Global Specimen Digitization and Sharing Development Trends Postprint

Authors: Chen Jianping, Xu Zheping

Date: 2021-05-27T00:00:00+00:00

Abstract

Specimen digitization constitutes a crucial foundation for biodiversity conservation and utilization. Through integrated analysis of specimen data, it provides data support for taxonomy, ecology, bioengineering, biological conservation, food security, biodiversity assessment, teaching and education, and human social activities. To understand the current status of global specimen digitization efforts and the development trends of data sharing strategies and technologies, and to provide recommendations for China's specimen digitization work through comparative analysis, this study respectively investigated and compiled the status of specimen digitization and platform construction in North America, South America, Europe, Africa, Asia, and Oceania. The current status and trends of data sharing were analyzed from the perspectives of data use agreements, new technologies and methods, and citizen science. Recommendations for domestic specimen digitization work in China include: strengthening the construction of coordination mechanisms for specimen digitization, management, and dynamic updating to ensure synchronization between physical resources and digital resource information; enhancing data curation and publication, promoting data quality improvement, fully opening data use agreements, and reducing barriers to data utilization; strengthening the learning and adoption of new technologies, particularly the application of open-source software, machine learning, and artificial intelligence technologies, which can play a role in rapid label recognition, automatic identification, and attribute data extraction; strengthening regional and international cooperation to promote integrated data application; and promoting the development of citizen science projects to facilitate field collection, indoor curation, online error correction, data product development, and related activities.

Full Text

Preamble

Global Specimen Digitization and Sharing Trends

Jianping Chen[1], Zheping Xu[2* 1] Chenshan Botanical Garden, Shanghai 201602, China [2] National Science Library, Chinese Academy of Sciences, Beijing 100190, China

*Corresponding author: xuzp@mail.las.ac.cn

Received: 2021-05-13

Funding: Basic Work Special Project of the National Ministry of Science and Technology of China (2015FY110200); Talent Introducing Plan in the field of Library and Information Science in the Chinese Academy of Sciences; Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19050000)

Author Introductions: - Jianping Chen (1974-), M.S., Senior Engineer. Research focuses on biodiversity informatics and information platform development. Email: chenjianping@csnbgsh.cn - Zheping Xu, Ph.D., Associate Research Librarian. Research focuses on biodiversity informatics and scientific data management and utilization.

Abstract

Specimen digitization constitutes a critical foundation for biodiversity conservation and utilization. Through integrated analysis of specimen data, researchers can obtain crucial data support across multiple domains including taxonomy, ecology, bioengineering, biological conservation, food security, biodiversity assessment, education, and human social activities. To understand the current status of global specimen digitization efforts and emerging trends in data sharing strategies and technologies—and to provide recommendations for China' s specimen digitization initiatives—this paper systematically reviews specimen digitization and platform development across North America, South America, Europe, Africa, Asia, and Oceania. We analyze current status and future trends in specimen data sharing from three perspectives: data use agreements, emerging technologies and methods, and citizen science. Based on this comparative analysis, we propose several recommendations for China' s domestic specimen digitization efforts: (1) Strengthen coordination mechanisms for digitization construction, management, and dynamic updating to ensure synchronization between physical specimens and digital resource information; (2) Enhance data curation and publication to improve data quality, fully open data use agreements, and reduce barriers to data utilization; (3) Accelerate adoption of new technologies, particularly open-source software, machine learning, and artificial intelligence applications for rapid label recognition, automated identification,

and attribute data extraction; (4) Expand regional and international cooperation to promote integrated data applications; and (5) Advance citizen science initiatives to facilitate field collection, indoor curation, online error correction, and data product development.

Keywords: specimen digitization, data sharing, GBIF, citizen science, biodiversity

1. Introduction

Over the past 450 years, researchers have collected more than 381 million plant specimens housed in over 3,000 herbaria worldwide (Krishtalka et al., 2016; Thiers, 2017). Integrated analysis of specimen data provides critical support for taxonomy, ecology, bioengineering, biological conservation, food security, biodiversity assessment, human social activities, and education (Culley et al., 2013; Heberling et al., 2017; Soltis et al., 2017; Willis et al., 2017; Ma et al., 2018; Zhang, 2017). In 2012, GBIF released the Global Biodiversity Informatics Outlook (GBIO) report based on comprehensive expert consultation (Hobern et al., 2012). This report envisioned future biodiversity data research across four dimensions—culture, data, evidence, and knowledge understanding—and identified specimen collection data as one of five fundamental data sources (alongside published materials, field observations, genetic sequencing, and automated remote sensing), charting a clear direction for specimen data integration and application.

Over the past decade, biodiversity data initiatives such as GBIF (Global Biodiversity Information Facility), CoL (Catalogue of Life), EOL (Encyclopedia of Life), and BHL (Biodiversity Heritage Library) have rapidly accelerated the aggregation and sharing of biological specimen data. As of 2021, GBIF—the world's largest biodiversity observation data platform—hosts 1.697 billion occurrence records, including 185 million digitized herbarium specimens (10.9% of total platform data). These comprise 86.37 million animal records (46.5%), 85.86 million plant records (46.2%), and 7.07 million fungal records (3.8%). The top ten countries by specimen data volume are: United States (35.46 million), Brazil (12.58 million), Australia (12.16 million), Mexico (8.92 million), Canada (7.96 million), Japan (6.09 million), Costa Rica (5.26 million), Norway (4.52 million), Spain (4.04 million), and Sweden (3.66 million). To further strengthen global collection integration, GBIF launched the Global Registry of Scientific Collections (GRSciColl) in 2019, which aggregates data on research institutions, collections, and personnel across all relevant disciplines including earth and space sciences, anthropology, archaeology, biology, biomedicine, agriculture, veterinary medicine, and technology applications. This initiative supplements specimen metadata and contextual information, thereby enhancing data quality.

This paper systematically reviews global specimen digitization efforts by continent, investigates current status and trends in specimen data sharing, compares

these with China's development, and proposes recommendations for digitization construction, sharing services, coordination mechanisms, international cooperation, and citizen science.

2. Digitization Construction

2.1 North American Specimen Digitization

North American specimen digitization is exemplified by the iDigBio platform, a comprehensive biodiversity data portal and the regional hub for specimen digitization (iDigBio, 2021). iDigBio has digitized 128 million specimen records (47% plants) and 39.17 million multimedia files (82.5% plants) across 1,688 datasets. The most rapid growth occurred before 2017, with steady increases since then. Organizationally, iDigBio divides participant institutions (data providers) into four stages: preparation, negotiation, action, and data aggregation, enabling stable and sustained project advancement.

Regarding data standards and specifications, iDigBio maintains detailed documentation for specimen digitization, image storage, image processing, and image usage, providing clear operational guidelines. These open-access documents are available on the iDigBio website. Core principles include: capturing images at maximum device resolution to ensure quality; archiving images in lossless compression formats permanently; avoiding excessive artificial manipulation; basing all processing on original images to prevent error accumulation; and providing users with optimal quality.

Technically, iDigBio divides digitization into five core task sets: (1) Predigitization: physical specimen repair and standardization; (2) Image capture: using professional DSLR cameras or high-resolution scanners; (3) Image processing: including quality control, barcode retrieval, format conversion, color/brightness adjustment, cropping, image stacking enhancement, editing, file transfer, and OCR text recognition; (4) Electronic data capture: extracting or inputting label data into databases via automated or manual methods including OCR, voice input, and keyboard entry; and (5) Georeferencing: converting textual locality descriptions into precise latitude/longitude coordinates with error ranges and coordinate system parameters.

2.2 European Specimen Digitization

European herbaria boast long histories, rich collections, and numerous academic institutions with strong inter-institutional cooperation. The Consortium of European Taxonomic Facilities (CETAF) represents the largest taxonomic research network, comprising over 5,000 members and preserving 80% of described global biodiversity specimens and data. BioCASe (Biological Collection Access Service) is a data standard and software infrastructure system developed under CETAF guidance, directing specimen digitization and sharing. BioCASe primarily promotes the ABCD standard and provides comprehensive software solutions including the BioCASe Portal for data platforms, BioCASe Provider

Software (BPS) for data providers, and tools for website monitoring and data quality checking. As a GBIF node, BioCASe significantly influences GBIF' s data standards, applications, and sharing policies, with European national data ultimately shared through GBIF. Beyond CETAF, BioCASe, and GBIF, European herbaria and botanical gardens have established numerous specialized web databases, many of which remain outside GBIF' s sharing scope. Some early-established databases have become de facto international standards and foundational platforms, such as the International Plant Names Index (IPNI).

In Russia, Moscow State University launched a specimen digitization project in 2014 (Alexey P. Seregin, 2018), developing the "National Biological System Storage Bank" initiative (Seregin A. P. (Ed.), 2021). This comprises two subprojects: the Moscow Digital Herbarium and the Russian Plant Distribution Atlas, maintained through the "Flora of Russia" project on iNaturalist. To date, the project has accumulated 1.13 million specimens, 1.11 million images, 39,000 species, 660,000 georeferenced records, 460,000 labels, and 660,000 OCR records.

In France, the National Museum of Natural History' s herbarium (code P) has extensive digitization experience (Le Bras et al., 2017). The specialized Vaillant database was developed in the mid-1980s, followed by the current Sonnerat database since 1993, which now stores both the museum' s collections and serves as a network system for Francophone herbaria (e-ReColNat project). Large-scale digitization began with the Renobota project in 2008. Table 1 summarizes the museum' s herbarium digitization projects.

2.3 African Specimen Digitization

African plant resource surveys originated during colonial periods, with numerous African specimens housed in European herbaria. As European collections are digitized and shared online, specialized databases have emerged such as those from Belgium' s Royal Museum for Central Africa and Kew Gardens' African Plants Initiative, with data primarily shared through GBIF. African nations' own herbarium construction and network platforms remain nascent, though collaborative projects with developed countries are underway, indicating substantial potential. The South African National Biodiversity Institute (SANBI) has established independent websites and specialized databases (SANBI, 2021), participating in international projects including Plants of the World Online, the Millennium Seed Bank, the African Plants POSA project, and the National Vegetation Database (NVD), promoting digitization of its collections with most data shared via GBIF. Kenya' s East African Herbarium at the National Museum houses the largest botanical collection in tropical Africa with over 700,000 specimens and serves as the region' s most important national data center, focusing on taxonomy, distribution, utilization, and conservation of East African flora. Beyond GBIF, the BRAHMS system is widely applied in Africa, providing technical support for institutions including Kenya' s National Museum and South Africa' s BLFU herbarium (East African Herbarium, 2021; BLFU, 2021).

In summary, Africa' s specimen digitization relies on foundational materials in Europe, primarily shared through GBIF, with strong external dependence and numerous gaps, indicating significant future potential.

2.4 South American Specimen Digitization

Most historical South American specimens are preserved in institutions across the United States and Europe. Apart from Brazil, digitization levels remain low, with available data primarily from GBIF. By plant specimen records, Brazil has approximately 7 million, Colombia 1.392 million, Peru 807,000, Argentina 722,000, and Bolivia 526,000. Brazil has developed a relatively complete and distinctive independent information system. Its digitization efforts center on the speciesLink platform (speciesLink, 2021), which as of April 15, 2021, includes 534 datasets with 15.219 million online records, of which 11.295 million have geographic coordinates, 3.785 million include images, and 532,000 are type specimens. Algae, fungi, and plant specimens total 10.97 million online records. Digitization has shown stable growth since 2002. The platform' s sophisticated real-time management technology is particularly notable: its indicator system displays daily metrics on new datasets, specimen records, and georeferenced records; its data cleaning system identifies and categorizes errors such as missing required fields, absent coordinates, georeferencing errors (e.g., locations in the ocean), duplicate numbers, suspicious taxon names, and locality errors, often providing automated suggestions. Problems can be traced to specific records for revision. The platform offers over a dozen professional software services covering data management, georeferencing and mapping, species database management, species distribution modeling, browser plugins, network platform management, and specialized metrics systems.

2.5 Oceanian Specimen Digitization

The Australasian Virtual Herbarium (AVH) is an online repository providing access to over 6.66 million plant specimen records from 23 Australian and New Zealand herbaria (AVH, 2021). With the development of the Atlas of Living Australia (ALA) project, AVH was integrated into ALA (ALA, 2021). ALA' s data partners at the same level include the Online Zoological Collections of Australian Museums (OZCAM), the Australian Seed Bank Partnership (ASBP), and the Murray-Darling Basin Authority (MDBA).

New Zealand's total plant collection exceeds 1.4 million specimens, including the world's largest Antarctic plant collection of approximately 640,000 specimens. In 2011, the New Zealand Virtual Herbarium (NZVH) launched as a collaborative network of 11 herbaria, providing online access to 700,000 specimens. The system received software and technical support from AVH and was subsequently merged into AVH and ultimately integrated into ALA.

2.6 Asian Region

Asian digitization efforts face significant challenges due to ethnic diversity, linguistic complexity, and relatively lagging economic and scientific research development. Although mainland China, Taiwan, India, Japan, and South Korea have established relatively sound biodiversity databases, most Asian countries lack comprehensive systems and remain behind in specimen digitization. Table 2 shows Southeast Asian countries' specimen volumes on GBIF and their primary data publishing countries as of April 15, 2021. The vast majority of specimens from these countries were published not by the countries themselves but by European and American nations after digitization; some countries have no shared digital specimens at all, urgently requiring internal and external collaboration to advance regional digitization.

In response, Chinese researchers have proposed the Mapping Asia Plants (MAP) initiative under the Asian Biodiversity Conservation Database Network (ABCDNet, 2021). MAP advances work across six regions—Southeast Asia, South Asia, West Asia, Central Asia, North Asia (Russia' s Asian portion), and Northeast Asia—beginning with literature and specimen data compilation to gradually promote biodiversity digitization and sharing collaboration across Asia.

3. Specimen Data Sharing Status and Trends

3.1 Data Sharing Status on GBIF

As the world' s largest specimen data sharing platform, GBIF provides insights into global sharing patterns through analysis of participating countries and data publishing volumes (Table 3). European and American countries contribute the vast majority of data, while developing nations' contributions remain limited but show substantial growth potential, representing the foundation for future data expansion.

3.2 Data Use Agreements and Licensing

Specimen data sharing requires clear use agreements for legal utilization and processing. In 2013, GBIF analyzed 416 million records across 12,000 datasets, finding only 10% had license declarations and identifying 432 distinct license types—severely hindering sharing and circulation (Peter Desmet, 2013). Following extensive consultation, GBIF' s Governing Board standardized all existing licenses into three categories: CC0, CC BY, and CC BY-NC. Current distribution is: CC0 1.0 (56.7%), CC BY 4.0 (27.6%), and CC BY-NC 4.0 (15.7%). Analysis of North American and Australian specimen platforms (Table 4) shows CC0 and CC BY are the most popular. However, many herbarium platforms still lack clear license identification, and some require cumbersome offline application and approval processes, constraining data reuse.

3.3 Emerging Technologies and Methods

Rapid IT advancement has revolutionized nearly all aspects of specimen digitization and sharing. For digital imaging, high-speed scanning systems enable mass rapid digitization. For taxonomic research needs, innovations include high-resolution scanners, side-light photography, and microscopic imaging combined with dissecting microscopes. Animal specimen digitization has introduced three-dimensional high-definition imaging. Field surveys now widely employ high-resolution digital cameras, smartphones, and handheld GPS devices, generating massive volumes of georeferenced plant images that provide rich contextual data for specimens and sometimes serve as the only vouchers.

For data management and publishing, beyond GBIF' s Integrated Publishing Toolkit (IPT), BRAHMS (Botanical Research and Herbarium Management System)—developed over decades by Oxford University' s Department of Plant Sciences—is widely used by herbaria, botanical gardens, arboreta, and seed banks for data management and online publishing.

For data mining and analysis, numerous tools and codes based on GBIF data enable specimen and observation data analysis. GitHub hosts 686 GBIF-related open-source repositories, including 85 in R, 54 in Python, 52 in Java, and 51 in JavaScript. Platforms including iDigBio, BioCASe, AVH, and NSII maintain dedicated application tool sections. Improved data access toolkits in various programming languages facilitate rapid integration with applications, providing flexible and efficient programming environments. For example, iDigBio' s Python package enables seamless data access, and combined with Python' s scientific computing and AI ecosystem, provides powerful technical support for developers. China' s software development is also flourishing, with practical and impactful tools such as herblabel for herbarium management (Zhang et al., 2017), the Taxonomic Tree Tool for classification tree construction and analysis (Taxonomic Tree Tool, 2021), and ipybd for biodiversity data cleaning, statistics, and analysis (Ipybd, 2021).

With breakthroughs in deep learning AI, image recognition apps have become practical. Tech giants including Google, Microsoft, Tencent, and Baidu offer specialized applications and open APIs, making AI recognition a public infrastructure service. In biology, this has spawned apps like XingSe and HuaBanLü for plant identification and Herbarium Companion for specimen recognition. While species coverage and accuracy remain nascent, these tools have gained societal attention and acceptance in citizen science and public education, achieving sufficient accuracy for common plant identification. Future improvements should incorporate geographic factors and taxonomic knowledge, particularly specimen data, to enhance training, expand recognition scope, and enable more applications. AI also holds promise for specialized fields including seed identification, pollen analysis, cultivar verification, and automated invasive species detection, enabling taxonomy to serve society through AI. Massive digitized specimen data will form an essential foundation for machine learning in the AI

era.

3.4 Development of Citizen Science

The convergence of specimen digitization and public scientific interest has spawned numerous citizen science projects. Kew Gardens offers projects including 19th-century letter transcription, fungal trait completion, plant and fungal label recognition, and mobile rare plant conservation. The Natural History Museum' s Orchid Observers project engages the public in data collection to study climate impacts on British flora using orchids as a model. North America' s iDigBio platform has proposed crowdsourcing and LiveScience projects. The most influential initiative is iNaturalist, which provides convenient field survey tools through a mobile app, creating online communities that organize taxonomists and enthusiasts by interest and collect massive biological image data via crowdsourcing for scientific research. These projects bridge the gap between the public and herbaria/databases, integrating research, outreach, and social service through flexible strategies that have gained widespread acceptance.

Similar to international trends, China has developed multi-level citizen science information platforms through years of development, including interest groups on forums, microblogs, and instant messaging platforms, as well as specialized tools including mobile apps, WeChat mini-programs, and web apps. Platforms such as the Chinese Natural History Museum (CFH), Chinese Plant Photo Bank (PPBC), Biotracks, XingSe, HuaBanLü, and LüTu have attracted numerous scientists and enthusiasts.

These citizen science platforms have accumulated substantial field observation data that importantly supplement traditional specimen data, with some materials serving as research vouchers. Through curation and verification, high-quality data can become a new data resource type within specimen databases, forming an integral component of digitized specimen resources.

4. Discussion and Recommendations

Compared with international developments, China' s specimen digitization and data sharing have achieved excellent results with distinctive advantages, yet several issues require resolution. These include: project-based centralized data management and sharing without subsequent distributed network node construction, creating a strong-center/weak-nodes pattern lacking mechanisms for continuous data quality updates; non-standardized sharing agreements and unclear identification, particularly in multilingual online environments where large-scale data use still requires offline communication, hindering reuse; and while foundations in new technology application, international cooperation, and citizen science are sound, flagship projects and applications remain lacking.

Based on comparative analysis with international trends and China' s actual conditions, we propose the following recommendations:

1. **Strengthen coordination mechanisms** for digitization construction, management, and dynamic updating to ensure synchronization between physical and digital resources. Enhance integration of specimen data with other biodiversity data to form essential resource components for the discipline.
2. **Improve data curation and publication** to enhance data quality, particularly for critical taxonomic and spatiotemporal information. Fully open data use agreements by adopting CC0 or CC BY licenses to reduce usage barriers. Utilize GBIF IPT for external data publishing and track literature citations to analyze specimen data applications across fields. Obtain data feedback and updates through sharing services to improve quality.
3. **Accelerate new technology adoption**, particularly machine learning and AI applications for rapid label recognition, automated specimen identification assistance, and attribute data extraction. Strengthen research on open-source code for specimen data.
4. **Expand regional and international cooperation** to enhance data aggregation and integration. Promote cross-national or cross-regional data construction and sharing through regional or international projects like MAP to advance digitization in less-developed countries.
5. **Advance citizen science collaboration and promotion** to engage professionals and enthusiasts in field collection, indoor curation, online error correction, and data product development.

References

ABCDNet: Asian Biodiversity Conservation Database Network. (2021-04-15). http://www.abcdn.org

Alexey P. Seregin. (2018). The Largest Digital Herbarium in Russia is now available online. *TAXON*, 67(2): 463–467.

AVH: The Australasian Virtual Herbarium. (2021-04-15). https://avh.chah.org.au

Culley, T. M. (2013). Why vouchers matter in botanical research. *APPL PLANT SCI*, 1(11).

East African Herbarium. (2021-04-15). http://eaherbarium.museums.or.ke/

Herbarium Potts (BLFU). (2021-4-15). https://herbaria.plants.ox.ac.uk/bol/blfu/

Global Biodiversity Information Facility (GBIF). (2021). GBIF Registry of Scientific Collection (GRSciColl). (2021-04-15). https://www.gbif.org/en/grscicoll

Global Biodiversity Information Facility (GBIF). (2021). Occurrence Search. (2021-04-15). https://www.gbif.org/occurrence/search

Hobern D, Apostolico A, Arnaud E. (2012). *Global Biodiversity Informatics Outlook: Delivering biodiversity knowledge information.* https://doi.org/10.15468/6jxa-yb44

Integrated Digitized Biocollections (iDigBio). (2021-04-15). https://www.idigbio.org/

Ipybd-Powerful Biodiversity. (2021-05-13). https://github.com/leisux/ipybd

J. Mason Heberling, Bonnie L. Isaac. (2017). Herbarium specimens as exaptations: New uses for old collections. *AMER J BOT*, 104: 1-3.

Krishtalka L, Dalcin E, Ellis S. (2016). *Accelerating the discovery of biocollections data.* Copenhagen: GBIF Secretariat. http://www.gbif.org/resource/83022

Le Bras G, Pignal M, Jeanson ML. (2017). The French Muséum national d' histoire naturelle vascular plant herbarium collection dataset. *Sci Data*, 4: 170016. doi: 10.1038/sdata.2017.16. https://pubmed.ncbi.nlm.nih.gov/28195585/

Ma KP. (2017). Mapping Asia Plants: A cyberinfrastructure for plant diversity in Asia. *BIODIV SCI*, 25(1): 1-2.

Ma KP, Zhu M, Ji LQ. (2018). Establishing China Infrastructure for Big Biodiversity Data. *BULL CAS*, 33(8): 838-845.

Peter Desmet. (2013). Analyzing the licenses of all 11,000+ GBIF registered datasets. (2021-04-15). http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html

Seregin A. P. (Ed.). (2021). *Moscow Digital Herbarium: Electronic resource.* Moscow State University, Moscow. https://plant.depo.msu.ru/

Soltis, P. S. (2017). Digitization of herbaria enables novel research. *AMER J BOT*, 104: 1-4.

speciesLink. (2021-04-15). http://www.splink.org.br/

Taxonomic Tree Tool (TTT). (2021-04-15). http://ttt.biodinfo.org/

The South African National Biodiversity Institute (SANBI) Website. (2021-04-15). https://www.sanbi.org/

Thiers, B. (2017). *The World' s herbaria 2016: A summary report based on data from Index Herbarium.* http://sweetgum.nybg.org/science/ih/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. A. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. https://doi.org/10.1038/sdata.2016.18

Zhang J. (2017). Biodiversity science and macroecology in the era of big data. *BIODIV SCI*, 25(4): 355-363.

Zhang JL, Zhu HL, Liu JG. (2016). Principles behind designing herbarium specimen labels and the R package 'herblabel' . *Biodiv Sci*, 24(12): 1345-1352.

Note: Figure translations are in progress. See original paper for figures.