JARQ 50 (4), 387 - 393 (2016) http://www.jircas.affrc.go.jp

Development of a National Center of Genetic Resources Passport Database: Managing Agriculture, Forestry, Livestock, Microbial, and Aquatic Genetic Resources with an Integrated Schema

Fukuhiro YAMASAKI[1], Ernesto BORRAYO[1,2], Ma. Elena CASTRO-CORTES[3], M. Daniel MARTÍNEZ-PEÑA[3] and Masaru TAKEYA[1]*

[1] Genetic Resources Center, National Agriculture and Food Research Organization (Tsukuba, Ibaraki 305-8602, Japan)
[2] Gene Research Center, University of Tsukuba (Tsukuba, Ibaraki 305-8572, Japan)
[3] Centro Nacional de Recursos Genéticos, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (Tepatitlán de Morelos, Jalisco CP 47600, México)

Abstract

Although databases for genetic resources have been developed, a comprehensive schema to manage such multiple subsystems as forestry, livestock, microbial, and aquatic germplasm has yet to be fully developed. Mexico's National Center of Genetic Resources faces the challenge of managing all of these subsystems; therefore, we projected an integral system to meet such demands. In this work, we present the first stage of implementing this system. We have developed a new schema and an input/output system for passport data that will improve the data management of multi-subsystem germplasm. This schema can unify all subsystems with a comprehensive identifier at the accession level. By establishing a sub-identifier, elements that were traditionally allocated as independent accessions would be set as a subgroup of the same accession. The sub-identifier allows variability in element treatments, facilitates diverse sample management, and avoids redundancy. The passport database is complemented by an effective insertion and retrieval system that will facilitate transition from the previous data management of individual subsystems to a subsystem-comprehensive database system.

Discipline: Information technology
Additional key words: batch insertion, gene bank, web-based retrieval system

Introduction

Agricultural biodiversity has been defined as the diversity within biological elements that play a fundamental role in agricultural ecosystem structure and processes (Dulloo et al. 2010). The narrow genetic basis of breeding can be easily subject to catastrophic loss caused by natural disasters, drastic climate change, or epiphytotic disease (Altieri & Merrick 1987, Blackburn 2009). And because the risk of a significant loss of diversity has increased in recent years, animals, plants, and microorganisms have attracted global attention relative to food, agriculture, and the preservation thereof (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 germplasm material alive as long as possible as a strategy to avoid the agrobiological loss of such genetic resources (Dulloo et al. 2010). Adequate genetic resource management in gene banks largely depends on the management of materials, human resources, budgets, and information about the materials (Clark et al. 1997). Such management should be conducted in accordance to the magnitude of a gene bank and the diversity of stored materials.

Mexico is considered to be among the world's most mega-diverse countries (Mittermeier 2005). In 2012, the

This research is supported in part by the SATREPS project conducted by JST and JICA entitled, Diversity Assessment and Development of Sustainable Use of Mexican Genetic Resources, and in part by JSPS Grant-in-Aid 25257416.
*Corresponding author: e-mail katu@affrc.go.jp
Received 26 October 2015; accepted 1 February 2016.

Mexican government established the 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 activities, an appropriate database system for genetic resources is essential. Although several database systems have been developed and implemented by gene banks 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.

CNRG database model

A gradual stage implementation plan was developed under a joint Mexico-Japan 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 projected. The first database would contain all genetic resource passport data (Fig. 1; Passport); the second would contain the locations of genetic resources within the Center (Fig. 1; Storage); the third would contain all other available information regarding the characteristics of each accession such as evaluation, phenotype, genotype, and images (Fig. 1; Characteristics); and the fourth would consist of publicly retrievable data (Fig. 1; Web).

1. Passport Data

(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 resource 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). The National Institute of Agrobiological Sciences (NIAS) Genebank in 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 multi-type particularity; therefore, a new schema was developed to facilitate the development of a five-subsystem germplasm inclusive passport database.

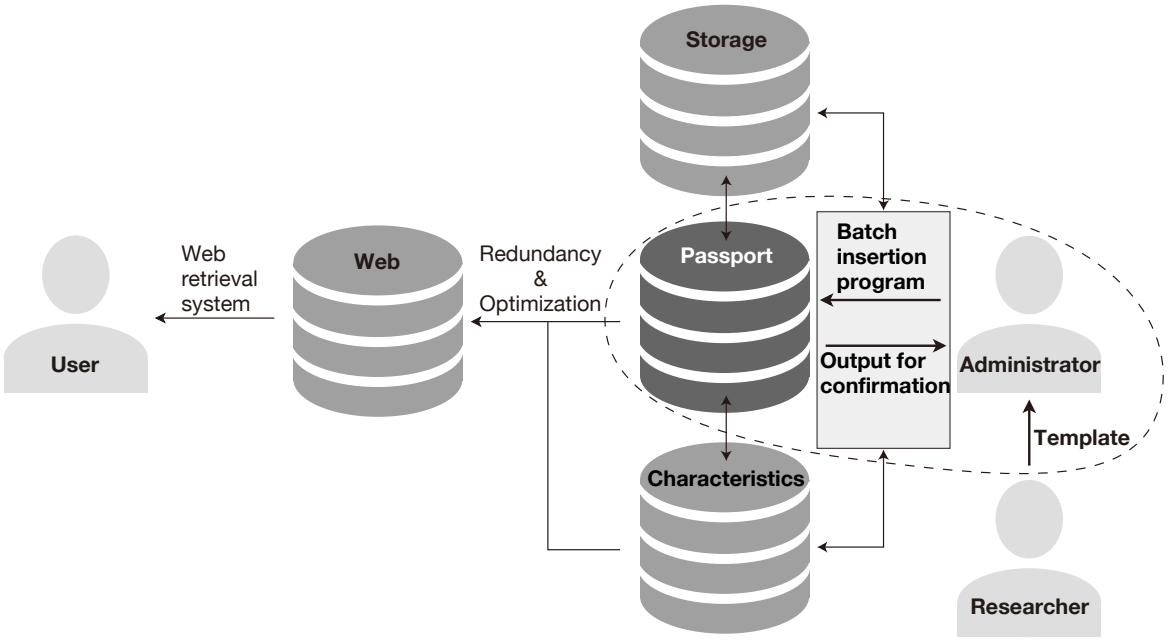


Fig. 1. Projected workflow for the CNRG's genetic resources management database system. The completed passport database and input/output system is encircled by the dashed line.

There are different concepts regarding what types of genetic resource information should be considered passport data. The CNRG adopted EURISCO's list (EURISCO 2002), in addition to data that met the needs of its particular genetic resource 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 shows a conceptual overview of the passport database.

(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

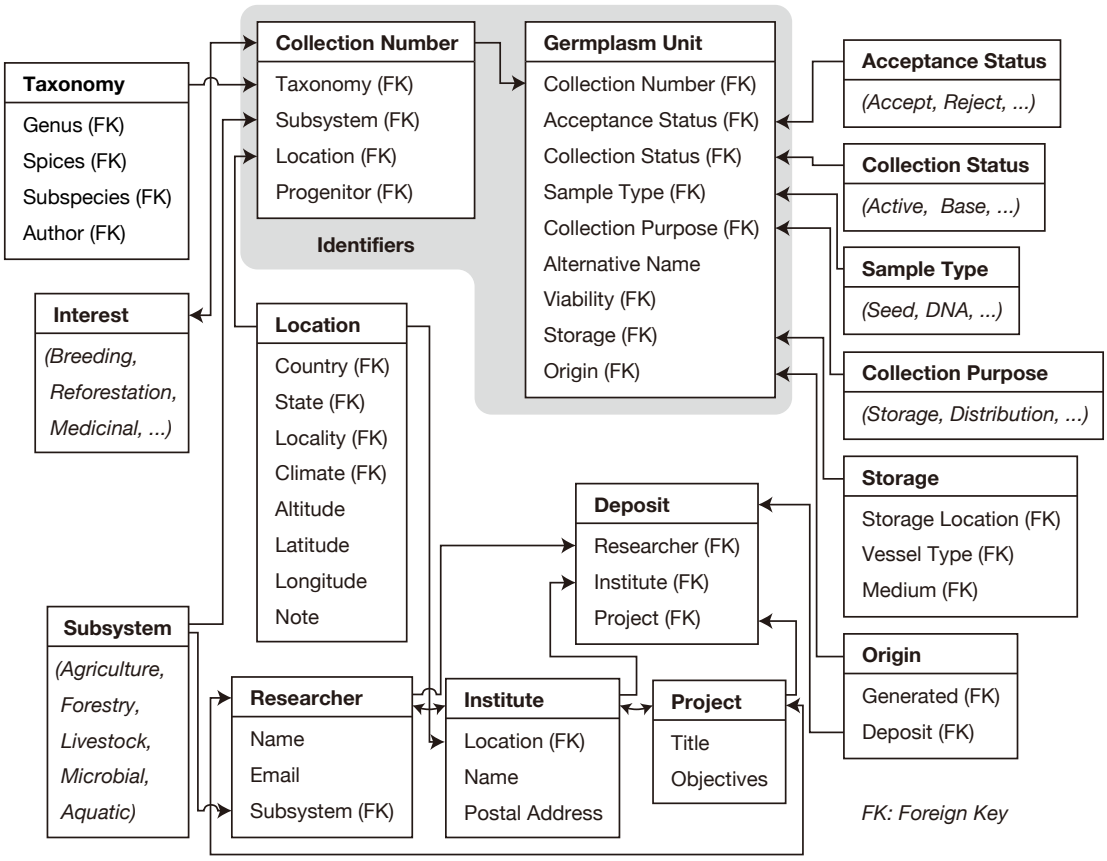


Fig. 2. Conceptual overview of the passport database

this identifier.

(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. 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 relational database management system (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. The NIAS Genebank adopted MySQL for its web retrieval system (Takeya et al. 2013) 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. And because a rollback feature is important for database integrity, the storage engine that we adopt must also include this feature. Given its adequate transaction support, InnoDB was selected as the storage engine (Kruckenberg et al. 2005).

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.

(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 germplasms 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) Batch Insertion Program

A batch insertion PHP (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 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.

4. Data Output

We have already begun to develop a web retrieval system. Currently, the system is not eligible for disclosure and has very 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. 4-6).

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

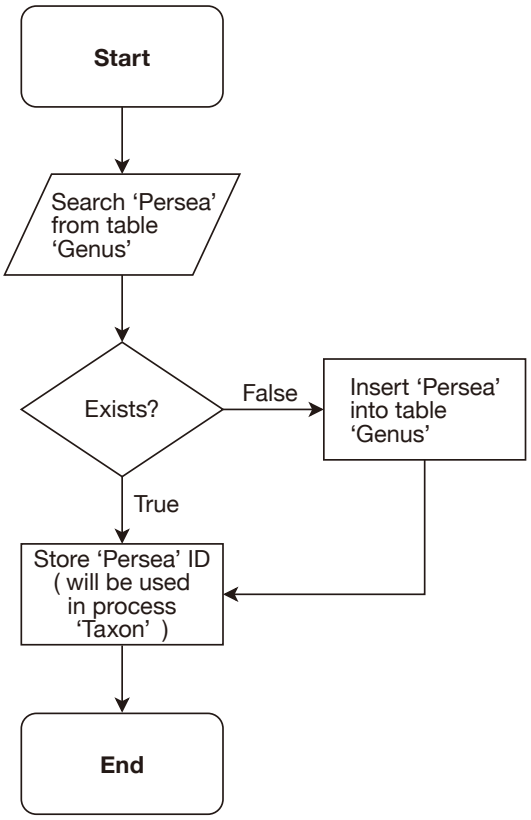


Fig. 3. Example of a batch insertion flow

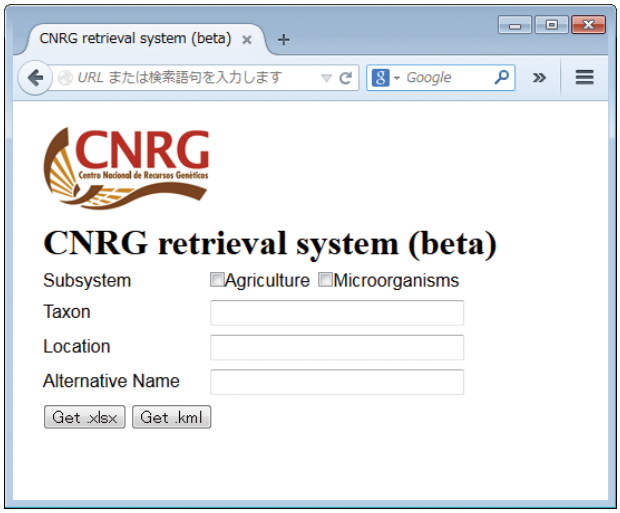


Fig. 4. Output system

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.

Although most genetic resource-related institutions do not have multiple subsystem problems and therefore do not require a comprehensive database, they tend to develop more specific systems (Blackburn 2006). Nevertheless, some institutions have faced this problem and solved it in efficient ways (Takeya et al. 2010).

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 the CNRG in fulfilling its purpose to safeguard and utilize any subsystem genetic resource.

Acknowledgments

We would like to thank CNRG Director José Fernando De la Torre Sánchez for his invaluable support, as well as CNRG researchers R. I. Arteaga-Garibay, J. M. Pichardo-González, C. R. Castillo-Martínez, H. Álvarez-Gallardo, M. A. Cortes-Cruz, L. F. Guzmán-Rodríguez, and A. M. Hernández-Ibánẽz for their feedback on passport data information.

References

Agrawal, R. C. et al. (2007) Genebank Information Management System (GBIMS). *Computers and Electronics in Agriculture*, **59(1)**: 90-96.

Altieri, M. A. & Merrick, L. C. (1987) In situ Conservation of Crop Genetic-Resources through Maintenance of Traditional Farming Systems. *Economic Botany*, **41(1)**: 86-96.

Blackburn, H. D. (2006) The National Animal Germplasm Program: challenges and opportunities for poultry genetic resources. *Poultry science*, **85(2)**: 210-215.

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

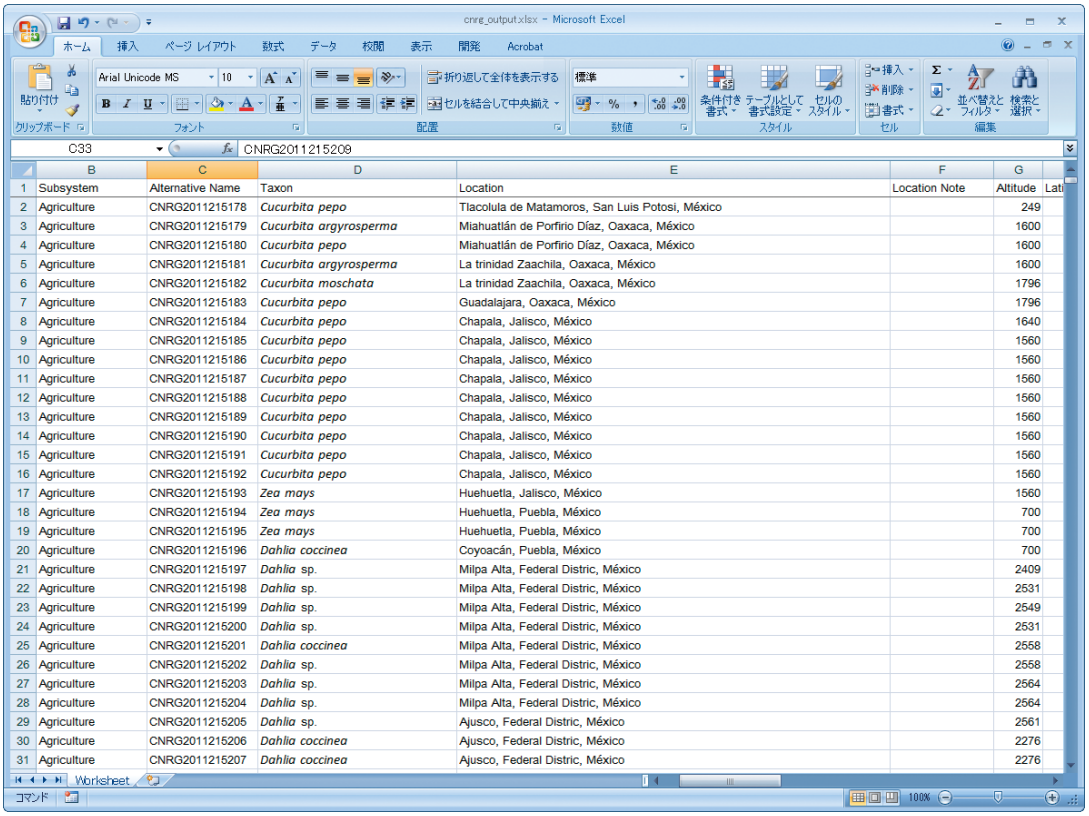
Clark, R. L. et al. (1997) Managing Large Diverse Germplasm Collections. *Crop science*, **37(1)**: 1-6.

Dulloo, M. et al. (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.

EURISCO (2002) Eurisco descriptors. http://www.ecpgr.cgiar.org/fileadmin/templates/ecpgr.org/upload/MISC/EURISCO_Descriptors.pdf.

Kruckenberg, M. et al. (2005) *Pro MySQL*, pp.614.

Machida-Hirano, R. et al. (2014) Diversity Assessment and Development of Sustainable Use of Mexican Genetic Resources: Prospects of a SATREPS Project. *Trop. Agr. Develop.*, **1(58)**:



	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. 5. Output as an .xlsx file

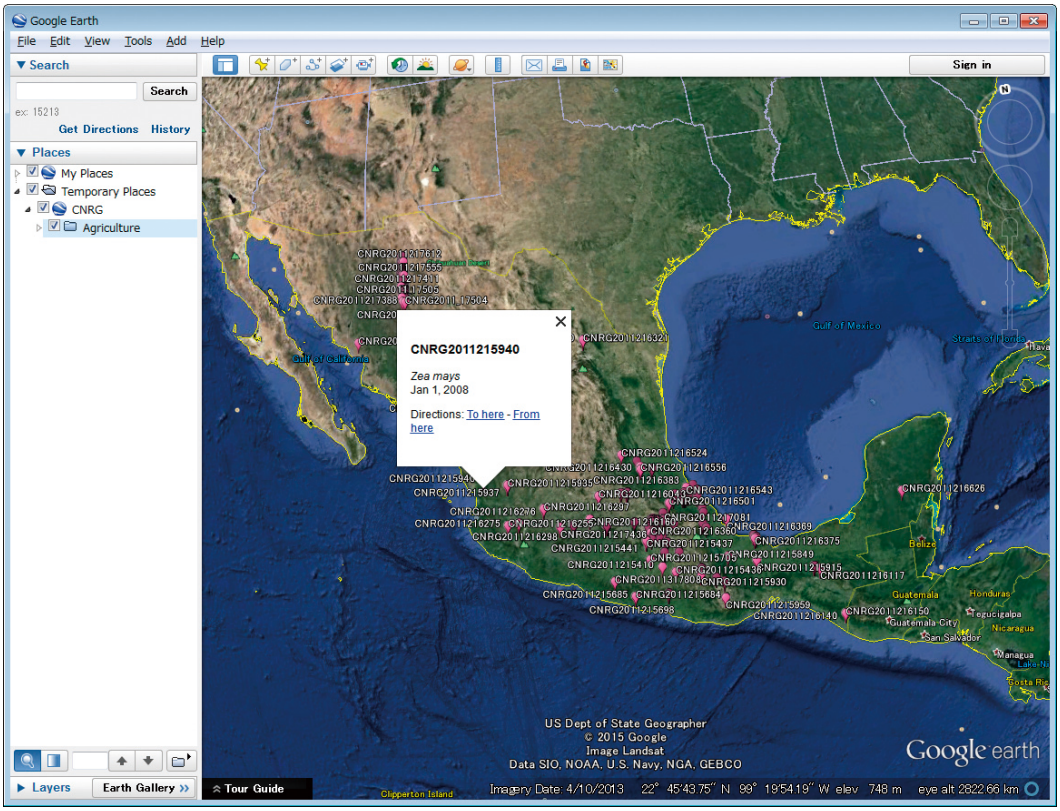


Fig. 6. Output as a .kml file

37-41.

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

Oppermann, M. et al. (2015) GBIS: the information system of the German Genebank. *The Journal of Biological Databases and Curation*, **2015**: bav021.

Postman, J. et al. (2010) Grin-global: An international project to develop a global plant genebank information management system. *Acta Hortic*., **859**: 49-55.

Rands, M. R. W. et al. (2010) Biodiversity conservation: challenges beyond 2010. *Science*, **329(5997)**: 1298-1303.

solid IT (2015) Db-engines ranking. http://db-engines.com/en/ranking.

Takeya, M. et al. (2010) Development of Data Processing System for NIAS Genebank. *The IEICE transactions on information and systems*, **93(10)**: 1926-1933 [In Japanese].

Takeya, M. et al. (2011) NIASGBdb: NIAS Genebank databases for genetic resources and plant disease information. *Nucleic Acids Research*, **39(SUPPL. 1)**: 1108-1113.

Takeya, M. et al. (2013) Genebank data management software incorporating seed-viability test results. *Plant Genetic Resources*, **11**: 217-220.

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