MID-TERM ACTIVITY REPORT

BIFA3_014 - Data mining of historical herbarium specimens from the Korean peninsula using the Brahms database: A look at the biodiversity informatics project of North Korea

Contents

Executive summary

Despite their abundance in collections, the needed data in the Korean peninsula including North Korea are inaccessible or insufficiently integrated to foster query-based inquiries, and are not applicable to studies at global scale. With nearly 50,000 specimens including the data about specimens stored at foreign herbaria, we have a comprehensive chronological, historical, taxonomic, and geographic coverage of Korean plants including those from inaccessible areas, such as North Korea. Those were expressed as digitized maps in two efloa websites. Especially since all media are strictly owned and controlled by the North Korean government, restrictions on communication between North and South Korea leave obscure information about biodiversity. We are not only deprived of the chance to learn about North Korean biodiversity, but also are suppressed from exchanging biodiversity information

Many foreign herbaria (A, E, TI, KYO, and others) constitute a large fraction of South and North Korean botanical collections more than 100 years. Nevertheless, local data is often inaccessible from outside countries and the available data in both countries are not well managed and formatted thus far.

The great majority of vascular plant species data (ca. 607,514 primary occurrence data) which is currently available as occurrence records in GBIF, has not been georeferenced (less than 1%) even in South Korea. This dataset is presumed to be orphaned, because its data host no longer has the resources or desire to host it online. Moreover, completeness and taxonomic and geographic precision of primary occurrence data is not quite satisfactory. In some cases the data host even loses the dataset (e.g. due to server failure), and no other backup exists

T.B. Lee Herbarium (SNUA) has been devoted to the study of the Korean peninsula flora since 2011. Our specimens at SNUA have been collected as vouchers in floristic studies of South Korea since 1950. This estimate includes about 70,000 housed at the SNUA institute.
The first phase and scope of this project integrates the BRAHMS software to allow queries of foreign herbaria historical records, generate specimen georeferenced data, and photo images about the north and South Korean vascular plants. The reason why we use the Brahms database has to do with information on use to collect systematically and to use consistent names for accurate data management. The other reason is to utilize the Darwin Core Standard for a stable and flexible framework for compiling data.

Also, this project will be limited to 15,000 SNUA specimens about woody plants collected from South Korea first year and available in our BRAHMS database and provide data. This project will handle occurrence records and makes them available through the GBIF web services and download files. After this year's project, we plan to provide additional data about 40,000 specimens collected by T. Nakai, J. Ohwi, S. Kitamura, V.L. Komarov and others and about 55,000 herbaceous specimens at SNUA that provide access to a wealth of information on biodiversity and the spatial and temporal distribution of plants.

Contact information

Chin Sung Chang, Department of Forest Science, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea, Coordinator

Introduction

The first phase and scope of this project integrates the BRAHMS software to allow queries of foreign herbaria historical records, generate specimen georeferenced data, and photo images about the north and south Korean vascular plants.

A dataset composed of 15,000 SNUA specimens records was compiled. Considering the attribute fields except recent collections within 10 years, completeness, such as adequate and sufficient collector names and consistent gezetteers, and georeferenced information, was quite low. Only about 5,000 records (33%) complied with georeferenced information. Also, herbarium datasets (ca. 11,500) are composed by legacy collections, such as E.H. Wilson (ca. 1,000), Faurie/Taquet collections(10,000 and 4,000 duplicate specimens).

To be able to use this dataset for the purpose a process of validation and data cleaning, including a retrospective georeferencing process, will be conducted till March, 2019.

To the best of our knowledge, only this Korean database has been created using the DwC standard since we developed our database. After this year's project, we plan to provide additional data about 40,000 specimens collected by T. Nakai, J. Ohwi, S. Kitamura, V.L. Komarov and others and about 55,000 herbaceous specimens at SNUA that provide access to a wealth of information on biodiversity and the spatial and temporal distribution of plants. This project will handle occurrence records and makes them available through the GBIF web services and download files when this project will be completed.

The project and its objectives

This proposed project has three primary objectives:

1) To establish a database, a network of specimens collected by Faurie/Taquet and Wilson deposited at foreign herbaria about North Korea that will develop and share tools,

2) To provide photos of specimens deposited at E, KYO, and A and to make the data accessible through the GBIF website,

and 3) To provide data of relatively recent specimens housed at SNUA about woody plants in Korea.

The intellectual merit of this project is in the types of occurrence data with georeferencing.

Activities

Description of activity	Partners involved	Contribution of activity to goals listed in table 4.3	Status of activity as of mid-term reporting
Digitizing and publishing georeferenced species occurrence data based on specimens held in Asian collections			
Provide data about 14,000 specimens collected by E. H. Wilson & Faurie/Taquet through the Korean peninsula (including South and North Korea) *1	Dr. H. Kim	Mobilize existing knowledge within the Korean Peninsula	Completed to input the records in the database and will load the data in the GBIF within three

			months
Provide data about 15,000 specimens (of woody plants) collected from 1953 to 2017 in South Korea	Dr. H. Kim	Mobilize existing knowledge within the Korean Peninsula	Records input completed and data cleaning process will be made
Compiling inventories of biodiversity data holdings (for example, by implementing metadata catalogues)			
Produce local annotated checklist*2	Dr. H. Kim	Deal with biodiversity data at the appropriate scale in Asia	revise the checklist using our herbarium records after the project
Develop the online flora using the scratchpad*3	Drs. H. Kim & Haining Qin	Deal with biodiversity data at the appropriate scale in Asia	Will be conducted after this project
Preparing data papers			
E.H. Wilson collections from 1917 to 1918 in Korea	Dr. H. Kim	Mobilize existing knowledge within the Korean Peninsula	Prepare the paper about Wilson data and this paper will be submitted early 2019
U.J. Faurie/E. Taquet from 1901 to 1914 in Korea	Dr. H. Kim	Mobilize existing knowledge within the Korean Peninsula	Prepare two papers about Faurie and Taquet and two papers will be submitted early 2019
SNUA collections about woody plants from 1953 to 2016 in Korea	Dr. H. Kim	Mobilize existing knowledge within the Korean Peninsula	Prepare the paper about SNUA woody plant data and will be submitted early 2019
Other activity types			

A checklist of North Korea vascular plants (revised version of a provisional checklist of KPF published in 2014) *4	Dr. H. Kim	Mobilize existing knowledge within the Korean Peninsula	Will be conducted after this project

Provide data about 14,000 specimens collected by E. H. Wilson & Faurie/Taquet through the Korean peninsula (including South and North Korea) *1

1. **E.H. Wilson data - 1,100 records: published due on Dec. 31, 2018.**
2. **Faurie/Taquet data - 10,500 records: published due on Jan. 31 2019.**
3. **SNUA data - 15,000 records: published due on March 31, 2019.**

Some of photo images are currently presented in our website (http://hosting03.snu.ac.kr/~quercus1/virtualherbarium.htm) will be connected with the GBIF network (https://www.gbif.org/occurrence/). We will use KBIF IPT as the hosting facility.

Produce local annotated checklist*2

Four years ago we produced a provisional checklist of Korean Peninsula flora. We would like to expand the checklist into a more updated listing of names including introduced and naturalized taxa, distribution, and conservation status of some taxa. We would like to record the most recent taxonomic assignment and provide annotations as to whether or not we accept recently proposed changes, and for some taxonomic placements we show alternative arrangement after this project. The checklist will be first published through the GBIF in 2019.

Develop the online flora using the scratchpad*3

We will prepare for publishing data through GBIF first and we learn about the process and understand more about the data sharing method theough the GBIF network. After this project we plan to present the asian distribution maps of all taxa recorded in Korea with the cooperation of Institute of Botany in Beijing in the scrathpad.

A checklist of North Korea vascular plants (revised version of a provisional checklist of KPF published in 2014) *4

We would like to provide separate checklist about North Korean flora and detailed distributional data for all taxa, listing occurrence by provinces and many smaller geographi subunits. The distribution information has not ben yet organized and updated thus far. For listings of all taxa with extensive annotations regarding areas we need more time to complete this job after this project. This new checklist will be also published via the GBIF.

Deliverables

a. Data

Details of datasets expected to be mobilized as an outcome of the project:

Title of dataset	Taxonomic/geographic scope	Approximate number of records (specimens)	Current format (e.g. undigitized, digitized)
Faurie/Taquet data	Vascular plants/South Korea including some parts of North Korea	10,500	photos (5,900) – low quality of digitized
E.H. Wilson data	Vascular plants/ South-North Korea	1,100	Photos (1,000) – low quality of digitized
SNUA data	Vascular plants/South Korea	15,000	Undigitized

The following data shows the number of photos about E.H. Wilson's, U. Faurie's, and E. Taquet's collections from four major herbaria.

Collector	Herbarium				Total
	E	A	TI	KYO	
Wilson	-	948	-	-	948
Fauri&Taquet	3859	674	346	1040	5919
Total	3859	1622	346	1040	6867

Summary of historical collections

collector	count	herbarium	count	georeferecing	count
Taquet, E.J.	5547	E	2637	mappable	8494
Faurie, U.J.	4885	KYO	1981	unmappable	3025
Wilson, E.H.	1087	A	1341		
	11519	TI	1101		
		E,KYO	890		
		P	753		
		LE	430		
		E,TI	343		
		A,E	324		
		KYO,TI	316		
		E,LE	209		
		KYO,P	144		
		A,KYO	140		
		N/A	137		
		TNS	82		
		P,TI	79		
		BM	66		
		E,P	49		
		KYO,LE	46		
		E,TNS	42		
		LE,P	39		
		A,TI	37		
		K	36		
		E,K	33		
		LE,TI	31		

Other deliverables

Data papers will be prepared at the end of the first year project

1) E.H. Wilson collections from 1917 to 1918 in Korea
2) U.J. Faurie/E. Taquet from 1901 to 1914 in Korea
3) SNUA collections about woody plants from 1952 to 2016 in Korea

We are preparing three papers, Gazetteers of Faurie and Taquet, Data description of E.H. Wilson, and Data description of Frurie and Taqeut and will submit these into journals no later than early 2019.

Calendar of activities

Proposed dates	Activity	Lead partner	Notes
June 2018	Attendance of project team member at BIFA Capacity Enhancement Workshop	Hui Kim	Project leader took part in the meeting and gave a presentation, but Dr. Kim attended the workshop for one week.
April-Sept 2018	Data input/Bar coding of SNUA specimens	C.S. Chang	April-Sept 2018. Completed the input
July 2018	Visiting TI(University of Tokyo Herbarium) and KYO(Kyoto University), five-day work ***1**	C.S. Chang/S.Y. Kwon/ San Kang (undergraduate student)	July 2018. Visited Kyoto University and the University of Tokyo for five days
Sept. 2018	Georeferencing and data cleaning***2**	C.S. Chang/H. Kim	Sept 2018
Jan-March 2019	Information dissemination and preparation of data papers	C.S. Chang/H. Kim/Haining Qin	Jan-March 2019 (plan)

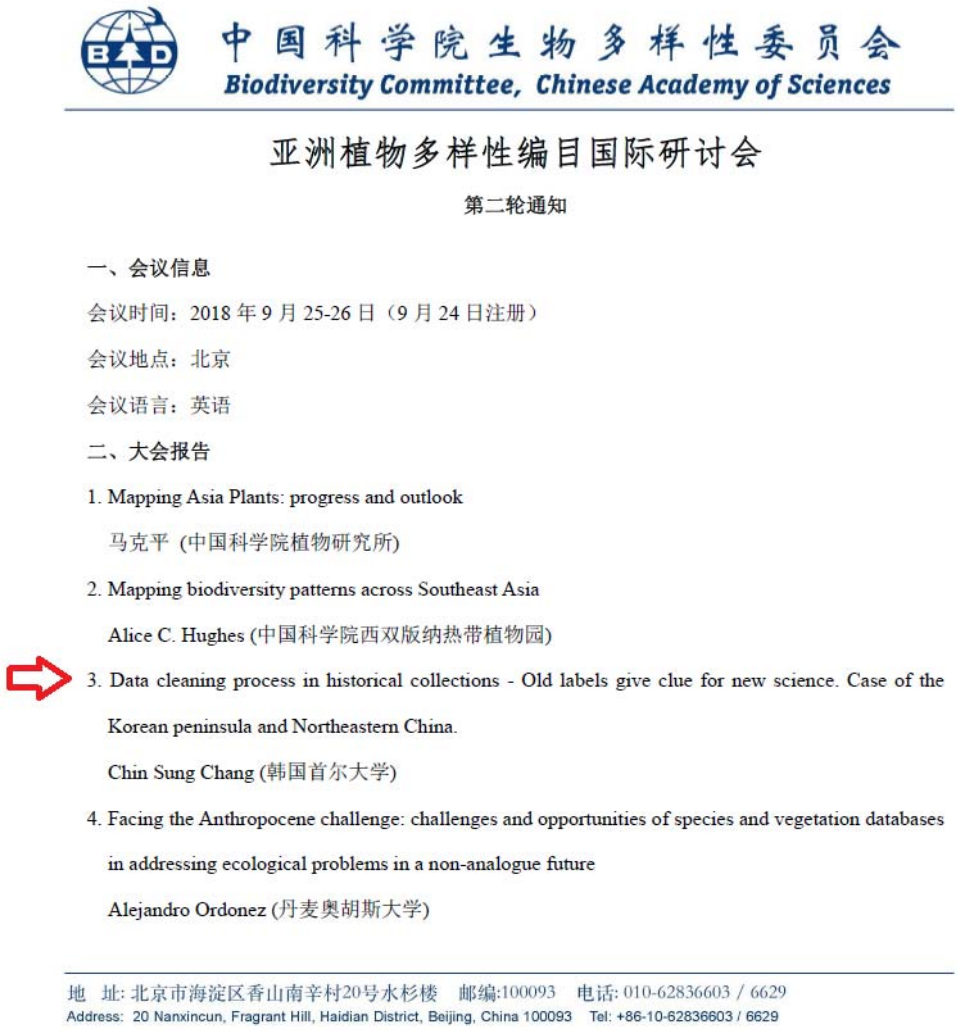
Visiting TI and KYO *1

Collection number of most specimens collected by Faurie/Taquet at E, A, P, K and KYO (some missing) can be confirmed easily using our acquired photos, but majority information on specimens (except type collections) deposited at TI (University of Tokyo, herbarium) and KYO

(Kyoto University) were obtained without collection number ("s.n.") and stored in our database. Many additional photos of these specimens (ca. 1,100 additional photos of Faurie/Taquet's specimens) were collected with our current visit to two Japanese herbaria.

1st MAP symposium *2

Presentation title: Data cleaning process in historical collections -Old labels give clue for new science. Case of the Korean peninsula and Northeastern China.
Chin Sung Chang (韩国首尔大学)



中国科学院生物多样性委员会
Biodiversity Committee, Chinese Academy of Sciences

亚洲植物多样性编目国际研讨会

第二轮通知

一、会议信息

会议时间：2018 年 9 月 25-26 日（9 月 24 日注册）

会议地点：北京

会议语言：英语

二、大会报告

1. Mapping Asia Plants: progress and outlook

马克平 (中国科学院植物研究所)

2. Mapping biodiversity patterns across Southeast Asia

Alice C. Hughes (中国科学院西双版纳热带植物园)

3. Data cleaning process in historical collections - Old labels give clue for new science. Case of the Korean peninsula and Northeastern China.

Chin Sung Chang (韩国首尔大学)

4. Facing the Anthropocene challenge: challenges and opportunities of species and vegetation databases in addressing ecological problems in a non-analogue future

Alejandro Ordonez (丹麦奥胡斯大学)

地 址: 北京市海淀区香山南辛村20号水杉楼 邮编:100093 电话: 010-62836603 / 6629
Address: 20 Nanxincun, Fragrant Hill, Haidian District, Beijing, China 100093 Tel: +86-10-62836603 / 6629

Project communications

1) The following Internal communication will be conducted coming Feb, 2019.

- present & publish the results in academic societies/journals (The Korean Society of Plant Taxonomists/Korean Society of Forest Sciences).

2)

- The part of our project will be given as a presentation in the international symposium, which will be held by Institute of Botany, Beijing, China on Sept. 25th and 26th . The title is “Data cleaning process in historical collections - Old labels give clue for new science. Case of the Korean peninsula and Northeastern China”.

Mid-term evaluation findings and recommendations for the remaining project implementation period

An evaluation of the project activities

The herbarium specimens in SNUA have been collected over the past 50 years from south Korea geographical areas. We try to make the data globally available and paid attention to use both the Korean and the English geographical names. A thorough input was conducted in order to compile digital occurrence records corresponding to ca. 500 species. As we planned, a total of 15,000 data was stored in the Brahms database within six months (till March 31, 2019). A process of validation and data cleaning including a retrospective georeferencing process will be counducted within six months.

The collective work about historical collections and knowledge of the experts across the foreign herbaria provide access to a wealth of information on the Korean peninsula. We visite two Japanese herbaria, TI and KYO and obtained additional photos about Faurie/Taquet collections. We completed labeling species name and collectin number for each photo taken. Completeness about Faurie/Taquet is currently medium. Only 60% complied with completeness in terms of scientific name and gazzetters. On regard of the quality of records, spatial data will be improved using georeferecing work. It is importan to manage carefully this datasets, promoting constant quality improvements.

The present study aims to assess quality of the dataset and records and make us keep all appropraiate documents that proves veracity of data.

Through the project we were able to obtain more information about Faurie/Taquet collections from Japanese herbaria . This BIFA project experience makes us to publish the remainders of SNUA dataset and other historical dataset about North Korea in the near future.

Comments on the project implementation

The National Science Museum by regulations of the Ministry of Science, ICT and Future Planning of the Korean government takes part in the GBIF global network as a national member. Although a considerable amount of research funding are operated by these Korean national institutions including Korea National Arboretum of Korea Forest Services, Korea Institute of Science and Technology information, KBIF national data repository, and Korean Natural History Research Information System, data gaps and data immobilization in digital accessible information in the Korean peninsula have hampered prospects of safeguarding biodiversity. Unfortunately the Korean government sector is not actively engaged in database disclosure and data creation.

A process of validation and data cleaning, including a retrospective georeferencing process, has not quite performed yet by the KBIF. A dataset composed of more than 1,000,000 records was compiled in Korea, but was considered of poor quality because none of the records had proper georeferenced data. In order to complement the activity of private sector with such BIFA fund, the GBIF is necessary to make recommendations for development and networking on this matter to the KBIF following the Japan's or Taiwan.' case. The GBIF website has already indicated major problems of taxonomic precision, geographic precision, temporal precision about data published by KBIF. The KBIF is necessary to present how to manage orphaned data (over 90% of the published data), and how to process data cleaning, such as detecting and correcting corrupt or inaccurate records in order to solve these problems within a certain time period. The GBIF should recommend that the KBIF needs to establish linkages with the databases of the GBIF and high qualified data made available will not be subject to limitation on dissemination to other countries. At present, the KBIF is internationally isolated and lacks data creation and sharing.

Areas of success

Although a large variety of tools is available to support data transformation and data cleaning tasks, we learned about the openrefine tool through the workshop. We use both the Brahms database and this openrefine about duplicate elimination, specific cleaning phase, and data analysis.

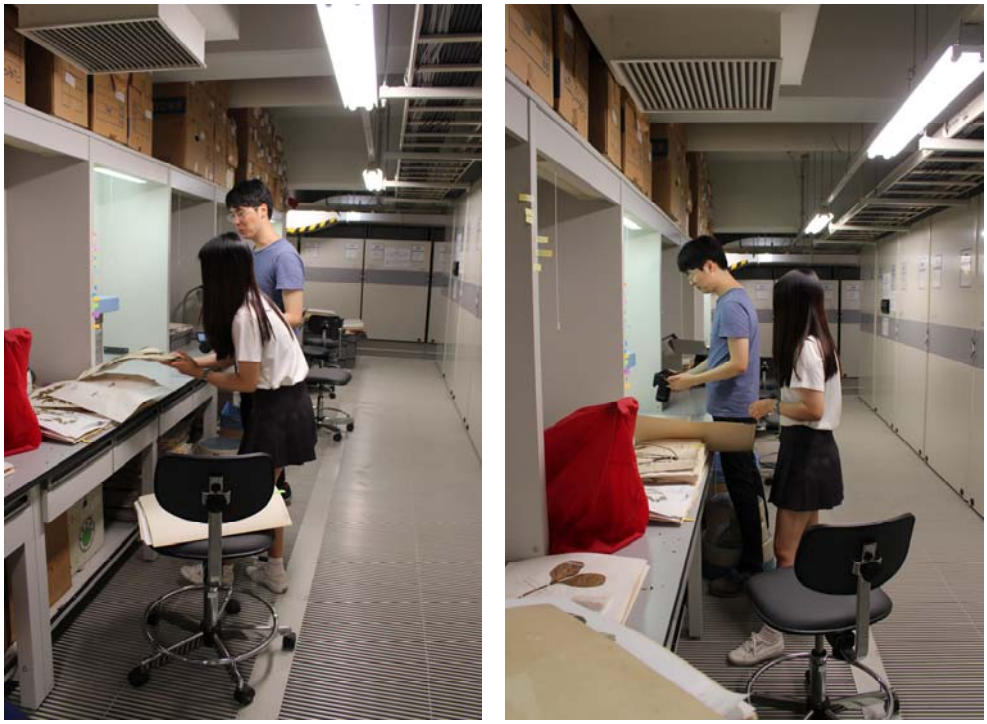


Fig.1. Working at KYO on June 28, 2018.

Fig. 2. Data input at SNUA on Sep 18, 2018.

Fig 3. Join the workshop on June 5th at PE

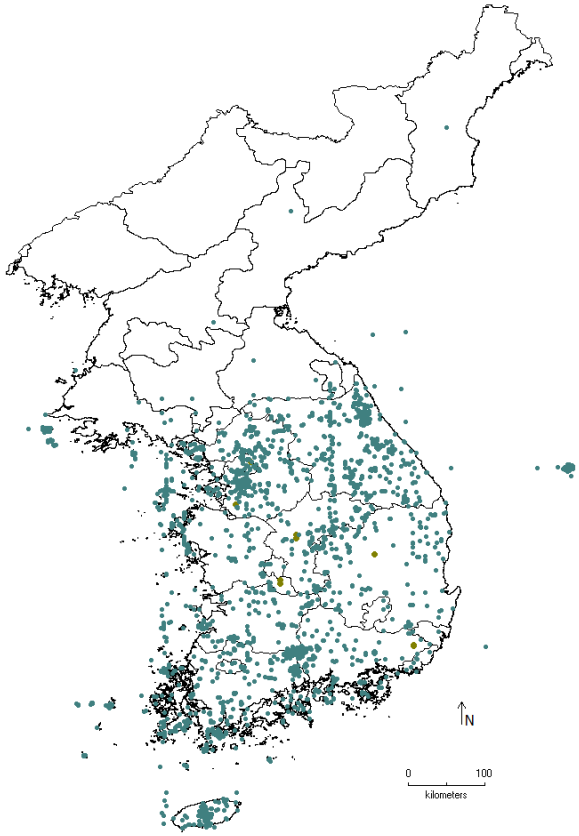


Fig 4. Collection sites represented by 15,000 specimens deposited at SNUA from 1951 to 2017 with a map.

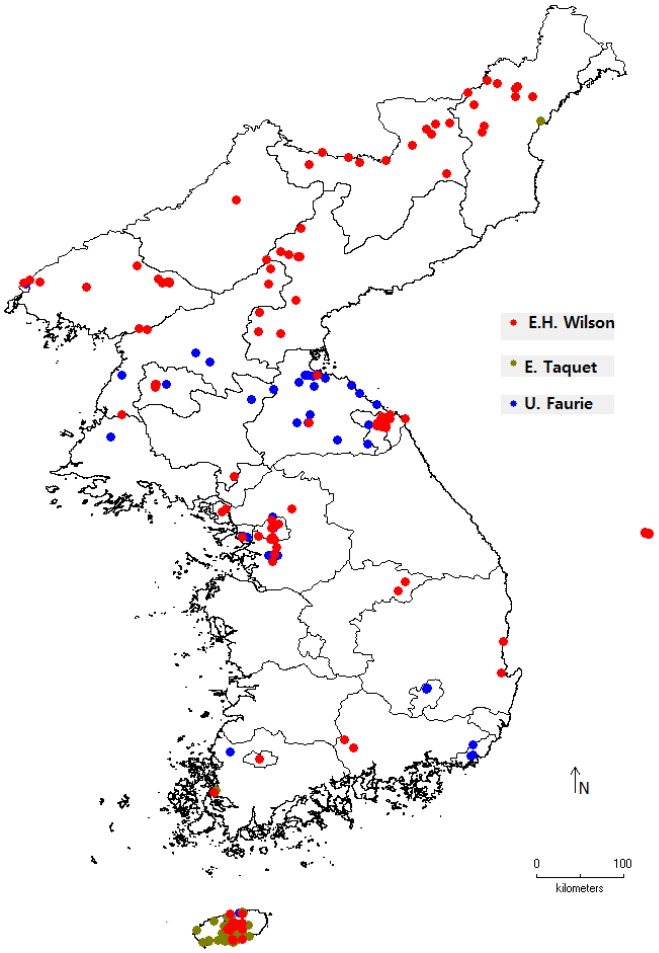


Fig. 5. Historical records of E.H. Wilson, U.J Faruie, and E. Taquet deposted at several foreign herbaria

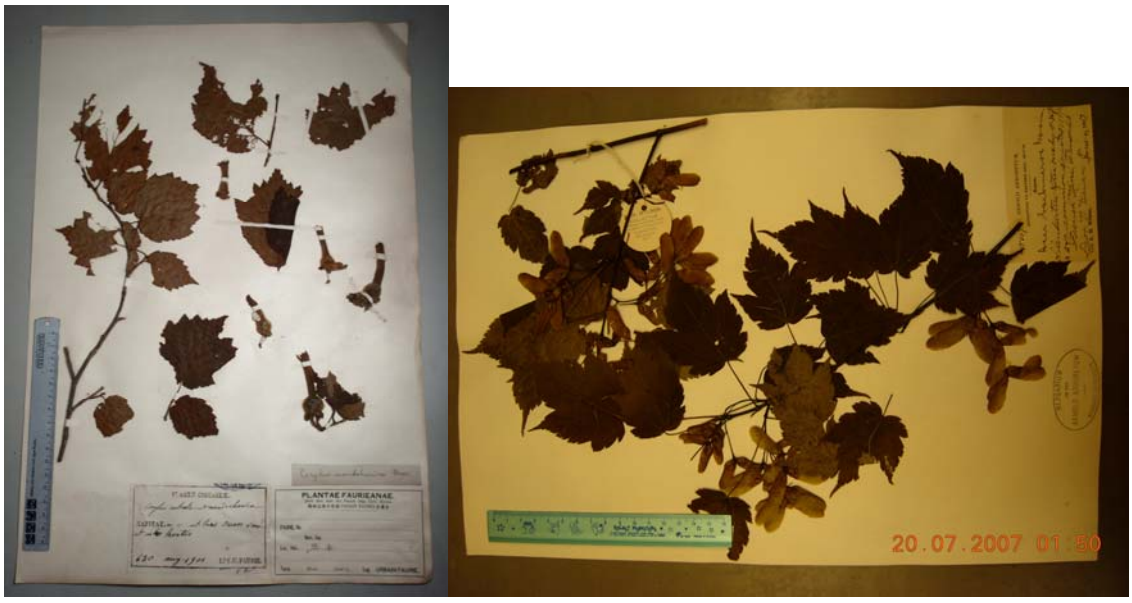


Fig. 6. Some examples of ca. 6,900 photos taken from several herbaria