PlantCollections – An Efficient Distributed Database System for Plant Collections

Min HENDERSON[1], Boyce TANKERSLEY[1], David VIEGLAIS[2], Greg RICCARDI[3], Christopher DUNN[4], Dan STARK[5], Pamela ALLENSTEIN[5] and Mike O'NEAL[6]

[1] Chicago Botanic Garden, Glencoe, IL 60022, U.S.A. mcai, btanker@chicagobotanic.org
[2] University of Kansas Natural History Museum and Biodiversity Information Center, Lawrence, KS 66045, U.S.A. vieglais@ku.edu
[3] School of Computational Sciences, Florida State University, Tallahassee, FL 32306, U.S.A. riccardi@ci.fsu.edu
[4] University of Hawaii, Honolulu, HI 96822, U.S.A. cpdunn@hawaii.edu
[5] American Public Gardens Association, Wilmington, DE 19801, U.S.A. dstark, pallenstein@publicgardens.org
[6] BG-BASE, Inc. Topsham, ME 04086, U.S.A. bg_base@bg-base.com

ABSTRACT

PlantCollections – A Community Solution is a hybrid-distributed database system developed to enable curators, research scientists, educators and the general public to access information previously held in inaccessible databases of botanic gardens and arboreta. Information from each institution defined by a federated schema is forwarded to Google Base and Morphbank where it can be accessed by the PlantCollections portal to synthesize reports in tabular, digital and geospatial formats reflecting data originating from living plant collections, herbaria (pressed and dried plants), DNA, images, long term seed storage and pickled or preserved collections. Eighteen institutions with an excess of 47,000 taxa were used to create and test the system.

1. INTRODUCTION

The growing recognition of global climate change and its impact upon natural and cultivated ecosystems has increased the urgency with which stewards of biodiversity collections, like botanic gardens and arboreta, have collaborated to exchange information in an effort to: identify what is in cultivation; create target plant collection lists for those species of conservation concern not in cultivation; study the responses of specific plants across a wide geographic range to the environmental changes; identify genotypes tolerant of the changing climatic patterns for use in ornamental horticulture, restoration ecology and agriculture [2, 4, 5].

Beyond the impact of climate change, plant scientists are in need of tissue samples to study evolutionary relationships. Unlike other organisms, no specific region of DNA has been identified that will differentiate one organism from another at the species or lower levels.

Computer technologies and distributed systems aim to unify multiple databases to enable them to access resources and share information. With the growth of plant collections, implementing and developing efficient computer-based systems that allow data to be accessed and shared becomes a significant concern. In this project, we developed a hybrid-distributed database system, PlantCollections, that allows information from different botanic gardens and arboreta to be accessed and integrated into comprehensive inventories.

The data for the system is provided by 16 American botanic gardens and arboreta, Beijing Botanical Garden in China, and National Trust in United Kingdom. The information of databases is given in Table 1. In the system, (1) Databases are distributed at different locations (and countries); (2) Databases are in six different formats: Access, BG-Base, PICK, FoxPro, Java, and MySQL; (3) Databases contain different fields of data; and (4) A field could have different field names in different databases.

Table 1. Databases of eighteen institutions

Institution	Location	Size	Database	Tech. Support Level
Chicago Botanic Garden	Glencoe, IL	Large	Access 97	Strong
The Huntington Botanic Garden	San Marina, CA	Large	BG-Base 6.4	Strong
Ganna Walska Lotusland	Santa Barbara, CA	Mid-Size	Access 2000	Medium, contractor
University of California at Davis	Davis, CA	Small	BG-Base 6.4	Medium
Missouri Botanical Garden	St. Louis, MO	Large	Access	Strong
The Morton Arboretum	Lisle, IL	Large	PICK	Strong
Mt. Cuba Center, Inc.	Greenville, DE	Large	BG-Base 6.4	Medium, contractor
United States National Arboretum	Washington, DC	Large	BG-Base 6.4	Strong
Arnold Arboretum, Harvard University	Cambridge, MA	Large	BG-Base 6.4	Strong
Landis Arboretum	Esperance, NY	Small	BG-Base 4.0	Weak, volunteer
Norfolk Botanical Garden	Norfolk, VA	Large	BG-Base 6.4	Medium, contractor
The North Carolina Arboretum, University of North Carolina	Ashville, NC	Large	BG-Base 6.4	Medium, contractor
Santa Barbara Botanic Garden	Santa Barbara, CA	Mid-Size	Access 2002	Medium, contractor
San Francisco Botanical Garden	San Francisco	Small	FoxPro	Weak
Scott Arboretum, Swarthmore College	Swarthmore, PA	Small	BG-Base 6.4	Strong
University of Washington Botanic Garden	Seattle, WA	Mid-Size	BG-Base	Strong
Beijing Botanical Garden	Beijing, China	Large	MySQL	Strong
The National Trust	Tiverton, Devon, England	Large	Java	Strong

In the PlantCollections system, we have:

- Defined a federated schema to share information of plants
- Applied URLs to retrieve images stored in Morphbank
- Used Google Base to link databases and publish data online
- Developed a Portal to "pull" data from Google Base and provide fast and correct search

In reporting our work, we present technology and procedures of the system, the results and a summary in the following sections.

2. TECHNOLOGY

PlantCollections is a hybrid-distributed database system for web-based querying that allows information from different botanical gardens and arboreta in a variety of incompatible database formats to be accessed and integrated into comprehensive reports. It is developed by Chicago Botanic Garden [12], the University of Kansas Biodiversity Research Center and Natural History Museum [13], the North American Plant Collections Consortium of the American Public Gardens Association (APGA) [14], and Morphbank [10], Florida State University School of Computational Sciences. The system is funded by a $666,326.00 grant from the Institution of Museum and Library Services National Leadership Grant in the Building Digital Resources category [15].

2.1 DATA MODEL – THE FEDERATED SCHEMA

We defined a common data model, federated schema, to share data of fields among different institutions. Based on surveys from eight audiences comprised of curators, taxonomists, educators, horticulturists, ecologists, weed scientists, conservation scientists and gardeners, our schema consists of 161 fields that were represented in at least four gardens' databases. No institution has all of the fields in their database. In the schema, records in a database are represented as items; the set of attributes available for an item corresponds to an item type; and common attribute values are relationships between items. All fields are categorized into six types of data items: Botanic Garden, Environment Summery, Garden Accession, Sampling Location, Commercial Source, and Plant Propagation. Botanic Garden provides institutional level information about a Botanic Garden; Environment Summary gives general information about the climate at a location; Garden Accession represents detail about a particular accession at a garden; Sampling Location offers detail about the location a specimen was sampled from; Commercial Source is a simple record providing an indication of commercial sources for an accession; and Plant Propagation shows information about how to propagate a particular taxon. The relationships of six item types are shown in Figure 1. More information of the schema is provided in [8].

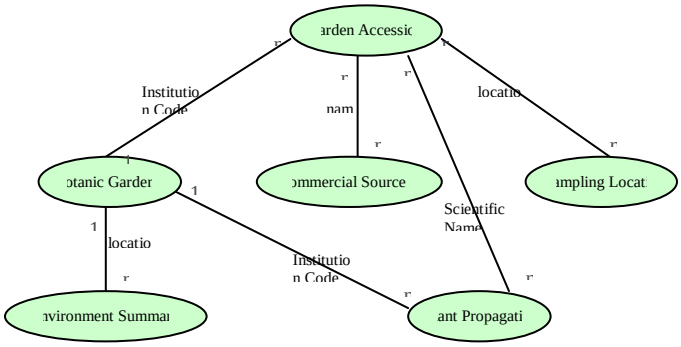

Figure 1: An overview of six item types

2.2 IMAGES

Data in the PlantCollections are represented as numbers, texts, or image URLs. Images provide important information in plant collections that can not easily be transmitted by textual data. Morphbank is one of the premiere repositories for biological images [10]. It holds more than 63,000 public images of more than 4500 different species from all over the world. In the system, images are provided by Morphbank and stored in the database of Morphbank. To improve the system performance, images are represented as URLs to be included in the records and seamlessly linked to the Morphbank.

2.3 GOOGLE BASE(GB)

Distributed information retrieval systems have been used in bibliographic, museum and natural history collections [6]. APGA conducted a survey of technology options to determine which option would be the best for linking many divergent systems in plant collections [14]. This survey identified the Distributed Generic Information Retrieval (DiGIR) system [9], developed by the Biodiversity Research Center at the University of Kansas, as the optimum method to link databases from botanical gardens and arboreta. Tests of a large numbers of simultaneous users of DiGIR systems revealed a significant time lag between entry of a data request and a response. Fortunately, computer technology has progressed rapidly and more efficient search algorithms and software applications have been implemented recently [1, 3, 7]. We have therefore moved to better software applications and found that a more effective, simpler to maintain, much cleaner, and more secure repository for our data, Google Base that currently is provided by Google. Data sets from each institution are uploaded to GB on a monthly schedule permitting faster data transfer from the cached data in GB to the PlantCollections Portal user than was possible using a traditional distributed system.

GB is open source. It brings offline content online and provides us functions for data uploads and dynamic updates. "Per-item" and "bulk upload" are two basic approaches to upload data to GB and it has been shown that "per-item" is useful for maintenance tasks but not effective for large collections' uploads. Google offers a "data API" to query data and manage items programmatically [11].

GB plays an important role to link databases from different botanical gardens and arboreta and make effective searches for the data for PlantCollections. We utilize GB to perform data upload, data deletion and update, and data retrieval in PlantCollections. Data files uploaded to GB are in Atom1.0 format and created in Python2.5 with Sqlites2. Once Atom files are generated, they are uploaded to GB with "bulk upload". The bulk upload process utilizes an attribute "id" to locate items; but a recently detected bug in GB does not allow searching on this "id", so it is necessary to provide an available "id" for searching. Thus, a globally unique attribute "id" that combines "institution_code" (defined from the "Botanic Garden" institutional level data) and "accession_number" of the record is built to identify individual records in a bulk upload file and is used in subsequent uploads to decide if the Google Base record needs to be modified.

Data deletions and updates are performed by the "data API" in GB. Data retrieval in GB is a little difficult due to the limitation of the maximum number of items to be retrieved anonymously in response to a query. To go around this constraint, we assign a

sequential integer value ("indexer") as an identifier for records and use numeric range queries to retrieve blocks of records.

2.4 PORTAL

We developed a portal for "pulling" records from GB. The Portal is implemented in Python and Javascript. The Portal performs two functions: "simple search" and "expert search" (it is also called "advanced search"). "Simple search" uses the most common fields and returns records containing terms that were entered to search for. "Expert Search" utilizes a more complex range of fields to retrieves records that fulfill all of the conditions. The Portal delivers spreadsheets, images, and maps related to records.

3. PROCEDURES

In the system, firstly we collected data matching the federated schema from each individual institution; secondly, we confirmed the mapped fields in the database match the federated schema; thirdly, we created data feeds that are acceptable by GB; fourthly, we uploaded data containing appropriate Morphbank URLs for the images to GB and updated GB with items inserted, modified, or deleted; lastly, we transferred data cached in GB to the Portal and achieve information sharable and searchable online.

4. RESULTS

The following experiments illustrate the performance of the PlantCollections system using datasets consisting of over 47,100 plant taxa in six different database formats: Access, BG-Base, PICK, FoxPro, Java, and MySQL.

First, the system has linked the databases of 18 institutions (16 American botanical gardens and arboreta, Beijing Botanical Garden in China, and National Trust in United Kingdom.) together and published more than 360,000 records of plants in GB. An example is shown in Figure 2. More results can be found at http://base.google.com/base/s2?a_n0=Garden+Accession&a_y0=9&hl=en&gl=US)

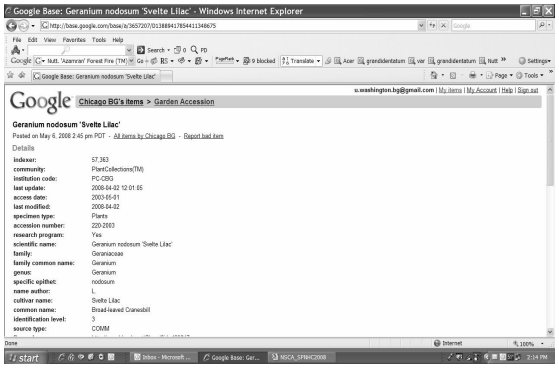

Figure 2. A garden accession item in Google Base

Second, the system provides effective searches for plants. We make a search for all plant with "genus = Geranium" from 18 botanic gardens and arboreta in the Portal and found 1101 records. Moreover, the system enables users to access spreadsheets, images and maps from different institutions. An example is shown in Figures 3-5.

Figure 3. Data and spreadsheets in the Portal

Figure 4. Images in the Portal

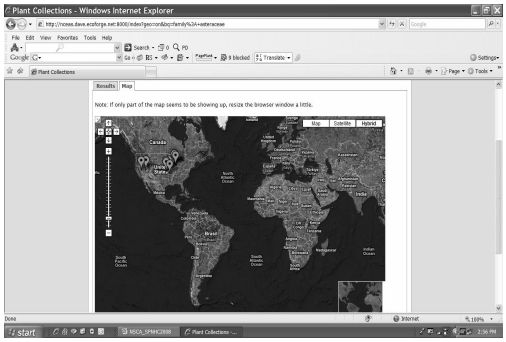

Figure 5. Maps in the Portal

Third, the system makes information sharable. Figure 6 provides a preliminary map created by exporting locations (denoted as latitudes and longitudes) of plants from eight institutions from GB and importing data into ESRI ArcGIS 9.2. It shows that the system provide information sharable among institutions.

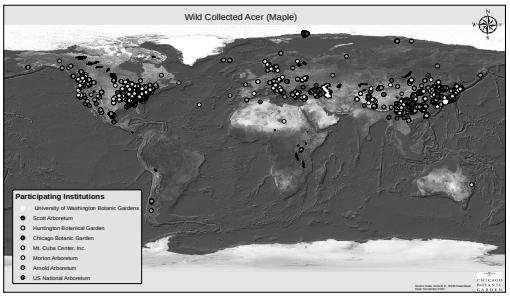

Figure 6. Locations of plants exported from Google Base

5. CONCLUSIONS

This paper presents an effective hybrid-distributed database system, PlantCollections, for web-based querying that allow information from different botanical gardens and arboreta in a variety of incompatible database formats to be accessed and integrated into comprehensive reports using tabular, geospatial and digital formats. The results show that, the system effectively links eighteen databases throughout the world together and provides efficient, accurate, fast access to biological information.

Efforts are currently underway to expand the number of institutions participating in the project, increase the number of foreign language Portals (Mandarin Portal currently under construction), increase the number of images stored in Morphbank, and to implement linkages to other national and international biodiversity databases using ATOM or RSS feeds.

These activities reflect the growing interest in a greater understanding of plants and the natural world. Botanic gardens and arboreta records reflect hundreds of years of data collection; information that will greatly improve the quantity and quality of data available to scientists, curators, educators and the general public in this era of global climate change.

6. REFERENCES

[1] Brin, S. and Page, L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", **Computer Networks and ISDN Systems**, Volume 30, Issue 1-7 (April 1998), Pages: 107 - 117. 1998

[2] Dessler, A., and Parson, E., **The Science and Politics of Global Climate Change: A Guide to the Debate**, Cambridge University Press, 2006.

[3] Hicks, J., Govindaraju, M., and Meng W., "Search Algorithms for Discovery of Web Services", **ICWS 2007**, **IEEE International Conference on Web Services**, Page(s):1172 – 1173. 9-13 July 2007.

[4] Huntley, B., "How plants respond to climate change: Migration rates, individualism and the consequences for the plant communities", **Annals of Botany** 67: 15– 22. 1991.

[5] Morgan, G. and Smuts, T., "Global Warming and Climate Change", **Carnegie Mellon University**, 1994.

[6] Wittkämper, M., Braun, A., Herbst, I., and Herling, J., "A Distributed System for Augmented Reality Experiences in Science Centers and Museums", **Lecture Notes in Computer Science (4469), Springer Berlin Heidelberg**, 2007.

[7]http://www.google.com/intl/en/press/pressrel/universalsearch_20070516.html

[8] http://plants.ecoforge.net/wiki/schema

[9] http://digir.net/prov/prov_manual.html

[10] http://www.morphbank.net/

[11] http://base.google.com/

[12] http://www.chicagobotanic.org/

[13] http://www.nhm.ku.edu/

[14] http://www.publicgardens.org/

[15] http://www.imls.gov/index.shtm