Development of Information Systems for Promoting Utilization and Expanding the Value of Genetic Resources

January 2017

YAMASAKI Fukuhiro

Development of Information Systems for Promoting Utilization and Expanding the Value of Genetic Resources

A Dissertation Submitted to
the Graduate School of Life and Environmental Sciences,
the University of Tsukuba
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
(Doctoral Program in Bioindustrial Sciences)

YAMASAKI Fukuhiro

Table of Contents

Table of Contents i

Abbreviations iv

CHAPTER 1 General introduction

1.1 Genetic resources for food and agriculture 1

1.2 Data management and publication in genebanks 1

1.3 Representative genebank management system 2

1.3.1 NIASGBdb 2

1.3.2 GRIN-Global 2

1.3.3 GBIS 3

1.3.4 GBIMS 3

1.4 Web retrieval system for genetic resources 4

1.4.1 Search by passport data 4

1.4.2 Search by characterization and evaluation data 4

1.4.3 Cross search system for PGR 4

1.5 Current problems 5

1.6 Objective of this dissertation 7

CHAPTER 2 Development of an integrated passport database schema for multiple genetic resource subsystems

2.1 Introduction 13

2.2 CNRG database model 14

2.2.1 Passport data 14

2.2.2 Relational database management system 17

2.2.3 Data input 17

2.2.4 Data output 19

2.2.5 Validation 19

2.3 Perspectives 19

CHAPTER 3 Advancement of information on genetic resources and improvement of user interface at the NARO genebank

3.1 Introduction 27

3.2 Improvement of user interface for C&E data queries by visualization of the data distribution 27

3.2.1 Background 27

3.2.2 Search by clicking histogram 28

3.2.3 Box plot like icon for individual C&E data 29

3.3 Association of genetic resources with geographic information and climate data 40

3.3.1 Using geographic information system 40

3.3.2 Using climate data 40

3.4 Publication of photographs of microorganism genetic resources 47

3.4.1 Background 47

3.4.2 Implementation 47

3.5 Perspectives 51

CHAPTER 4 Development of an agricultural field study database: For sharing multidisciplinary information from photographs of genetic resources collected *in situ*

4.1 Introduction 52

4.2 Implementation 53

4.2.1 Preparation of photographs 53

4.2.2 Database implementation 54

4.2.3 Web-based retrieval system 55

4.3 Perspectives 56

CHAPTER 5 General discussion and conclusions

5.1 General discussion and conclusions 67

Acknowledgments 71

References 72

List of Publications 78

Conference Presentation 78

Summary (in Japanese) 79

Abbreviations

Ajax	Asynchronous JavaScript and XML
C&E	Characterization and evaluation
CBD	Convention on Biological Diversity
CIPA	Camera & Imaging Products Association
CNRG	National Center of Genetic Resources (Mexico)
EURISCO	European Search Catalogue for Plant Genetic Resources
Exif	Exchangeable image file format
FAO	Food and Agriculture Organization of the United Nations
GIS	Geographic information system
GPS	Global positioning system
GRIN	Germplasm Resources Information Network
IPTC	International Press Telecommunications Council
KML	Keyhole Markup Language
MAFF	Ministry of Agriculture, Forestry and Fisheries (Japan)
MLS	Multilateral system
MSIE	Microsoft Internet Explorer
NARO	National Agriculture and Food Research Organization (Japan)
NIAS	National Institute of Agrobiological Sciences (Japan)
PGR	Plant genetic resources
PHP	PHP: Hypertext Preprocessor
RDBMS	Relational database management system
SVG	Scalable Vector Graphics
WWW	World Wide Web
XML	Extensible Markup Language

CHAPTER 1 General introduction

1.1 Genetic resources for food and agriculture

Genetic resources for food and agriculture are the basis of world food security. They are the raw materials for agricultural development such as breeding new cultivars with tolerance to biotic and abiotic stresses. Genetic resources refer to genetic materials that are conserved in genebanks or grow *in situ* have potential value as well as those currently in use (Frankel & Bennett, 1970). Genetic uniformity increases disease vulnerability (Ullstrup, 1972; Schumann, 1991), however, about 75 percent of plant genetic resources (PGR) have been lost during the last hundred years and one third of today's diversity could disappear by the year 2050 (FAO, 2011). Genetic resources are almost impossible to regain once lost (NRC, 1993), therefore conservation of genetic resources is an urgent task.

1.2 Data management and publication in genebanks

Genebanks play a significant role in both conservation and distribution of genetic resources, particularly PGR for crop improvement that sustain global food security. To fulfill genebank duties, an appropriate data management system and publication of data is crucial. If there is no effective documentation system, genebanks will have problems to manage their collections and cannot make appropriate action policies (Painting *et al*., 1995). Without published information on the World Wide Web (WWW), it is virtually impossible for breeders and researchers find genetic resources they really need.

Before 1980, genebank data were usually recorded on paper and managed manually (Ruredzo, 1992). Due to improved information technology, today Genebank data is generally compiled on computer-based databases and a number of national

genebanks have published collection data on the WWW (FAO, 2010). In this chapter the main features and differences of some representative Genebank databases are summarized, and future topics for improving information systems for genetic resources are discussed.

1.3 Representative genebank management system

1.3.1 NIASGBdb

NARO Genebank (renamed from NIAS Genebank due to an organization restructuring in April, 2016) is the coordinating project in Japan for the conservation of plant, microorganism, and animal genetic resources related to food and agriculture. Informix Dynamic Server was chosen for the database management system and MySQL is also used for web retrieval database. Most of the data management software has been developed using Informix-4GL programming language and its character-based user interface allows expert users to manipulate data quickly (Takeya *et al*., 2013a). A particular feature of NIASGBdb is managing information not only of plants, but also microorganisms and animals.

1.3.2 GRIN-Global

GRIN-Global is a database system for plant genebanks that has been developed by Agricultural Research Service of the United States Department of Agriculture (USDA-ARS), Bioversity International, and Global Crop Diversity Trust (Postman *et al*., 2010). They claim that many of the current genebanks lack effective technology for documenting and managing PGR information. This project's objective is to provide a flexible and easy-to-use plant genetic resource information management system for the world's genebanks. The Global Crop Diversity Trust has provided support to install and operate the GRIN-Global system for national genebanks in

developing countries such as Instituto Nacional de Innovación Agraria (Peru), Plant Resources Center (Vietnam), SADC Plant Genetic Resources Centre (Zambia), Centre for Pacific Crops and Trees (Fiji), Genetic Resources Research Institute (Kenya), Instituto de Investigaciones Agropecuarias (Chile), Instituto Nacional de Innovación Agropecuaria y Forestal (Bolivia), Instituto de Ciencia y Tecnología (Guatemala), and Centro Agronómico Tropical de Investigación y Enseñanza (Costa Rica) (Crop Trust, 2016). This system runs on ".NET Framework" and supports several open-source relational database management systems (RDBMS) including MySQL, Oracle Express, MS SQL Express, and PostgreSQL.

1.3.3 GBIS

The Leibniz Institute of Plant Genetics and Crop Plant Research, Germany (IPK) has largest collection of agricultural and horticultural crops among European Union countries. GBIS has been developed for managing the data of more than 151,000 accessions in the IPK collection. The database schema is implemented on Oracle 11g and program is written in Java. They also developed the application that runs on mobile devices equipped with a barcode scanner to support recording phenotypic data from field trials (Oppermann *et al*., 2015).

1.3.4 GBIMS

The National Bureau of Plant Genetic Resources, India (NBPGR) has more than 360,000 accessions belonging to 1,495 species and is ranked globally the third biggest collection (FAO, 2010). GBIMS has been developed to reduce their paper record and filing work. The database is implemented on Microsoft SQL Server and client application for data entry is written in Visual Basic. They also developed web-based query system for non-administrative users (Agrawal *et al*., 2007).

1.4 Web retrieval system for genetic resources

1.4.1 Search by passport data

Passport data is the basic data accession such as identifier number, taxonomic data, common name, accession name, country of origin, collecting institute, collecting location, breeding institute, biological status (e.g. wild or landrace), and storage type (e.g. seed or *in vitro*). Today there are some web retrieval systems that support search by passport data. For example, U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search.aspx), EURISCO (http://eurisco.ipk-gatersleben.de/), NARO Genebank Plant Search (Simple Queries) (http://www.gene.affrc.go.jp/?db_pl_e), and Svalbard Global Seed Vault (http://www.nordgen. org/sgsv/) are representatives.

1.4.2 Search by characterization and evaluation data

Characterization and evaluation (C&E) data such as morphological traits, yield potential, and disease resistance are of the greatest value to plant breeders seeking particular genetic traits for crop improvement (Koo & Wright, 2000), therefore serving search-query interactive databases containing C&E data is recommended in genebank standards (FAO, 2014). However, there are not many genebanks that provides web retrieval system using C&E data. U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx), EURISCO (http://eurisco.ipk-gatersleben.de/apex/f?p=103:47:::NO:::), and NARO Genebank Plant Search (Evaluation Data Queries) (http://www.gene.affrc.go.jp/?db_pl_ev_e) are some examples (Figs. 1.1-1.3).

1.4.3 Cross search system for PGR

Today there are more than 1,750 genebanks worldwide (FAO, 2010). Nevertheless, the reality is that few are successful in the development of a genetic

resources web retrieval system, especially in developing countries, there is no case of success. Considering that most of the crop diversity is to be found in developing countries (Rao, 2004) and the difficulty of access to the genetic resources in developing countries after Convention on Biological Diversity (CBD) came into force in December of 1993, this situation should be remedied. Moreover, even if genebanks in developing countries can construct web retrieval system, fragmentation of information should be avoided because it is inconvenient for users and will reduce the impact of the information.

Genesys PGR (https://www.genesys-pgr.org/) is a portal site that enables one stop search for more than 2,600,000 genetic resources conserved in over 440 genebanks including mega genebanks such as USDA-ARS, IPK, Centro Internacional de Mejoramiento de Maíz y Trigo (CIMMYT), International Centre for Agricultural Research in Dry Areas (ICARDA), International Rice Research Institute (IRRI), N.I. Vavilov Research Institute of Plant Industry (VIR), International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) (Fig. 1.4). This system can gather eccentrically located information from small genebanks worldwide and bring them to research community. Given the number of PGR listed in the multilateral system (MLS) that aims to facilitate open access is increasing (http://www.fao.org/plant-treaty/areas-of-work/the-multilateral-system/collections/en/?q=inclusions), Genesys PGR will be of increasing significance in the future.

1.5 Current problems

Firstly, only a few genebanks have effective technology for documenting and managing information on genetic resources they conserve. Some systems such as GRIN-Global have been developed, however, it is not applicable to all genebanks. In one example, a genebank that manages genetic resources other than plants has to

develop their own systems.

Secondly, there few PGR databases that allow search of C&E data and where they exist the user interface of retrieval can be improved. To realize a cross search system for genebanks worldwide, standardization of descriptor is necessary. With respect to passport data, Multi-Crop Passport Descriptors (Alercia *et al*., 2012) have been defined and adopted by Genesys PGR. While the standardization of C&E of PGR is presently being studied (Krajewski *et al*., 2015), still there are no absolute de facto standard. Conversion of existing data to a standard system will require a lot of time and effort. From the perspective of maintenance and scale, it is not reasonable to store all data in one portal site, therefore genebanks that already have their own web retrieval system should give priority to their own user interface for C&E data queries.

Lastly and most importantly, that cultural information is not associated with genetic resources. Nakao (1966), a Japanese ethnobotanist argued that agriculture not only produces food for human beings but also serves as the root of culture. Sakamoto (1992) claimed that researcher should collect and record cultural information as far as possible such as local name, cultivation style, and utilization method thus recognizing that genetic resources are not only useful material for breeding but also as “cultural resources”. However, genebanks are not currently providing enough cultural information on genetic resources. Local and traditional foods are not only a nutrient source but also have important cultural, ceremonial or religious meanings. They are also valued by farmers because of the imbued cultural values such as symbolism in religious ceremonies (Soleri & Cleveland, 1993).

Just as we cannot understand the importance of glutinous rice without understanding rice cake and wine culture, many genetic resources lose their value without cultural information. Although FAO (2014) states “Field management processes and cultural practices should be recorded and documented” in Genebank standards 5.8.2,

recording and documentation of cultural information is insufficient. Genebanks should proactively make cultural information available to the wider research community and promote conservation of genetic resources by interdisciplinary approaches.

1.6 Objective of this dissertation

The objective of this thesis is to promote utilization and expand the value of genetic resources through the development of information systems. The remainder of the dissertation is organized as follows: Chapter 2 discusses development of an integrated database system for managing various genetic resources. Chapter 3 describes improvement of the web database and its retrieval system, including visualization of C&E data. Chapter 4 presents agricultural field study database for sharing *in situ* photographs that contains cultural information as well as morphology of collected material (Fig. 1.5). The concluding chapter 5 summarizes the whole dissertation.

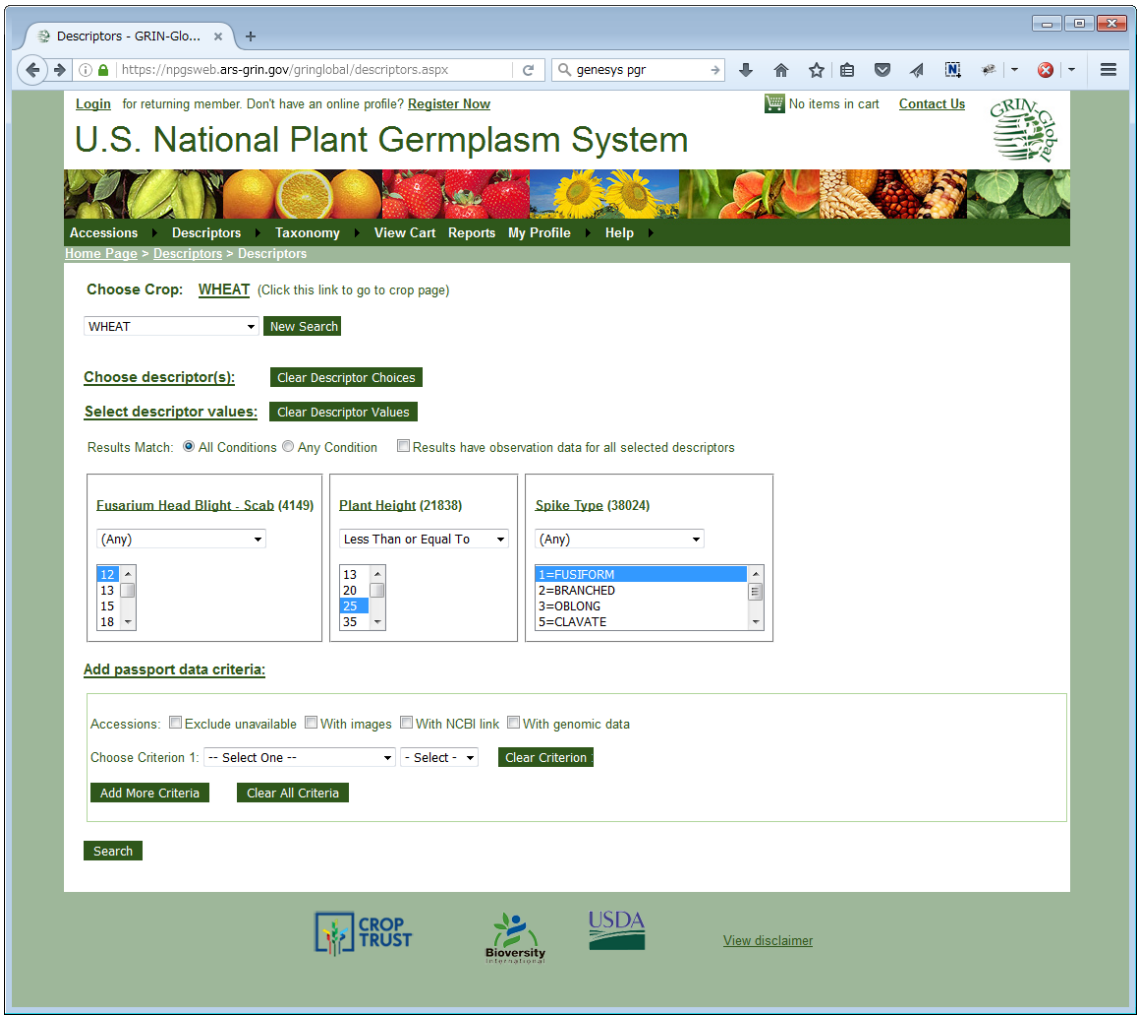


Fig. 1.1. Screenshot of C&E data queries in U.S. National Plant Germplasm System. By selecting a crop and descriptor(s) to use, users are able to configure search criteria. Genetic resources that meet all/any conditions will be displayed in search result. To refine search conditions, it is required to back to previous page.

https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx

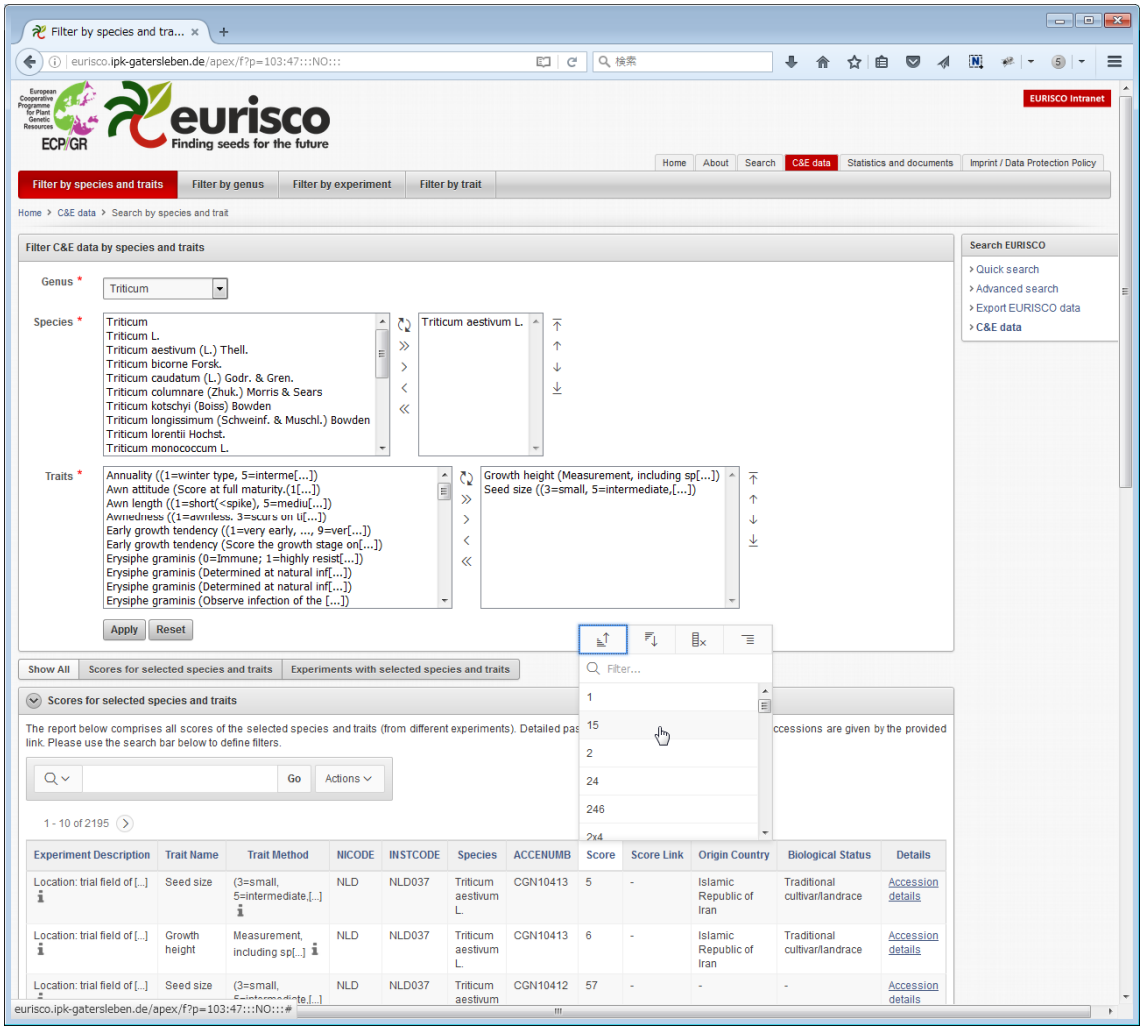


Fig. 1.2. Screenshot of EURISCO C&E data queries.

By selecting Genus, Species, and Traits, OR search will be executed and related genetic resources are displayed. Although users are able to narrow down search results by using filters, however, executing intersection search by more than two descriptors is difficult. C&E data is not shown in “Accession details” page.

http://eurisco.ipk-gatersleben.de/apex/f?p=103:47:::NO:::

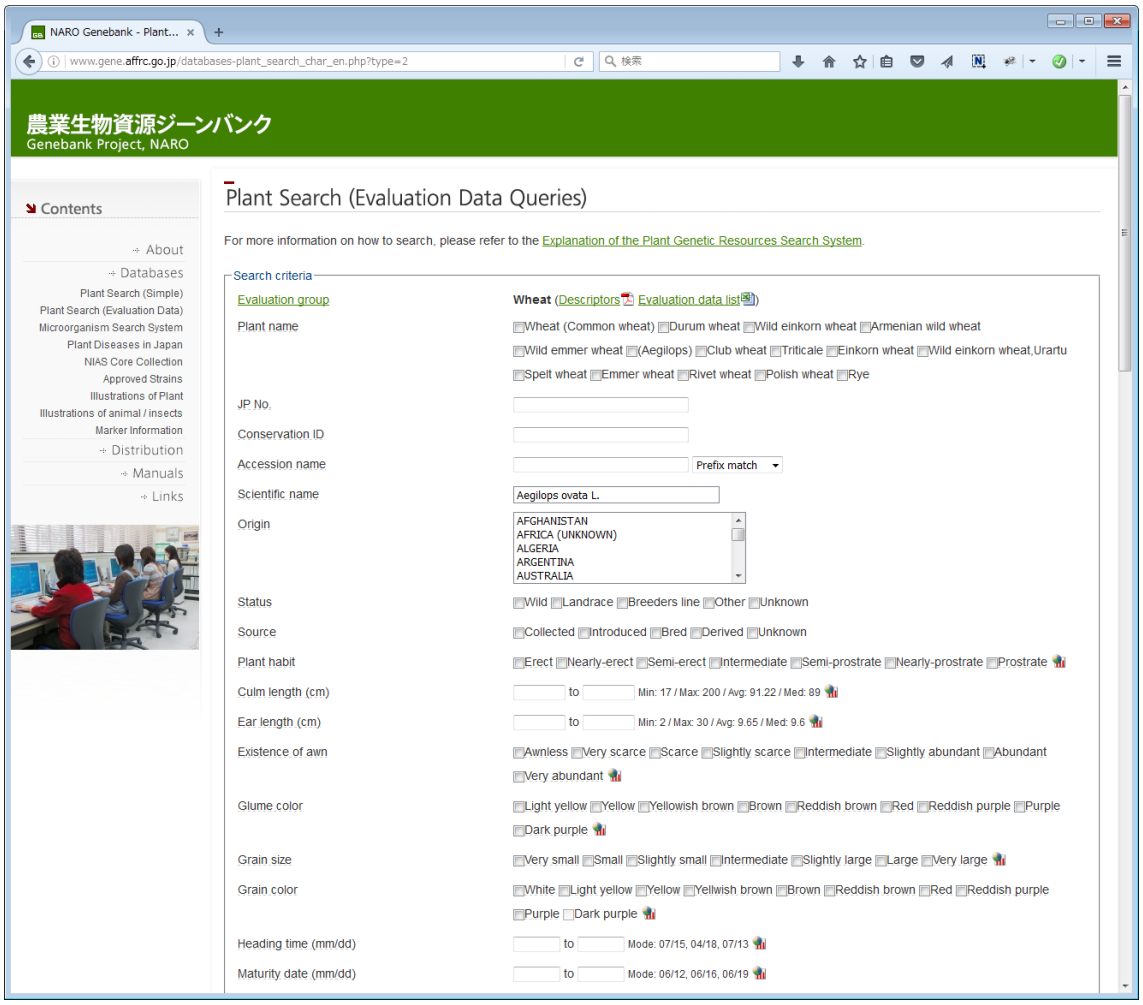


Fig. 1.3. Screenshot of NARO Genebank Plant Search (Evaluation Data Queries).

By selecting Evaluation group, all descriptors will be shown at once. Genetic resources that meet all conditions will be displayed in search result. Users are able to refine search conditions, download details of descriptor as a PDF file, and confirm data distribution seamlessly.

https://www.gene.affrc.go.jp/?db_pl_ev_e

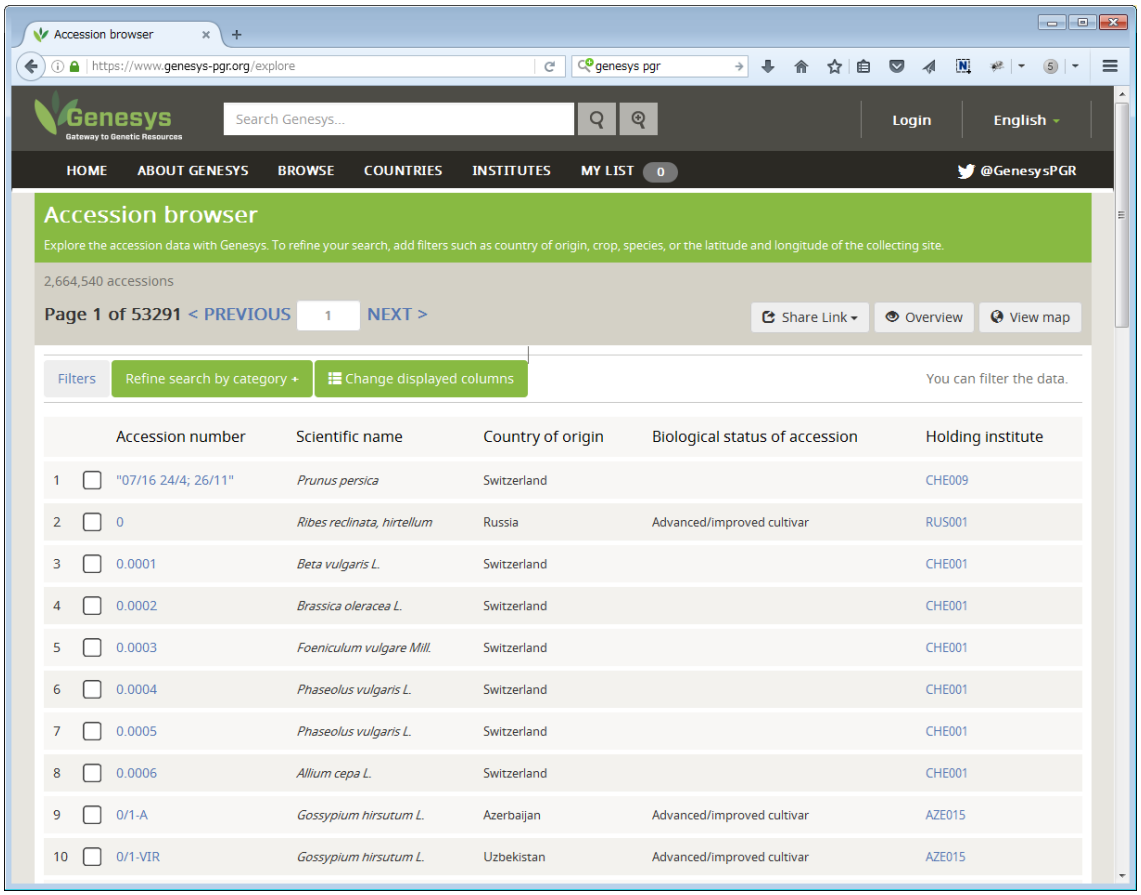


Fig. 1.4. Screenshot of Accession browser from Genesys PGR.

It is a portal site that enables one stop search for more than 2,600,000 genetic resources conserved in over 440 genebanks.

https://www.genesys-pgr.org/

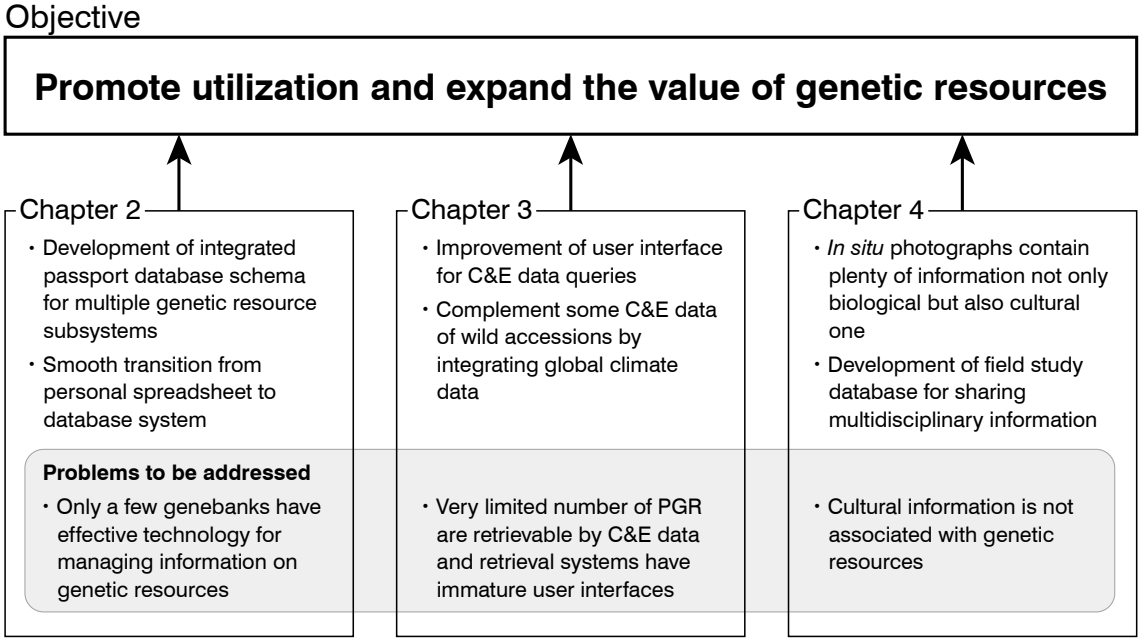


Fig. 1.5. Objective of this dissertation. Chapter 2 discusses development of an integrated database system for managing various genetic resources. Chapter 3 describes improvement of the web database and its retrieval system, including visualization of C&E data. Chapter 4 presents agricultural field study database for sharing *in situ* photographs that contains cultural and biological information.

CHAPTER 2 Development of an integrated passport database schema for multiple genetic resource subsystems

2.1 Introduction

Agricultural biodiversity has been defined as the biological diversity that play a fundamental role in agricultural ecosystems structure and processes (Dulloo *et al*., 2010). Breeding can result in a narrow agricultural genetic basis that can result in catastrophic agricultural loss caused by natural disasters, climate change/extreme weather events, or epiphytotic disease (Altieri & Merrick, 1987; Blackburn, 2009). The risk of a significant loss of diversity has increased in recent years, so conservation of agriculturally related animals, plants, and microorganisms have attracted global attention (Rands *et al*., 2010).

The sustainable use and development of genetic resources are essential for stabilizing the world's food supply. Consequently, *ex situ* conservation aims to keep the biological integrity of germplasm as long as possible as a strategy to avoid agrobiodiversity loss (Dulloo *et al*., 2010). Sustainable genetic resources management in genebanks largely depends on the handling of materials, human resources, budgets, and information about the materials (Clark *et al*., 1997). Such management should be conducted in accordance to the size of a genebank and the diversity of stored materials.

Mexico is considered to be among the world's most mega-diverse countries (Mittermeier, 2005). In 2012, the Mexican government established National Center of Genetic Resources (CNRG) as a part of a national strategy to safeguard the country's food supply and environment (Machida-Hirano *et al*., 2014). The CNRG is designed to manage agriculture, forestry, livestock, microbial, and aquatic germplasm subsystems, and serve as the reference center for various related research institutes in Mexico.

In order to achieve the CNRG's expected goals, an appropriate database system for genetic resources is essential. Although several database systems have been

developed and implemented by genebanks worldwide (Postman *et al*., 2010; Oppermann *et al*., 2015; Agrawal *et al*., 2007; Blackburn, 2009), as well as portal sites that facilitate searches for genetic resources across such databases (Telenius, 2011), only a few have addressed the management of multiple genetic resource subsystems (Takeya *et al*., 2011).

This study presents an overview of the projected database system for the CNRG, and describes the development of the passport database system as a first step toward the integrated management of five different types of genetic resources.

2.2 CNRG database model

The implementation plan was developed under a joint Japan-Mexico project in order to ensure the sufficient management of information related to CNRG genetic resources (Machida-Hirano *et al*., 2014). Under this plan, a system with four principal databases was envisaged. The first database would contain all genetic resource passport data (Fig. 2.1; Passport); the second would contain the locations of genetic resources within the Center (Fig. 2.1; Storage); the third would contain all other available information regarding the characteristics of each accession such as evaluation, phenotype, genotype, and images (Fig. 2.1; Characteristics); and the fourth would consist of publicly retrievable data (Fig. 2.1; Web).

2.2.1 Passport data

2.2.1.1 Schema

In developing the database system, top priority was placed on establishing the passport data database as it contains fundamental information related to each germplasm managed at the CNRG. Databases for genetic resources passport data have been developed by different institutions. For example, GRIN-Global provides a plant genetic resource information management system as open source software (Postman *et al*.,

2010). NARO Genebank, Japan has a well-developed schema for the management of plant, microbial, and animal genetic resources, as well as their interactions and plant disease information (Takeya *et al*., 2011). However, these systems are not suitable for the CNRG's wider scope. Therefore, a new schema was developed to facilitate the development of a five-subsystem germplasm inclusive passport database.

There are different concepts regarding what types of genetic resource information should be considered passport data. The CNRG adopted EURISCO's list (http://www.ecpgr.cgiar.org/fileadmin/templates/ecpgr.org/upload/MISC/EURISCO_Descriptors.pdf), in addition to data that met the needs of its particular genetic resources management.

Passport data was classified according to its nature into the following conceptual fields, although this classification was not implemented in the database schema.

- Taxonomy: This includes all information regarding biological classification. After genus, three sub taxa are allowed, and their nature depends on their particular conventional systematics. In particular cases, progenitor and individual identifiers may also be assigned.
- Location: This includes every detail regarding where the original germplasm was collected.
- Storage: This includes all information related to how the germplasm is preserved, as well as its location within the CNRG. In the future, this data will be linked to a storage database, which will provide traceability of each germplasm unit at any moment.
- Collection Status: Logistic details such as current availability and purpose are grouped here.
- Sample Status: This includes the type of sample (e.g. seed, DNA, tissue, cell),

viability, and qualitative/quantitative information about the germplasm unit.

- Origin: This is information detailing how the germplasm was collected, as well as additional information about its isolation, breed, or acceptance as a donation from another institute.
- Interest: This details the relevance of the germplasm and its conservational importance.
- Institutional Information: This is information about affiliated institutions, researchers, projects, etc.

Figure 2.2 shows a conceptual overview of the passport database.

2.2.1.2 Collection number

One of the principal achievements of this schema is that it can unify all subsystems with a comprehensive identifier at the accession level. This identifier complies with the internal definition of accession: as long as the taxonomy and location remain the same, every germplasm will share this identifier.

2.2.1.3 Germplasm unit number

The germplasm unit number concept was implemented to achieve a single identifier for any subsystem accession. This identifier contains data relevant to storage, collection status, sample type, sample status, and origin. Under the germplasm unit number concept, one accession may also be stored in the form of seed, tissue and DNA, and will be given different germplasm unit numbers in case one tissue sample is stored in liquid nitrogen and another sample is stored in medium-term preservation. Thus, such samples will be managed with two unique germplasm unit numbers. Consequently, elements that were traditionally allocated as independent accessions would be allocated as a subgroup of the same collection number. This concept allows different sample types

with different treatments to be managed independently, even in the laboratories for different subsystems, thereby providing an opportunity for accession to enrich different ongoing projects within the Center. Although the concept of any sort of sub-accession level for different treatments is not unique, to the best of our knowledge, this marks the first implementation of a database schema that can comprehensibly manage any subsystem accession.

2.2.2 Relational database management system

As previously stated, the first step toward realizing the CNRG database system was the design and implementation of the passport data database. Although the complete database system is expected to implement RDBMS software that integrates input, retrieval, application, and administration functions, at the moment, priority is placed on meeting the minimum requirements so that the CRNG can begin relying on the database for management of its expanding genetic resources.

MySQL (http://www.mysql.com/) was chosen for this project because it is a very popular RMDBS (solid IT, 2015). When considering continuity and portability, market share is an important factor. NARO Genebank adopted MySQL for its web retrieval system (Takeya *et al*., 2013a) and MySQL is listed as the RDBMS supported by GRIN-Global (Postman *et al*., 2010). This will ensure an effective exchange of information between institutions in the future. A rollback capability is important for database integrity so the storage engine must also include this feature. Given its adequate transaction support, InnoDB was selected as the storage engine (Kruckenberg & Pipes, 2005).

2.2.3 Data input

Input efficiency is a key element in any newly developed database. The previous information management system at the CNRG relied on each subsystem's

individual administration, which, due to the nature of the Center, resulted in redundant, inconsistent, and non-unified criteria. In order to ease the transition to the new system, an intuitive spreadsheet application that does not require a deep understanding of the database schema was determined to be the optimum input tool.

2.2.3.1 Template

The input application primarily consists of an .xlsx template. This template contains a capture field for all information considered passport data. Note that not all information is required for all germplasm in each subsystem because some clearly differ in nature, and the template is grouped according to each subsystem's common requirements. According to the particular data characteristics, the input system may be either a drop-down menu for finite fields or a free description field.

2.2.3.2 Batch insertion program

A batch insertion PHP (PHP: Hypertext Preprocessor; https://php.net/) program was developed for statement insertions. This program defines the association between the input template and passport database by translating captured information in the template's fields to the corresponding table and column in the database. In an iterative sequential execution, the program processes the input at its lowest hierarchical level to determine its existence. If the input does not exist in the table, a new record will be generated. The corresponding key is stored as the input's new value. The same process continues to higher hierarchical levels until reaching the highest level. Figure 2.3 shows an example of this process. The PHP data input system allows any researcher to easily submit information to database administrators. This results in a fast transition between previous CNRG information to the new database system. Moreover, this input system can be distributed to depositor researchers or institutions so as to maintain an information standard and promptly enrich the database content.

2.2.4 Data output

The developed a web retrieval system, is still at an initial stage. Currently, the system is not eligible for disclosure and has limited functionality; however, the .xlsx file as output is useful for data confirmation. In addition, it provides a .kml output function to visualize the location data with georeferenced maps (Figs. 2.4-2.6).

2.2.5 Validation

To determine the consistency of the newly designed passport database schema, arbitrary values were manually provided to the system. This allowed virtual testing of the design relative to collection number and germplasm unit assignation. Once the schema was perfected, implementation of the input/output system was tested with real CNRG data from all subsystems. When the system demonstrated the expected outcomes, implementation was conducted on a temporary server at CNRG facilities. Although in an elementary stage, the CNRG database process has begun, and in its current state, the implemented passport database is capable of managing all accessions in the Center regardless of subsystem.

Most genetic resource-related institutions do not have multiple subsystem problems and therefore do not require a database of such sophistication; they tend to develop more databases specific for their rather limited needs (Blackburn, 2006). Nevertheless, some institutions have faced this problem and solved it in efficient ways (Takeya *et al*., 2010).

2.3 Perspectives

This marks the first step among many toward realizing a fully integrated database system that can adequately manage the expected amounts of information to be generated by a genebank that handles different types of genetic resources in fulfilling its

purpose to safeguard and utilize any subsystem of genetic resource. The batch insertion program enables smooth transition from personal spreadsheet to database system.

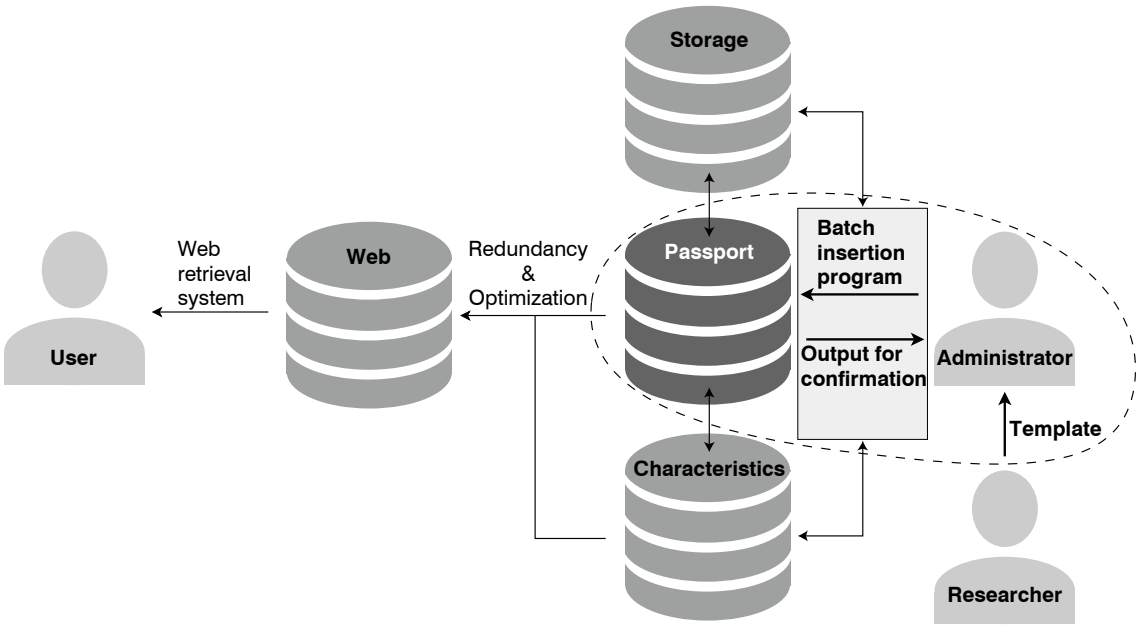


Fig. 2.1. Projected workflow for the CNRG's genetic resources management database system.

The completed passport database and input/output system is encircled by the dotted line.

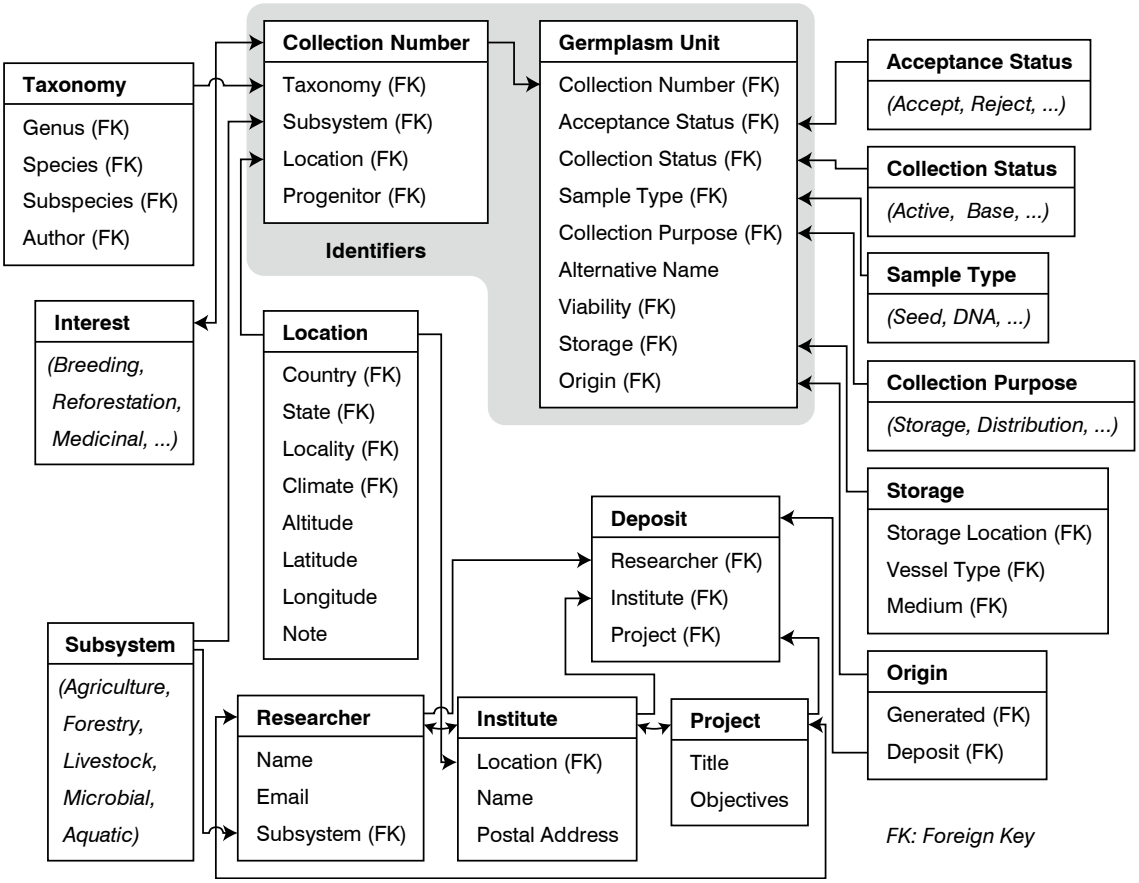


Fig. 2.2. Conceptual overview of the CNRG passport database. As long as the taxonomy and location remain the same, germplasm will share the same collection number.

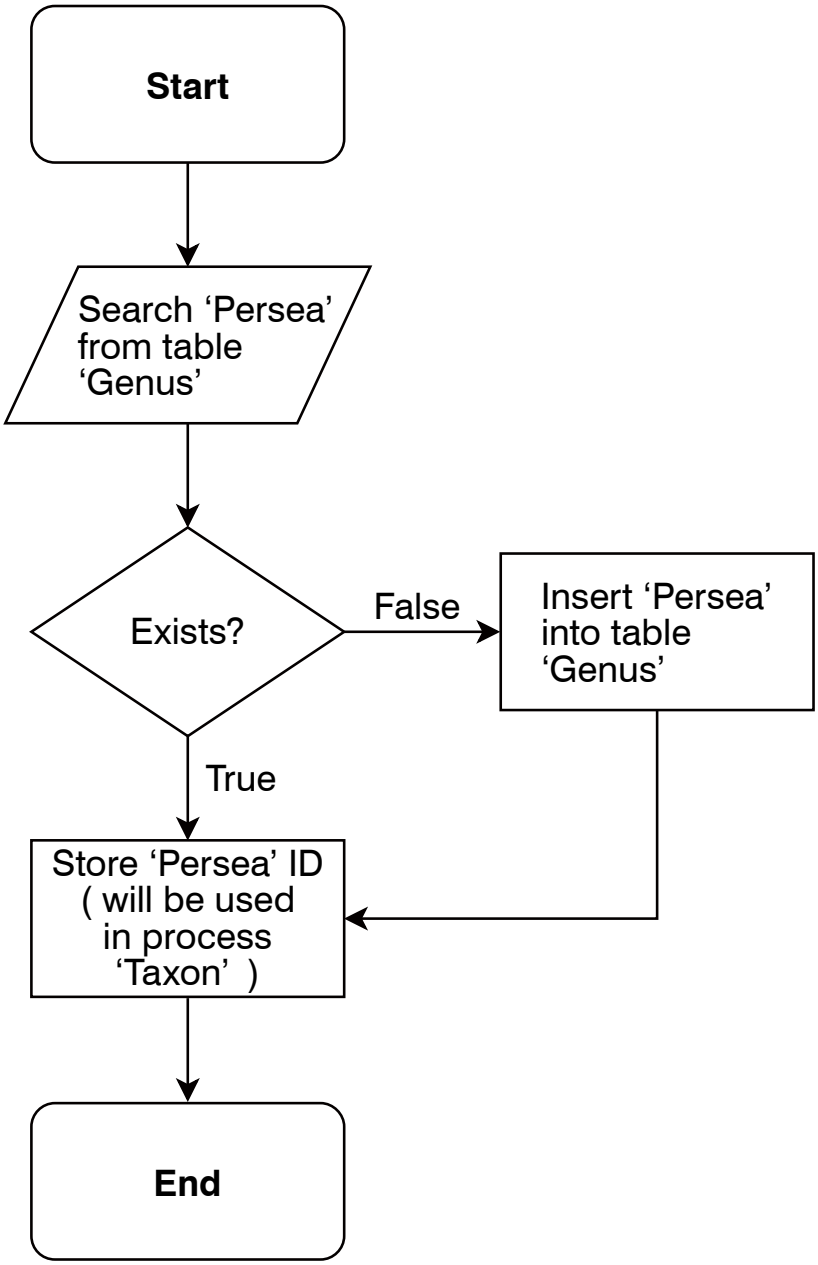


Fig. 2.3. Example of a batch insertion flow. If the inputted value does not exist in the corresponding table, a new record will be generated automatically.

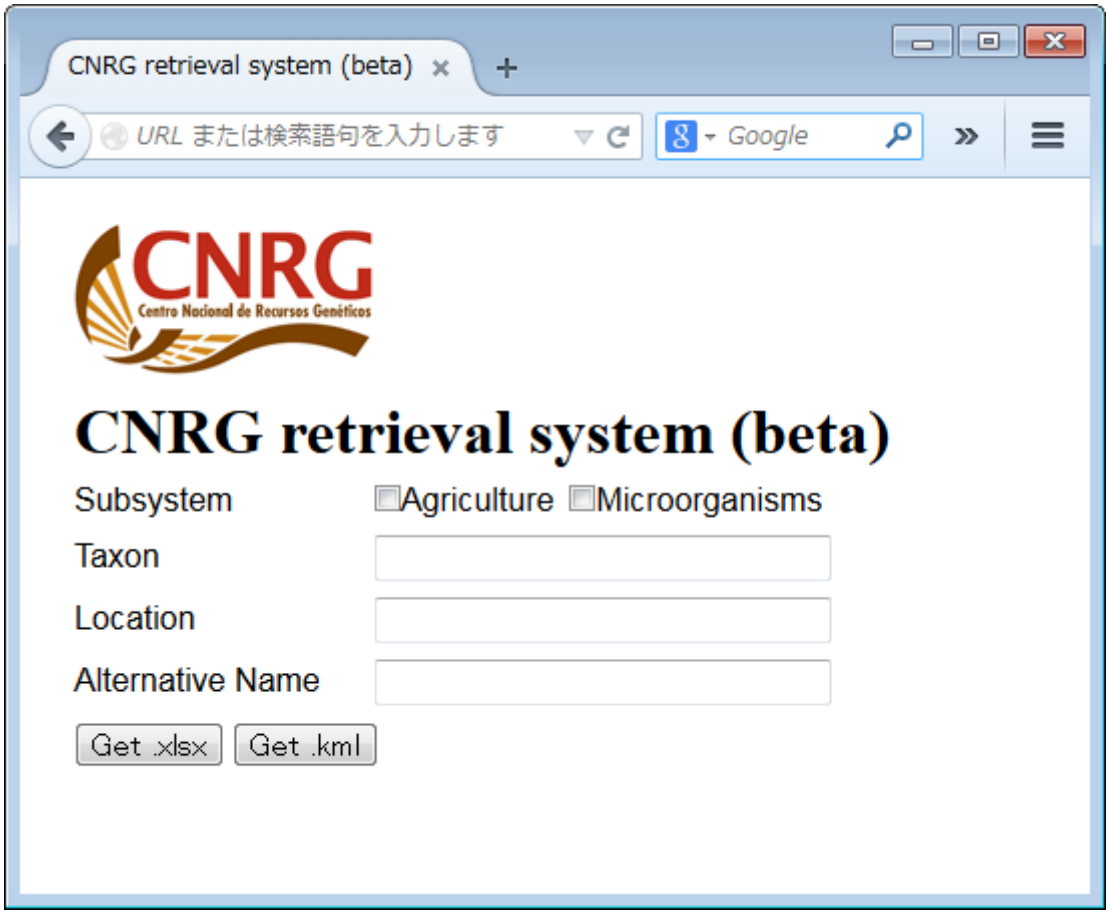
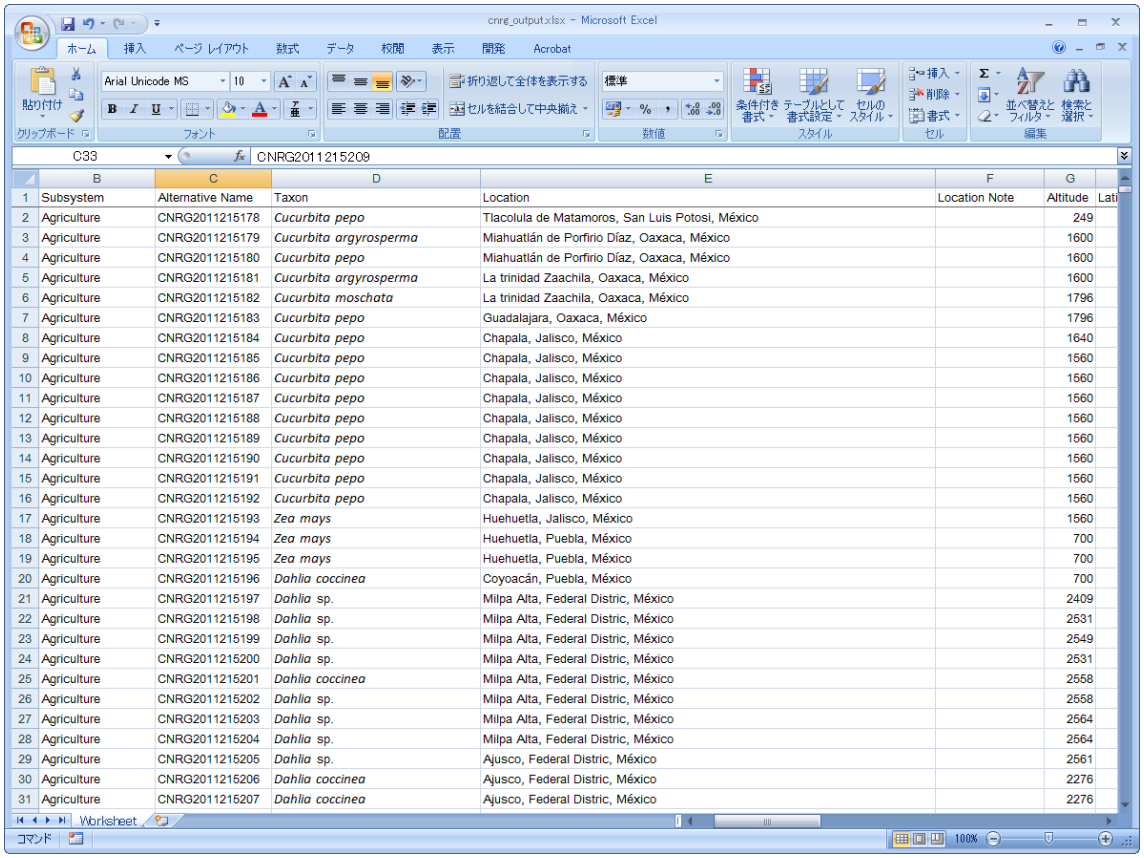


Fig. 2.4. Screenshot of output system for data confirmation.



	B	C	D	E	F	G
1	Subsystem	Alternative Name	Taxon	Location	Location Note	Altitude
2	Agriculture	CNRG2011215178	*Cucurbita pepo*	Tlacolula de Matamoros, San Luis Potosi, México		249
3	Agriculture	CNRG2011215179	*Cucurbita argyrosperma*	Miahuatlán de Porfirio Díaz, Oaxaca, México		1600
4	Agriculture	CNRG2011215180	*Cucurbita pepo*	Miahuatlán de Porfirio Díaz, Oaxaca, México		1600
5	Agriculture	CNRG2011215181	*Cucurbita argyrosperma*	La trinidad Zaachila, Oaxaca, México		1600
6	Agriculture	CNRG2011215182	*Cucurbita moschata*	La trinidad Zaachila, Oaxaca, México		1796
7	Agriculture	CNRG2011215183	*Cucurbita pepo*	Guadalajara, Oaxaca, México		1796
8	Agriculture	CNRG2011215184	*Cucurbita pepo*	Chapala, Jalisco, México		1640
9	Agriculture	CNRG2011215185	*Cucurbita pepo*	Chapala, Jalisco, México		1560
10	Agriculture	CNRG2011215186	*Cucurbita pepo*	Chapala, Jalisco, México		1560
11	Agriculture	CNRG2011215187	*Cucurbita pepo*	Chapala, Jalisco, México		1560
12	Agriculture	CNRG2011215188	*Cucurbita pepo*	Chapala, Jalisco, México		1560
13	Agriculture	CNRG2011215189	*Cucurbita pepo*	Chapala, Jalisco, México		1560
14	Agriculture	CNRG2011215190	*Cucurbita pepo*	Chapala, Jalisco, México		1560
15	Agriculture	CNRG2011215191	*Cucurbita pepo*	Chapala, Jalisco, México		1560
16	Agriculture	CNRG2011215192	*Cucurbita pepo*	Chapala, Jalisco, México		1560
17	Agriculture	CNRG2011215193	*Zea mays*	Huehuetla, Jalisco, México		1560
18	Agriculture	CNRG2011215194	*Zea mays*	Huehuetla, Puebla, México		700
19	Agriculture	CNRG2011215195	*Zea mays*	Huehuetla, Puebla, México		700
20	Agriculture	CNRG2011215196	*Dahlia coccinea*	Coyoacán, Puebla, México		700
21	Agriculture	CNRG2011215197	*Dahlia* sp.	Milpa Alta, Federal Distric, México		2409
22	Agriculture	CNRG2011215198	*Dahlia* sp.	Milpa Alta, Federal Distric, México		2531
23	Agriculture	CNRG2011215199	*Dahlia* sp.	Milpa Alta, Federal Distric, México		2549
24	Agriculture	CNRG2011215200	*Dahlia* sp.	Milpa Alta, Federal Distric, México		2531
25	Agriculture	CNRG2011215201	*Dahlia coccinea*	Milpa Alta, Federal Distric, México		2558
26	Agriculture	CNRG2011215202	*Dahlia* sp.	Milpa Alta, Federal Distric, México		2558
27	Agriculture	CNRG2011215203	*Dahlia* sp.	Milpa Alta, Federal Distric, México		2564
28	Agriculture	CNRG2011215204	*Dahlia* sp.	Milpa Alta, Federal Distric, México		2564
29	Agriculture	CNRG2011215205	*Dahlia* sp.	Ajusco, Federal Distric, México		2561
30	Agriculture	CNRG2011215206	*Dahlia coccinea*	Ajusco, Federal Distric, México		2276
31	Agriculture	CNRG2011215207	*Dahlia coccinea*	Ajusco, Federal Distric, México		2276

Fig. 2.5. An example of output result in MS-Excel format.

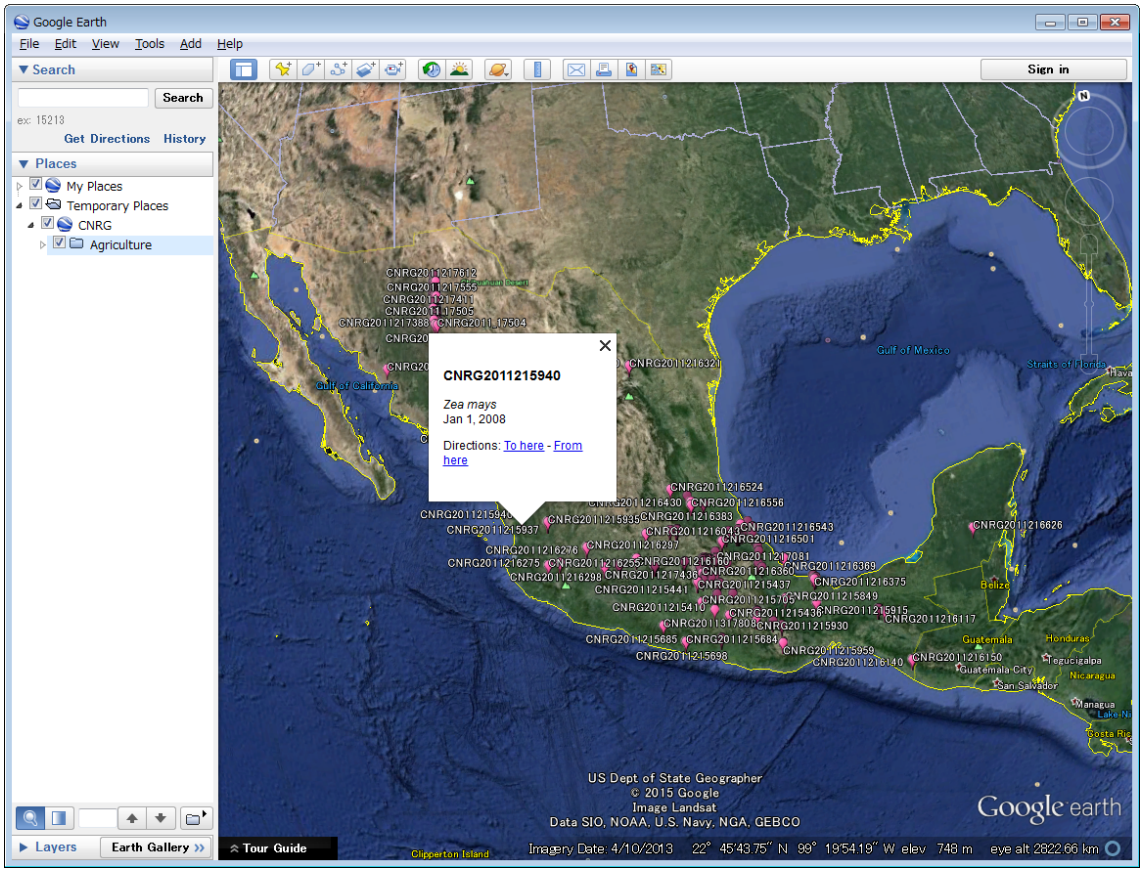


Fig. 2.6. An example of output result in KML format. It visualizes the location data with georeferenced maps.

CHAPTER 3 Advancement of information on genetic resources and improvement of user interface at the NARO genebank

3.1 Introduction

For users to efficiently find genetic resources a web retrieval system is crucial. Genebank standards (FAO, 2014) recommends a search-query interactive databases containing C&E data, however, currently existing systems have some drawbacks at the user interface. Moreover, only a few websites are providing photographs that capture the morphological characteristic of microorganism genetic resources. In this chapter, improvements at the user interface and improved information in the NARO Genebank databases are discussed.

3.2 Improvement of user interface for C&E data queries by visualization of the data distribution

3.2.1 Background

A web retrieval system for C&E data in the NARO Genebank database system was previously developed and its user interface has been adopted by some other genebanks (Fig. 3.1). However, the system could be improved. According to an analysis of the access log, evaluation data queries account for 63.2% of access to the entire PGR search system between April 2013 and March 2016. On the other hand, evaluation data queries account for a mere 24.8% of the number of executions. This imbalance implies that some users had difficulty in the use of the C&E query system. From the analysis of executed SQL, only 5.3% used more than three descriptors for search criteria and no descriptor was used in 52.0% (Table 3.1). Given this fact, more intuitive user interfaces were developed using histograms.

3.2.2 Search by clicking histogram

3.2.2.1 Drawing histograms

Scalable Vector Graphics (SVG) were chosen for drawing histograms, rather than raster format. SVG is an XML-based vector that can contain hyperlinks and can be interactive and dynamic (W3C, 2011). SVG did not become popular initially because it was not supported by Microsoft Internet Explorer (MSIE) version 8 and earlier versions, however, considering now Microsoft is a member of SVG Working group (Dengler, 2010) and the usage share of MSIE 6/7/8 has declined to less than one percent in May 2016 (W3Counter, 2016), there is no reason not to use SVG.

3.2.2.2 Number of bins

Although there have been discussions about the optimal histogram bin width (Sturges, 1926; Scott, 1979; Freedman & Diaconis, 1981; Wand, 1997; Shimazaki & Shinomoto, 2007), no method is perfect and each has some drawbacks and advantages. If users are able to change the number of bins, seeking optimal bin size in advance does not have much meaning. Therefore the Sturges' formula was adopted for the default bin size because of its low computational cost.

3.2.2.3 How to search

The system will display histogram drawn dynamically using SVG when evaluation group and descriptor are selected (Fig. 3.2). Previously, all descriptors, were displayed just after choosing evaluation group (Fig. 3.1; upper), however this is not rational because 94.7% of searches are done by using less than two descriptors. Users are able to set search criterion by clicking a bin. The histogram can be redrawn without screen transition when descriptor or number of bins is changed (Fig. 3.3), and this feature helps the user to grasp the data distribution more appropriately.

Moreover, users are able to choose one more descriptor. When a second

descriptor is chosen, the system will display a two dimensional histogram (Fig. 3.4). Frequency is expressed by opacity, and opacity is adjusted between 0.05 and 1. Users are able to set search criteria by clicking a square, and it guarantees at least one record in the search results. As the number of bins increases, the histogram becomes like a scattering diagram and this is useful for grasping a correlation between two descriptors (Fig. 3.5). Users are also able to add more descriptors or refine the conditions using traditional forms (Fig. 3.6).

3.2.3 Box plot like icon for individual C&E data

Although C&E data is valuable, its worthiness is not completed without data distribution of the parent population (Fig. 3.7). To solve this problem, box plot like icon for individual C&E data was invented. This icon expresses minimum, maximum, median, 25th percentile, and 75th percentile of parent population and where C&E data stands simultaneously. C&E data does not always exhibit the normal distribution pattern, therefore showing minimum and maximum values is not enough. From the example shown (Fig. 3.8), users will understand that this sorghum cultivar has a characteristic of high dry matter yield and is suitable for biomass use at a glance, because its value is considerably more than the 75th percentile. There are different types of boxplot such as violin plot (Hintze & Nelson, 1998) and Beanplot (Kampstra, 2008), however, they are not suitable for small spaces, therefore these features were not included. For the same reason, inter-quartile range and outliers are not shown in the icon. The computing method for quartile is shown (Fig. 3.9). This function returns 20.25 as first quartile for dataset {7, 15, 36, 39, 40, 41}. This result is different from methods presented on Wikipedia (https://en.wikipedia.org/wiki/Quartile) and same as return value of `quantile(c(7, 15, 36, 39, 40, 41))` in R (https://www.r-project.org/) version 3.3.1.

Table 3.1. Number of requests from evaluation data queries (from April 2013 to March 2016).

Number of descriptors used	Number of requests	Percentage
0	16,490	52.0%
1	12,225	38.5%
2	1,350	4.3%
3	663	2.1%
4	291	0.9%
5	145	0.5%
6	85	0.3%
7	64	0.2%
8	32	0.1%
9	27	0.1%
10	8	0.0%
11	210	0.7%
12	6	0.0%
13	6	0.0%
14	1	0.0%
15	1	0.0%
16	3	0.0%
17	2	0.0%
18	1	0.0%
19	2	0.0%
≧20	121	0.4%

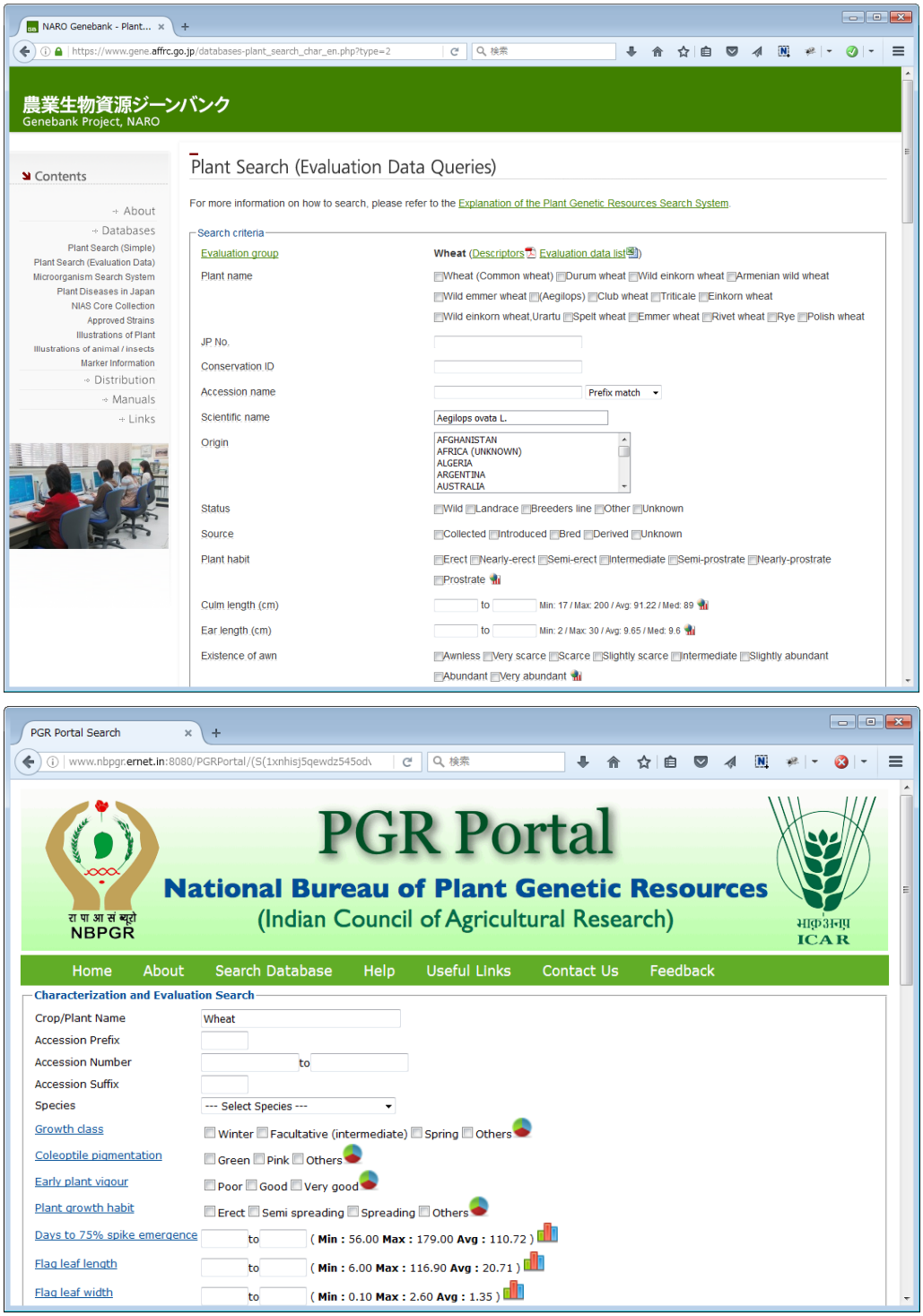


Fig. 3.1 Screenshots of web retrieval systems for C&E data.

(Top: NARO Genebank, Bottom: PGR Portal, India. http://www.nbpgr.ernet.in:8080/PGRPortal/(S(ndxpd5nvzmkufonawbipdi55))/AdvancePlantSearch.aspx?crop=wheat)

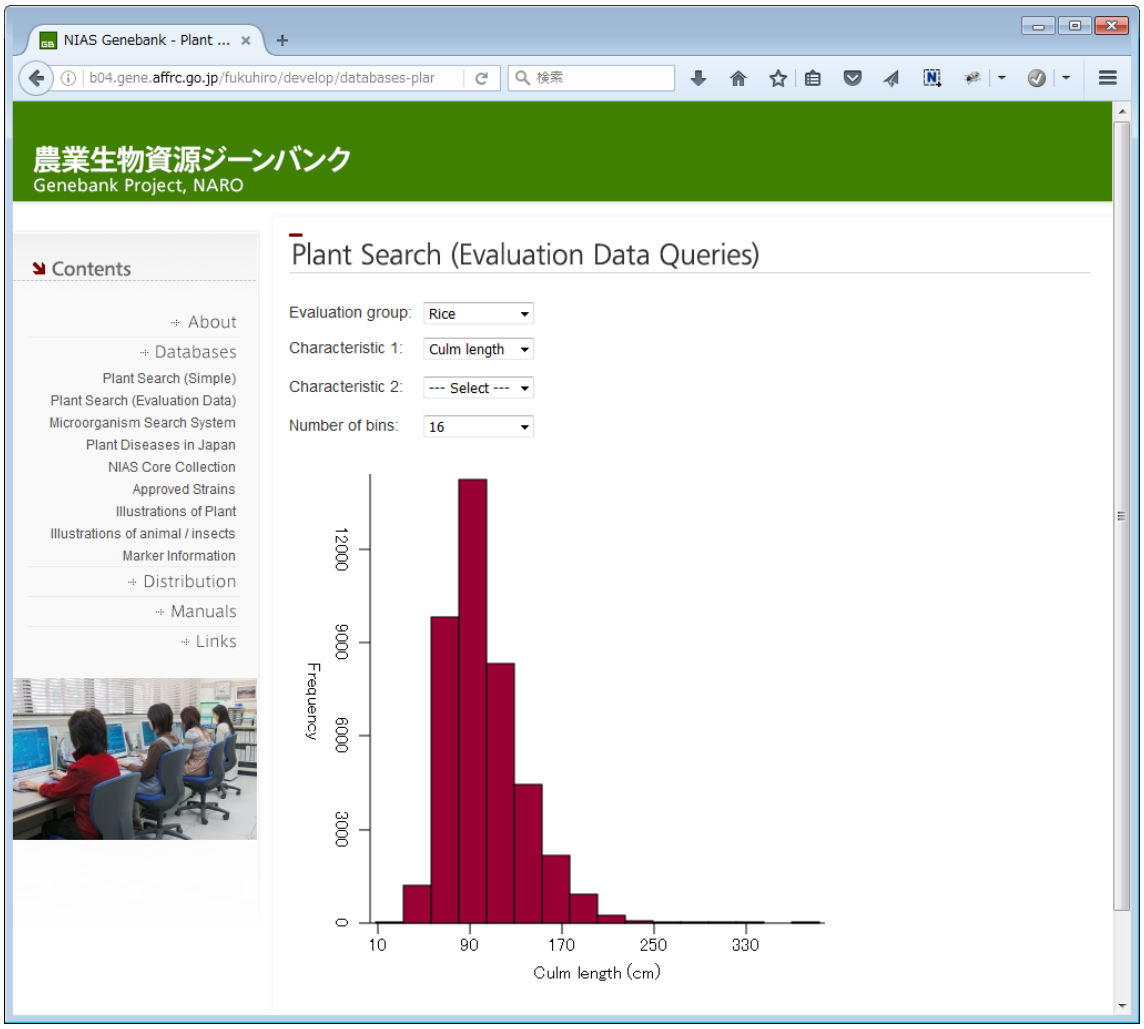


Fig. 3.2. Example of dynamically generated clickable histogram (Culm length of Rice, number of bins = 16).

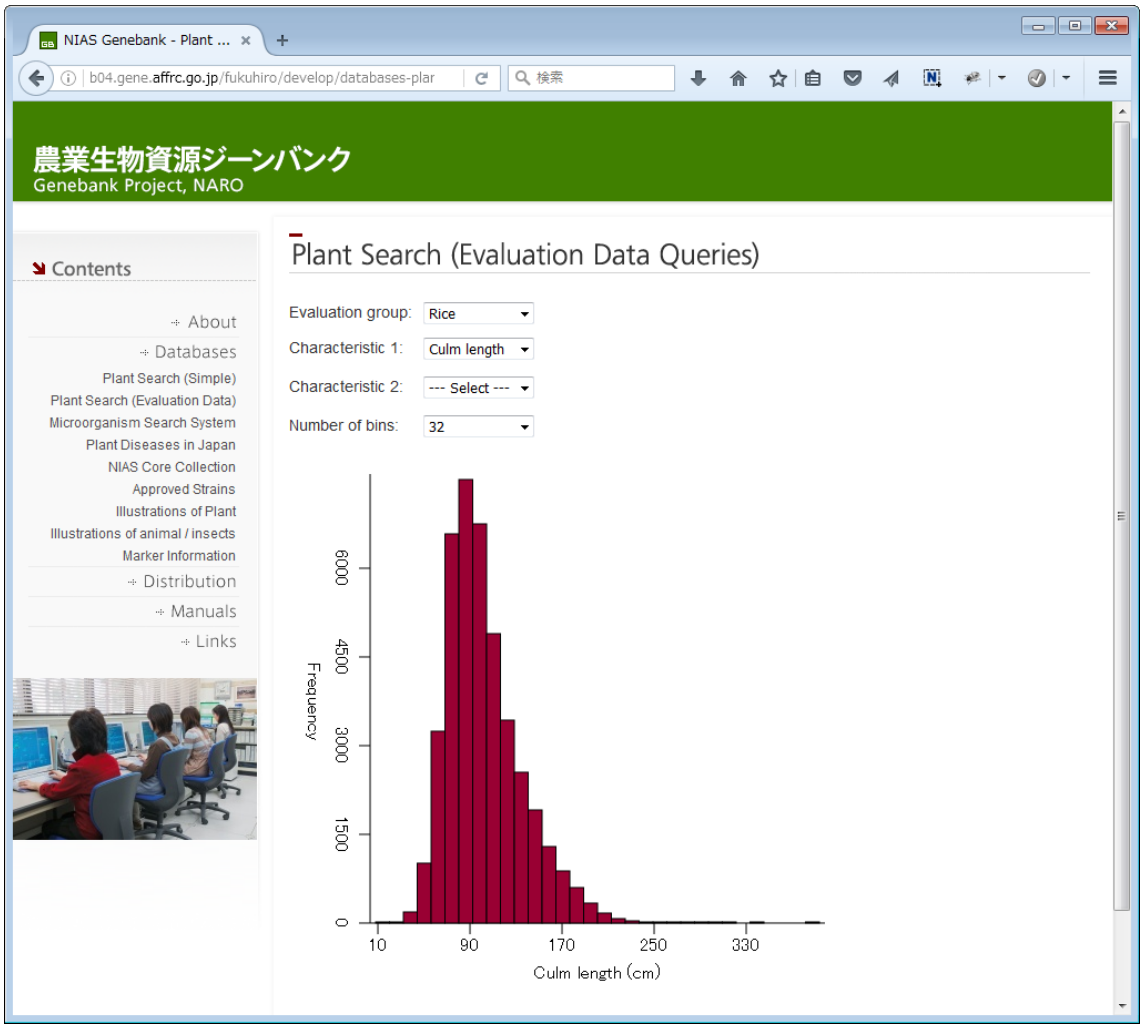


Fig. 3.3. Example of dynamically generated clickable histogram (Culm length of Rice, number of bins = 32). Users are able to change the number of bins.

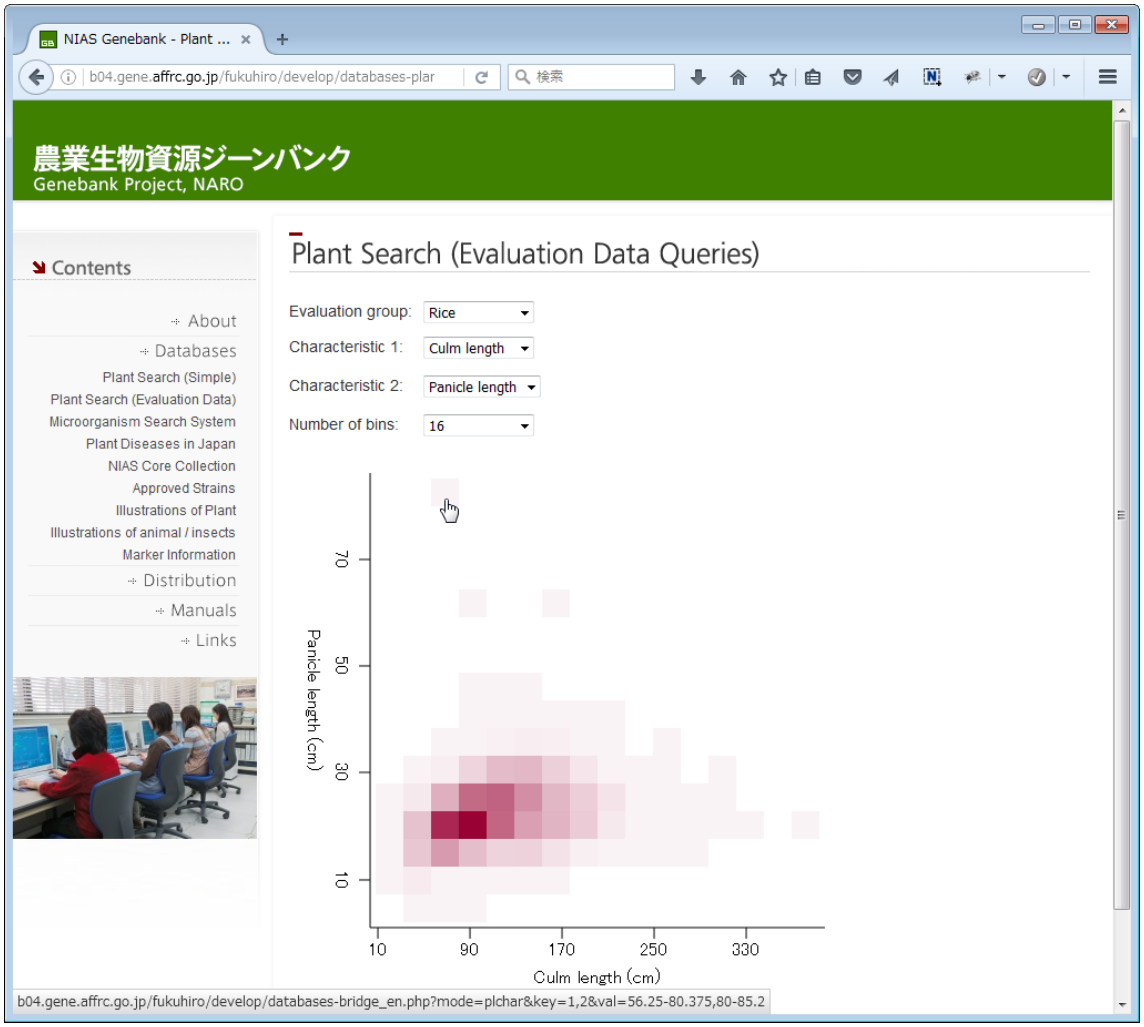


Fig. 3.4. Example of dynamically generated clickable two dimensional histogram (Culm length and Panicle length of Rice, number of bins = 16). Frequency is expressed by opacity instead of height of bins. By clicking on the colored area, accession with traits of interest will be shown.

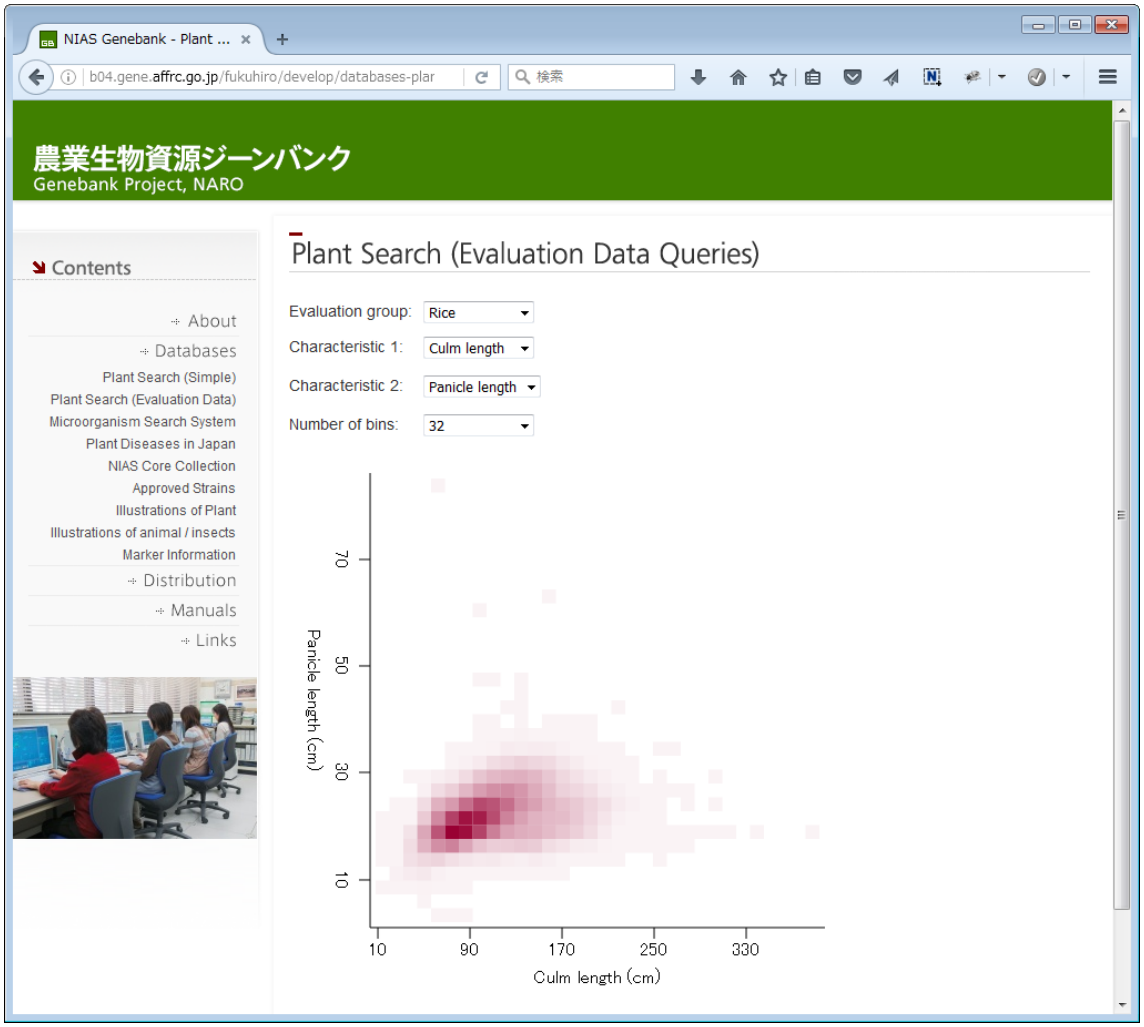


Fig. 3.5. Example of dynamically generated clickable two dimensional histogram (Culm length and Panicle length of Rice, number of bins = 32). As number of bins increase, histogram becomes like a scatter diagram and this is useful for grasping correlations between two descriptors.

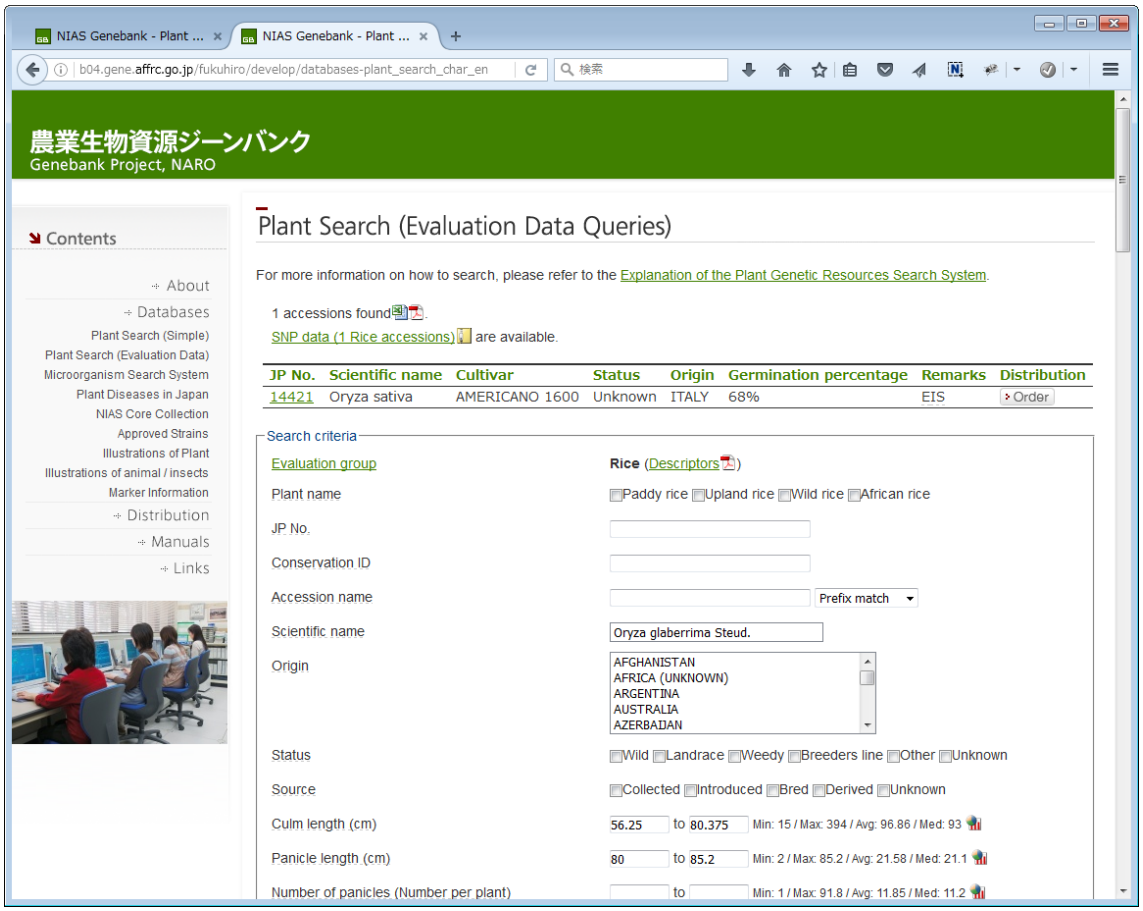


Fig. 3.6. Search results and forms for refined searches. Values are pre-filled on descriptors specified by clicking histogram.

Fig. 3.7. Screenshot of C&E data output in U.S. National Plant Germplasm System. It is not easy to determine if 2.041 is light, average, or heavy for seed weight of sorghum.

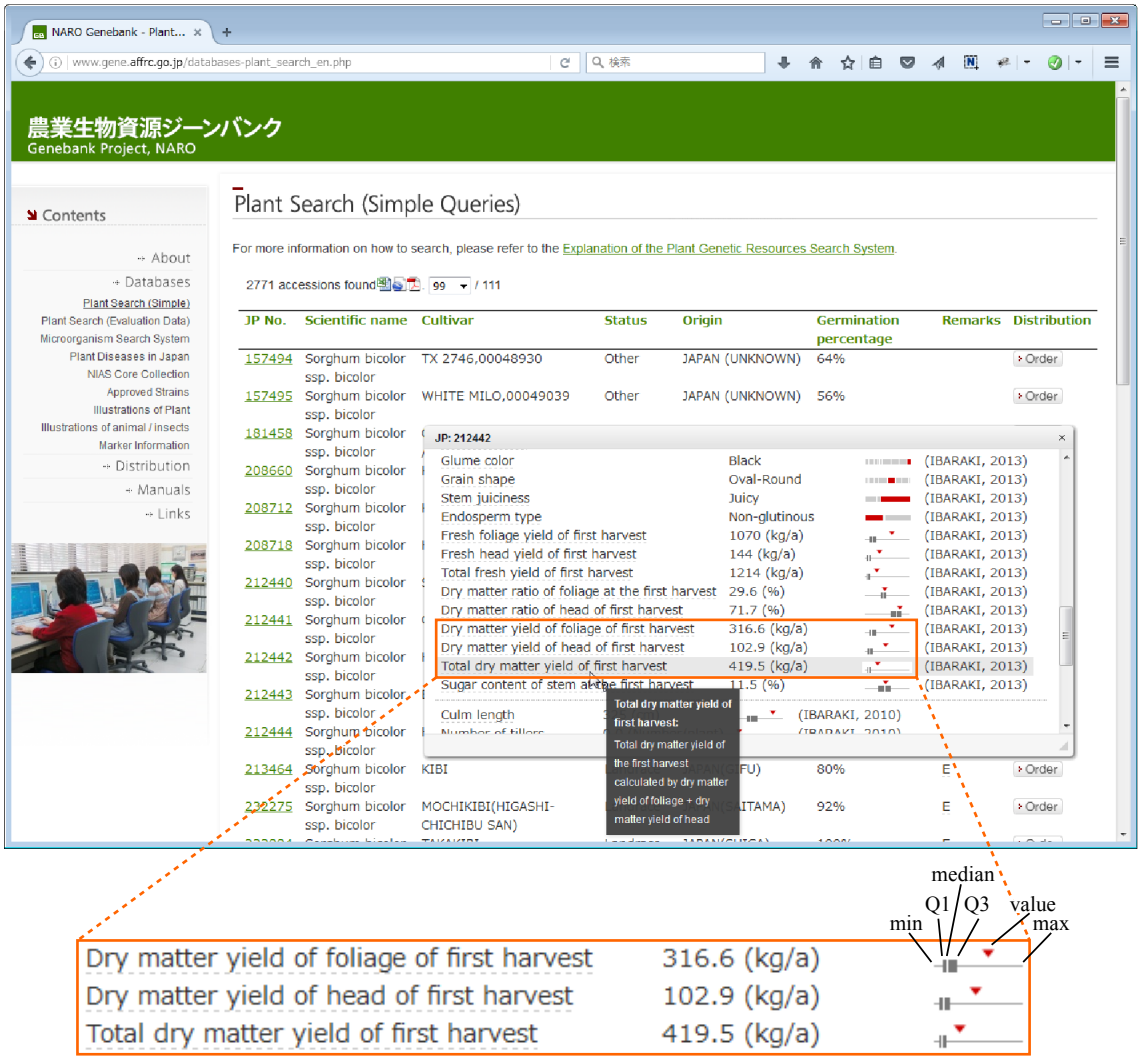


Fig. 3.8. Example of C&E data representation with box plot like icon. Users can understand at a glance that its dry matter yield value is considerably more than its 75th percentile.

```
function get_quartiles($values) {
  $q = array();
  $c = count($values) - 1;
  sort($values, SORT_NUMERIC);

  $q['min'] = min($values);
  $q['max'] = max($values);

  if ($c === 1) {
    $q[1] = $q['min'];
    $q[2] = ($q['min'] + $q['max']) / 2;
    $q[3] = $q['max'];
  } else if ($c % 4 === 0) {
    $q[1] = $values[$c / 4 * 1];
    $q[2] = $values[$c / 4 * 2];
    $q[3] = $values[$c / 4 * 3];
  } else if ($c % 4 === 2) { //average
    $q[1] = ($values[floor($c / 4 * 1)] + $values[ceil($c / 4 * 1)]) / 2;
    $q[2] = ($values[floor($c / 4 * 2)] + $values[ceil($c / 4 * 2)]) / 2;
    $q[3] = ($values[floor($c / 4 * 3)] + $values[ceil($c / 4 * 3)]) / 2;
  } else { //weighted average
    $q[1] = ($values[floor($c / 4 * 1)] * 0.75) + ($values[ceil($c / 4 * 1)] * 0.25);
    $q[2] = ($values[floor($c / 4 * 2)] * 0.5) + ($values[ceil($c / 4 * 2)] * 0.5);
    $q[3] = ($values[floor($c / 4 * 3)] * 0.25) + ($values[ceil($c / 4 * 3)] * 0.75);
  }

  return $q;
}
```

Fig. 3.9. PHP code for computing quartiles.

3.3 Association of genetic resources with geographic information and climate data

3.3.1 Using geographic information system

Recording and publishing geolocation data of collecting site is important because it is an essential element of passport data, and also helps genetic resources collectors to make plans for exploration. Some attempts have been made to use GIS (Geographic Information System) for genebank activities (Hijmans *et al*., 1999; Takeya *et al*., 2009), however, it is most effective when integrated to a versatile web retrieval system. Accordingly, a function to export search results of PGR to Keyhole Markup Language (KML) file was developed. Now users are able to find PGR using Google Earth (https://www.google.co.jp/intl/en/earth/) (Fig. 3.10). In addition, a function to display inline google maps in the detailed page of PGR was provided (Fig. 3.11). It will help users to check collection site of the PGR and its ecological condition by aerial photograph without starting applications separately.

3.3.2 Using climate data

3.3.2.1 Acquisition of climate data and conversion to MySQL

C&E of PGR is expense and time consuming (Frankel, 1989; FAO, 2014), though if climate data of habitat are available, we are able to predict some traits such as cold tolerance and drought resistance. Collecting site and climate data was associated using Worldclim (http://www.worldclim.org/) version 1.4 (2.5 minutes resolution: 21.6225km^2 at the equator). It was generated by interpolation of average monthly climate data for the period 1960-1990 from a large number of weather stations and covers the global land mass except Antarctica (Hijmans *et al*., 2005). As data is distributed in a generic grid data format, it was necessary to convert into a CSV format to import to MySQL. How to obtain and convert using R and the raster package is shown (Fig. 3.12). Serial number is assigned at the beginning of each row, and data is stored in order from north (N90) to south (S60, note that Antarctica is not included), and

west (W180) to east (E180). How to convert from latitude and longitude to serial number is shown (Fig. 3.13).

3.3.2.2 Implementation into web sites

Annual Mean Temperature, Temperature Annual Range (Max Temperature of Warmest Month - Min Temperature of Coldest Month), Annual Precipitation, and Precipitation Seasonality are shown in the detailed page to complement information on collecting site (Fig. 3.11; highlighted part). Now users are also able to search PGR that are collected from arid regions and thus maybe a source of drought resistance (Fig. 3.14).

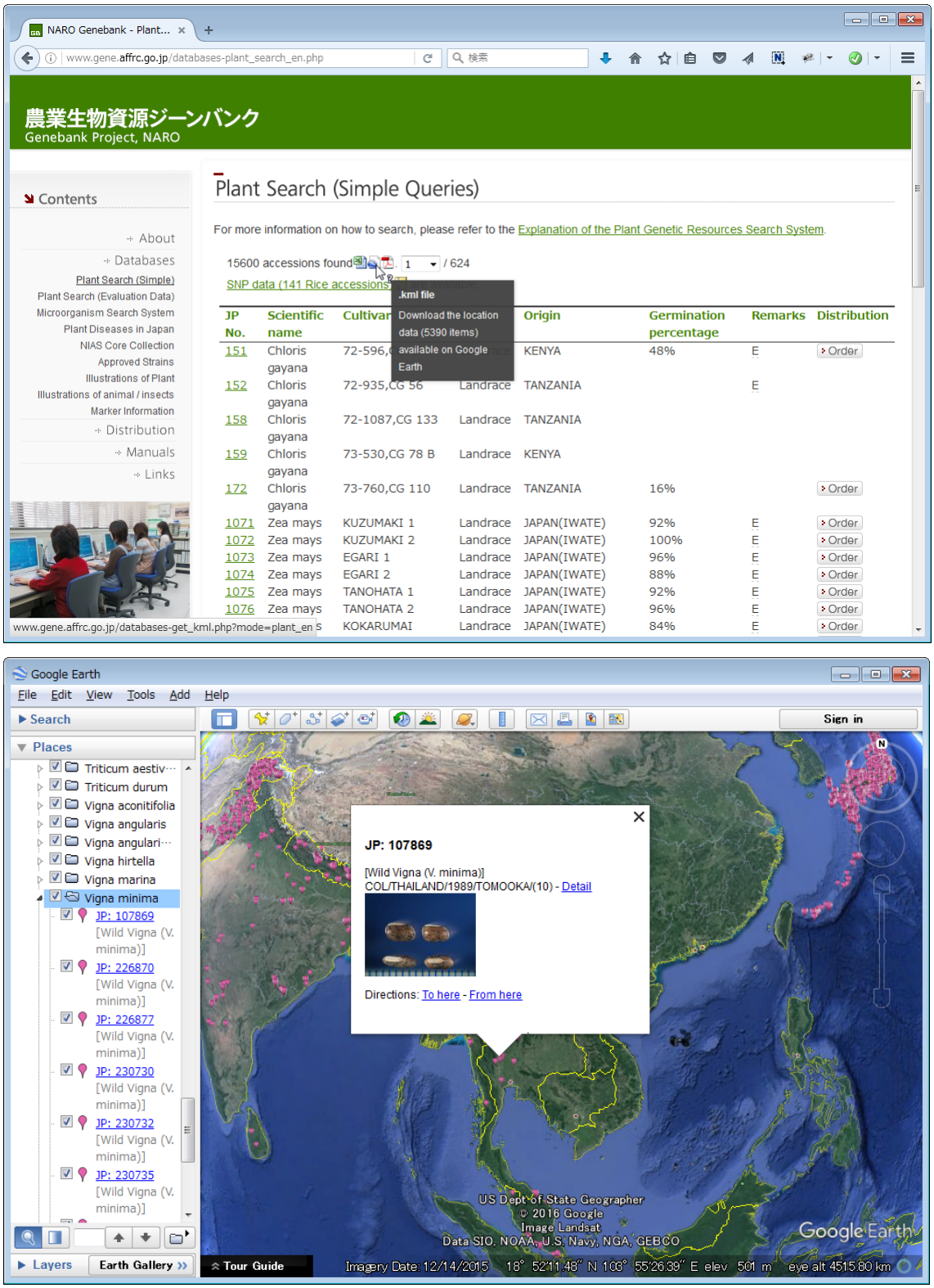


Fig. 3.10. Example of KML download. (Top: Download page, Bottom: Downloaded file)

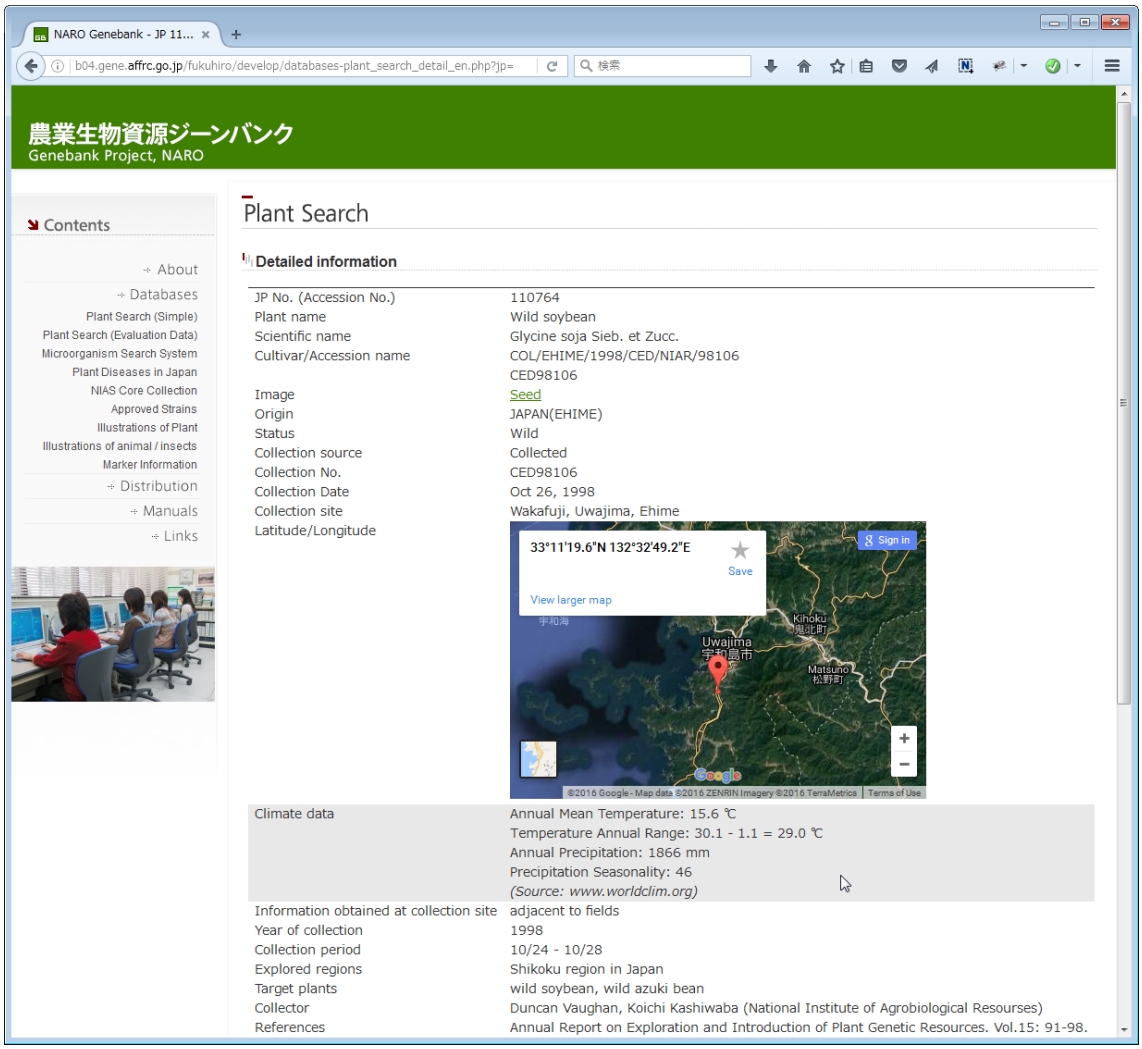


Fig. 3.11. Screenshot of detailed information of PGR with an inline google maps and climate data. Annual Mean Temperature, Temperature Annual Range (Max Temperature of Warmest Month - Min Temperature of Coldest Month), Annual Precipitation, and Precipitation Seasonality (Coefficient of Variation) are displayed.

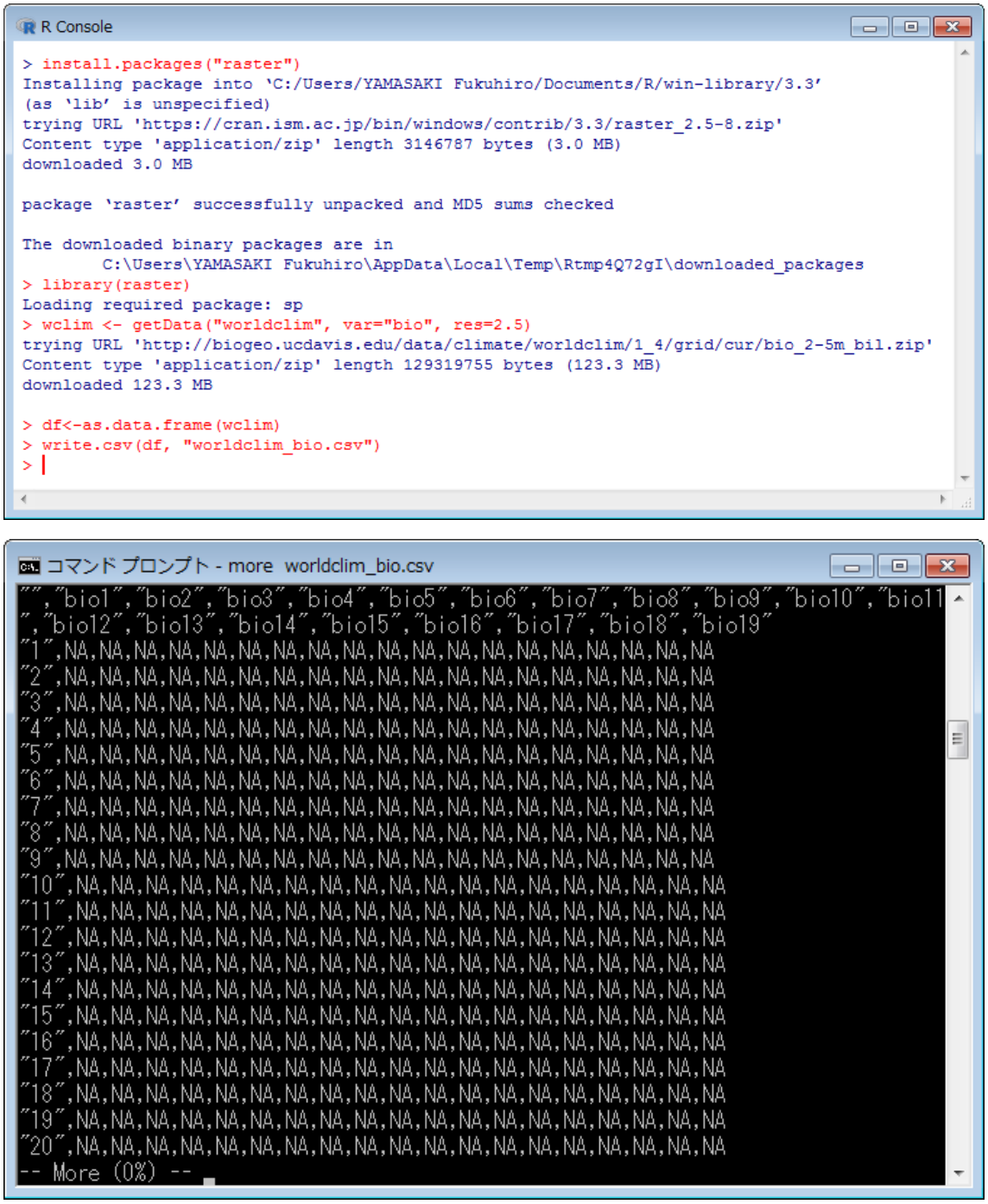


Fig. 3.12. How to obtain and convert worldclim data using R and raster package (top). Content of converted CSV file (bottom). Sequential NA (have to be null) appears in the top of the file because data starts from the Arctic.

```
function latlong_to_meshcode($lat, $long, $res = 2.5) {
   $dpd = 60 / $res;
   $offset_lat = floor((90 - $lat) * $dpd) * 360 * $dpd;
   $offset_long = floor((180 + $long) * $dpd);
   return $offset_lat + $offset_long + 1;
}
```

Fig. 3.13. PHP code for finding meshcode from the latitude and longitude.

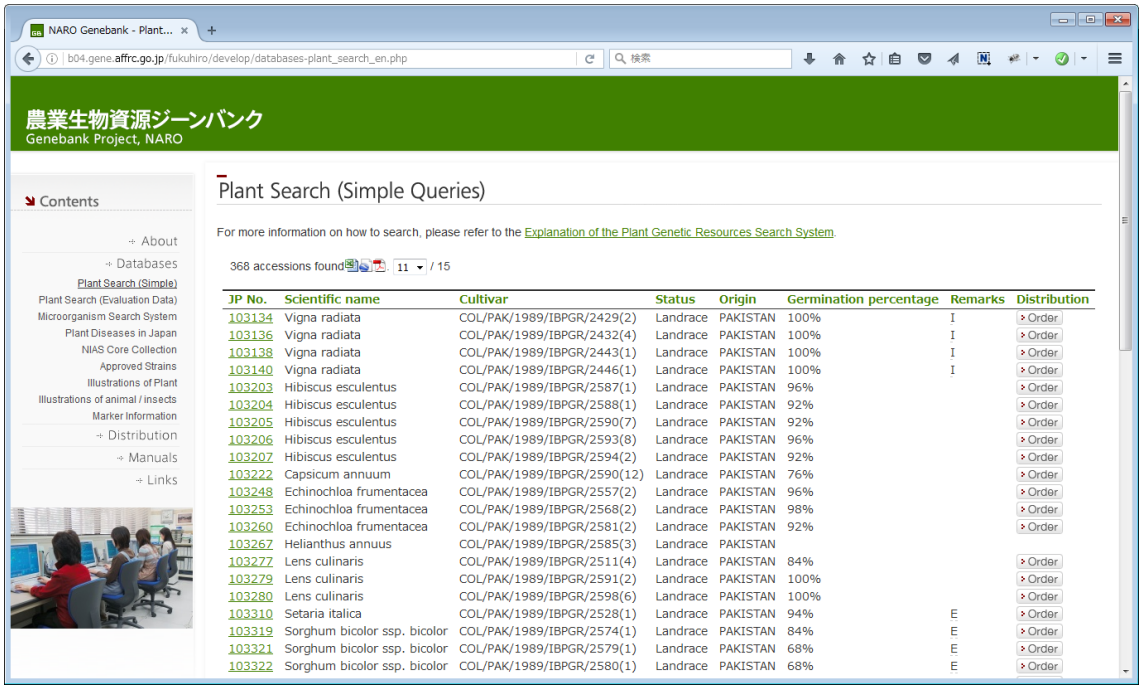


Fig. 3.14. Example of search results by annual precipitation for collecting sites of less than 200 mm. (Top: Search results, Bottom: Detailed page)

3.4 Publication of photographs of microorganism genetic resources

3.4.1 Background

Microorganisms genetic resources are important materials for food processing and study of the plant protection. NARO Genebank is providing information on more than 25,600 accessions of microorganism genetic resources through its website as of October 2016. About 7,900 strains are open to the public with DNA sequence data and about 10,700 strains have links to a plant disease database for improving plant protection and food security (Takeya *et al*., 2012). With the aim of further improving user convenience of the database, photographs of microorganism genetic resources were added that are taken during regular inspection and DNA sequencing.

3.4.2 Implementation

New tables for the photographs of microorganism genetic resources are created in the databases. Image data is stored as binary object, not as the actual file in the file system. In order to maintain performance, binary objects are separated from the main table (Fig. 3.15). A batch insertion program was developed to avoid decrease in efficiency and mistakes caused by manual labor. MAFF number (ID of strain) and type of image is determined automatically from the filename.

Actual display examples are shown (Figs. 3.16, 3.17). MAFF number and scientific name in the image are composited dynamically. This enables updating of information on the image if necessary (i.e. image and text data are separate objects). The dual naming system for fungi will be replaced with one scientific name for each species (McNeill *et al*., 2012), however, the image re-editing operation will not be required for this system.

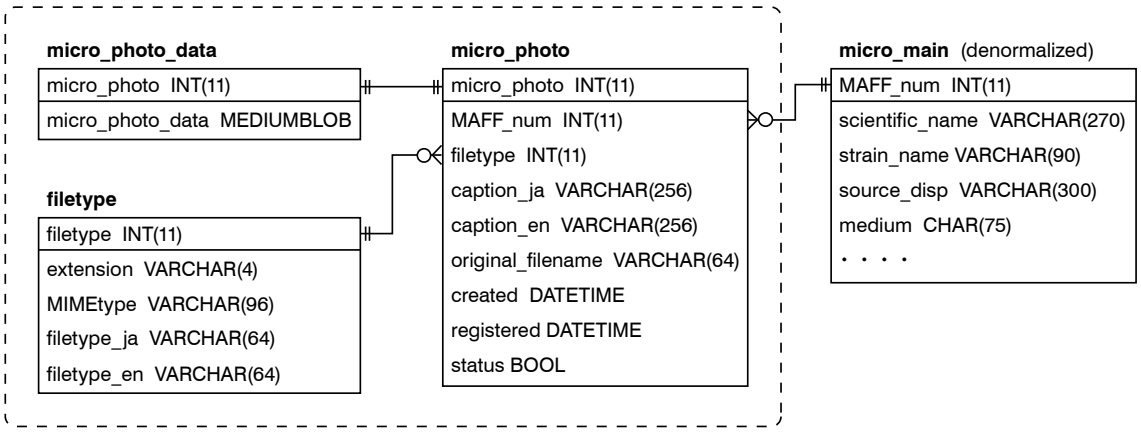


Fig. 3.15. Current entity-relationship diagram for the microorganism photographs. Newly developed parts are encircled by the dotted line.

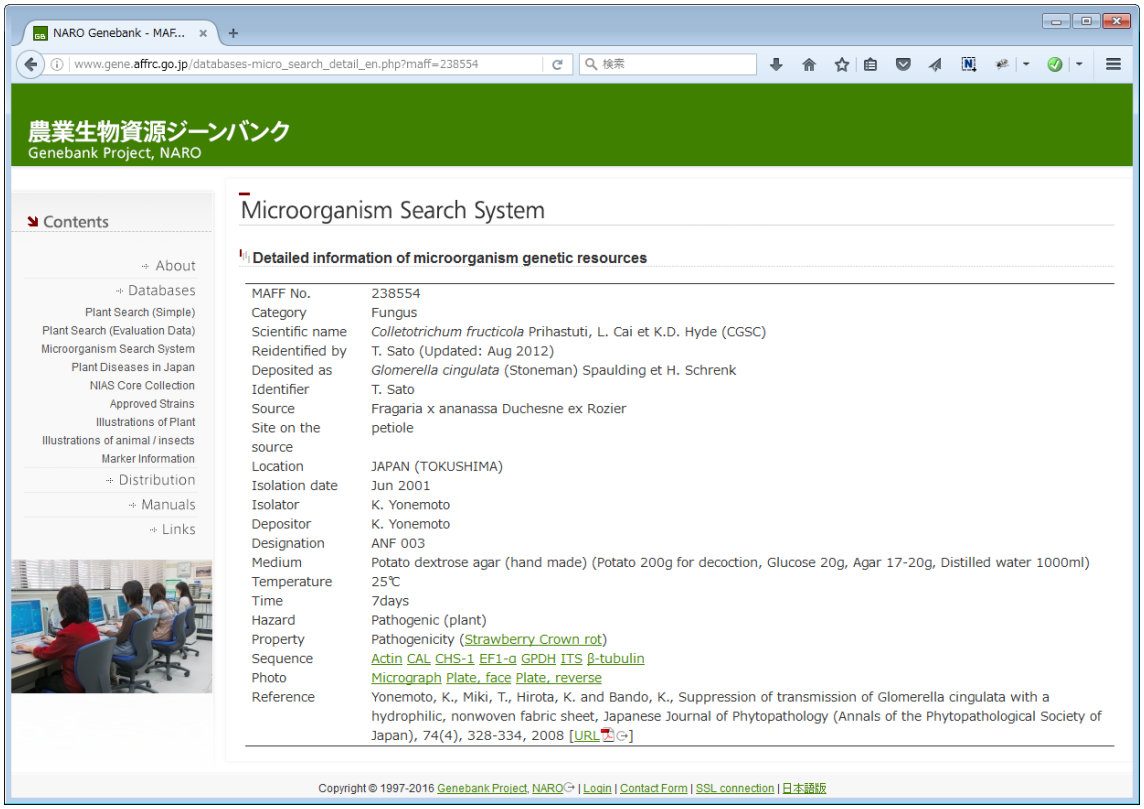


Fig. 3.16. Screenshot of detailed information of microorganism genetic resources. Passport data and links to plant disease database, sequence data, and photographs are available now.

Fig. 3.17. Example of available photograph (Top: photographs taken through the microscope. Bottom: fungal colony on the medium). Recent photographs are taken under stable light conditions.

3.5 Perspectives

By improving user interface including visualization of C&E data, the amount of effort required to find PGR having desired characteristics was reduced. Integrated global climate data will complement some traits such as cold tolerance and drought resistance. Steve Fick and Robert Hijmans are now preparing Worldclim version 2 (http://worldclim.org/version2) that covers period 1970 to 2000, it will be useful not only for updating but also to find the areas with specific climatic conditions.

For microorganisms genetic resources, it has become easier to get morphological information for identification by publishing photographs taken under the microscope and photographs of colonies on the medium. Number of photographs are constantly increasing, more than 5,800 strains are open to public with photographic data as of October 2016. This new feature will promote the diagnosis of fungal plant disease and be of educational use as well. This system can apply to PDF or other non-image files, therefore results of experiments will be also published in the database in the future.

CHAPTER 4 Development of an agricultural field study database: For sharing multidisciplinary information from photographs of genetic resources collected *in situ*

4.1 Introduction

The NARO Genebank is the central coordinating project in Japan for the conservation of plants, microorganisms, animals, and DNA materials related to agriculture. It provides various information on genetic resources including passport, C&E data via its web retrieval database (Takeya *et al*., 2013b) to facilitate user access to conserved genetic resources. A large number of photographs are also available on its website, although most show close-ups of seeds and plants taken during experiment fields or in the laboratories, while very few were taken showing *in situ* conditions in the fields of farmers. This is because the primary nationwide responsibility of the NARO Genebank is to focus on *ex situ* conservation of genetic resources as breeding and research materials for agricultural development. However, *ex situ* conservation removes traditional cultivars from their original habitats (Jarvis *et al*., 2000). The NARO genebank project has been dispatching several overseas exploration teams annually for collecting genetic resources (Okuno *et al*., 2005). Exploration for collecting plant materials from farms and *in situ* conditions is one of the important first steps for introducing genetic resources into genebanks. Agriculture not only produces food for human beings but also serves as the root of culture (Nakao, 1966). Genebanks cannot conserve farming culture directly, but offer a general strategy for collecting in fields that entails questionnaires and photographs (Hawkes *et al*., 2000). Photographs taken *in situ* contain plenty of information, such as the morphology of collected material, geographical and ecological information about the collection site, farming practices, ethnic costumes, people's lifestyles, and cooking procedures. Such photographic records can be beneficial to ethnological studies, as well as agricultural research. By making

genetic resources *in situ* photographs available through the Internet it is hoped that the genetic resources database will be useful to a wider range of research fields. A good but simple precedent is “Sasuke Nakao's Slide Database” on a website linked with the Library and Science Information Center of Osaka Prefecture University (Kojima *et al*., 2001) that publishes numerous photographs taken by Nakao, a Japanese botanist and ethnobotanist during his field trips from 1955 through 1984. The present study goes one step further by designing an “Agricultural Field Study Database” featuring a more user-friendly interface and providing links to relevant genetic resources and the GIS.

4.2 Implementation

4.2.1 Preparation of photographs

About 13,200 photographs taken in South and Southeast Asia during the exploration of genetic resources by Dr. M. Kawase (currently University of Tsukuba) were used as the materials for developing the Agricultural Field Study Database. Given Kawase’s participation in several plant exploration missions to collect crop landraces, observe cultivation practices, and note traditional uses in the areas visited, those photographs are considered ideal materials for constructing this multidisciplinary database related to PGR. As film cameras were used from 1985 to 2006, about 11,850 photographs were digitized as high resolution tiff files with a film scanner for protection against aging degradation. The digitized photographs and about 1,350 images taken recently with digital cameras were appended to the Adobe Lightroom catalog, along with attached keyword tags that represent the photographic subject(s) for enabling efficient search. The shooting date of each film was also recorded as a keyword tag with reference to a date stamp and field notes. Photographs were also exported to smaller web-optimized jpeg files after a quick correction.

4.2.2 Database implementation

A structured query language, MySQL was selected for the RDBMS in our new field study database because it is fast, multi-threaded, multi-user, and robust, and was adopted by the NARO Genebank for its web retrieval system (Takeya *et al*., 2013a).

4.2.2.1 Schema

The current entity-relationship diagram of the Agricultural Field Study Database (Fig.4.1). Photo (independent entity) and tag (dependent entity) have a one-to-many relationship. The details of fs_photo_similarity will be discussed later.

4.2.2.2 Extract data from jpeg files and insert into tables

Keyword tags are stored in the IPTC block (IPTC, 2014). Using PHP with the GD graphic library (http://php.net/manual/en/book.image.php) makes it possible to parse the binary IPTC block into single tags by code (Fig. 4.2).

A jpeg file created by a digital camera usually contains information in the Exif header (CIPA, 2012). For example, when PHP code `$exif = exif_read_data ($source, 'ANY_TAG');` runs, `$exif` will have the keys and values listed in Table 4.1. Latitude and longitude data are stored in degrees, minutes, and seconds (DMS) format, in which decimal degrees are more suitable for treating GIS data; thus, latitude and longitude data were converted into fractional values as follows: `1 * ((21/1) + (145479/10000)/60 + (0/1)/3600)`. It is important to note that when GPSLatitudeRef is “S” or GPSLongitudeRef is “W”, it must be multiplied by -1 to make it negative. Keyword tags, latitude, longitude, altitude, and DateTimeOriginal data were extracted from all files, and then the data inserted into respective tables.

4.2.2.3 Image similarity

To reinforce arbitrary classification by tagging, a computational image comparison was used algorithm—libpuzzle (http://www.pureftpd.org/project/libpuzzle)—based on the relative brightness of image regions (Wong *et al*., 2002) for computing the distance between all photograph pairs, and then inserted the data into table fs_photo_similarity. In almost all cases where a pair of photographs show a quite low distance value close to 0, both are shots of the same subject under the same conditions. It helps in finding duplicated photographs and reduce redundancy in the database. As some photographs have a similar overall composition of a relatively small distance, their computational results will also prove useful for suggesting related photographs (Fig. 4.3).

4.2.3 Web-based retrieval system

4.2.3.1 Top page

The top page shows ten thumbnails to make it easier for users to see what kinds of photographs are available (Fig. 4.4). Samples are chosen randomly, though the frequency of appearance is weighted by SQL (Fig. 4.5). The system tends to choose photographs that have keyword tags associated with many photographs.

Major keyword tags are also shown at the bottom of the Top page as “Popular tags.” The most popular tags appear in bold type (Fig. 4.4). Users are able to move to a “details page” from sample photographs, and may also search for photographs by inputting or selecting a tag.

4.2.3.2 Search results page

The search results page shows a list of thumbnails matching the search criteria. There are ten thumbnails per page, though the asynchronous JavaScript and XML (Ajax) pagination feature will help to see an arbitrary page with efficient data communication. Moreover, once a tag has been selected, a new list is displayed as

“Related tags” instead of “Popular tags” (Fig. 4.6). This new list only displays tags related to the current selection and allows the search to be narrowed down. Selected tags can be deselected by clicking on them.

4.2.3.3 Details page

The Details page shows a large-sized photograph with its additional information and related photographs (Fig. 4.7). Additional information includes keyword tags, shooting date, shooting location, and relevant genetic resources. In cases where latitude and longitude data are available, Google maps appear inline. Related photographs are selected based on common keyword tags or low distance values (Fig. 4.7). Users can narrow down their search by adding a keyword tag, or may move to another details page from related photographs by clicking a thumbnail. When relevant genetic resources exist, a link is generated to detailed pages of the plant search system (https://www.gene.affrc.go.jp/?db_pl_e) that provides passport and C&E data, and an online application form for distribution.

4.3 Perspectives

A web-accessible Agricultural Field Study Database has been developed to share photographs taken during explorations of PGR to the wider research community. The photographs are searchable with keyword tags, and users can obtain information about the shooting date, shooting location, and relevant genetic resources. The major keyword tags and the number of currently published photographs related to those tags are shown (Table 4.2).

This database will be expanded, while giving due consideration to portrait rights, copyrights, religious rules, and other issues. As this system features the efficient insertion of data, many photographs can be added efficiently. Field notes once digitized are expected to be published in a form that allows cross-referencing to photographs in

the next version of the Agricultural Field Study Database. In the future this database will be cross referenced, as much as possible to collecting reports, passport data and genebank accession (JP) number. The enrichment of text information is also important from the perspective of search engine optimization. In terms of image similarity, there is room for improvement because the current version does not consider color information and only finds photographs having a similar overall composition. Adding the analysis of color histograms or a comparison of regions-of-interest could be an effective way to achieve improvement. Giving a high relevance rank to frequently viewed photographs using access analysis may also be useful.

Many indigenous cultivars are threatened as a result of the move away from traditional cultivation systems, therefore conserving them is an urgent task (Castañeda-Álvarez *et al*., 2016). More efforts should be paid to interviewing farmers during field exploration, as sharing cultural information that contributes to conservation of genetic resources. It is expect that this database will become a common platform for sharing information related to field studies by adding content and improving functions.

Table 4.1. Sample output of the exif_read_data function.

Key	Value
ExposureTime	1/125
FNumber	13/1
DateTimeOriginal	2014:07:12 09:26:50
GPSLatitudeRef	N
GPSLatitude	Array([0] => 21/1, [1] => 145479/10000, [2] => 0/1)
GPSLongitudeRef	E
GPSLongitude	Array([0] => 95/1, [1] => 8472/10000, [2] => 0/1)
GPSAltitude	7515/100

Table 4.2. Major tags and the number of related photographs.

Tag	Number of photographs
PAKISTAN	2,757
INDIA	2,087
Landscape	1,159
Mountain	824
KARNATAKA	506
Sesamum indicum	448
River	363
Setaria italica	355
Cattle	354
ANDHRA PRADESH	285
BALTISTAN	241
Oryza sativa	213
Zea mays	201

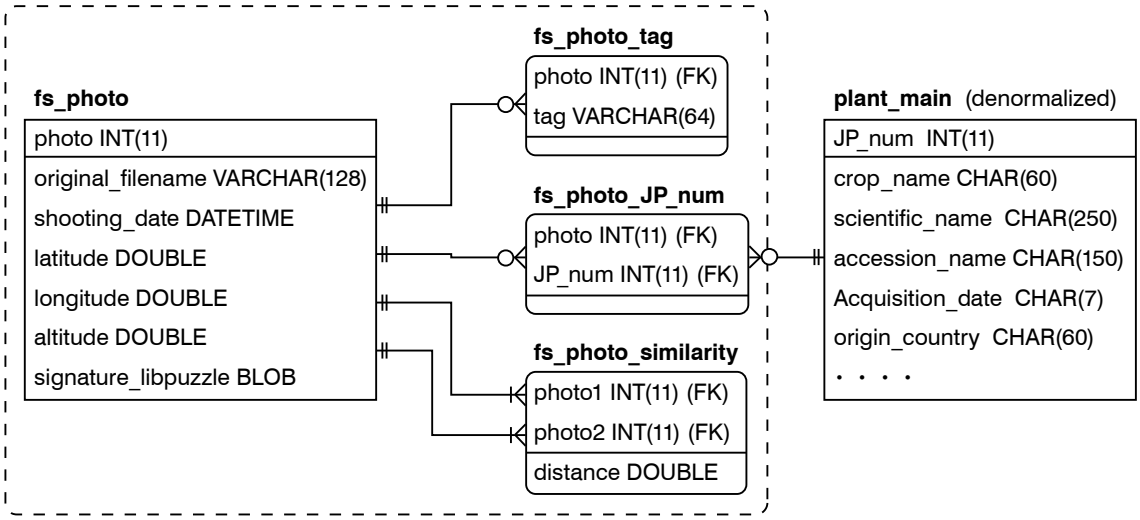


Fig. 4.1. Current entity-relationship diagram of the Agricultural Field Study Database. Newly developed parts are encircled by the dotted line.

```
$source = "example.jpg";
$size = getimagesize($source, $info);
if (isset($info['APP13'])) {
  $iptc = iptcparse($info['APP13']);
  if (isset($iptc['2#025'])) {
    foreach ($iptc['2#025'] as $tag) {
      //insert $tag into DB
    }
  }
}
```

Fig. 4.2. Example PHP code to parse the binary IPTC block into single tags.

Distance: 0.10561578927066
Same subject, high similarity
(Both: Myanmar 2000)

Distance: 0.27077971952578
Different subject, higher similarity
(Top: India 1992, Bottom: Myanmar 2004)

Distance: 0.90451816904023
Different subject, low similarity
(Top: Pakistan 1991, Bottom: Myanmar 2005)

Fig. 4.3. Examples from the computational results of image similarity.

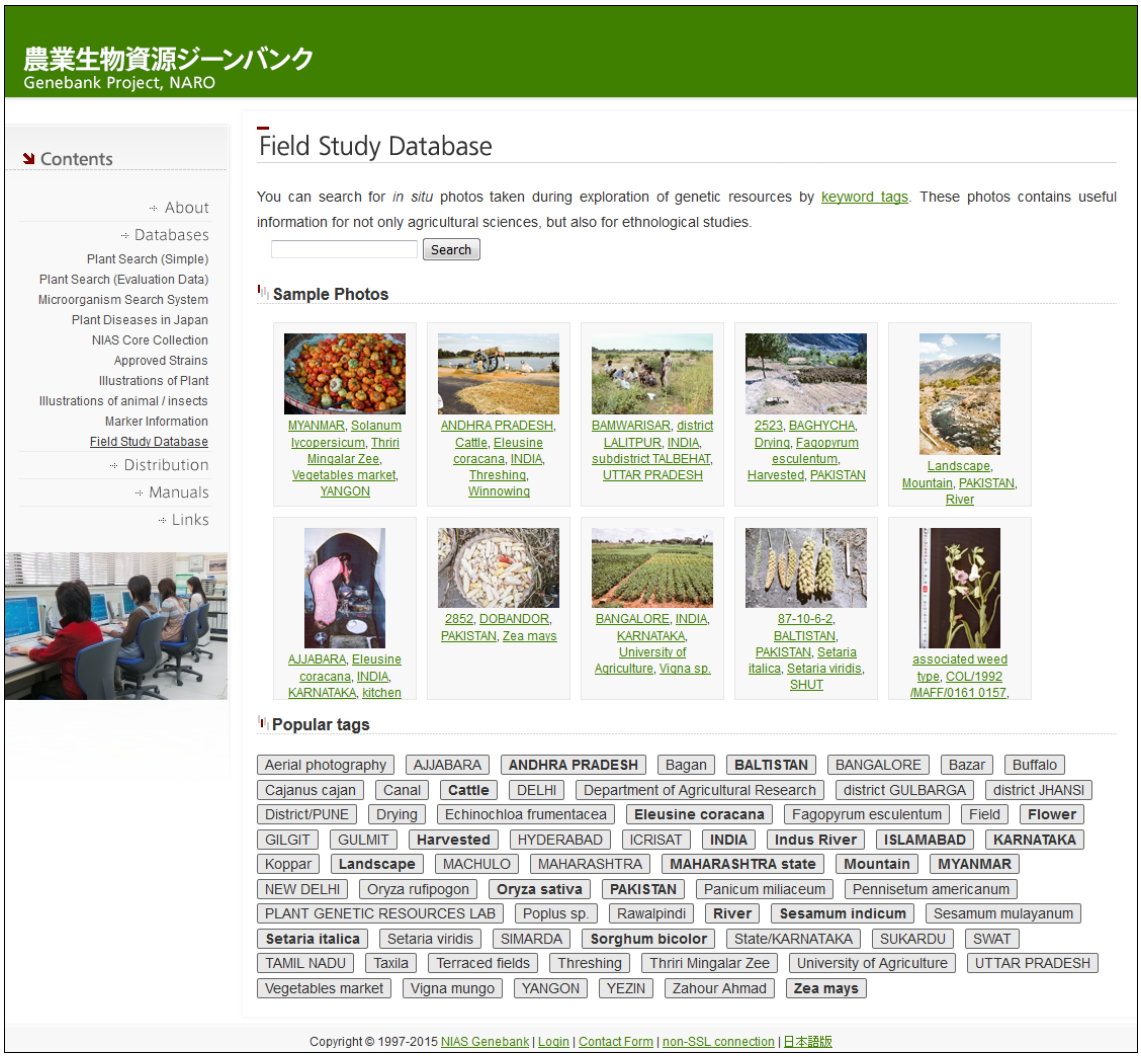


Fig. 4.4. Top page of the Agricultural Field Study Database. Sample photographs and a list of popular tags are shown.

```
  SELECT substring_index(
           group_concat(photo order by rand() separator '\t')
         , '\t', 1) as sample_photo
    FROM fs_photo_tag
GROUP BY tag
ORDER BY count(*) * rand() desc
   LIMIT 10
```

Fig. 4.5. SQL code for sample photograph selection. Ten photographs will be chosen by weighted random sampling.

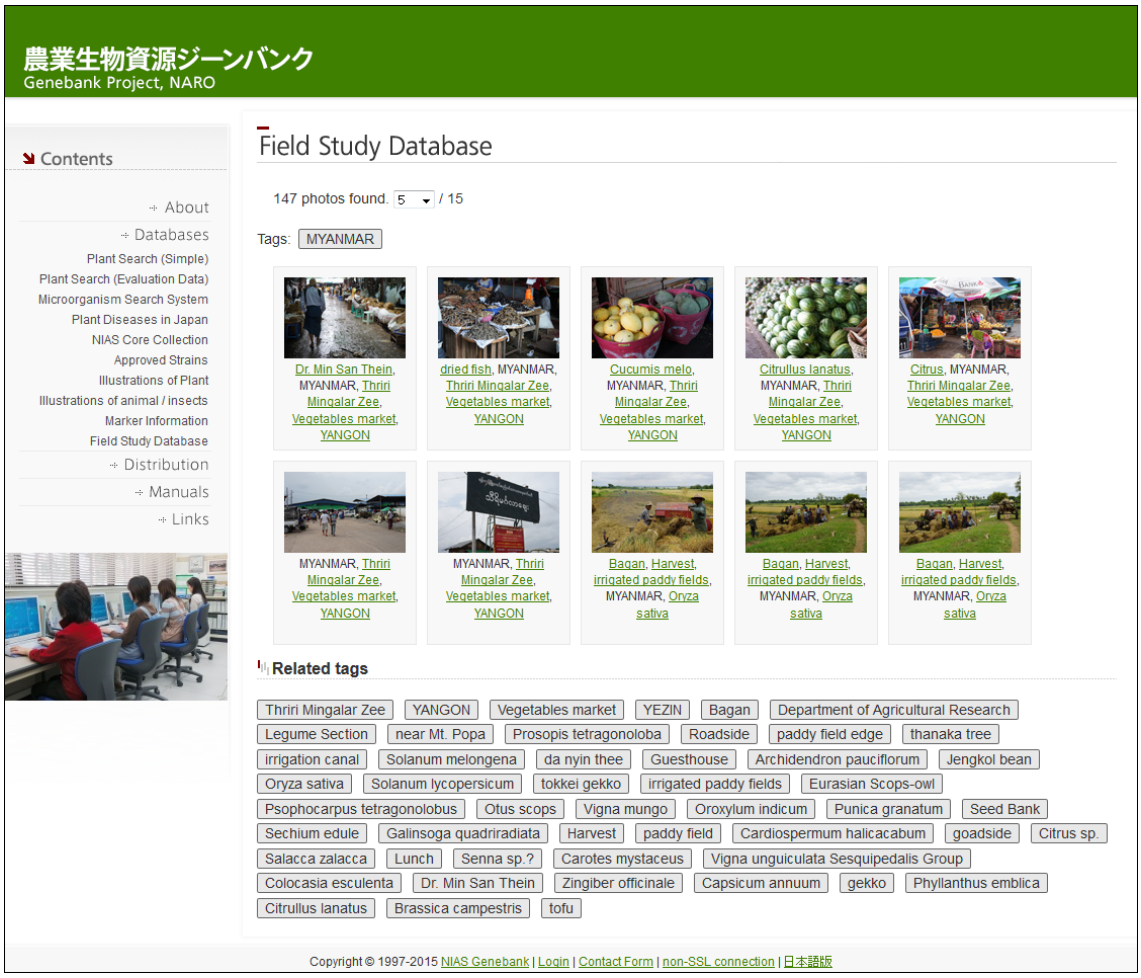


Fig. 4.6. Selected tags, matched photographs, and related tags are shown on the search results page.

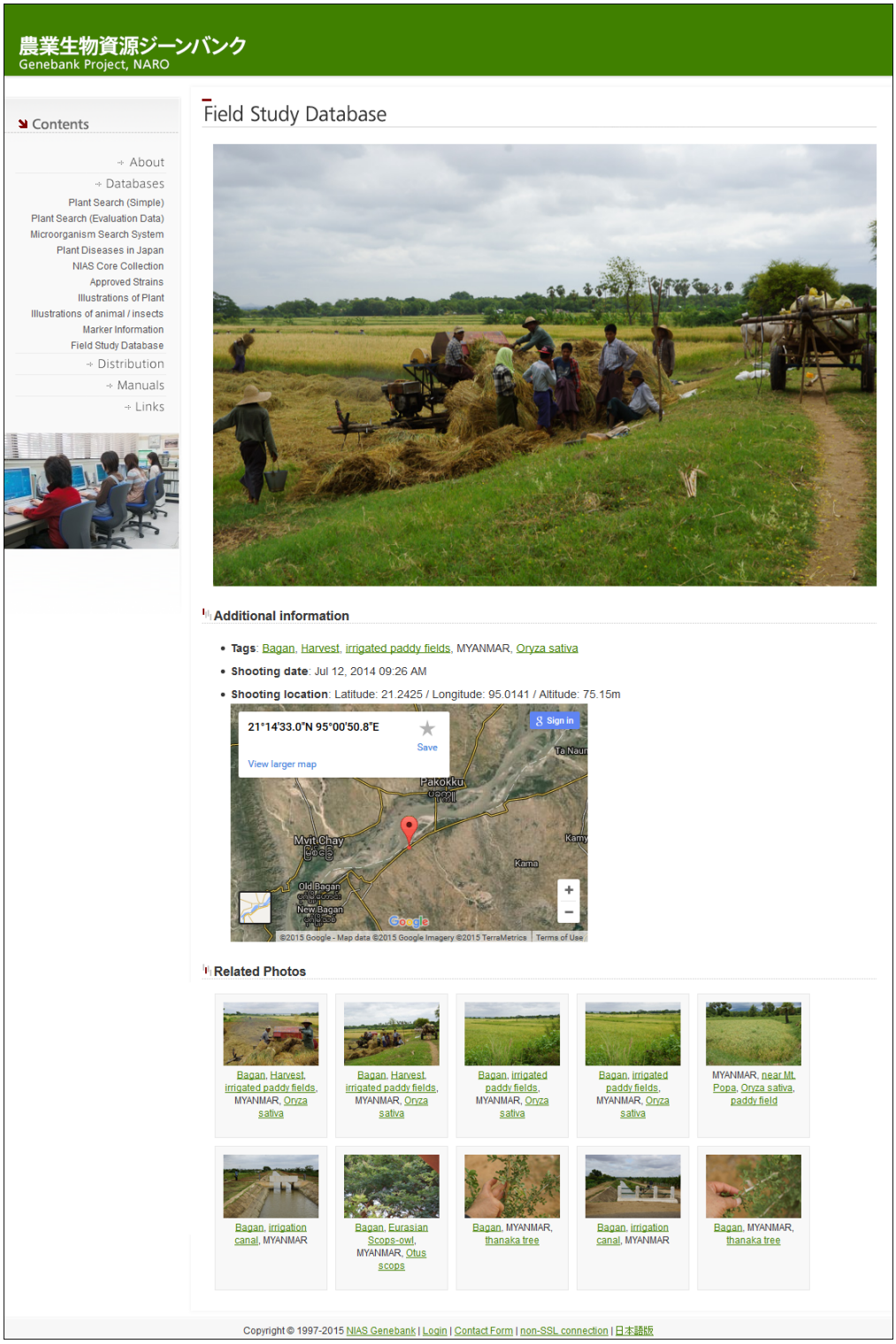


Fig. 4.7. Details page shows a large-sized photograph, additional information also includes inline Google maps (when latitude and longitude are available), and related photographs.

CHAPTER 5 General discussion and conclusions

5.1 General discussion and conclusions

The objectives of this study was promoting utilization of genetic resources and increasing their value as cultural resources through database system improvements. The existing database system for genetic resources lack user friendly searching for C&E data and lacks information on cultural aspects of genetic resources.

In chapter 2, the development of a passport database to manage agriculture, forestry, livestock, microbial, and aquatic genetic resources with an integrated schema was discussed as in the case of CNRG. Only a few genebanks have effective technology for documenting and managing information. Some solutions for managing PGR have been developed, however, CNRG, Mexico, faces the challenge of managing all mentioned subsystems, therefore a new schema and data transition system was needed.

In chapter 3, improvement of the web database and its retrieval system was described. On the grounds that only 5.3% of queries used more than three descriptors for C&E search criteria, a clickable histogram was developed to provide interactive user interface. Users are able to set cultivar specific search criterion to produce a histogram. By clicking on the bin, accession with traits of interest will be shown. A two dimensional histogram is useful to grasp the correlation between two descriptors and it guarantees at least one record in search results. The box plot like icon for individual C&E data has been invented. The icon expresses minimum, maximum, median, 25th percentile, and 75th percentile of parent population and where C&E data stands spatially in a histogram.

Genetic resources have been associated with geographic information and climate data in this study. PGR search results can be exported to a KML file and inline google maps is available in the detailed page. These features will help users to check collection site of the PGR and its ecological condition based on aerial photographs. By

integrating global climate data, some traits will be tentatively identified such as cold tolerance and drought resistance. Users are able to search PGR that are collected from arid regions and thus expected to have drought resistance.

Further improvement of the microorganisms genetic resources database has been achieved. Photographs such as conidiospore and fungal colony on the medium that are taken during regular inspection and DNA sequencing have been added to the microorganism genetic resources database. Number of photographs are constantly increasing, already 35% of fungi have at least one photograph. This feature will promote the diagnosis of fungal plant disease and can be of educational use.

In chapter 4, the lack of *in situ* photographs of PGR in genetic resources databases was addressed. *In situ* photographs contain plenty of information, such as the morphology of collected material, geographical and ecological information about the collection site, farming practices, ethnic costumes, people's lifestyles, and cooking procedures. A web-accessible multidisciplinary database has been developed to share photographs taken during collecting missions with the wider research community. Most genetic resources are found in developing countries (Nabhan, 1989) and those areas are the subject of ethnological studies. Genetic resources can be used as cultural resources as well, therefore, photographs can be used by the wider research community. The achievements of this dissertation are shown in Fig. 5.1.

The advancement of databases for genetic resources is an international challenge. Article 17 of International Treaty on Plant Genetic Resources for Food and Agriculture states that the contracting parties shall cooperate to develop and strengthen a global information system to facilitate the exchange of information, based on existing information systems, on scientific, technical and environmental matters related to plant genetic resources for food and agriculture. It seems cultural aspects are outside the scope of global genetic resources information systems. However, sharing cultural information is also crucial for the conservation of genetic resources. Agriculture and

culture is inseparable. Therefore, we have to continue to advance genetic resources databases for multidisciplinary purposes.

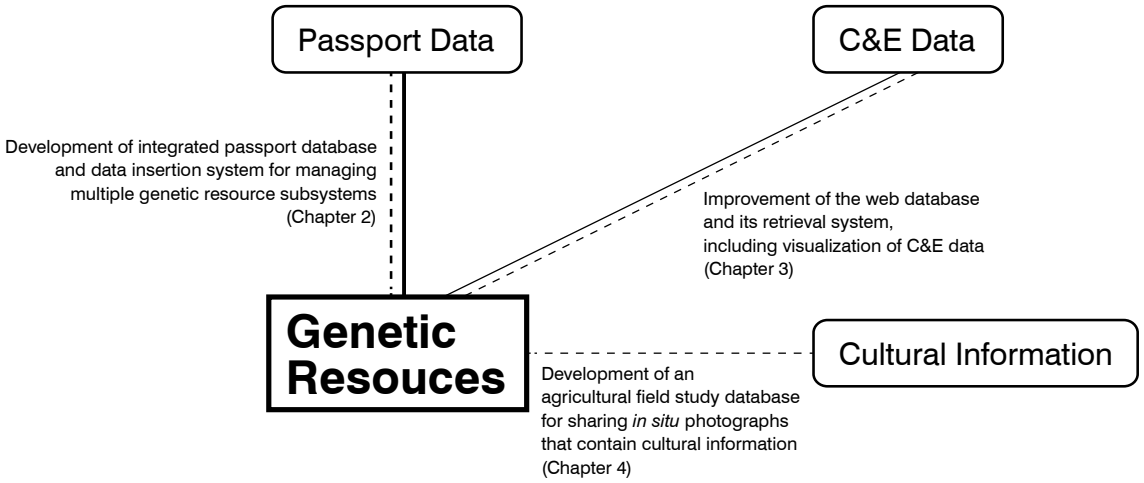


Fig. 5.1. The achievements of this dissertation. The dotted line shows newly established relationships resulting from this study.

Acknowledgments

I would like to express my deepest appreciation to my supervisor, Professor Kazuo N. Watanabe, for his kind guidance and tremendously insightful comments.

I wish to express equal appreciation to Professor Makoto Kawase for his collaboration on the work of Chapter 4 and thoughtful guidance throughout my research and writing.

I wish to thank the Co-Research Advisors, Professor Akira Kikuchi and Associate Professor Motoo Utsumi for their insightful and detailed comments which helped me improve this thesis.

I also thank Dr. Ernesto Borrayo for his collaboration on the work of Chapter 2, with daily encouragement, and challenging questions.

It is a pleasure to thank all the staff at Genetic Resources Center, NARO. In particular: Dr. Hiroshi Nemoto for giving me opportunity to attend the Ph.D. program. Dr. Takeya for his collaboration throughout the past decade and patience. Dr. Toyozo Sato for his collaboration on chapter 3.4. Dr. Toshirou Nagai for giving me initial introduction to agricultural sciences. All the other people who involved genebank project including part-time workers have helped my research and the writing of this thesis.

I would like to pay respect to the relevant parties in Mexico, especially CNRG-INIFAP staff for their collaboration via SATREPS project “Diversity Assessment and Development of Sustainable Use of Mexican Genetic Resources”.

Last but not least, special thanks to my family for their understanding, incessant support and love.

References

Agrawal RC, Behera D and Saxena S (2007) Genebank information management system (GBIMS). Computers and Electronics in Agriculture 59(1): 90-96.

Alercia A, Diulgheroff S and Mackay M (2015) FAO/Bioversity multi-crop passport descriptors v.2.1. Retrieved from http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/

Altieri MA and Merrick LC (1987) In situ conservation of crop genetic-resources through maintenance of traditional farming systems. Economic Botany 41(1): 86-96.

Blackburn HD (2006) The national animal germplasm program: Challenges and opportunities for poultry genetic resources. Poultry Science 85(2): 210-215.

Blackburn HD (2009) Genebank development for the conservation of livestock genetic resources in the United States of America. Livestock Science 120(3): 196-203.

Castañeda-Álvarez NP, Khoury CK, Achicanoy HA, Bernau V, Dempewolf H, Eastwood RJ, Guarino L, Harker RH, Jarvis A, Maxted N, Müller JV, Ramirez-Villegas J, Sosa CC, Struik PC, Vincent H and Toll J (2016) Global conservation priorities for crop wild relatives. Nature Plants 2: 16022.

CIPA (2012) Exchangeable image file format for digital still cameras: Exif version 2.3. Retrieved from http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf

Clark RL, Shands HL, Bretting PK and Eberhart SA (1997) Managing large diverse germplasm collections. Crop Science 37(1): 1-6.

Crop Trust (2016) GRIN-Global: A data management tool for all genebanks. Retrieved from https://www.croptrust.org/blog/grin-global-a-data-management-tool-for-all-genebanks/

Dengler P (2010) Microsoft joins W3C SVG working group. Retrieved from https://blogs.msdn.microsoft.com/ie/2010/01/05/microsoft-joins-w3c-svg-workin

g-group/

Dulloo ME, Hunter D and Borelli T (2010) *Ex situ* and *in situ* conservation of agricultural biodiversity: Major advances and research needs. Notulae Botanicae Horti Agrobotanici Cluj-Napoca 38(2): 123-135.

FAO (2010) The second report on the state of the world's plant genetic resources for food and agriculture. FAO, Rome, 370pp.

FAO (2011) Save and grow: A policymaker's guide to sustainable intensification of smallholder crop production. FAO, Rome, 102pp.

FAO (2014) Genebank standards for plant genetic resources for food and agriculture. Rev. ed. FAO, Rome, 166pp.

Frankel OH and Bennett E (1970) Genetic resources in plants - Their exploration and conservation. Oxford, Blackwell, 575pp.

Frankel OH (1989) Principles and strategies of evaluation. in The use of plant genetic resources (pp. 245-260). Cambridge University Press, Cambridge.

Freedman D and Diaconis P (1981) On the histogram as a density estimator: L_2 theory. Probability Theory and Related Fields 57(4): 453-476.

Hawkes JG, Maxted N and Ford-Lloyd BV (2000) The *ex situ* conservation of plant genetic resources. Kluwer Academic, Dordrecht, 250pp.

Hijmans RJ, Schreuder M, De la Cruz J and Guarino L (1999) Using GIS to check co-ordinates of genebank accessions. Genetic Resources and Crop Evolution 46(3): 291-296.

Hijmans RJ, Cameron SE, Parra JL, Jones PG and Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Hintze JL and Nelson RD (1998) Violin plots: A box plot-density trace synergism. The American Statistician 52(2): 181-184.

IPTC (2014) IPTC Photo Metadata Standard (Revision 3). Retrieved from

https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata

Jarvis DI, Myer L, Klemick H, Guarino L, Smale M, Brown AHD, Sadiki M, Sthapit B and Hodgkin T (2000) A training guide for *in situ* conservation on-farm. International Plant Genetic Resources Institute, Rome, 190pp.

Kampstra P (2008) Beanplot: A boxplot alternative for visual comparison of distributions. Journal of Statistical Software 28(1): 1-9.

Kojima A, Miyamoto T, Tomisaka T, Koike R, Ishii K and Yamano M (2001) Construction of web-based image retrieval and sharing system for digital-archive of valuable cultural material. IEICE Technical Report 101(180): 61-68 [in Japanese].

Koo B and Wright BD (2000) The optimal timing of evaluation of genebank accessions and the effects of biotechnology. American Journal of Agricultural Economics 82(4): 797-811.

Krajewski P, Chen D, Ćwiek H, van Dijk ADJ, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B and Weise S (2015) Towards recommendations for metadata and data handling in plant phenotyping. Journal of Experimental Botany 66(18): 5417-5427.

Kruckenberg M and Pipes J (2005) Transaction processing. in Pro MySQL (pp. 69-103). Apress, Berkeley.

Machida-Hirano R, Cortés Cruz M, Shirata K, Castillo Martínez CR, Niino T, De La Torre Sánchez JF, Kawase M, Fernández RS and Watanabe K (2014) Diversity assessment and development of sustainable use of Mexican genetic resources: Prospects of a SATREPS project. Tropical Agriculture and Development 1(58): 37-41.

McNeill J, Barrie FR, Buck WR, Demoulin V, Greuter W, Hawksworth DL, Herendeen PS, Knapp S, Marhold K, Prado J, Prud'homme van Reine WF, Smith GF,

Wiersma JH and Turland N (2012) International code of nomenclature for algae, fungi, and plants (Melbourne Code). Retrieved from http://www.iapt-taxon.org/nomen/main.php

Mittermeier R (2005) Hotspots revisited. University of Chicago Press, Chicago, 392pp.

Nakao S (1966) Saibai shokubutsu to noukou no kigen. Iwanami-shoten, Tokyo, 192pp [in Japanese].

National Research Council (1993) Managing global genetic resources: Agricultural crop issues and policies. National Academy Press, Washington, D.C., 480pp.

Okuno K, Shirata K, Niino T and Kawase M (2005) Plant genetic resources in Japan: Platforms and destinations to conserve and utilize plant genetic diversity. Japan Agricultural Research Quarterly 39: 231-237.

Oppermann M, Weise S, Dittmann C and Knüpffer H. (2015) GBIS: The information system of the German genebank. Database 2015: bav021.

Painting KA, Perry MC, Denning RA and Ayad WG (1995) Guidebook for genetic resources documentation. IPGRI, Rome, 296pp.

Postman J, Hummer K, Ayala-Silva T, Bretting P, Franko T, Kinard G, Bohning M, Emberland G, Sinnott Q, Mackay M, Cyr P, Millard M, Gardner C, Guarino L and Weaver B (2010) GRIN-GLOBAL: An international project to develop a global plant genebank information management system. Acta Horticulturae 859: 49-55.

Rands MRW, Adams WM, Bennun L, Butchart SHM, Clements A, Coomes D, Entwistle A, Hodge I, Kapos V, Scharlemann JPW, Sutherland WJ and Vira B (2010) Biodiversity conservation: Challenges beyond 2010. Science 329(5997): 1298-1303.

Rao NK (2004) Plant genetic resources: Advancing conservation and use through biotechnology. African Journal of Biotechnology 3(2): 136-145.

Ruredzo TJ (1992) Documentation and use of data in genebank management. Dinteria 23: 55-59.

Sakamoto S (1992) Exploration, collection and conservation of plant genetic resources. Journal of Pesticide Science 17: S213-S219 [in Japanese].

Scott DW (1979) On optimal and data-based histograms. Biometrika 66(3): 605-610.

Schumann GL (1991) The Irish potato famine and the birth of plant pathology. in Plant diseases: Their biology and social impact (pp. 1-24). American Phytopathological Society, St. Paul.

Shimazaki H and Shinomoto S (2007) A method for selecting the bin size of a time histogram. Neural Computation 19(6): 1503-1527.

Soleri D and Cleveland DA (1993) Hopi crop diversity and change. Journal of Ethnobiology 13(2): 203-231.

solid IT (2015) DB-engines ranking. Retrieved from http://db-engines.com/en/ranking

Sturges HA (1926) The choice of a class interval. Journal of the American Statistical Association 21(153): 65-66.

Takeya M, Yamasaki F and Tomooka N (2009) A web-based search and map display system for the integration of collection sites of plant genetic resources with geographic, climatic, and plant characteristic data. Agricultural Information Research 18(2): 82-90 [In Japanese].

Takeya M, Yamasaki F, Tsubokura M and Hattori S (2010) Development of data processing system for NIAS genebank. The IEICE Transactions on Information and Systems 93(10): 1926-1933 [In Japanese].

Takeya M, Yamasaki F, Uzuhashi S, Aoki T, Sawada H, Nagai T, Tomioka K, Tomooka N, Sato T and Kawase M. (2011) NIASGBdb: NIAS genebank databases for genetic resources and plant disease information. Nucleic Acids Research 39(Suppl. 1): 1108-1113.

Takeya M, Yamasaki F, Uzuhashi S, Kumagai M, Sawada H, Nagai T, Tomioka K, Sato T, Aoki T and Kawase M (2012) Development of a database of plant diseases in Japan and a system for making microorganism genetic resources and their DNA

sequence data available to the research community. Japan Agricultural Research Quarterly 46(2): 193-198.

Takeya M, Yamasaki F, Hattori S, Oyanagi C, Chibana T and Tomooka N (2013a) Genebank data-management software incorporating seed-viability test results. Plant Genetic Resources 11(3): 217-220.

Takeya M, Yamasaki F, Hattori S, Kaga A and Tomooka N (2013b) Systems for making NIAS core collections, single-seed-derived germplasm, and plant photo images available to the research community. Genetic Resources and Crop Evolution 60(7): 1945-1951.

Telenius A (2011) Biodiversity information goes public: GBIF at your service. Nordic Journal of Botany 29: 378-381.

Ullstrup AJ (1972) The impacts of the southern corn leaf blight epidemics of 1970-1971. Annual Review of Phytopathology 10(1): 37-50.

W3C (2011) Scalable vector graphics (SVG) 1.1 (second edition). Retrieved from https://www.w3.org/TR/SVG/

W3Counter (2016) Web browser usage trends. Retrieved from https://www.w3counter.com/trends

Wand MP (1997) Data-based choice of histogram bin width. The American Statistician 51(1): 59-64.

Wong HC, Bern M and Goldberg D (2002) An image signature for any kind of image. Proceedings of International Conference on Image Processing 2002: 409-412.

List of Publications

Yamasaki F, Kawase M and Takeya M (2017) Development of an agricultural field study database: For sharing multidisciplinary information on *in situ* photographs. Japan Agricultural Research Quarterly 51(1): 91-97.

Yamasaki F, Borrayo E, Castro-Cortes ME, Martínez Peña MD and Takeya M (2016) Development of a national center of genetic resources passport database: Managing agriculture, forestry, livestock, microbial, and aquatic genetic resources with an integrated schema. Japan Agricultural Research Quarterly 50(4): 387-393.

Yamasaki F and Takeya M (2016) Advancement and provision of information on genetic resources. JATAFF Journal 4(5): 26-30 [in Japanese].

Conference Presentation

Yamasaki F, Kaneko S, Kobayashi Y, Nakajima H, Sato T, Takeya M, Aoki T and Nemoto H (2016) Fungal pictures have been linked to catalogue information of microbial strains in the genebank project, NARO. 23rd Meeting of Japan Society for Microbial Resources and Systematics. Chiba University, Chiba, Japan, July 5, 2016. (Poster)

Summary (in Japanese)

遺伝資源はさまざまな特長を持った新品種を開発するための元となる素材であり，世界の食料供給を支える極めて重要な基盤である。遺伝資源の保存と配布を担うジーンバンクにとって，適切なデータ管理と公開はその業務を遂行するために必要不可欠な要素といえる。世界には現在約 1,750 のジーンバンクが存在するとされ，その内のいくつかは電子的なデータベースを構築し，WWW を通した情報公開を行っているものの，改善の余地は依然として大きい。本研究は，複数のサブシステムに属する遺伝資源を統合的に管理するためのデータベーススキーマの開発，特性評価データの視覚化と検索システムのユーザインターフェイスの改良，微生物遺伝資源の画像データおよび植物遺伝資源の生息域に関する地理・気象・文化情報の蓄積と公開を通して遺伝資源の利用促進と価値拡大を図るものである。

第 1 章では，代表的な遺伝資源データベースの比較を行うとともに，特性評価データを用いて目的にかなった遺伝資源を探すことが難しいこと，および遺伝資源が本来の生息域で持っていた文化的情報が保存されていないことの問題点を指摘した。餅文化の背景なしにモチ米の重要性を語ることが出来ないように，文化的情報が無ければ価値が失われてしまう遺伝資源は多い。

第 2 章では，農業・森林・畜産・微生物・水産といった多様な分野に属する遺伝資源を統合的に管理するための来歴データベースシステムおよび効率的なデータ移行を行うためのプログラムの開発について，メキシコ国立遺伝資源センターにおける実例を基に論じた。植物遺伝資源のみを扱うためのデータベースは既に開発・提供されているが，扱う対象の広さによって適切なスキーマは異なる。

第 3 章では，農業生物資源ジーンバンクにおける WWW 公開用データベースおよびその検索システムの改良について述べた。まず，現行の特性データ検索

システムにおいて3つ以上の特性項目を組み合わせた検索が占める割合が5%程度であることを根拠に，ユーザが選択した任意の特性項目の 1 次元/2 次元ヒストグラムからのインタラクティブな検索機能を実装した。特に，2 次元ヒストグラムは 2 項目間の相関関係の程度を確認しながらの検索が可能になるという点で有用である。また，遺伝資源の詳細情報を表示する画面において，それぞれの特性データ値と母集団の関係が視覚的に把握できるアイコンを考案した。次に，探索により収集した遺伝資源の情報を補完するために開発した地理情報システム(GIS)と連携したシステムについて説明した。GPS により取得した緯度経度を単に GIS 上にプロットするだけではなく，全世界の気象情報などを統合することで，特性調査が完了していない遺伝資源に対しても耐暑性・耐乾燥性などを想定することが可能になった。さらに，微生物遺伝資源の形態的特長を捉えた画像を撮影し公開することにより，菌類病の診断への貢献が期待できるほか，大学・専門学校などにおける図鑑的な利用が考えられることを提示した。

第 4 章では，従来の植物遺伝資源の画像データはそのほとんどが研究室や圃場などの生息域外で撮影されたものであり，本来の生息域で持っていたさまざまな情報が失われていることを指摘し，探索収集の過程で撮影された写真を活用して，生育環境はもちろん，農耕器具や民族衣装，調理方法など幅広い情報を含んだ学際的なデータベースを開発したことを報告した。ジーンバンクのような生息域外保全は在来品種を元のコンテクストから切り離してしまい，農耕文化を直接保存することはできないが，本データベースを通して農学分野のみならず，民族学や文化人類学への貢献も期待できる。これらの学問が主な研究対象としてきた地域には遺伝資源が豊富に存在することが多く，遺伝資源の探索・保存において協働していくことが望ましい。

遺伝資源データベースの充実は国際的な課題であり，食料および農業に用いられる植物遺伝資源に関する国際条約(ITPGRFA)において，締約国は食料農業植物遺伝資源に関連する科学・技術・環境に係る問題に関する情報交換を促進

する世界情報システムの構築に協力することが求められている。しかしながら，とりわけ在来品種については文化的背景を併せて保存しておくことが必要であり，より学際的な視点からデータベースの構築を進めていくことを提案した。