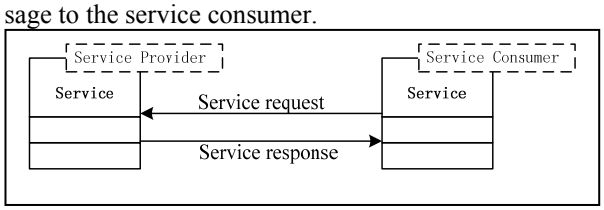

Figure 3.NIPCGR Data Collection Toolkit

The key concepts attached to a Service are:

- **Visibility** Exposure of capabilities to be found by an entity with needs, typically done through Service Descriptions
- **Interaction** The activity of using a capability, typically through the exchange of Messages between consumer and provider in a particular Execution Context(Real World)
- **Effect** Purpose of using a capability, result of an interaction.

Based on the SOA concepts, NIPCGR architecture was estabished.Fig.4 illustrates the Service-oriented architecture of NIPCGR.

The system management services, human interactive services, processing services and information management services as well as the integration services are provided in NIPCGR platform. The detailed contents will be interpreted in next subsection.

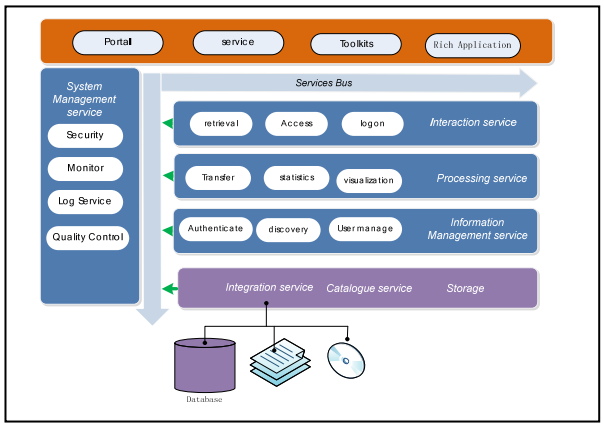

Figure 4. Service-oriented Architecture

4.2. NIPCGR Capabilities

The system management services, human interactive services, processing services and information management services and integration services make up the NIPCGR Services Capabilities. The following list (Tab.2) gives an overview of the services classification of NIPCGR.

Table 2. Classification of NIPCGR Services

Service	Description
System management service	
User Management Service	Creates and maintain subjects including groups of principals as entities that need authentication.
Authentication Service	Verifies genuineness of principals using a set of given credentials
Authorization Service	Gives a compliance value as response to a given authorization context
Service Monitoring Service	Provides an overview about Service Instances currently running
Log service	Record the server system problems, failure, and the server to track the efficiency.
Processing Services	
Communication service	Harmonized access to direct user-to-user communication means based on information technology and data-exchange between users
Processing service	Common interface for services offering processing operations by initiating the calculation and managing the outputs to be returned to the client.
Visualization service	Enables clients to interactively visualize tabular and statistic data by providing a intuitionistic representation of the data
Interaction Services	
Access service	Allows interoperable read and write access on service instances available in NIPCGR
Catalogue Service	Supports the ability to publish, query, and retrieve descriptive information for resources data, independent of a specific meta-information standard
Upload/Download service	Download or upload data use the interface provided by NIPCGR

This list of basic services shall be incrementally updated during the part of the optimizing phase. The instances of core services currently is:

- **Retrieve service**

NIPCGR provides a versatile query interface allowing intuitive access to the resources which contains data compiled from multiple heterogeneous resources. Two kinds of patterns for retrieval data are provided: Criteria Query: User input parameters, NIPCGR generates the standard SQL sentence which can interaction with database. and Classified Query: with the defined classification and taxonomy, user can access the sequence dataset under a certain classification.Fig.5 and Fig.6 show the user interface of Criteria Query and Classified Query respectively.

Figure 5. Criteria Query

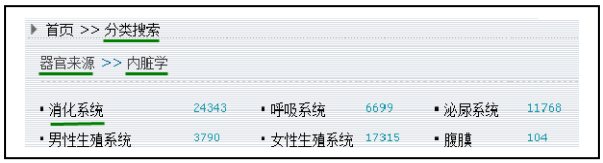

Figure 6. Classified Query

- **Processing service**

NIPCGR platform has the capability of statistical analysis and the function of statistical figures, using the fusion chart open sources, customized statistical processing service instance has been done. Users can get these services directly. Fig.7 shows the user interface and an example about statistical analysis services.

- **Visualization service**

The infrastructure has the capability of presentation for both attribute data and the multimedia data, the attribute data expressed as the tabular form in client browser and the medical images shown dynamicly in flash viewer tools [15]. The users can access to the general and the details information about the resources using the web browser like Fig.8. Fig.9 illustrates the medical image representation service.

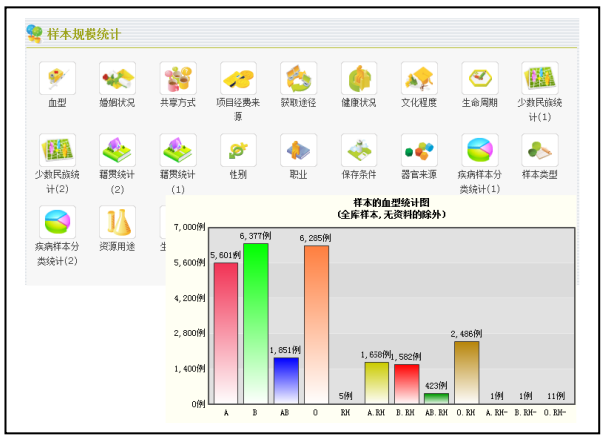

Figure 7. Statistical Analysis Example

978-1-935068-43-3 © 2010 SciRes.

Figure 8. Reprsention of the Attribute Data

Figure 9. Medical Image Resources Dynamic Represention

- **Download service**

Download service is the function that user can export the data or information from server to client, it includes the metadata, catalogue data, documents, reports download services etc. The context can be output as new data format which user has selected before such Adobe PDF format， Microsoft Excel format or Microsoft Access format for local application.

4.3 Information Sharing Environment

The objective of the NIPCGR is to promote the human genetic resources information sharing using the technical framework [16-18]. The capabilities developed have increased the services qualities but it's not enough for sustainable sharing and the scope extended for more resources. At the same time, following question must be answered: 1) How to ensure the data security and resources intellectual property right (IPR)? 2) How to ensure the sharing benefits for providers, manager and the users?

- **Data Security and IPR**

The metadata specification supports extensive metadata on intellectual property rights (IPR) and other rights, thus ensures that data providers can make their claims as to copyright, proper accreditation, and utilization of their data. NIPCGR extends IPR information to all kinds of data, requiring it for primary data as well as generating it for derived data. Such information may be represented as meta-information related to a globally unique identifier (GUID) for the data.

The providers and the preservation organization are the owner of data, the legal entity possess the right resulting from the act of creating a digital record in metadata. The record may be a product derived from another, such as the digital watermarking,

For sensitive data or confidential data, NIPCGR not make available public, only the users or participants who were authorized can access these types of data.

- **Sharing Agreements**

In according to the information sharing principle that the process of and agreements for making data freely are universally available on the web. The participants will signed the Memorandum of Understanding (MoU) to express their willingness to make human genetic resources data available through NIPCGR and to support the public access of these data (metadata). Data providers offer the primary data and the metadata on specimens and are responsible for updating the information and data quality. Non-registered users use the internet to access data can get the metadata information. The registered users can access the special data information.

5. Conclusion

The research results demonstrate that the human genetic resources and the potential knowledge in the diversity of information integrated into a collaborative network platform will lead new research opportunities for human genetic resources application. It's very critical to build the information sharing environment and the infrastructure for human genetic resources information sharing. The following contents can be summary: 1) Data is the fundament and core resources, the standards and specifications system establishment make it possible for different participants with diversity resources integration. A large-scale and resource-oriented database has been constructed for the first time, 2) The manage regulations and the sharing agreement documents have improved the sharing status, and ensured the data providers' IPR and the data security. This can increase the material resources sharing future.3) the technological capabilities and the infrastructure with generic and thematic services have realized the human genetic resources information like metadata, special data sharing on-line, the retrieve, representation service, the statistic analysis services promoted the information sharing feasibility. In future, the capabilities can guide the sharing services of the resources and strengthen the e-science environment development of

human health research.

Acknowledgment

The research presented in this paper was supported by National Infrastructure Program of Chinese Human Genetic Resources (No.2006DKA21301). MA Xu supervised the design of the study of the article. MA Li-Guang draft the manuscript and developed the toolkits. CAO Yan-Rong was participated in the work of standards and the data collection and integration. All authors read and approved the final manuscript.

References

[1] Liguang, M., C. Yanrong, and H. Jianbang. Chinese Human Genetic Resources Sharing service infrastructure. 2008. Sanya, Hainan, China: Institute of Electrical and Electronics Engineers Computer Society.

[2] Peter AC 't Hoen., What? Where? Which? WWW Resources for Geneticists: Bioinformatics for geneticists: a bioinformatics primer for the analysis of genetic data. Eur J Hum Genet, 2007. 15(12): p. 1280-1280.

[3] AURELLE, D., et al., Permanent Genetic Resources added to the Molecular Ecology Resources Database March 2010. Molecular Ecology Resources, 2010. 10(4): p. 751-754.

[4] Garte, S., Human population genetic diversity as a function of SNP type from HapMap data. American Journal of Human Biology, 2010. 22(3): p. 297-300.

[5] Hayashizaki, Y. and J. Kawai, A new approach to the distribution and storage of genetic resources. Nat Rev Genet, 2004. 5(3): p. 223-228.

[6] Kanthaswamy, S. Resources for genetic management and genomics research on non-human primates at the National Primate Research Centers (NPRCs). Journal of Medical Primatology, 2009. 38(s1): p. 17-23.

[7] J. Roger Guard, Ralph F. Brueggemann, William K. Fant, John J. Hutton, John R. Kues, Stephen A. Marine, Gregory W. Rouan, Leslie C. Schick: Integrated Advanced Information Management Systems:a twenty-year history at the University of Cincinnati. J Med Libr Assoc 92(2) April 2004

[8] Shimbo, I., Y. Ito, and K. Sumikura, Patent protection and access to genetic resources. Nat Biotech, 2008. 26(6): p. 645-647.

[9] CAO Zong-Fu,CAOYan-Rong,MALi-Guang. Standardization for sharing and utilization of Chinese genetic resources[J], HEREDITAS ,2008, 30(1): 51-58

[10] Svetlana Pacifico, Guozhen Liu, Stephen Guest, Jodi R Parrish, Farshad Fotouhi, Russell L Finley Jr: A database and tool, IM Browser, for exploring and integrating emerging gene and protein interaction data for Drosophila. BMC Bioinformatics 2006, 7:195.

[11] Yi Lu, Adrian E Platts, G Charles Ostermeier, Stephen A Krawetz: K-SPMM: a database of murine spermatogenic promoters modules& motifs. BMC Bioinformatics 2006, 7:238

[12] Kikuya Kato, Riu Yamashita, Ryo Matoba, Morito Monden, Shinzaburo Noguchi, Toshihisa Takagi, Kenta Nakai: Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues. Nucleic Acids Research, 2005, Vol. 33, Database issue

[13] Brandon W Higgs, Michael Elashoff, Sam Richman, Beata Barci: An online database for brain disease research. BMC Genomics 2006,7:70

[14] Liguang, M., C. Yanrong, and H. Jianbang. Biomedical image storage, retrieval and visualization based-on open source project. 2008. Sanya, Hainan, China: Inst. of Elec. and Elec. Eng. Computer Society.

[15] T. Erl, Service-Oriented Architecture: Concepts, Technology, and Design, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2005.

[16] Current status and prospects of studies on human genetic alleles associated with hepatitis B virus infection. World Journal of Gastroenterology, 2003. v.9(04): p. 641-644.

[17] Markel, D.S. and B.M. Yashar, The Interface Between the Practice of Medical Genetics and Human Genetic Research: What Every Genetic Counselor Needs to Know. 2004(5).

[18] Najmabadi, H., et al., The Iranian Human Mutation Gene Bank: A data and sample resource for worldwide collaborative genetics research. Human Mutation, 2003. 21(2): p. 146-150.