

세계생물다양성정보기구(GBIF)에 출판된 동아시아 관속식물 생물다양성 정보 현황과 자료품질 분석

장진성¹ · 권신영¹ · 김 휘^{2*}

¹서울대학교 산림과학부, ²목포대학교 식의약자원개발학과

Status and Quality Analysis on the Biodiversity Data of East Asian Vascular Plants Mobilized through the Global Biodiversity Information Facility (GBIF)

Chin-Sung Chang¹, Shin-Young Kwon¹ and Hui Kim^{2*}

¹Department of Forest Sciences and The Arboretum, Seoul National University, Seoul 08826, Korea

²Department of Pharmaceutical Resources, Mokpo National University, Muan-gun 58645, Korea

요약: 생물다양성정보학(Biodiversity Informatics)은 정보과학을 생물다양성정보에 접목한 분야로 정이명으로 구성된 학명을 비롯한 중정보를 기초로 일차종발생자료를 구축하고 이를 활용한다. 본 연구에서는 생물다양성 정보의 이용적합도를 기준으로 세계생물다양성정보기구(GBIF)에 출판된 동아시아 자료의 품질을 BRAHMS 프로그램을 이용하여 평가하고 이를 통해 생물다양성자료 정제의 필요성을 확인하였다. 국립생물자원관, 국립생태원, 국립수목원 등의 국내 생물다양성 관련기관과 더불어 일본, 중국, 대만의 출판 자료는 자료정제과정의 문제로 학명, 지리정보, 채집자, 날짜 등에 대한 오류가 확인된다. 기본적인 속성자료에서 오류가 발생하는 원인은 동아시아의 생물다양성관리기관들이 구조화되지 않은 데이터베이스를 사용하고 평면적인 스프레드시트형 정보를 사용하기 때문이다. 생물다양성 정보 특성상 다양한 정보가 구조화가 되지 않을 경우 학명, 인명, 지명, 문헌, 생태정보에 대한 데이터 무결성을 해결하지 못한다. 동아시아 생물다양성정보 관리 문제를 극복하기 위해서는 자료의 구조화와 함께 자료정제에 대한 이해도를 높이고, 오류 수정을 위한 지속적인 자료 관리자인 전문 분류학자 양성이 필요하다. 생물다양성 정보관리자는 오류 원인분석을 통해 문서화된 관리 지침을 수정, 추가하는 등 향후 오류 예방을 위한 대책이 필요하며 시스템에 적용시켜야 한다. 이런 모든 과정은 데이터베이스를 기반으로 진행되고 기록되어야 한다. 동아시아의 생물다양성 출판자들은 현재 수준의 단순한 자료구조보다는 생물다양성 정보 관리를 위해 전문적인 선진 프로그램의 사용 혹은 이에 준하는 수준의 고도화된 데이터베이스의 개발이 필요하다.

Abstract: Biodiversity informatics applies information technology methods in organizing, accessing, visualizing, and analyzing primary biodiversity data and quantitative data management through the scientific names of accepted names and synonyms. We reviewed the GBIF data published by China, Japan, Taiwan, and internal institutes, such as NIBR, NIE, and KNA of the Republic of Korea, and assessed data in diverse aspects of data quality using BRAHMS software. Most data from four Asian countries have quality problems with the lack of data consistency and missing information on georeferenced data, collectors, collection date, and place names (gazetteers) or other invalid data forms. The major problem is that biodiversity management institutions in East Asia are using unstructured databases and simple spreadsheet-type data. Owing to the nature of the biodiversity information, if data relationships are not structured, it would be impossible to secure the data integrity of scientific names, human names, geographical names, literature, and ecological information. For data quality, it is essential to build data integrity for database management and training systems for taxonomists who are continuous data managers to correct errors. Thus, publishers in East Asia play an essential role not only in using specialized software to manage biodiversity data but also in developing structured databases and ensuring their integration and value within biodiversity publishing platforms.

Key words: biodiversity informatics, data cleaning, data integrity, fitness-for-use, GBIF, georeferencing, scientific name

* Corresponding author
E-mail: huikim@mokpo.ac.kr

ORCID
Hui Kim  <https://orcid.org/0000-0002-7765-6812>

서론

생물학은 분자부터 지구생태계까지 연구 대상의 범위와 크기가 다양함에도 불구하고 정보학과 생물학의 융합인 생물정보학 (bioinformatics)이 너무 과도하게 강조되고 있다 (Berendsohn, 2009). 생물정보학은 분자 수준에 초점을 맞추고 있는 반면 (Sarkar, 2007), 생물다양성정보학(Biodiversity Informatics)은 유기체 수준 이상의 범위에서 자료의 관리, 발표, 발견, 탐구 및 분석과 같은 정보 해석을 시도한다. 생물다양성정보는 생물개체의 종정보를 중심으로 지리 정보를 비롯한 다양한 정보와 연계를 통해 폭넓은 활용도를 갖게 된다. 초기 생물다양성정보 구축은 자료를 모으고 관리하는 것에 핵심 기술이 집중되지만, 점차 자료의 응용 측면이 강조되면서 디지털 자료를 활용하고자 하는 창의적인 개념과 균형이 더 중요하게 되었다 (Peterson et al., 2010). 생물다양성정보의 이용은 분포도, 지리분포, 종목록, 동식물상, 계통수, 분포모델링, 보전정책 등 다양하다. 다방면의 응용에도 불구하고 생물다양성정보학의 핵심은 정이명의 관계, 분류체계 등에 기초한 분류 정보이다. 결국 생물다양성정보학의 발전은 정제된 분류학 정보를 기반으로 구축되는 생물다양성 자료의 질과 양이 결정하게 된다. 일반적인 생명과학의 발견원리는 가설설정과 통제된 실험을 통해 가설을 검증하는 구조를 갖지만, 생물다양성정보학은 대량의 정보를 플랫폼을 통해 정제된 자료를 축적하고 활용한다는 점에서 차이가 있다 (Peterson et al., 2010). 생물다양성 자료는 대상 분류군이 다양하고 표본, 문헌, 관찰자료, 샘플링데이터 등의 다양한 자료구성으로 인해 공통된 플랫폼에 이를 구현하는 것이 까다로우며 이를 극복하기 위해서는 관련 자료의 통합 기술이 중요하다. 이용 가능한 생물다양성자료를 효과적으로 활용하기 위해서는 플랫폼 구축과 관련된 기술 개발이 핵심이지만 (Hardisty et al., 2013; Peterson et al., 2010) 국내의 경우 생물정보학과도 구별하지 않고 생물다양성정보학 고유의 전문성도 고려되지 않고 있다.

전 세계적으로 수 많은 표본기록과 관측 자료를 모아 생물다양성 데이터를 축적하고 이를 공유하기 위한 플랫폼에 대한 경쟁은 오래전에 시작되었다. 관련 조직으로는 OECD중심의 GBIF(the Global Biodiversity Information Facility, 세계생물다양성정보기구)와 북미중심의 iDigBio (Integrated Digitized Biocollections, 통합디지털 생물수집)와 신대륙인 남북중미 지역을 포괄한 BIEN(Botanical Information and Ecology Network) 등이 있다. 생물다양성 자료로서 문헌은 학명과 관련된 정보 질 향상뿐만 아니라 일차종발생 자료(primary occurrence data)의 활용으로 중요하다. 국제적으로 BHL(Biodiversity Heritage Library, 생물다양성전통도서관)을 통해 86개 주요 기관 및 360여개 이

상의 관련 기관이 참여하여 무료로 생물다양성 문헌자료를 공개하고 있다 (Gwinn and Rinaldo, 2009). 궁극적으로 구조화되고 통합된 자료가 완성된다면 각자의 플랫폼을 통한 통합 자료가 각 목적을 위해 활발하게 활용이 될 수 있지만 자료의 품질 향상과 관리가 주요 장애 요인이 된다 (Peterson et al., 2010; Wen et al., 2015).

생물다양성자료는 1차 종발생자료가 핵심으로 이를 활용하는 분야는 유해침입생물종의 분포현황에 대한 파악, 기후변화로 인한 생물종의 분포 변화에 대응하고 종분포 모델링, 멸종위기종의 개체, 집단, 군집에 대한 보전, 작물로 이용되고 있는 유전자원의 야생상태에서의 보전, 인류 보전에 영향을 미치는 다양한 매개생물종의 분포 등 다양하다 (Bebber et al., 2010; Chavan and Krishnan, 2003; Fuentes et al., 2013; Kier and Barthlott, 2001). 이용도가 높은 1차 종발생자료는 많은 연구자들이 자료를 구축한 플랫폼 자료를 그대로 이용하여 해당 자료의 품질에 대한 명확한 판단없이 사용하였으나 연구자들이 정보의 품질을 점차 상세하게 요구하고 있다. 본 연구에서는 표본자료의 오동정에 의한 일차적 오류보다는 자료의 이용적합도(data fitness for use, Andersson et al., 2016)를 기준으로 동북아시아 국가들이 등록한 GBIF 출판 자료를 상호 비교하고 평가하여 자료의 문제점을 파악하고, 자료 정제의 필요성에 대해 검토하고 해결할 방법론을 제시하고자 한다.

재료 및 방법

동북아시아 지역내 생물다양성 정보자료의 중요 척도인 세계생물다양성정보기구(Global Biodiversity Information Facility; GBIF)에 등록된 자료를 이용하여 자료의 품질을 분석하였다. GBIF에 등재된 동북아 지역의 4개 국가 즉, 대만, 대한민국, 일본, 중국의 국가별 노드(node)에서 발생 자료 중 표본과 관련된 관속식물자료를 국가별로 분석하였다 (GBIF.org, 2020, 2021a, 2021b, 2021c).

분석을 위해 R의 GBIF분석 package인 “rgbif”를 이용하였다 (Chamberlain, 2021). GBIF의 자료의 4가지 주요 구분인 종발생자료(occurrence), 종목록(checklist), 샘플링데이터(sampling data) 및 메타데이터 중 종발생자료를 주요 분석 대상으로 정하였다. 종발생자료의 이용적합도는 3가지의 구성요소 식물동정(identification), 지리참조연산에 의한 좌표정보(georeferencing) 자료 품질에 집중하여 분석하였다 (Anderson et al., 2016; Chapman et al., 2020). GBIF는 개별 종발생자료의 품질을 확보하기 위하여 위의 세 가지 주요 자료품질에 대한 문제가 발생한 자료를 issue flagging을 이용하여 이를 표시하는데, rgbif의 occ_count() 및 occ_search() 함수를 이용하여 개별 자료의 자료적합도에

따라 구별하여 각 국가별 자료 품질을 제시하였다.

심층적인 이용적합도 분석을 위하여 개별 출판자들의 자료를 직접 GBIF자료를 Darwin-core형식으로 받아 영국의 옥스퍼드대학과 큐식물원에서 공동개발 한 식물데이터베이스 프로그램인 브라암스(BRAHMS: Botanical Research And Herbarium Management System)를 이용해서 자료를 분석하였다. 집중분석의 대상은 국내자료의 경우 산림청 국립수목원 산림생물표본관(Korea National Arboretum, KH)과 환경부 국립생물자원관(National Institute of Biological Resources, KB)에서 소장된 식물표본자료, 국립생태원(National Institute of Ecology)에서 발표한 전국자연환경조사 자료를 활용하였다.

결 과

1. GBIF 기초자료분석

GBIF는 생물다양성 자료를 생성, 관리 및 정보제공하고 인프라를 공동으로 구축하고 네트워크 활성화를 위해 지정한 국가별 단위를 노드라고 한다. 한 국가의 GBIF 종발생자료는 해당 국가노드에서 출판한 자료와 해당 국가에서 발생한 자료를 다른 국가 노드에서 출판한 자료로 나눌 수 있다. 현재 동북아시아 4개국의 노드를 통한 GBIF출판의 규모는 관속식물의 경우 전체 590만 건으로 일본 255만

건, 중국 160만 건, 대한민국 116만 건, 대만 57만 건이 확인된다(Figure 1). 절대적인 종발생자료의 크기는 일본과 중국이 높으나 해당국가의 국토면적당 종발생량 즉 km²당 종발생량은 대만이 15.88건, 대한민국 11.64, 일본 6.77, 중국 0.17로 단위면적당 종발생량의 비율은 대만과 중국의 단순비교로 100배 가까운 차이를 보인다(Figure 1). GBIF에 실제 등재된 자료의 양은 차이가 있으나 자료의 이용적합도 측면에서 분포정보를 활용할 수 있는 자료의 양은 일본이 153만 건, 중국이 139만 건, 대한민국이 50만 건, 대만이 41만 건을 보유하고 있으나 전체 종발생량에 대한 이용 가능한 자료의 비율은 한국이 43.59%로 가장 낮다(Figure 2).

지리정보의 정확도와 분류학 정보의 정확성은 각 국가별로 이용 가능한 자료의 수준을 판정할 수 있으며, GBIF는 개별 정보에 대한 품질을 판정할 수 있도록 문제(issue)가 있을 경우 이를 표시(flagging)하는 기능을 갖고 있다(Anderson et al., 2016; Chapman et al., 2020). 동아시아 종발생자료 중 좌표의 품질을 확인한 결과 주로 null 값이 아닌 0으로 기록된 좌표의 양(zero coordinate, Figure 3A)과 국가경계를 넘어간 좌표를 보유한 자료의 양(country coordinate mismatch)을 조사해 보면 일본이 가장 높은 수치를 보였다(Figure 3B). 분류정보에 있어 제시된 학명이 GBIF의 기준학명자료(backbone data)와 일치하지 않거나(taxon match fuzzy, Figure 3C), 종소명이 불일치할 경우

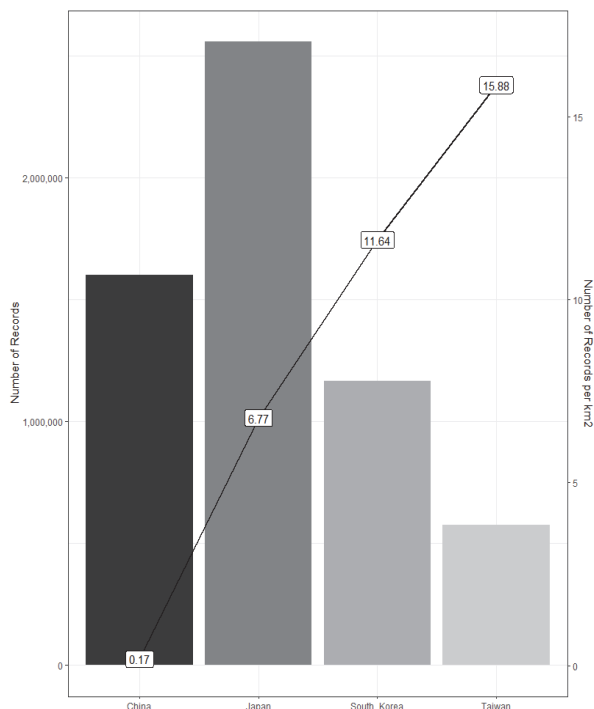


Figure 1. Total number of occurrence data and the number of records per km² which were published our East Asian countries, China, Japan, South Korea and Taiwan.

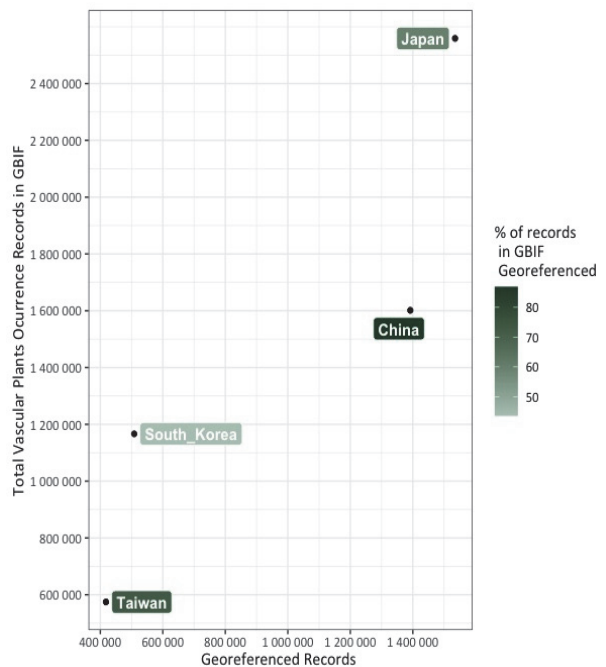


Figure 2. Total number of occurrence data and the number of georeferenced records and ratio which were published by our East Asian countries, China, Japan, South Korea and Taiwan.

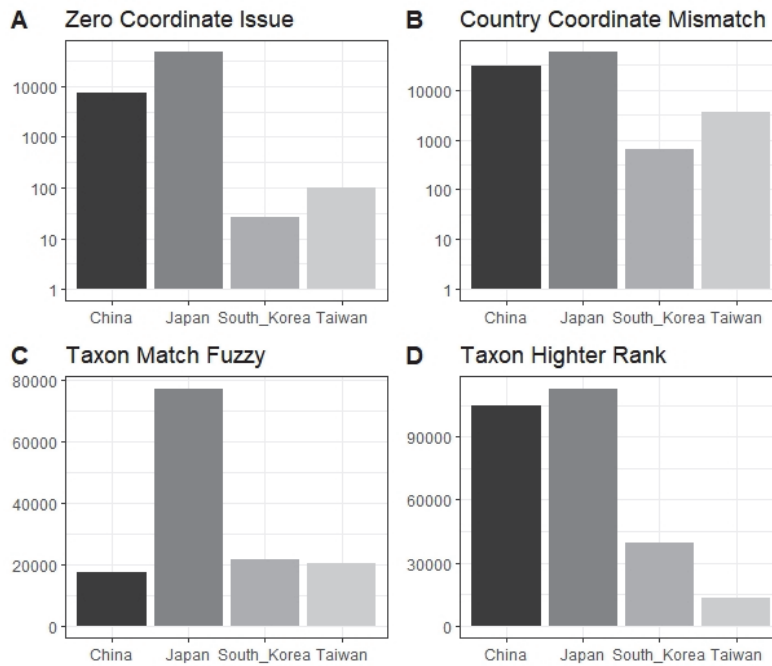


Figure 3. Total number of certain issues of flags in GBIF occurrence data published by four East Asian countries, A. Coordinates are exactly 0/0, often indicating an actual null coordinate. B. The interpreted occurrence coordinates fall outside of the indicated country. C. A match with a different spelling was found. D. No match was found at the same taxonomic rank but one was found for a higher rank.

속명에 그 정보를 연결시키는 종발생자료의 양(taxon higher rank, Figure 3D)이 일본이 가장 높아 양에 비해 질적인 문제는 크다는 것을 확인할 수 있다.

2. 국내GBIF 자료현황

GBIF내에서 국내 식물관련 종발생 자료의 대표적 출판 기관으로는 국립생물자원관, 국립생태원, 국립수목원 등이 있다. 국립생물자원관의 경우 관속식물자료가 607,514건이 확인되나 모든 자료가 좌표가 없이 발표됐으며 이는 268,815건을 발표한 국립수목원의 경우도 모든 자료가 좌표가 없어 이용적합도에 있어 활용도가 거의 없다. 2020년 12월에 국립생태원은 387,863건의 관속식물 관찰(observation) 자료를 발표했으며 모두 좌표가 부여되었고 0.5% 자료만 GBIF의 기준학명자료만 일치하지 않을 정도의 정밀 자료를 발표하였다. 국립생물자원관과 국립생태원의 자료는 국립중앙과학관에서 운영하고 있는 GBIF 종합출판도구(IPT; Integrated Publishing Toolkit) 서버를 통해 자료를 관리하고 있으나 국립수목원 출판자료의 경우 해당자료의 IPT가 연결되어 있지 않고 있어 해당자료를 ‘고아자료(orphaned data)’로 취급받는다.

3. 동아시아GBIF 자료현황

일본내 기관에서 출판한 관속식물종의 종발생은

1,721,319건으로 동아시아에서는 가장 많은 발생자료를 발표하였다. 해당자료는 국립, 도립 혹은 현립의 18개 박물관을 중심으로 자료가 발표되었다. 표본관별로 발표한 자료는 TNS (National Museum of Nature and Science)가 378,321건 (22.0%), KPM (Kanagawa Prefectural Museum of Natural History)은 277,990 (16.1%), HYO (Museum of Nature and Human Activities)는 174,085건 (10.1%), OSA (Osaka Museum of Natural History)는 114,066 (6.6%)건으로 4개 기관이 전체 발표자료의 절반 이상(54.9%)을 차지한다(Figure 4). 좌표가 없는 자료는 579,821건으로 전체의 50.8%를 차지하고 있고 동정이 되지 않은 분류군은 22,182건으로 1.3%가 확인된다. 가장 많은 자료를 발표한 TNS와 KPM의 경우 좌표정보가 결여된 것이 18.4%, 9.5%가 있어 비교적 충실한 자료 관리가 되고 있음을 확인할 수 있다. 반면 자료수에 비해 이용적합도가 떨어지는 기관은 HYO, OSA로 44-68%의 좌표 부재의 질적 관리에서 차이를 보인다. 연도별 채집은 가장 많은 표본자료를 확보한 TNS는 주요 채집이 1980년대에 집중되어 있는 반면, 다른 일본내 기관은 1980년 말에서 2010년까지 최근의 채집이 진행되었다. 일본자료의 분류 정보의 품질은 명명자와 자동명에 대한 문제점 등이 있었고 채집자의 이름, 채집날짜나 기타 정보에 대한 자료정제의 결여, 채집지역에 대한 자세한 지명 정리에 대한 일관성이나 균일성이 다른

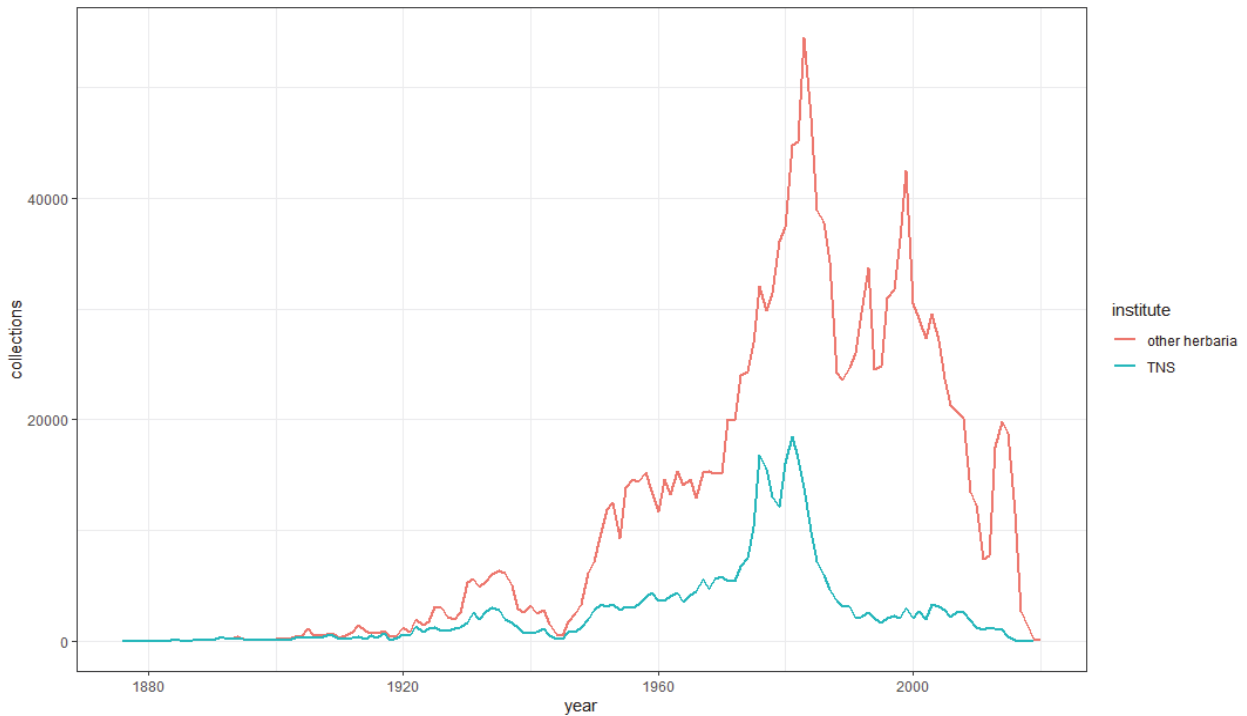


Figure 4. Occurrences per year graphs based on the number of herbarium records published through the GBIF by TNS (National Museum of Nature and Science) and other 17 herbaria with in Japan.

나라에서 확인되는 문제점과 유사하다.

2001년부터 생물다양성 데이터베이스 통합관리를 시작한 대만은 4개의 기관에서 GBIF에 출판한 식물 표본 자료는 210,720건이며 이중 전체의 53.6%는 TAI (National Taiwan University, 112,960), 45.7%는 TAIF (Taiwan Forestry Research Institute, 96,349)가 발표하여 양분하고 있다(Shao et al., 2013). 표본 중에는 동정이 완료되지 못한 종류 및 채집지가 없는 자료는 전체 13,220개로서 6.2%를 차지하고 있으며 대부분 대만대학 식물표본관에서 나온 자료(73.6%, 9736개)가 된다. 채집년도별 정리를 보면 1945년 이전 표본과 1960년대 그리고 1980년대 채집은 주로 대만대학의 표본이 주를 이루며 이후 90년대와 2000년대 초반은 대만임업시험소가 대부분을 차지한다(Figure 5). 특히 1895년에서 1945년까지 일본의 대만 강점기 시기의 채집품은 47,903건으로 주로 일본인 S.Sasaki (佐佐木舜一), T.Suzuki (鈴木時夫), S.Suzuki (鈴木重良), Y.Yamamoto (山本由松) 등의 채집품이 주요 수집이다. 대만의 경우 자료 내용에 있어 학명에서는 명명자와 자동명에 대한 일관성이 없는 정리, 채집자의 이름의, 채집날짜나 기타 정보에 대한 자료정제의 결여, 채집지역에 대한 자세한 지명 정리에 대한 일관성이나 균일성이 부족하다. 최근 표본 자료에서도 이런 자세한 지명 정보 부족이 동일해서 이와 관련된 지리정보 관리에 대한 개선이 필요해 보인다. 다른 국가에

비해 대부분 좌표 정보를 거의 모두 제시하여 자료 활용도에서는 동아시아에서 국가나 혹은 지역 단위에서는 가장 잘 정리되어 있다.

중국은 23개의 표본관에 소장된 280 만 개의 표본 자료를 2008년까지 중앙에서 정보화를 구축하였다. 이중 약 절반이 넘는 160만 개의 자료가 GBIF에 출판되었으나, 자료 정제에서 흔하게 확인되는 단순오류는 20여 개 서로 다른 표본관에서 독립적인 기록을 하면 예를 들어 월과 일을 바꿔 기록하거나 혹은 숫자 3을 5로, 8을 6, 또는 3으로 기록하거나 7을 1로, 9를 7로 기록된 오류가 빈번하게 확인된다. 1945년 이전 표본의 경우에는 북한처럼 빈번한 행정개편이 있었던 것은 아니지만 2차세계대전 이후 14개 성이 3개 성으로 정리된 지명 때문에 북한의 경우처럼 지명에 대한 혼란이 다수 존재한다. 다른 형태의 오류는 채집자가 3-4명 혹은 그 이상임에도 불구하고 표본레이블에 대표자 이름만이 기록되어 赵大昌, 巴拉诺夫, 朱有昌을 “赵大昌等”으로 기록하는 부분등이 다수 존재한다. 데이터정제를 못해 동일 인물임에도 불구하고 러시아 채집자였던 B.W.Skvortsov (1896-1980)는 司克窝尔错夫나 Skvortzov, B.W. 등으로 기록하여 동일 인물을 알지 못하면 채집자명을 통일하기 어렵게 되어 있다(Williams et al., 2002). 중국의 자료는 표본에 근거한 모든 자료를 비교적 충실하게 기록하여 자료정제과정을 거쳐 수정이 가능하다는 장점이 있다.

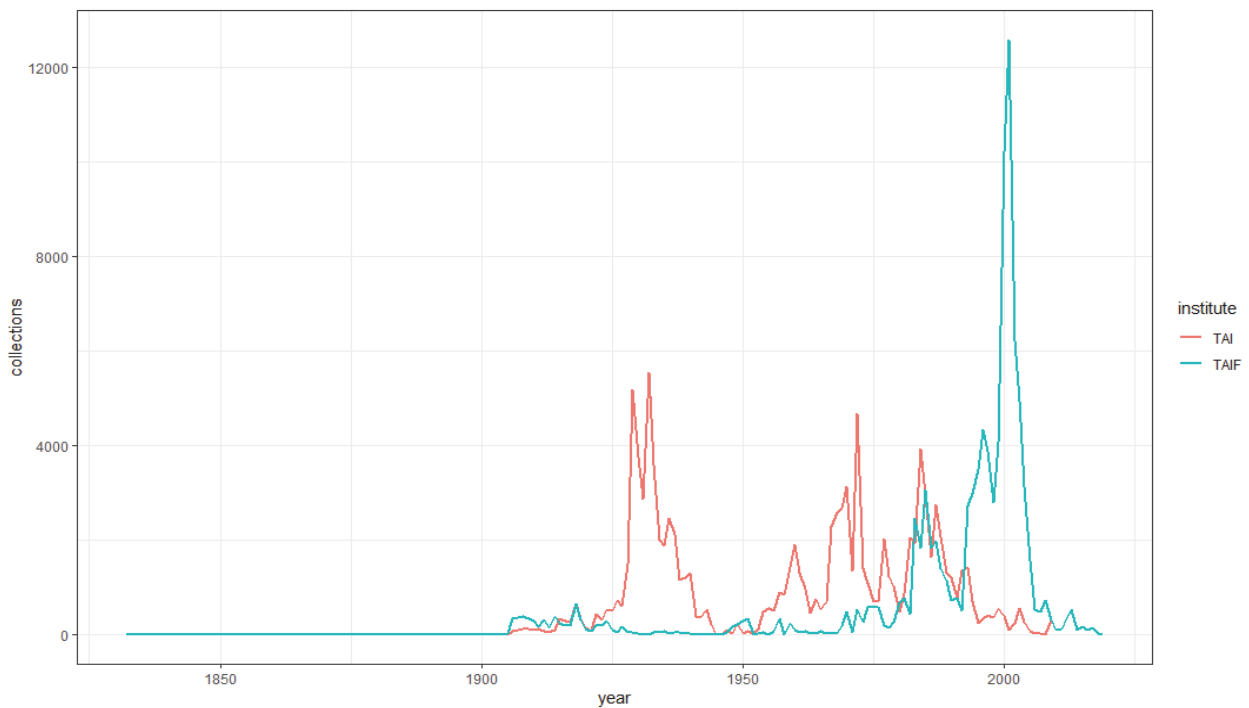


Figure 5. Occurrence per year graphs based on the number of herbarium records published through the GBIF by TAI (Taiwan National University) and TAIF (Taiwan Forestry Research Institute) in Taiwan.

고찰

생물다양성정보 데이터교환은 오래전부터 이루어져 왔으며 초기에는 특정 분류군 연구(revision)에 대한 연구의 기본 자료 분양에서 사용되다가 점차 분포 연구와 종분포 모델링, 생물지리학분석, 계통지리학 분석 및 보전과 관련된 계획으로 점차 분야를 넓혀가고 있다. 종분포 모델링 분야로의 확장은 1차 종발생 자료의 품질, 특히 종분포 모델링에 직접 이용 적합한 자료인지에 대한 근본적인 질문을 하게 된다(Chapman, 2005a). 자료의 직접적인 구축에서 GBIF와 같은 대규모의 플랫폼에서 자료를 받는 형식이 다양화 하면서 이러한 이용적합도(fitness for use)에 대한 분석이 자료의 이전에 이미 제공되는 방향으로 점차 발전하고 있다. 국내의 종분포모델링 연구에서 GBIF자료의 경우 해당 자료의 분류정보의 정확도, 좌표 정보의 정밀도에 대한 파악 없이 기계적으로 사용되는 경우가 대부분이라서 자료에 대한 정밀한 검토를 통한 연구의 예를 찾기 어렵다(Do et al., 2017). 따라서, GBIF에 대한 자료출판시 다양한 정보의 활용에 적합한 수준으로 데이터를 가공하고 정밀도에 대한 정보를 제공하는 것이 필요하다(Chapman et al., 2020).

동아시아의 생물다양성 정보에 대한 공통된 문제점은 구조화되지 않은 데이터베이스를 사용하고 평면적인 스

프레드시트형 자료구조를 사용한다. 자료 생성 및 관리단계에서 이렇게 구조화된 데이터베이스를 사용하지 않을 경우 생물다양성 정보 특성상 학명, 인명, 지명, 문헌, 생태 정보에 대한 데이터 무결성(data integrity)을 해결하지 못한다. 동아시아에서는 국가별로 데이터베이스를 자체 개발하여 자료를 관리하고 있다고 주장하지만, 명명자의 균일함, 학명의 관리, 지명에 대한 통일성 등에 대한 데이터 무결성을 해결하지 못하는 수준의 데이터베이스를 사용하고 있는 것으로 확인된다. 문제점은 외국에서 통용되는 보편화된 관리 프로그램을 채용하면 자료 정제의 단순원인인 위에 구조화된 데이터베이스를 사용하여 개선될 수 있음에도 불구하고, 각 국가의 자체 데이터베이스프로그램의 개발에 몰두하고 있어 자료의 정제와 관련된 후진성은 지속되고 있다. 국제적으로 생물다양성자료 특히 분류학자료를 표준화, 구조화하려는 노력이 있어 왔으며 분류 데이터실무그룹(TDWG; Taxonomic Data Working Group)의 주도로 관련정보를 유형별로 표준화하고 있다. TDWG에 의해 제안된 다윈코어(Darwin-core)로 알려진 주요 데이터 항목의 표준화는 국내 주요기관들이 이를 전혀 적용하고 있지 않아 자료 통합성이나 호환성이 떨어지며 이런 이유로 국내 자료의 등재도 어렵고 자료의 신뢰도도 낮음과 동시에 국제적인 플랫폼에서 검색되지 않아 국제적으로 자료 활용성이 낮다(Berendsohn et al., 2011).

자료 입력의 과정이 아무리 효율적이라 하더라도 오류는 발생하는데(Goodwin et al., 2015), 데이터 유효성 및 수정 자체를 무시할 수 없기 때문에 유럽과 미국에서는 데이터 정제는 정보 관리에서 중요한 부분으로 인식한다. 즉 오류를 수정하는 과정보다는 오류 방지를 우위에 두고 오류 예방 및 데이터 수정을 자료 관리 정책의 중요 부분으로 판단한다(Chapman, 1999; Chapman, 2005a, Chapman, 2005b). 생물다양성정보학의 자료는 종 단위의 학명을 중심으로 정리하는데 데이터베이스에서 학명의 정제와 통계 문제는 국제적으로 큰 고민거리이다. Catalogue of Life나 GBIF에서 학명의 정이명에 대한 문제와 올바른 이름 사용에 대해 Plant list와 IPNI와 같은 웹사이트가 정보를 제공하지만 해당 국가나 지역에서 대표되는 정이명 목록, 즉 종목록(체크리스트 checklist)의 확보는 가장 중요한 핵심 기술에 해당된다(Berendsohn et al., 2011). 현재 동아시아 국가들이 출판한 자료에서 제시된 학명은 국제적 수준에 미흡하며 (Figure 3C, D) 그 원인은 잘 정제된 학명기반 자료가 없는 상태에서 종발생자료를 다루고 있어 통일성이 결여된다.

자료오류에는 단순원인과 복합원인으로 구분하는데 (Rahm and Hong, 2000), 단순원인은 자료입력의 오류로 철자오류, 중복입력, 상충되는 자료입력이며 특히 자료 통합의 구조가 형성되지 않는 경우를 지칭한다. 반면에 복합원인은 자료가 상충되면서 일관성이 결여된 자료 자체가 이질적이거나 일관성이 없는 내용으로 구성되는 경우를 말한다. 자료의 질적 관리를 위해서는 정확하고 일관성이 있는 자료를 제공해야 하며 이런 목적을 위해 중복되거나 반복되는 자료는 제거해야 한다. 특히 자료의 질은 유효성, 정확성, 완성도, 일관성 및 동질성을 포함하는 부분이기 때문에 단순 오류를 수정하는 것을 자료 정제라는 잘못된 인식뿐만 아니라 자료의 무결성을 이해하지 못하는 상황이 지속적으로 발생한다. 국내에서는 국가 연구기관(예, 국립수목원)이나 일반 연구에서 자료 정제 혹은 자료정제를 표본 오동정을 수정하거나 좌표수정 하는 수준으로 인식하여 오남용하는 사례가 있다(Shin, 2014). 실제 표본관에서 확인되는 오동정의 비율은 기존 Kim(2017)의 연구에 의하면 물푸레나무과, 피나무과, 녹나무과를 대상으로 국립수목원과 국립생물자원관의 17,517점 표본을 검토한 결과 0-67%로 분류군별로 오동정의 빈도는 차이가 많이 난다. 연구된 3개과중 물푸레나무과가 가장 오동정 빈도가 높았고 이중 들메나무(48.3%)와 물들메나무(39.2%)와 같은 종의 오동정은 매우 높은 분류군으로 인식된다. 단순 오류인 분류학적인 오동정의 경우 조사대상 전체의 평균 오동정률은 10.4%정도이지만 오동정이 전혀 없는 분류군부터 67.1%인 매우 높은 오동정 분류군들이 섞여 있어 분류군별로 표준화하기가 어렵고 오동정률도 분류군별 경

향을 찾기가 어렵다. 이는 표본 자료(종발생자료)를 사용함에 있어 신뢰도를 크게 떨어뜨리며 자료 분석의 결과에 대한 혼란을 야기할 수도 있다. 평균 10%정도의 오동정이 확인되지만 이런 평균값은 개괄적인 자료의 속성에 대해 이해도를 높일 수 있으나 자료 관리에는 단순 참고의 수치에 불과하다.

생물다양성자료의 품질을 떨어뜨리는 단순원인으로는 학명 입력의 오타, 채집자의 서로 다른 이름의 입력(영명/국명, 영명의 각기 다른 철자, 약어로 입력된 경우 등), 동일 채집임에도 다른 학명으로 동정된 경우, 행정구역이 변화되어 동일 지역임에도 서로 다른 지명으로 기록된 경우, 월과 날짜를 바꿔 입력한 경우, 채집자가 여러 명임에도 불구하고 대표채집자 이름 1명이 기록된 경우 혹은 동일 지명임에도 입력자가 기록자에 의해 모두 다르게 기록된 경우이다. 예로서 복제표본은 동일 채집이지만 기관별로 소장된 상태에서 서로 다른 동정이 되어 있어 데이터베이스를 통한 이런 자료의 복구나 수정은 비교적 쉽게 진행된다. 표본 자료를 발생자료로 이용하는 목적으로 일본이나 중국의 경우 채집자의 정보나 채집번호 그리고 좌표정보에 대한 자료만을 제시하면서 채집지에 대한 정보를 누락시키는 경우가 대부분을 차지하는데 자료 정제의 필수적인 정보인 채집번호와 채집지역에 대한 정보 부재는 일차적인 자료정제의 가능성을 배제하여 자료 수준 향상의 장애요인이 된다.

복합 원인은 오류 2-3개가 중복되는 경우이거나 혹은 입력된 자료가 엑셀이나 다른 프로그램에서 데이터베이스로 전환하면서 구조적 문제가 발생하는 경우이다. 대만에 위치한 老佛山은 Laofoshan 혹은 Mt. Laofu로 기록되는데 이에 해당되는 지명은 Taiwan, Pingtung, Manzhou (屏東縣 滿州鄉)이거나, Taiwan, Pingtung, Hengchun (屏東縣 恆春鎮)의 동일 지명이 존재한다. 이런 예는 북한에 대한 자료에서도 쉽게 확인이 되는데 1950년 미군과 중공군의 전투로 유명한 ‘장진호’의 지역은 일본식 발음과 한자와 다른 영문표기로 Chosin Reservoir, Changjin-ho, Jangjin-ho, 장진호 등 여러 이름이 존재한다. 과거 자료의 가장 큰 문제점은 지명에 대한 서로 다른 방식의 기록과 자주 바뀌는 지명으로 인해 정확한 지리정보 좌표의 확보가 어려웠다. 특히, 1952년 이후 북한의 행정개편에 의해 경성군으로 통합된 주을온면 보상동(浦上洞)은 Hojodo, Hojyodo, Hohado로 각기 기록되어 있는데 이런 예는 북한 지명에서 매우 흔하게 접하는 문제점이다. 1945년 이전의 채집품에는 일본발음으로 기록된 지명이 가장 큰 장애가 되며 지리 참조(georeference)라 해서 각 채집자의 채집경로를 확인해서 정확한 좌표화를 시도하는 작업이다. 지명 표기에 대해서는 각 국가별로 지역별(예 제주도)로 지명사전(gazetteer)

의 전문성도 필요하지만 자료관리 차원에서 좌표에 대한 정보가 이런 복합 원인의 문제점을 해소할 수 있다.

생물다양성자료의 최근 이용 패턴에서 주목되는 것은 좌표 오류로 자료 입력시 수치를 변환하는 과정이나 지명에 좌표를 부여하는 지리참조연산 과정에서 오류가 필연적으로 발생한다. 동아시아 자료의 경우 일부 지리참조연산을 통해 좌표를 부여한 예가 확인되지만 국내의 경우 국립생물자원관, 국립수목원의 경우처럼 아예 모든 자료의 좌표를 누락함으로써 자료 이용적합도 기준으로 쓸모 없는 자료를 출판하는 상황이 불행히도 대부분이다. 국가기관에서 자료 품질관리가 안되는 이유는 자료를 관리하는 우수한 데이터베이스 프로그램이 없는 상태에서 자료 입력시 균질한 자료관리가 불가능하고 지리적 오류, 단순 좌표의 입력오류에도 있지만 지리참조연산시 정밀도를 부여하는 등의 수준관리가 결여되어 있다(Chapman, 2005b). 실제 지역에 대한 지명의 제시는 각 국가의 지명을 로마자로 표기하는 것이 가장 바람직하지만 과거의 수 십 만개 표본을 모두 데이터베이스에 정리하지 않기 때문에 각 기관에서는 어려운 작업으로 인식하고 있다. 개별 지명에 지리참조연산을 시도한다면 단순 대표 좌표만을 제시하는 수준보다는 정확한 지리정보를 제시할 수 있는데, 현재 동아시아 국가기관에서 일정한 수준의 좌표의 정확성을 표시하고 좌표를 제공하는 곳은 단 한군데도 없기 때문에 유럽, 북미에서 시행하는 수준과는 현격한 격차를 보인다(Chapman, 2005a, 2005b).

위의 단계에 오류의 점검과 정제 및 정제한 내용에 대한 문서화, 그것을 추후 관리의 단계가 추가할 필요가 있다. 자료의 정제 과정은 오류의 원인을 밝히는 차원에서 중요하며 그 결과를 통해 같은 오류가 다시 발생하지 않도록 오류의 정제와 오류의 예방은 반드시 같은 시점에 진행되어야 한다(Chapman, 2005a). 관리자는 이런 오류 탐지와 정제 과정, 그 결과의 기록, 분석을 통해 관리 지침을 수정하거나 추가하는 등 향후 오류 예방을 위한 대책을 바로 마련하고 시스템에 적용시켜야 한다. 모든 과정은 데이터베이스를 기반으로 진행되고 기록되어야 하며 검증된 선진 프로그램인 BRAHMS (Pouwer et al., 2008), Specify (Beach, 2018), Symbiota (Gilbert et al., 2020) 사용이 필수적이다. 자료정제와 관련된 의사소통은 일종의 자료 동업자(partnership)의 기본적 공동체 의식이 필요하고 자료에 대한 책임성과 상호 검증기능이 포함되어 피상적인 자료의 방치나 관리보다는 책임소재에 대한 명확한 흐름이 명백하게 제시되어야 한다. 자료 생성이나 오류 수정 그리고 일관성이나 정밀, 정확성을 위해 늘 기록을 정제화하여 오류시 무엇이 원인이며 이를 어떻게 수정하여 관리할지 기본적인 관리 개념이 필요하고 자료정제의 문서화는 이런 일

련의 과정이 투명성을 가져야 한다.

GBIF에 출판된 각 국가별 자료에서 중국의 경우 연도만을 제시하고 날짜를 숨기거나 정확한 지명을 공개하지 않는 것이나 일본 역시 채집자나 채집번호 혹은 정확한 지명에 대한 정보의 비공개 그리고 대만의 균일화되지 않은 지명이나 학명의 사용 등은 자료 공개 및 관리의 국제적인 공유 목적과는 어긋난다. 국내에서도 여전히 자료 공개에 대한 것보다는 자료를 제한적으로 GBIF와 같은 플랫폼에 올리는 행위 역시 국제적인 시각에서 수준이 매우 낮다. 국내 최대 GBIF출판 기관인 국립생물자원관과 국립수목원은 표본 정보의 자료를 구축하고 온라인상에 공개만 개별적으로 시도하였을 뿐 적극적인 활용에 대해서는 결과물을 제시하지 못하고 있고, 따라서 표본관의 체계적인 관리와 활용의 부재에 원인이 있다. 디지털화하여 시스템을 구축하였다고 해도 실제로 이용하지 않거나, 혹은 쉽게 이용할 수 없다면 죽은 정보로서 정보의 축적과 디지털화와 더불어 표본의 관리(큐레이션) 방향과 방법의 변화를 추구하는 선진국과는 큰 격차를 보인다(Scoble, 2010). 외국의 전문가들의 요구는 사용자와의 소통을 통한 오류의 탐지와 정제의 필요성이며(Orr 1998, Stribling et al., 2003), 온라인 상에서 전문가와 비전문가간 사이 끊임없는 소통을 통한 자료 관리가 지속적으로 향상되어야 한다(Anderson et al., 2020). 자료 정제를 통해 데이터베이스와 실제 표본과의 정보의 격차를 줄여 신뢰도를 높이며 누구나 온라인에서 정보에 쉽게 접근하고 다운받을 수 있도록 하는 목표 지향성을 가져야 한다. 결론적으로 국내에서는 생물다양성 정보학에 대한 정확한 이해와 인식의 기준으로 보다 전문적인 자료관리가 요구된다.

References

- Anderson, R.P., Araújo, M., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T. and Soberón, J. 2016. Final report of the task group on GBIF data fitness for use in distribution modelling. Global Biodiversity Information Facility. Copenhagen. pp. 27.
- Anderson, R.P., Araújo, M.B., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T. and Soberón, J.M. 2020. Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Frontiers of Biogeography* 12(3): 1-14.
- Bebber, D.P., Carine, M.A., Wood, J.R.I., Wortley, A.H., Harris, D.J., Prance, G.T., Davidse, G., Paige, J., Pennington, T.D., Robson, N.K.B. and Scotland, R.W. 2010. Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences of the United States*

- of America 107(51): 22169-22171.
- Beach, J. 2018. Specify Collections consortium—building durable infrastructure. *Biodiversity Information Science and Standards* 2: e26860.
- Berendsohn, W.G. 2009. Data and information management and communication. pp. 253-272. In: Barthlott, W., Linsenmair, K.E. and Porembski, S. (Ed.). *Biodiversity: Structure and Function – Volume I*. EOLSS Publishers. Oxford, UK.
- Berendsohn, W.G., Güntsch, A., Hoffmann, N., Kohlbecker, A., Luther, K. and Müller, A. 2011. Biodiversity information platforms: From standards to interoperability. *ZooKeys* 150: 71-87.
- Chavan, V. and Krishnan, S. 2003. Natural history collections: A call for national information infrastructure. *Current Science-Bangalore* 84(1): 34-42.
- Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L. and Ram, K. 2021. RGBIF: Interface to the global biodiversity information facility API. R package version 3.5.2.93 <https://cran.r-project.org/package=rgbif>. (2021. 3. 15).
- Chapman, A.D. 1999. Quality control and validation of point-sourced environmental resource data. In *Spatial accuracy assessment: Land information uncertainty in natural resources*. K. Lowell and A. Jatton (eds.), Ann Arbor Press, Chelsea.
- Chapman, A.D. 2005a. Principles and methods of data cleaning: Primary species and species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility. <http://www.gbif.org/document/80528>. (2021. 3. 15).
- Chapman, A.D. 2005b. Principles of data quality. Global Biodiversity Information Facility. <https://doi.org/10.15468/doc.jrgg-a190>. (2021. 3. 15).
- Chapman, A.D. et al. 2020. Developing standards for improved data quality and for selecting fit for use biodiversity data. *Biodiversity Information Science and Standards* 4: e50889.
- Do, M.S., Lee, J. W., Jang, H. J., Kim, D. I., Park, J. and Yoo, J. C. 2017. Spatial distribution patterns and prediction of hotspot area for endangered herpetofauna species in Korea. *Korean Journal of Environment and Ecology*, 31(4): 381-396.
- Fuentes, N., Pauchard, A., Sánchez, P., Esquivel, J. and Marticorena, A. 2013. A new comprehensive database of alien plant species in Chile based on herbarium records. *Biological Invasions* 15(4): 847-858.
- GBIF.org. 2020. GBIF.org (24th Dec 2020) GBIF Occurrence Download (Taiwan) <https://www.gbif.org/occurrence/download/0172317-200613084148143>. (2020.12.24).
- GBIF.org. 2021a. GBIF.org(29th Jan 2021) GBIF Occurrence Download (China) <https://www.gbif.org/occurrence/download/0176738-200613084148143>. (2021.01.29).
- GBIF.org. 2021b. GBIF.org(29th Jan 2021) GBIF Occurrence Download (Korea) <https://www.gbif.org/occurrence/download/0176754-200613084148143>. (2021.01.29).
- GBIF.org. 2021c. GBIF.org(29th Jan 2021) GBIF Occurrence Download (Japan) <https://www.gbif.org/occurrence/download/0144048-200613084148143>. (2021.01.29).
- Gilbert, E., Franz, N. and Sterner, B. 2020. Historical overview of the development of the symbiota specimen management software and review of the interoperability challenges and opportunities informing future development. *Biodiversity Information Science and Standards* 4: e59077.
- Goodwin, Z.A., Harris, D.J., Filer, D., Wood, J.R.I. and Scotland, R.W. 2015. Widespread mistaken identity in tropical plant collections. *Current Biology* 25(22): R1066-R1067.
- Gwinn, N.E. and Rinaldo, C.A. 2009. The biodiversity heritage library: Sharing biodiversity with the world. *The International Federation of Library Associations and Institutions Journal* 35(1): 25-34
- Hardisty, A., Roberts, D. and The Biodiversity Informatics Community. 2013. A decadal view of biodiversity informatics: Challenges and priorities. *BMC Ecology* 13(1): 16-39.
- Kier, G. and Barthlott, W. 2001. Measuring and mapping endemism and species richness: A new methodological approach and its application on the flora of Africa. *Biodiversity & Conservation*, 10(9): 1513-1529.
- Kim, H.W. 2017. Status assessment and cause of herbarium database errors -Selected woody plants taxa stored in national herbarium of Korea- (Dissertation). Seoul. Seoul National University, MS.
- Orr, K. 1998. Data quality and systems theory. *Communications of the ACM* 41(2): 66-71.
- Peterson, A.T., Knapp, S., Guralnick, R., Soberó N, J. and Holder, M.T. 2010. The big questions for biodiversity informatics. *Systematics and Biodiversity* 8(2): 159-168.
- Pouwer, R., Willemsse, L.P.M., Mols, J.B. and Wieringa, J.J. 2008. Guidelines for collection data registration with BRAHMS 6. Nationaal Herbarium Nederland. Leiden, The Netherlands.
- Rahm, E. and Do, H.H. 2000. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin* 23(4): 3-13.
- Sarkar, I.N. 2007. Biodiversity informatics: Organizing and linking information across the spectrum of life. *Briefings in Bioinformatics* 8(5): 347-357.
- Scoble, M. J. 2010. Rationale and value of natural history

- collections digitisation. *Biodiversity Informatics* 7(2): 77-80.
- Shao, K.T., Lai, K.C., Lin, Y.C., Chen, L.S., Li, H.Y., Hsu, C.H., Lee, H., Hsu, H.W. and Mai, G.S. 2013. Experience and strategy of biodiversity data integration in Taiwan. *Data Science Journal* 12: WDS61-WDS69.
- Shin, C.H. 2014. Report on improvement of the Herbarium specimens infrastructure for forest biodiversity on the Korean Peninsula. Korea National Arboretum. <https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=TRKO201500014016&dbt=TRKO&rn=>. (2021. 03. 15).
- Stribling, J.B., Moulton, S.R. and Lester, G.T. 2003. Determining the quality of taxonomic data. *Journal of the North American Benthological Society* 22(4): 621-631.
- Wen, J., Ickert-Bond, S.M., Appelhans, M.S., Dorr, L.J. and Funk, V.A. 2015. Collections-based systematics: Opportunities and outlook for 2050. *Journal of Systematics and Evolution* 53(6): 477-488.
- Williams, P., Margules, C.R. and Hilbert, D.W. 2002. Data requirements and data sources for biodiversity priority area selection. *Journal of Biosciences* 27(4): 327-338.

Manuscript Received : April 25, 2021

First Revision : May 27, 2021

Accepted : May 28, 2021