생물종 데이터의 정제 원칙과 방법

옮긴이: 이주원, 박형선, 안성수

KiSTi
www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

생물종 데이터의 정제 원칙과 방법

KiSTi
www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

생물종 데이터의 정제 원칙과 방법

초판 인쇄: 2006 년 6 월 30 일
초판 발행: 2006 년 6 월 30 일

최종 갱신: 2007 년 3 월 27 일
번역 버전: v1.0

옮긴이 | 이주원, 박형선, 안성수
펴낸이 | 조영화

주소 | 대전시 유성구 어은동 52-11 번지 한국과학기술정보연구원
전화 | (042) 828-5067
팩스 | (042) 828-5179
www.kbif.re.kr

이 책은 Arthur D. Chapman 이 GBIF DIGIT 연구 프로그램의 산출물로 작성한 생물종 데이터의 정제 원칙과 방법(PRINCIPLES AND METHODS OF DATA CLEANING) 자료를 원저자의 허락을 받고 번역한 것입니다. 이 번역물이 국내의 생물다양성데이터를 인터넷상에서 공유하고 활용하려고 할 때 참고자료로 사용되고 도움이 될 수 있기를 바랍니다. 단, 이 책을 참조할 경우 참조한 사실을 반드시 인용해야 합니다.
원본 파일은 다음 URL 에서 다운로드할 수 있습니다.

- http://www.gbif.org/prog/digit/data_quality/URL1124374342
- http://www.kbif.re.kr/Download/DIGIT/data_cleaning.pdf
- http://www.kbif.re.kr/Download/DIGIT/data_cleaning_korean.pdf

Published by KISTI(Korea Institute of Science and Technology Information)
Printed in Republic of Korea

이 책에 대한 의견이나 조언을 주시고자 할 경우, 또는 오자, 탈자, 오류 등을 발견했을 경우 언제든지 다음의 저자에게 이메일로 연락주시기 바랍니다.

한국과학기술원 이주원 (b_corbomite@hotmail.com)
한국과학기술정보연구원 박형선 (seonpark@kisti.re.kr)
한국과학기술정보연구원 안성수 (ssahn@kisti.re.kr)

표지 디자인 | 박양숙 (greenish3@kisti.re.kr)

ISBN 89-5884-640-2 93470

Chapman, A. D. 2005. *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

This paper was commissioned from Dr. Arthur Chapman in 2004 by the GBIF DIGIT programme to highlight the importance of data quality as it relates to primary species occurrence data. Our understanding of these issues and the tools available for facilitating error checking and cleaning is rapidly evolving. As a result we see this paper as an interim discussion of the topics as they stood in 2004. Therefore, we expect there will be future versions of this document and would appreciate the data provider and user communities' input.

Comments and suggestions can be submitted to:

Larry Speers
Senior Programme Officer
Digitization of Natural History Collections
Global Biodiversity Information Facility
Universitetsparken 15
2100 Copenhagen Ø
Denmark
E-mail: lspeers@gbif.org

and

Arthur Chapman
Australian Biodiversity Information Services
PO Box 7491, Toowoomba South
Queensland 4352
Australia
E-mail: papers.digit@gbif.org

July 2005

Helophilus pendulus (Linnaeus, 1758)

목차

데이터 정제

데이터 정제(Data Cleaning)는 관련 문서인 *데이터 품질의 원칙 (Principles of Data Quality)* (Chapman 2005a)에서 언급된 것처럼 정보관리사슬(Information Management Chain)의 핵심적인 부분이다. 해당 문서에서 강조된 것처럼, 오류 예방은 오류 탐지 및 정제보다 훨씬 더 나은 것으로, 오류를 예방하는 것이 나중에 이러한 것을 찾아 수정하는 것 보다 비용이 더 적게 들고 효율적이기 때문이다. 데이터 입력 절차가 아무리 효율적이더라도, 오류는 여전히 발생할 것이고 따라서 데이터 검증과 교정은 등한시 될 수 없다. 오류 탐지, 검증, 그리고 정제는 특히 오래된 데이터(예를 들어, 지난 300 년 동안에 걸쳐 수집된 박물관 및 식물표본관 데이터)에 대해 중요한 역할을 하며, 따라서 오류의 방지 및 데이터 정제 둘 모두는 기관의 데이터 관리 정책에 포함되어야 한다.

데이터 정제의 중요한 산출물 하나는 탐지되는 오류의 근본 원인을 파악하는 것이고 이 정보를 이용하여 데이터 입력 절차를 개선하고 이러한 오류가 재발하는 것을 방지하는 것이다.

이 문서는 1 차 생물 수집물 데이터베이스에서 오류를 탐지하고 정제하는 방법 뿐만 아니라 예방하는 방법에 대해 조사할 것이다. 이 문서는 박물관과 식물표본관이 정보를 디지털화, 문서화, 그리고 검증할 때 최선의 실행사례(best practice)를 적용할 수 있도록 지원하는 여러 지침서, 방법, 및 도구에 대해 논의한다. 하지만 그보다 먼저, 모든 데이터 정제 작업에서 따라야 할 일련의 간단한 원칙들을 소개할 것이다.

정의: 데이터 정제

부정확, 불완전, 또는 비논리적인 데이터를 결정하여 탐지된 오류와 누락된 부분의 수정을 통해 품질을 개선하는데 이용되는 과정. 이 과정은 형식 검사, 완전성 검사, 논리성 검사, 제한 검사, (지리적, 통계적, 시간적 또는 환경적인) 특이점(outlier) 또는 그 외 다른 오류를 동정하기 위한 관련 데이터의 검토, 그리고 주제 분야의 전문가(예, 분류학 전문가)에 의한 데이터 평가를 포함할 수 있다. 이러한 과정은 통상적으로 의심되는 레코드를 표시하고, 문서화하고, 그리고 이에 따른 검사와 수정 단계를 거친다. 검증 검사는 또한 적용할 수 있는 표준, 규칙, 그리고 관례에 대한 준수 검사와 관련될 수 있다.

데이터 정제의 일반적인 뼈대는 다음과 같다 (Maletic & Marcus 2000):

- 오류 유형을 정의하고 결정한다;
- 오류 사례를 검색하고 파악한다;
- 오류를 수정한다;
- 오류의 사례와 오류 유형을 문서화한다; 그리고
- 앞으로의 오류를 줄이기 위해 데이터 입력 절차를 수정한다.

서로 다른 사람들이 대체로 동일한 절차를 뜻하기 위해 사용하는 몇 개의 용어가 있다. 어느 것을 사용할 지는 개인의 선호도에 따라 결정된다. 다음과 같은 용어들이 있다:

- 오류 검사 (Error Checking);
- 오류 탐지 (Error Detection);
- 데이터 검증 (Data Validation);
- 데이터 정제 (Data Cleaning);
- 데이터 세척 (Data Cleansing);
- 데이터 세탁 (Data Scrubbing); 그리고

- 오류 수정 (Error Correction).

필자는 아래 세 개의 하위 과정을 포함하는 *데이터 정제 (Data Cleaning)* 용어의 사용을 선호한다. 즉,

- 데이터 검사와 오류 탐지;
- 데이터 검증; 그리고
- 오류 수정.

네 번째 요소가 있다면 오류 예방 절차의 개선이 아마도 포함될 수 있을 것이다.

데이터 정제의 필요성

데이터 정제의 필요성은 데이터의 오류를 줄이고 이것의 문서화 및 표현 방식의 개선을 통해 사용자에게 "이용에 적합"할 수 있도록 데이터 품질의 개선에 중심을 두고 있다 (관련 문서 참고: *데이터 품질의 원칙* – Chapman 2005). 데이터 내의 오류는 보통 존재하고 예상되는 일이다. 레드만 (Redman) (1996)은 극도의 노력이 취해지지 않았다면, 1-5% 정도의 (데이터베이스 테이블의) 필드 오류 비율이 예상된다고 제안하였다. 오류와 불확실성에 대한 통상적인 관점은 이것들은 나쁘다라는 것이지만, 오류와 오류 전이에 대한 올바른 이해가 있으면 전반적인 데이터 품질 측면에서 적극적인 품질 조정과 관리향상을 도모할 수 있다 (Burrough and McDonnell, 1998). 공간적인 위치 (geocoding) 그리고 동정 관련한 오류는 종-발생 데이터 오류의 두 가지 주요 원인이고, 이 논문에서 다루는 것은 이러한 오류의 정제이다. 데이터의 오류를 수정하고 품질이 좋지 않은 레코드를 제거하는 것은 시간이 많이 소요되고 지루한 작업이 될 수 있지만 (Williams *et al.* 2002) 이것이 무시될 수는 없다. 하지만 단순히 오류가 삭제만 되는 것이 아니라 관련된 수정사항이 문서화되고 변경사항이 기록되는 것이 중요하다. *데이터 품질의 원칙*에서 언급된 것처럼, 원래의 데이터를 분리된 하나 또는 여러 개의 필드에 유지를 하면서 데이터베이스에 수정사항을 추가하는 것이 가장 좋으며, 이것은 원래의 정보로 되돌릴 수 있는 기회를 항상 두기 위한 것이다.

오류는 어디에 있는가?

1 차 종 데이터는 전체 영역의 데이터를 포함한다 – 박물관과 식물표본관의 데이터부터, 관찰 데이터 (지점-기반, 권역 또는 지역-기반, 그리고 시스템 또는 그리드-기반), 그리고 체계적이거나 그렇지 못한 조사 데이터까지 포함한다 (Chapman 2005a). 많은 박물관과 식물표본관 수집물의 역사적인 성질때문에 (이것은 종종 레거시 데이터(legacy data)라고 불려진다), 많은 수의 레코드는 이것들이 수집된 장소에 대한 대략적인 서술 이외에 지리 정보를 거의 가지고 있지 않다 (Chapman and Milne 1998). 역사적인 데이터의 경우, 지리코드 좌표가 있다고 해도 종종 아주 정확하지 않으며 (Chapman 1999) 대개 수집자가 아닌 다른 사람들에 의해 나중에 추가되었다 (Chapman 1992). 이러한 많은 데이터는 종의 분포 연구에 사용될 때 단점이 된다. 관찰 및 조사 데이터 또한 많은 연구에 귀중한 레코드가 되고 지리참조연산(georeferencing) 정보가 꽤 정확할 수 있지만, 그러나 확증된 참조 자료가 거의 보관되지 않기 때문에, 이것의 명명 및 분류학적 정보는 문서화된 박물관 수집물보다 신뢰도가 일반적으로 떨어진다. 그리고 조사 및 관찰 레코드의 지리참조연산 정보는 여전히 오류 또는 모호성을 포함할 수 있는데 예를 들어, 지리코드가 그리드의 중심을 가리키는지 또는 그리드 기반 레코드에서 한쪽 구석을 가리키는지 명확하지 않을 수 있다.

많은 (박물관 및 관찰) 데이터는 체계적이라기보다 임기응변적으로 수집되었고 (Chapman 1999, Williams *et al.* 2002) 이것은 커다란 공간적인 편향을 초래할 수 있다 – 예를 들면, 도로 또는 강 네트워크와 연관이 높은 수집물이 있다 (Margules and Redhead 1995, Chapman 1999,

Peterson *et al.*2002, Lampe and Riede 2002). 박물관 및 식물표본관 데이터 그리고 대부분의 관찰 데이터는 일반적으로 특정한 시간에 실물의 존재 정보만을 제공하고 다른 임의의 장소 또는 시간에 부재 관련하여서는 어떠한 것도 제공하지 않는다 (Peterson *et al.*1998). 이것으로 인해 이러한 것들의 이용이 일부 환경 모델에서만 제한되고 있지만, 이러한 데이터는 지난 200 년 이상에 걸쳐 이제까지 우리가 가질 수 있는 가장 방대하고 완벽한 생물학 정보 수집물로 남아있다. 새로운 조사로 이러한 데이터를 대체하는 비용은 엄청날 것이다. 한번의 조사를 수행하기 위해 1 백만 달러를 초과하는 것은 흔치 않은 일이 아니다 (Burbidge 1991). 더구나, 오랜 시간에 걸친 수집때문에, 이것들은 인간이 이러한 다양성에 거대한 영향을 끼친 시간동안 생물학적 다양성에 대한 대체할 수 없는 기준선 데이터를 제공한다 (Chapman and Busby 1994). 이것들은 환경 보전을 위한 임의의 노력에 필수적인 자원인데, 그 이유는 이것들이 농지 개척, 도시화, 기후 변화로 인해 서식지 변화를 겪었거나 다른 방식으로 변화되었을 지역에 대해 유일하게 온전히 문서화된 종 발생 레코드를 제공하기 때문이다.

오류를 예방하기

앞에서 강조된 것처럼, 오류의 예방은 나중에 수행하는 오류의 수정보다 낫고, 오류를 예방하는 기관들을 지원하기 위한 새로운 도구들이 개발되고 있다.

데이터베이스화된 수집물에 지리참조연산 정보를 추가하는 절차를 지원하기 위한 도구들이 개발되고 있다. 이와 같은 도구들은 eGaz (Shattuck 1997), geoLoc (CRIA 2004a), BioGeomancer (Peabody Museum *n.dat*.), GEOLocate (Rios and Bart *n.dat*.) 그리고 Georeferencing Calculator (Wieczorek 2001b)가 있다. 2005 년 고돈베티무어재단(Gordon and Betty Moore Foundation)이 지원하고 세계적인 공동 작업을 필요로 하는 한 프로젝트가 현재 많은 이러한 도구를 하나로 통합하는 시도를 하고 있으며 이러한 도구를 독립형 오픈 소스 소프트웨어 도구와 웹 서비시스(Web Services) 둘 모두로 이용 가능할 수 있게 하는 것을 목표로 하고 있다. 그렇지만 더 심도 있고 다양한 검증 도구들의 필요성은 부정할 수 없다. 이러한 도구들은 이 논문의 뒤에서 상세하게 논의될 것이다.

분류학적 및 명명 데이터의 오류를 감소시키는 지원 도구들이 또한 개발되고 있다. 이러한 데이터에 오류의 두 가지 주요 원인이 있다. 이것들은 (분류의 경우) 부정확한 동정 또는 잘못된 동정이고, (명명법의 경우) 철자가 틀린 경우이다. 분류군 동정을 돕는 도구들은 개선된 분류 방법들, (권역과 지역 모두의) 동식물상, 분류군에 대한 자동화된 컴퓨터 기반의 열쇠, 유형 및 다른 표본에 대한 디지털 사진을 포함한다. 이름의 철자 관련해서 지구적, 권역적 그리고 분류학적 이름-리스트가 개발되고 있으며, 이것들은 데이터를 입력할 때 오류를 감소시키는 전거 파일(authority file)과 데이터베이스 입력 체크리스트의 개발을 가능하게 하고 있다.

아마도 많은 오류를 예방하는 가장 좋은 방법은 제일 처음 데이터베이스를 올바르게 설계하는 것일 것이다. 관계형 데이터베이스 철학과 설계를 올바르게 실현함으로써, 종의 이름, 지역 정보 및 기관과 같이 자주 반복되는 정보는 단지 한번만 입력되고 처음 시작할 때만 검증될 필요가 있을 것이다. 그 후 참조 무결성(referential integrity)은 향후 입력 작업의 정확성을 보호하게 된다.

공간 오류

공간적인 관점에서 종 데이터의 품질을 결정할 때, 검사할 필요가 있는 몇 가지 쟁점사항이 있다. 이러한 것들은 표본 또는 관찰 대상의 신원(잘못된 동정은 이 레코드를 해당 분류군의 기대 영역 밖에 위치시킬 수 있고 따라서 공간적인 오류로 보일 수도 있다), 지리코딩상의

오류 (위도와 경도), 그리고 수집물 데이터에서의 공간적인 편향을 포함한다. 분류군이 지리적으로 잘못 배치되거나 잘못 동정될 수도 있는 것들을 찾을 때 공간적인 기준을 사용하는 것은 도움이 될 수 있다. 공간적인 편향의 문제는 – 식물표본관과 박물관의 데이터에서 아주 명확하며 (예: 도로를 따라 수집된 수집물), 개별적인 식물 또는 동물 표본 레코드들의 정확성과 관련되어 있기 보다는 앞으로의 수집물과 미래의 조사 설계에 대한 문제이다 (Margules and Redhead 1995, Williams *et al*. 2002). 실제로 수집물의 편향성은 임의의 어느 하나의 개별 수집물보다는 개별 종 수집물의 전체와 더 관련이 있다. 한 지역 내에서 생물학적 수집물의 전반적인 공간적 및 분류학적인 적용 범위를 향상시키고, 그리하여 공간적인 편향성을 줄이기 위해서는, 생태학적으로 가장 가치 있는 향후 조사의 장소 결정에, 예를 들어 지리적인 기준 (기후 등) 뿐만 아니라 환경적인 기준을 이용함으로써, 기존의 역사적인 수집물 데이터가 이용될 수 있다 (Neldner *et al.* 1995).

명명 및 분류학적인 오류

이름은 생물 종 데이터베이스의 정보를 접근할 때 핵심 열쇠가 된다. 이름이 틀렸다면 사용자가 원하는 정보를 접근하는 것이 불가능하지는 않더라도 어려울 것이다. 약 100 년 동안 생물학적 명명 규칙이 있었음에도 불구하고, 데이터베이스에서 명명 및 분류학적 정보는 (the *Classification Domain* of Dalcin 2004) 오류를 탐지하고 정제할 때 때때로 가장 어려운 부분이다. 일차 종 데이터베이스 사용자들에게 가장 많은 고뇌와 자신감 상실을 초래하는 분야가 또한 이 분야이다. 이것은 때때로 사용자들이 분류학적 변경사항과 명명법적인 변경사항의 필요성을 알고 있지 못한 때문이기도 하지만, 부분적으로 사용자들에게 이러한 변경사항을 온전히 문서화하고 설명하지 않은 분류학자들, 이름과 분류군 간의 관계에서 복잡성, 그리고 일차 종 데이터베이스에서 종종 잘 다루어지지 않는 분류학적 개념의 혼동 때문이기도 하다 (Berendsohn 1997).

이러한 오류 가운데 정제하기 더 쉬운 것은 명명적인 데이터 – 틀린 철자에 관한 것이다. 이름(과 동의어)의 목록은 이 작업을 돕는 핵심 도구이다. 많은 목록이 여러 권역과 분류 그룹에 대해 이미 존재하고, 이러한 것은 지구적인 목록으로 점진적으로 통합되고 있다 (Froese and Bisby 2002). 그렇지만 신뢰할 수 있는 목록을 갖지 못한 세계의 많은 권역과 분류학적 그룹들이 여전이 존재한다.

분류학적 오류 – 수집물의 부정확한 동정 또는 잘못된 동정으로 인한 것은 탐지하고 정제하기에 가장 어려운 오류이다. 박물관과 식물표본관은 전통적으로 분류학 그룹에서 종사하는 전문가들이 가끔 표본들을 검사하고 이러한 것의 범위를 결정하거나 동정하는 것을 지원하는 *결정(determinavit)* 시스템을 운영하여 왔다. 이것은 검증된 방법이지만 시간이 많이 걸리고 대체로 무작위로 행해졌다. 그렇지만 가까운 또는 심지어 장기적인 미래에도 자동화된 컴퓨터 동정이 하나의 선택사항(option)이 될 것 같지 않기 때문에 이것 이외에 어떤 다른 방법이 있을 것 같지 않다. 하지만 이러한 과정을 돕는 이용 가능한 많은 도구가 있다. 이것들은 우리 모두가 친숙한 전통적인 분류학 출판물과 최신의 전자 도구이다. 전통적인 도구는 분류학 개정판, 국가와 지역의 동물과 식물상, 그리고 삽화가 있는 점검표(checklists)와 같은 출판물 등이 있다. 더 새로운 도구로는 분류군에 대한 자동화되고 컴퓨터에 의해 생성된 열쇠(keys); 그림, 설명문, 열쇠, 그리고 그림이 있는 어휘 목록을 가진 대화식 전자 출판물, 문자-기반 데이터베이스; 사진 편집 도구; 기준(모식) 표본의 사진을 포함하는 과학적인 사진 데이터베이스, 수집물에 대한 체계적인 사진들, 그리고 (과학적 및 다른 방법으로 모두 검증된) 쉽게 접근할 수 있는 온라인 사진들이 있다.

데이터베이스를 통합하기

두 개 또는 그 이상의 데이터베이스를 통합하는 것은 (두 데이터베이스 간에 차이가 있는 경우) 오류를 찾아내고 새로운 오류를 (즉, 중복되는 레코드) 만들어 낼 것이다. 중복되는 레코드들은 통합할 때 표시를 해두어 이것들이 분석을 한쪽으로 편향되게 할 경우 이러한 것이 동정되고 분석에서 제외될 수 있도록 하여야 하지만, 일반적으로 삭제되지는 않아야 한다. 중복된 것으로 보이지만 많은 경우 두 데이터베이스의 관련 레코드들은 각각에 고유한 어떤 정보를 포함할 수도 있으므로, 중복된 것에서 단지 하나를 삭제하는 것('병합과 제거'라고 알려져 있음 (Maletic and Marcus 2000))은 귀중한 데이터 손실을 야기할 수 있기 때문에 항상 좋은 선택사항은 아니다.

데이터베이스를 통합할 경우 일어날 수 있는 추가적인 문제는 서로 다른 분류학적 개념, 측정치에 관한 서로 다른 가정 또는 단위의 사용, 그리고 서로 다른 품질 제어 메커니즘과 같은 각기 다른 기준에 기반을 둔 데이터의 통합이다. 이러한 통합은 개별 데이터 단위의 출처를 항상 문서화하여 데이터 정제 과정이 서로 다른 출처의 데이터에 대해 서로 다른 방식으로 수행될 수 있도록 하여야 한다. 이것을 하지 않으면, 임의의 변경이 있을 때 데이터베이스에 효과적으로 정제하고 문서화하는 일이 더욱 어려워 질 것이다.

데이터 정제의 원칙

데이터 정제 원칙의 많은 부분이 관련 문서인 *데이터 품질의 원칙* (Chapman 2005a)에서 다룬 일반적인 데이터 품질 원칙과 중복이 된다. 핵심 원칙들은 다음과 같다:

계획은 필수적이다 (비전, 정책, 그리고 전략을 수립하기)

올바른 계획은 훌륭한 데이터 관리 정책의 필수적인 부분이다. 정보관리사슬 (figure 1) (Chapman 2005a)은 데이터 정제를 기관의 데이터 품질 비전과 정책에 반영될 필요가 있는 핵심 부분으로서 포함하고 있다. 해당 기관의 문화 속에 데이터 정제와 검증을 구현하는 전략은 해당 기관의 전반적인 데이터 품질을 개선시키고 이용자와 공급자 모두에게 이 기관의 위상을 높일 것이다.

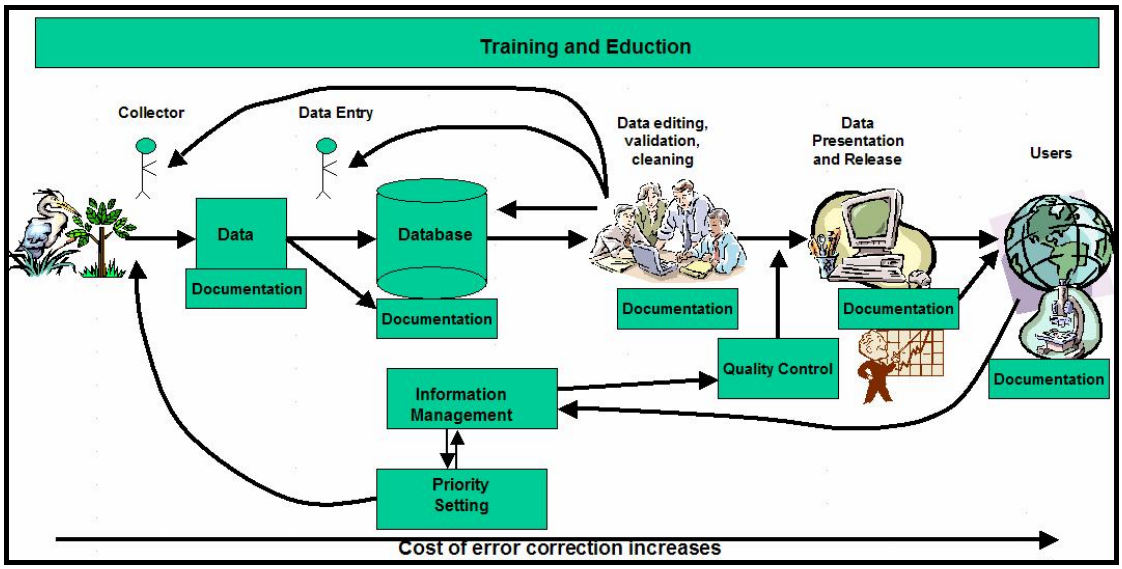

Fig. 1. *오류 수정의 비용이 사슬을 따라 이동할수록 증가한다는 것을 보여주는 정보관리사슬. 교육, 훈련, 그리고 문서화가 모든 단계에 필수적이다 (Chapman 2005a).*

데이터를 구조화하는 것은 효율성을 향상시킨다

데이터 검사, 검증 그리고 수정 전에 데이터를 구조화하는 것은 효율성을 향상시키고 데이터 정제의 비용과 시간을 상당히 줄일 수 있다. 예를 들어, 데이터를 장소에 따라 정렬함으로써, 열쇠 참조에 따라 앞뒤로 왔다갔다하기 보다 한 장소와 관련된 모든 데이터를 동시에 검사하여 효율성 증가를 이룰 수 있다. 이와 비슷하게, 레코드를 수집자와 날짜로 정렬함으로써, 특정 일자에 특정 수집자가 방문하지 않았을 장소에 대한 레코드의 오류를 탐지하는 것이 가능하다. 다양한 필드에서의 철자 오류도 또한 이런 방식으로 찾을 수도 있을 것이다.

예방이 치료보다 더욱 낫다

이전에 강조된 것처럼 (Chapman 2005a), 오류를 예방하는 것이 나중에 오류를 찾아서 수정하는 것보다 훨씬 비용이 덜 들고 효율적이다. 오류가 탐지되었을 때, 관련 피드백 메커니즘으로 데이터 입력동안 해당 오류가 다시 발생하지 않도록 하거나 재발 가능성이 매우 낮아지도록 확실히 하는 것이 또한 중요하다. 올바른 데이터베이스 설계로 데이터 입력이 확실히 제어될 수 있도록 분류군 이름, 지역 정보, 그리고 사람과 같은 개체가 단

한번만 입력되고 입력될 때 검증될 수 있도록 해야 할 것이다. 이것은 드롭-다운(drop-down) 메뉴를 이용하거나 필드 내에 기존 개체의 자동 완성 기능을 통해 수행될 수 있다.

책임은 모두에게 있다 (수집자, 관리자, 그리고 이용자)

데이터 정제에 대한 책임은 모두에게 있다. 정보관리사슬(figure 1)에서 데이터 정제 부분의 1 차 책임은 분명히 데이터 관리자(데이터 관리 및 저장의 주요 책임을 가진 개인 또는 조직)에게 있다. 수집자 또한 책임이 있고 수집자가 제공한 원본 정보와 관련된 오류 또는 모호함을 관리자가 발견했을 때 데이터 관리자의 질문에 대답할 필요가 있다. 이러한 것은 라벨상의 모호함, 날짜 또는 장소에 대한 오류 등이 관련되어 있을 수 있다. 이 문서의 후반부에서 명확해지겠지만, 이용자 또한 데이터와 연관된 문서내의 오류를 포함하여, 이용자들이 접할 수도 있는 오류 또는 누락 부분에 대한 정보를 관리자에게 피드백 해야 할 중요한 책임이 있다. 해당 데이터를 다른 데이터의 문맥에서 분석 또는 조사할 때 그냥 지나칠 수도 있었을 데이터에서 오류와 특이점을 발견하게 되는 것은 종종 이용자 자신이다. 하나의 박물관은 전체 이용 가능한 데이터 (예를 들어 하나의 주(State) 또는 권역) 중에서 단지 일부분만을 가지고 있을 수 있고, 이 데이터가 다른 출처의 데이터와 통합될 때 오류가 명백히 드러날 수도 있다. 이 문서에서 설명되는 많은 도구들이 종, 수집자 또는 탐험의 부분보다는 전체를 조사할 때 더 좋은 성능을 발휘한다.

협력 관계는 효율성을 향상시킨다

협력 관계는 데이터 정제를 관리하는 효율적인 방법이 될 수 있다. 언급된 것처럼, 이용자는 종종 데이터의 오류를 동정하기에 가장 좋은 입장에 있다. 만약 데이터 관리자들이 이러한 핵심 이용자들과 협력 관계를 발전시키면 이러한 오류들은 무시되지 않을 것이다. 협력 관계를 구축함으로써, 많은 데이터 검증 과정이 중복될 필요가 없을 것이고, 오류가 더 잘 문서화되고 수정될 가능성이 높아지며, 오류는 아니지만 의심이 가는 레코드에 대해 무의식적인 수정으로 새로운 오류들이 추가되지 않을 것이다. 관련 논문인 *데이터 품질의 원칙*에서 논의되는 것처럼 조직 외부뿐만 아니라 내부에서도 이용자들과 이러한 협력 관계를 구축하는 것이 중요하다.

우선순위 선정은 중복을 감소시킨다

구조화 및 정렬과 함께, 우선순위 선정은 경비를 감소시키고 효율성을 향상시키는데 도움이 된다. 최소의 비용으로 광범위한 데이터를 정제할 수 있는 그러한 레코드에 집중하는 것이 종종 가치 있는 일이다. 예를 들어, 좀 더 복잡한 레코드를 처리하기 전에, 일괄 처리 또는 자동화된 방법을 이용하여 검사할 수 있는 것들이 있다. 이용자에게 가장 가치가 있는 그러한 데이터에 집중함으로써, 오류를 탐지하고 수정할 가능성이 또한 한층 더 높아질 수 있다. 이것은 이용자/공급자의 관계와 명성을 향상시키고, 데이터 공급자와 이용자 모두에게 데이터의 품질을 향상시킬 수 있는 더 큰 자극제를 제공하는데 왜냐하면 데이터가 곧 바로 이용될 수 있기 때문이다.

목표 설정과 성과 측정수단

성과 측정수단은 품질 관리 절차의 가치 있는 부분으로 공간적인 메타데이터에 널리 이용된다. 이것들은 또한 한 기관이 자체의 데이터 정제 작업을 관리하는데 도움이 되기도 한다. 이와 같은 측정수단은 이용자에게 데이터와 이것의 품질에 대한 정보를 제공해 줄 뿐만 아니라, 관리자와 큐레이터가 데이터베이스에서 주의를 필요로 하는 부분을 추적할 때 사용할 수 있을 것이다. 성과 측정수단은 데이터상의 통계적인 검사 (예를 들어, 전체

레코드의 95%는 보고된 위치에서 5,000 미터 이하의 정확성을 가진다), 품질 관리 수준상의 통계적인 검사 (예를 들어, 전체 레코드의 65%는 지난 5 년 동안 검증된 분류학자가 검사하였다; 90%는 지난 10 년 동안 검증된 분류학자가 검사하였다), 완전성(예: 모든 10-도 그리드 격자가 샘플링되었다)을 포함할 수 있을 것이다 (Chapman 2005a).

중복과 데이터 재처리의 최소화

대부분의 기관에서 중복은 데이터 정제에 영향을 끼치는 주요 요소이다. 많은 기관들은 데이터베이스에 레코드를 입력할 때 지리참조연산을 동시에 수행한다. 레코드가 지리적으로 거의 정렬이 되지 않기 때문에, 이것은 같거나 비슷한 장소들이 많이 검색될 것이라는 것을 의미한다. 텍스트 위치 정보만 있고 좌표 정보가 없는 수집물에 대해 특별 연산으로서 지리참조연산을 수행함으로써, 비슷한 위치를 가진 레코드는 정렬될 수 있고 적합한 지도 또는 지명사전(gazetteer) 위에 표시될 수 있다. 일부 기관들은 데이터베이스 자체를 이용하여 해당 장소가 이미 지리참조연산이 수행되었는지 알아내기 위한 검색을 수행함으로써 데이터 중복을 또한 줄이고 있다 (아래 *데이터 입력과 지리참조연산* 참고).

(되도록 표준화된 형식으로) 검증 과정을 문서화하는 것은 또한 데이터의 재작업을 줄이는데 중요하다. 예를 들어, 이용자가 데이터에 대해서 수행한 데이터 품질 검사는 많은 의심이 가는 레코드를 발견할 수도 있을 것이다. 이러한 레코드들을 검사하게 되면, 정상적인 레코드들과 실제 특이점을 가진 것으로 나뉘게 될 것이다. 이 정보가 해당 레코드에 문서화되지 않으면, 한참 후에 다른 어떤 사람이 와서 더 많은 데이터 품질 검사를 수행하여 같은 레코드들을 의심가는 것으로 동정할 수도 있다. 그러면 이 사람은 이 정보를 다시 검사하고 데이터에 대해 다시 작업하면서 귀중한 시간을 소비해야 할 수도 있다. 데이터베이스를 설계할 때, 데이터를 누가, 언제 검사하였는지 그리고 그 결과가 어떠한 것인지를 나타내는 하나의 필드 또는 여러 개의 필드가 포함되어야 한다.

실무 경험에 의하면 정보사슬관리(figure 1 참고)를 활용하는 것은 데이터의 중복과 재작업을 줄일 수 있고 오류율을 최대 50% 감소시킬 수 있으며 품질이 낮은 데이터의 사용으로 인해 발생하는 비용의 최대 2/3 까지 감소시킬 수 있다 (Redman 2001). 이것은 주로 데이터 관리와 품질 제어에 대한 명확한 책임을 할당하고, 병목 현상 부분과 대기 시간을 최소화하고, 서로 다른 직원이 품질 관리 검사를 다시 하는 중복을 줄이고, 개선된 업무 방법 파악을 통한 효율성 증가 때문이다 (Chapman 2005a).

피드백은 양방향이다

데이터의 이용자들은 불가피하게 오류 탐색을 수행할 것이고, 이들이 그 결과를 관리자에게 피드백하는 것이 중요하다. 이미 언급하였듯이, 개별 데이터 관리자가 단독으로 일하는 것보다, 이용자들은 종종 여러 범위의 소스에서 오는 데이터의 결합을 통해 특정 오류 유형을 더 자주 탐지할 기회를 가지게 된다. 데이터 관리자들이 자신들의 데이터에 대해 이용자들의 피드백을 촉진시키고 그들이 받은 피드백을 구현하는 것이 중요하다 (Chapman 2005a). 표준 피드백 장치(mechanisms)가 개발되고, 데이터 관리자와 이용자 간에 피드백을 받는 절차가 합의될 필요가 있다. 데이터 관리자들은 또한 관련된 수집자와 데이터 공급자에게 오류에 관한 정보를 제공할 필요가 있다. 이런 방식으로 향후 오류의 발생빈도가 감소되고 전반적인 데이터의 품질이 개선될 가능성이 더욱 높아질 것이다.

교육과 훈련은 기법을 향상시킨다

불충분한 훈련은, 특히 정보품질사슬(Information Quality Chain)의 데이터 수집과 입력 단계에서, 1 차 종 데이터에서 오류의 많은 부분을 차지하는 원인이다. 데이터 수집자는 데이터 관리자와 데이터 이용자의 요구사항에 대해 교육을 받을 필요가 있으며, 이것은 올바른 데이터가 수집되고 (즉, 모든 관련된 부분과 성장 단계), 수집물의 문서화가 잘 되고 – 즉, 관련 장소 정보가 잘 기록되는 것 (예를 들어, 10km NW of Town 'y'가 길을 따라 10km 인가 또는 직선상의 10km 인가?), 표준이 관련된 곳에 적용되고 (예, 동일한 그리드 크기가 관련된 조사에 사용되기), 라벨이 명확하고 읽을 수 있으며, 되도록 일관된 방식으로 배치되어 있어 데이터 입력자들이 편리하게 작업할 수 있도록 하기 위한 것이다.

데이터 입력자들에 대한 훈련도 또한 MaPSTeDI 지리참조연산 지침서에 파악된 것처럼 중요하다 (University of Colorado Regents 2003a). 데이터 입력자들을 올바르게 훈련시키는 것은 데이터 입력과 관련된 오류를 상당부분 줄일 수 있고, 입력 비용을 줄일 수 있으며, 그리고 전반적인 데이터 품질을 향상시킬 수 있다.

책임성, 투명성, 그리고 감시성

책임성(accountability), 투명성(transparency), 그리고 감시성(audit-ability)은 데이터 정제의 필수적인 요소들이다. 즉흥적이고 계획적이지 않은 데이터 정제 활동은 매우 비효율적이며 일반적으로 비생산적이다. 데이터 품질 정책과 전략 내에서 – 데이터 정제에 명확한 책임 소재를 설정할 필요가 있다. 데이터의 "이용에 대한 적합성"과 그리하여 이것의 품질을 향상시키기 위해서, 데이터 정제 절차는 올바른 감사 추적 장치와 함께 투명하고 적합하게 문서화될 필요가 있으며 중복을 줄이고 오류가 한번 수정된 경우는 다시 발생하지 않도록 확실히 해야 한다.

문서화

문서화는 좋은 데이터 품질의 핵심 부분이다. 올바른 문서화가 되어있지 않으면, 이용자들이 데이터에 대한 이용의 적합성을 판단하기 힘들고 관리자들은 누가 어떠한 데이터 품질 검사를 수행하였는지 알기 어렵다. 문서화에 일반적으로 두 가지 종류가 있고, 이것에 대한 규정이 데이터베이스 설계에 반영되어야 한다. 첫 번째 것은 각 레코드 및 레코드들과 관련이 있으며 누구에 의해 어떠한 데이터 검사가 수행되었고 어떠한 변경이 있었는지에 대한 것이다. 두 번째 것은 데이터 집합 수준에서 정보를 기록하는 메타데이터이다. 둘 모두 중요하며, 이러한 것들이 없다면, 좋은 데이터 품질을 유지하는 것은 어렵다.

데이터 정제의 방법

서론

전 세계의 박물관과 식물표본관은 자신들의 수집물에 대해 증가되는 비율로 데이터베이스 구축을 시작하고 있고, 적어도 이 정보의 일부를 인터넷을 통해 이용할 수 있도록 하는 일을 하고 있다. 수집물의 데이터베이스화 비율은 이 과정을 지원할 수 있는 여러 도구와 방법론의 개발로 최근 증가하였고 "인터넷을 통해 세계의 생물다양성에 관한 일차 데이터를 자유롭게 그리고 보편적으로 이용할 수 있게 하는 것"(GBIF 2003a)을 목표로 하는 세계생물다양성정보기구(Global Biodiversity Information Facility, GBIF)의 태동 이후 출판이 증가하였다.

좋은 실행사례(관련 문서 참고 – *데이터 품질의 원칙* 그리고 *데이터 정제의 원칙,* 이 문서) 뿐만 아니라, 데이터 정제 작업을 자동화하는, 즉 이것을 폭 넓게 지원하는 유용하고 강력한 도구에 대한 요구가 있다. 자동화된 방법들은 이 절차의 일부분만이 될 수 있고 이 과정을 지원하기 위한 새로운 도구 개발과 이러한 이용이 최선의 실행사례 (best practice) 업무에 통합되도록 하는 것에 대한 지속적인 요구가 있다. 데이터를 수동으로 정제하는 것은 손이 많이 가고 시간이 많이 소요되며, 그 자체가 오류를 발생하기 쉽지만 (Maletic and Marcus 2000), 이것은 1 차 종-발생 데이터베이스에서 계속해서 중요한 위치를 차지할 것이다. 가능하다면, 이것은 소규모 데이터 집합, 즉 다른 어떤 방식으로도 검사될 수 없는 단지 약간의 오류에 대해 마지막 수단으로서 수행되어야 한다.

개발된 기술 중의 일부는 기후 공간(Chapman 1992, 1999, Chapman and Busby 1994)과 지리 공간(CRIA 2004b, Hijmans *et al*. 2005, Marino *et al.* in prep.)에서 특이점을 동정하기 위한 기후 모델의 이용, 자동화된 지리참조연산 도구의 이용 (Beaman 2002, Wieczorek and Beaman 2002) 그리고 다른 많은 것들이 있다. 대부분의 수집 기관은 데이터 관리 기법 또는 지리정보시스템(Geographic Information System, GIS)에 대한 고수준의 전문 지식을 보유하고 있지 않다. 이러한 기관이 필요로 하는 것은 지리코드 정보를 포함하여 데이터와 정보 모두의 입력을 도울 수 있는 간단하면서 값싼 일련의 도구와 데이터 검증을 위해 값비싼 GIS 소프트웨어와 연결하지 않고도 이용할 수 있는 유사하고 간단하지만 비싸지 않은 도구이다. 일부 도구들이 데이터 입력을 지원하기 위해 개발되었다 – Biota (Colwell 2002), BRAHMS (University of Oxford 2004), Specify (University of Kansas 2003a), BioLink (Shattuck and Fitzsimmons 2000), Biótica (Conabio 2002) 그리고 다른 도구들은 데이터베이스 관리 및 연관된 데이터 입력을 제공하고 있으며 (Podolsky 1996, Berendsohn *et al.* 2003), eGaz (Shattuck 1997), geoLoc (CRIA 2004a, Marino *et al.* in prep.), GEOLocate (Rios and Bart *n.dat*.) 그리고 BioGeoMancer (Peabody Museum *n.dat*.)는 수집물의 지리참조연산을 지원하는 도구들이다. 또한 여러 기관들이 데이터베이스 프로그램을 설정하고 관리할 때 각 기관을 지원할 수 있는 많은 지침 문서가 인터넷상에서 제공되고 있다. MaNIS 지리참조연산 지침서 (Wieczorek 2001a), MaPSTeDI 지리참조연산 지침서 (University of Colorado Regents 2003a) 그리고 HISPID (Conn 1996, 2000)가 그 예들이다.

1 차 종 및 종-발생 데이터베이스에서 오류 정제를 지원할 수 있는 많은 방법과 기법이 있다. 이것들은 수백 년 동안 박물관과 식물표본관에서 사용되었던 방법들부터 아직 많은 부분 검증되지 않은 자동화된 방법들까지 다양하다. 이 논문은 종 데이터베이스를 정제하기 위한 여러 가지 방법들을 자세하게 살펴보고, 가능하다면 관련 사례들을 소개할 것이다. 많은 기관들이 자신들만의 고유 기법과 방법을 개발하였기 때문에 이것들이 결코 포괄적인 목록은 아니다.

자연사 수집물 자체의 고유한 특성 때문에, 모든 지리코드 정보가 매우 정확하다든지, 또는 데이터베이스 내에 일정한 수준의 정확성이 있다고는 말할 수 없다. 하지만 극도로 낮은 정확성을 가진 데이터라고 해서 꼭 낮은 품질을 가지는 것은 아니다. 품질은 데이터가 실제로 사용되고 있을 때에만 존재하게 되고 데이터 자체의 특징은 아니다 (연관 논문인 *데이터 품질의 원칙* 내용 참고 – Chapman 2005a). 품질은 단순히 이용에 대한 적합성 또는 잠재적인 이용에 대한 하나의 요소이고 상대적인 용어이다. 중요한 것은 데이터의 이용자들이 데이터 자체를 보고 요구되는 응용 프로그램에서 데이터가 적합할 것인가를 결정할 수 있도록 하는 것이다. 따라서 각각 주어진 지리코드의 정확성이 데이터베이스 내부에 기록되어야 한다. 필자는 이것이 범주화 되지 않은 형태인 미터 단위로 기록되는 것을 선호하지만, 많은 데이터베이스들은 이 목적을 위해 범주화된 코드를 개발하였다. 이 정보가 이용 가능할 때, 예를 들어, 이용자는 특정 미터 값(5,000 미터)보다 나은 데이터만을 요청할 수 있을 것이다 (University of Colorado Regents 2003b 에서 데이터를 추출하기 위해 코드를 사용하는 예 참고). 지리코드 레코드의 정확성을 판단하는 많은 방식이 있다. 필자가 생각하기에 지점-반지름 방식(Wieczorek *et al.* 2004)이 가장 쉽고 가장 실용적인 방식이며 이전에 호주에서 이용 목적으로 권장되었다 (Chapman and Busby 1994). 자동화된 지리참조연산 도구들의 출력 결과에서 한 개의 필드에 계산된 정확성을 포함하는 것이 또한 중요하다. 아직 개발 중인 geoLoc(CRIA 2004a, Marino *et al.* in prep.)과 BioGeomancer (Peabody Museum *n.dat.*) 도구는 이 특징을 포함하고 있다.

여러 기관들이 새로운 레코드들의 위치 기록에 (GPS 와 같은) 더 정확한 도구를 사용하고 과거의 레코드들이 수정되고 개선됨에 따라 앞으로 종 수집물 데이터 자원들이 개선될 것으로 기대된다. 수집자들은 GPS 를 사용하기 전에 정보를 기록했던 가장 세분화된 방법이라는 역사적인 이유 때문에 데이터 기록을 위해 1 분 해상도를 가지는 GPS 를 사용하지 않기 보다는 자신들이 이용할 수 있는 가능한한 최상의 도구들을 모두 이용하는 것이 또한 중요하다. 만약 이렇게 된다면, 그들은 데이터베이스에 적합한 정확성이 추가될 수 있도록 확실히 해야 하며, 그렇지 않을 경우 GPS 가 사용되었을 당시 정확성이 실제 2,000 미터였을지라도 10 미터 정확성으로 가정될 수도 있을 것이다. 오류 예방은 오류 탐지보다 선호되지만, 오류 예방 자체만으로 모든 잠재적인 오류를 막을 수 있다고 보장할 수 없기 때문에 오류 탐지의 중요성은 여전히 강조되어야 한다.

분류 및 명명 데이터

이름은, 이러한 것이 2 항으로 이루어진 학명(scientific binomials)이든 일반 이름(common name)간에, 대부분의 종 및 종-발생 데이터베이스에 첫 번째로 입력되는 것이다. 이름의 오류는 여러 가지 방식으로 발생할 수 있다: 동정이 잘못될 수 있고, 이름의 철자가 틀리거나 또는 형식이 잘못될 수도 있다 (또는 사용자가 예상하고 있지 않은 것일 수 있다). 이러한 것들 중 첫 번째는 지루한 노력 없이 검사하거나 교정하기가 쉽지 않으며, 분류 전문가의 서비스를 필요로 한다. 그렇지만 나머지 것들은 이러한 오류가 발생하지 않도록 또는 최소화될 수 있도록 올바른 데이터베이스 설계와 데이터 입력을 지원하는 방법을 이용하여 조금 더 쉽게 처리될 수 있다.

A. 동정 확실성

전통적으로, 박물관과 식물표본관은 분류학 그룹에서 종사하는 전문가들이 때때로 표본을 검사하고 이것들의 동정을 결정하는 동정 또는 "*결정(determinavit)*" 시스템을 운영하여 왔다. 이것은 보다 큰 개정 연구의 일부로 수행되거나, 또는 다른 기관을 방문중인 전문가가 그곳에 머무르면서 수집물을 검사하면서 수행될 수도 있다. 이것은 검증된 방법이지만 시간이 많이 걸리고 대체로 무작위로 행해졌다. 자동화된 컴퓨터 동정이 가까운 또는 심지어 장기적인 미래에도 선택사항이 될 것 같지 않기 때문에 이것 이외에 다른 방법이 있을 것 같지 않다.

i. 데이터베이스 설계

첫 번째 옵션은 동정을 할 때 이것의 확실성에 대해 어떤 표시를 제공하는 데이터베이스의 필드를 추가하는 것이다. 이것이 수행될 수 있는 많은 방식이 있고, 표준 방법론의 개발 관련하여 아마도 약간의 토론이 필요할 것이다. 이것은 코드 필드일 수 있고, 아래와 같은 맥락에서 할 수도 있을 것이다:

- 해당 분류군의 세계적인 전문가에 의해 높은 확실성으로 동정됨
- 해당 분류군의 세계적인 전문가에 의해 타당한 확실성으로 동정됨
- 해당 분류군의 세계적인 전문가에 의해 약간 불확실하게 동정됨
- 해당 분류군의 지역적인 전문가에 의해 높은 확실성으로 동정됨
- 해당 분류군의 지역적인 전문가에 의해 타당한 확실성으로 동정됨
- 해당 분류군의 지역적인 전문가에 의해 약간 불확실하게 동정됨
- 해당 분류군의 비전문가에 의해 높은 확실성으로 동정됨
- 해당 분류군의 비전문가에 의해 타당한 확실성으로 동정됨
- 해당 분류군의 비전문가에 의해 약간 불확실하게 동정됨
- 수집자에 의해 높은 확실성으로 동정됨
- 수집자에 의해 타당한 확실성으로 동정됨
- 수집자에 의해 약간 불확실하게 동정됨

이러한 것을 어떻게 등급화해야 할지는 다소 논의가 필요하고, 같은 식으로 이러한 것이 가장 최선의 범주인지 아닌지도 논의가 필요하다. 일부 기관에서는 분명히 이런 성질의 필드를 이미 가지고 있다. HISPID 표준 버전 4 (Conn 2002)는 좀 더 단순화된 버전을 포함하고 있다 – 5 개의 코드가 있는 검증 수준 플래그 (Table 1).

많은 기관들은 또한 확실성 기록 형식을 이미 가지고 있으며 다음과 같은 용어들을 사용하고 있다 : "aff.", "cf.", "*s. lat.*", "*s. str.*", "?". 이러한 것의 일부(aff., cf.)는 엄격한 정의가 있지만,

각 개인이 이러한 것을 이용하는 것은 상당히 다양할 수 있다. *sensu stricto(좁은 의미로)*와 *senso lato(넓은 의미로)*의 사용은, 항상 이런 방식으로 사용되지는 않지만, 확실성의 수준보다는 분류학적 개념의 변이사항을 암시한다.

0	레코드의 이름은 어떤 권위자에 의해서도 검사되지 않았다
1	레코드의 이름은 명명된 다른 식물 이름과 비교하여 결정되었다
2	레코드의 이름은 식물표본관 그리고/또는 도서관 그리고/또는 문서화된 최신 자료를 이용하여 분류학자 또는 다른 능력 있는 사람들에 의해 결정되었다
3	식물의 이름은 체계적인 개정 그룹에 종사하는 분류학자에 의해 결정되었다
4	레코드는 무성 방법(asexual methods)에 의한 기준(모식) 자료의 일부이거나 기준(모식) 자료에서 파생되었다

Table 1. *HISPID (Conn 2000)의 검증 수준 플래그.*

다른 방법으로서, 이름이 분류학 전문지식이 아닌 다른 것에서 도출되었을 경우, 사용된 이름의 출처를 나열할 수 있다 (Wiley 1981 자료):

- 새 분류군의 기재사항
- 분류학적 개정판
- 분류방법
- 분류학적 열쇠
- 동물상과 식물상 연구 결과
- 도감
- 목록표
- 점검표
- 편람
- 명명법에 대한 분류학적 법칙/규칙
- 계통발생학적 분석

ii. 데이터 입력

데이터베이스에 데이터가 입력될 때, 전문가가 이름을 검사하였는지 아닌지, 그리고 위의 제시된 필드가 있다면 이러한 정보가 입력되었는지에 대해 검사를 할 수 있다. 이러한 필드가 사용된다면, 이용 가능한 옵션을 한정시키고, 따라서 오류가 추가될 가능성을 감소시키는 점검표 또는 전거 파일을 이용하여 이것들이 입력되어야 한다.

iii. 오류 검사

지리코드 검사 방법(아래, *공간 데이터* 참고)은 지리 또는 환경 공간에서 특이점 탐지를 통해 잘못된 동정 또는 부정확한 동정을 식별하는데 또한 도움이 될 수 있다. 일반적으로 지리코드 검사를 통해 발견되는 특이점은 위도나 경도의 오류일 수 있지만, 가끔 표본이 현재 연구 중인 분류군으로 잘못 동정되고 이에 따라 해당 분류군의 일반적인 기후, 환경 또는 지리적인 범위 밖에 위치된 것을 의미한다. 지리적인 특이점을 동정하는 기법에 관한 좀 더 상세한 논의는 아래를 참고하기 바란다.

그렇지만 수집물이 정확하게 동정되었는지 아닌지를 탐지하는 주된 방법은 전문가들이 현존하는 표본 또는 확증 수집물을 조사함으로써 그 동정을 검사하는 것이다. 지리코드 특이점 탐지 방법들은 수집물이 정확하게 동정되었는지 아닌지를 결정할 수는 없지만, 전문적인 분류군 검사를 위한 우선순위 수집물을 동정하는데 도움을 줄 수 있다. 관찰 데이터의 경우, 전문가들은, 개인적인 지식에 근거하여, 분류군이 해당 지역의 레코드일 가능성이 있다는 것이라는 것을 판단할 수 있다 (예, Birds Australia 2004); 그러나 일반적으로 확증 표본이 없을 경우 관찰 레코드의 부정확한 동정을 식별하는 것은 어려운 일이다. 많은 기관들은 확실치 않거나 의심이 가는 레코드들에 어떤 표시를 할 수 있고, 이것들이 이용에 적합한지 아닌지를 결정하는 것은 사용자의 몫이다.

B. 이름의 철자

이 논문은 일차 종 데이터베이스에 입력될 수 있는 모든 가능한 종류의 이름을 다루려고 하지는 않는다. 예를 들어, 식물 데이터베이스의 잡종과 변종, 다양한 종류의 동의어, 그리고 분류학적 개념들은 모두 특정한 쟁점사항들이 있으며 이러한 것들을 검사하는 것은 논란의 대상이 될 수 있다. 이러한 이름들이 어떻게 다루어지는가에 대한 사례는 HISPID 같은 TDWG 표준과 식물학데이터베이스의 식물이름(Plant Names in Botanical Databases) (Bisby 1994) 뿐만 아니라 다양한 국제명명규약(International Codes of Nomenclature)에서 (Conn 1996, 2000) 발견할 수 있다.

a. 학명

학명의 올바른 철자는 일반적으로 다양하고 관련된 명명법 기준중의 하나에 의해 결정된다. 그러나 오류는 타자 오류, 명명법의 모호성 등을 통해 여전히 발생할 수 있다. 이러한 오류들을 최소로 줄이는 가장 쉽고 확실한 방법은 데이터를 입력할 때 '전거 파일 (Authority File)'을 이용하는 것이다. 대부분의 데이터베이스는 변경할 수 없는 전거 파일 또는 입력하면서 갱신될 수 있는 전거 파일과 통합될 수 있도록 설정될 수 있다.

i. 데이터베이스 설계

분류군 이름에 대해 좋은 데이터 품질을 유지할 수 있는 핵심 방법 중의 하나는 해당 이름을 하나의 필드에 모두 넣기 보다는 여러 개별 필드로 나누는 것이다 (예, 속(genus), 종(species), 종이하 순위(infraspecific rank), 종이하 이름(infraspecific name), 저자 그리고 확실성). 이러한 것들을 하나의 필드에 관리하는 것은 올바른 데이터 검증과 오류 탐지의 가능성을 감소시키고, 오류가 발생될 수 있는 가능성을 상당히 증가시킬 수 있다. 예를 들어, 이름의 종 부분과 속 부분을 분리함으로써, 각각의 속 이름은 (전거 파일 또는 선택-리스트를 통해) 관계형 데이터베이스에 단지 한번만 입력될 필요가 있고, 따라서 타자나 철자 오류의 발생 가능성을 줄일 수 있다.

하나의 필드에 이름의 모든 부분을 포함하는 데이터베이스는 품질을 유지하거나 다른 데이터베이스와 통합하는 것을 아주 어렵게 할 수 있다. 이것은 또 다른 범위의 오류를 발생시킬 수 있으며 권장되지 않는다. 일부 데이터베이스는 결합된 필드와 세분화된 필드 모두를 사용하지만, 이러한 것들이 자동으로 생성되지 않고, 하나는 갱신되고 다른 것은 그렇지 않다면 이것은 추가적인 오류의 가능성을 또한 제공한다. 자동으로 생성되는 필드들은 이 위험성을 제거한다.

세분화된 데이터와 관련하여 고려될 필요가 있는 두 가지 사항은 데이터가 하나의 필드에 존재하는 데이터베이스와의 통합(예를 들어, 식물 또는 동물의 목록을 가져오는 것)과

세분화된 데이터베이스에서 데이터를 연결된 것처럼, 예를 들어 웹 사이트 또는 출판물에서, 보여주어야 할 필요성이다.

이러한 것의 첫 번째 – 연결된 필드에서 개별 필드로 데이터를 파싱하는 것은 보이는 것처럼 일반적으로 간단한 작업이 아니다. 이것은 물론 단지 이름과 관계된 문제 뿐만이 아니고, 아래에서 논의되는 것과 같이 참고 문헌과 장소 정보에서도 같은 문제가 발생한다. 필자가 알기로는 이 파싱에 사용될 수 있는 어떠한 간단한 도구도 존재하지 않지만, 많은 박물관과 식물표본관은 자체 기관 목적으로 이러한 것을 수행하였고, 따라서 이러한 기관에서 관계된 알고리듬을 구할 수도 있을 것이다.

두 번째의 경우, 웹 또는 보고서의 출력에서 연결된 결과를 보여주어야 하는 요구사항은 다양한 필드를 연결하는 데이터베이스 내의 추가 (생성된) 필드를 사용해서 또는 추출 과정동안 즉시 수행될 수 있다. 이것은 데이터베이스와 이것의 보고 메커니즘을 설계할 때 고려되어야 할 사항이다. 이것들은 분류학 커뮤니티가 간단한 도구 또는 방법론을 개발할 목적으로 토론할 필요가 있는 사항들이다.

ii. 전거 파일

전거 파일은 많은 분류학 그룹에 대해 존재하고, 여러 종류의 기관들에 의해 개발되고 있다. 많은 상위 분류군(Families, Orders, and Genera)에 대한 신뢰할 수 있는 전거 파일들이 이용 가능하고, 이것들은 이러한 필드에서 데이터 무결성을 보장하기 위해 사용될 수 있다. 모든 분류군, 특히 종 수준과 그 하위에 대해 상세한 전거 파일이 가까운 미래에 작성될 것 같지는 않지만, 현존하는 전거 파일(예, IPNI 1999, Froese and Bisby 2004)이 그 시작점으로 이용될 수 있다. 전거 파일들이 이용 가능하게 되면, 관련 데이터베이스들은 새로운 이름들이 이것들에 추가될 수 있는 방식으로 설정될 수 있다. 예를 들어, 데이터베이스가 풀다운(pull down) 리스트 기능이 있는 전거 파일을 가지고 있거나 하나의 타입으로서 필드를 자동 완성 (예를 들어 EXCEL 프로그램과 같이 앞 줄에 이미 이름이 존재할 경우, 그 이름을 입력할 때처럼) 한다고 가정하자.

1. 해당 이름 검색을 위해 풀다운 리스트를 사용한다
2. 이름이 목록에 존재하지 않는다
3. “새로운 이름” 버튼을 클릭한다
4. 새로운 이름을 추가한다.
5. 데이터베이스는 “이 이름은 <이름>과 유사합니다” 계속할까요? 라고 응답할 수 있다
6. 예
7. 목록에 이 이름이 추가되고, 다음에 이름을 추가할 때, 그 이름이 풀다운 리스트에 나타날 것이다.

이러한 방식으로, 여러분은 점차로 전거 파일을 추가하고 이것을 개선시킬 수 있을 것이다.

여분의 검사로서, 이러한 이름들은 관리자가 검증한 후에 승인 또는 거부할 수 있는 2 차 목록에 저장될 수도 있다. 데이터베이스의 복잡성 수준에 따라 이 목록은 동의어를 포함할 수 있고, 여러분이 이름을 입력하기 시작하면, 해당 이름이 <이름>의 동의어로 전거 파일에 등록되어 있는데 이것을 추가하기 원하는지를 물을 수 있을 것이다.

전거 파일은 이용될 수 있는 곳이면 어느 곳에나 이용되는 것이 권고된다. 좋은 시작점은 Species2000 & ITIS Catalogue of Life (Froese and Bisby 2004)가 있으며, 연례 점검표로서 CD 로 이용 가능하다. 이 문서의 형식은 차기 버전이 데이터베이스에 쉽게 통합될 수 있도록 개선되고 있다. 이 점검표는 개별 분류군의 점검을 위해 전자적으로 또한 이용할 수 있고,

정기적으로 갱신되고 있으며, 이것도 또한 온라인상에서 구할 수 있다. 또한, 많은 이름 데이터베이스가 존재하고 또는 개발되고 있으며 이러한 것들은 이름 전거 파일의 기초를 마련할 수 있다. 몇 가지 사례는 다음과 같다,

지구적인 목록:

- Species2000 & ITIS Catalogue of Life (Froese and Bisby 2002),
- Ecat (GBIF 2003b),
- International Plant Name Index (IPNI 1999);
- Global Plant Checklist (IOPI 2003).

지역적인 목록:

- Integrated Taxonomic Information System (Ruggiero 2001);
- Australian Plant Name Index (Chapman 1991, ANBG 2003);
- Proyeto Anthos – Sistema de información sobre los plantas de España (Fundación Biodiversidad 2005)
- Australian Faunal Directory (ABRS 2004);
- Med Checklist (Greuter *et al.* 1984-1989).

분류학적 목록:

- ILDIS World Database of Legumes (Bisby *et al.* 2002);
- Fishbase (Froese and Pauly 2004);
- World Spider Catalog (Platnik 2004);
- 다른 많은 것들.

전거 파일을 위의 것들 중의 하나와 같은 외부 소스에서 가져올 경우, 전거 소스의 버전 간에 만들어지는 여러 변경사항이 쉽게 데이터베이스에 반영되고, 데이터베이스가 쉽게 갱신될 수 있도록 데이터베이스에 소스-식별자(Source-Id)가 기록되어야 한다. 기대컨대, 조만간에 이것은 GUIDs(Globally Unique Identifiers)[1]의 사용으로 더욱 쉬워질 것이다.

iii. 중복된 입력자료

데이터베이스를 처음부터 설계하고 가능한한 최대로 정규화 하려고, 예를 들어 전거 테이블을 이용하여, 시도하더라도, 중복 레코드의 문제는 피할 수 없으며, 특히 2 차 소스(예, 이름 또는 참고 자료)에서 데이터를 가져올 때는 더욱 그러하다. 이와 같은 중복 레코드를 제거 (또는 표시)하기 위해서 특별한 인터페이스가 필요할 수 있다. 이 인터페이스는 특수 알고리듬을 이용하여 잠재적인 중복 레코드를 식별할 수 있어야 한다. 그 다음, 데이터 입력자 (또는 큐레이터, 전문가 등)는 중복 레코드의 목록에서 실제 중복이 되는 레코드들과 보관되어야 할 레코드를 판별해야 할 것이다. 그 이후, 이 시스템은 자체의 참조 무결성을 유지하면서 불필요한 레코드를 삭제하고 저장 또는 표시해야 할 것이다. 이것을 교정할 수 있는 범용 소프트웨어가 구현될 수 있지만, 파싱 소프트웨어의 경우처럼, 당장은 이러한 소프트웨어가 존재하지 않는 것처럼 보인다. 생물다양성 데이터베이스 설계자들은 이 문제를 인식하고 이러한 일에 적합한 범용 소프트웨어 도구의 설계를 고려해야 한다.

iv. 오류 검사

이름에 대해서 자동적인 검사가 어느 정도 가능하다. 완전한 종 이름 목록이 존재하지는 않지만, 특히 일부 분류군에 대해서 과 이름(family names)의 목록과 속 이름(generic names)은 (예, IAPT 1997, Farr and Zijlstra *n.dat.*) 완전성이 더욱 높다. 이러한 목록들에 대하여 관련 필드에 대한 검사가 수행될 수 있다. 종소명 (species epithet) (2 항 학명의 두 번째 부분)

[1] http://www.webopedia.com/TERM/G/GUID.html

관련하여 많은 검사를 수행할 수 있다. 예를 들어, 높은 유사도를 가지는 같은 속 내에서 이름을 찾기 – 한 개의 문자가 잘못 배치된 이름 또는 한 개의 문자 또는 여러 개의 문자가 누락된 이름 등이 있다. CRIA 데이터정제(Data Cleaning) 시스템(CRIA 2005)은 speciesLink (CRIA 2002)를 통해 얻은 분산 데이터상에서 이러한 많은 검사를 수행한다. 이 경우 최선의 실행사례는 자동화된 탐지, 그러나 자동화되지 않은 수정일 것이다. 다른 가능한 검사는 다음과 같다 (English 1999, Dalcin 2004 을 수정):

입력되지 않은 데이터 값 – 이것은 값이 존재해야 하는 공백 필드의 검색과 관련이 있다. 예를 들어, 식물학 데이터베이스에서 종이하 이름이 인용되면, 다음으로 해당되는 종이하 순위 필드에 값이 존재해야 한다; 또는 종 이름이 존재하면, 해당되는 속 이름이 또한 존재해야 한다.

부정확한 데이터 값 – 이것은 타자상의 오류, 철자 입력의 뒤바뀜, 잘못된 위치에 입력된 데이터 (예, 속 이름 필드에 있는 종소명) 그리고 값이 요구되는 필드(즉, 필수 필드)에 강제로 입력된 데이터 값, 즉, 데이터 입력자가 해당 값을 알지 못해, 임의의 값을 추가한 경우 등과 관련이 있다. 이러한 오류 가운데의 일부를 검사하는 몇 가지 방식이 있다 – 예를 들어, Soundex (Roughton and Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), 또는 Skeleton-Key (Pollock and Zamora 1984)가 있다. 이러한 각각의 방법들은 유사성을 탐지하기 위해 조금씩 다른 알고리듬을 이용한다. (언급된 것들을 포함하여) 종 이름과 많은 데이터 집합(Dalcin 2004)을 이용하여 여러 가지 방법에 대한 최근의 테스트는 테스트된 데이터 집합에서 Skeleton-Key 방식이 거짓 오류에 대해 가장 높은 참 오류 비율을 산출한다는 것을 보였다. 이러한 방식을 사용하는 온라인 사례를 브라질의 CRIA 사이트에서 볼 수 있다 (CRIA 2005). 이러한 것들은 아래에서 더 상세하게 설명된다.

세분화되지 않은 데이터 값 – 이것은 하나 이상의 사실이 입력된 필드들을 찾는 것과 관련이 있다. 예를 들어 종이하 (infraspecies) 필드에 입력된 “subsp. bicostata”는 두개의 필드로 나누어져야 한다. 데이터베이스 설계에 따라서 (윗부분 참고) 이것은 오류가 아닐 수도 있다. 세분화되지 않은 데이터 값은 많은 데이터베이스에서 발생하고 제거하기 어렵다. 이러한 값들은 데이터베이스에 아마도 새로운 필드를 생성할 필요가 있다는 것을 나타낸다. 그 이후 일부 세분화되지 않은 데이터는 자동화된 방법을 이용하여 적절한 필드로 분리될 수 있지만, 남아있는 많은 것들은 전문가의 감독 하에서 수동으로만 수정될 수 있다.

도메인 모순(Schizophrenia) – 이것은 의도되지 않은 목적으로 사용된 필드를 찾는 것과 관련이 있다. 이것은 확실성 필드가 데이터베이스에 포함되지 않았을 경우에 종종 발생하며, 물음표 표시, cf., aff.와 같은 불확실성 기호가 종소명과 같은 필드에 추가되거나 주석이 추가되는 경우이다 (Table 2). 이 ‘오류’의 성질은 데이터베이스의 설계에 또한 의존적일 수 있다.

Family	Genus	Species
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

Table 2. *도메인 모순의 예 (Chapman 2005a 자료).*

중복 발생 – 이것은 동일한 실제 값을 가리키는 이름을 찾는 것과 관련이 있다. 여기에서 발생할 수 있는 두 가지 주요 중복 유형이 있다 – 첫 번째는 잘못된 철자로 인한 오류이고, 두 번째는 한 대상에 대해서 하나 이상의 유효한 대체 이름이 있는 경우로 국제식물명명규약 (International Code of Botanical Nomenclature) (2000)은 대체할 수 있는 과(Family) 이름을 인정한다 (예, Brassicaceae/Cruciferae, Lamiaceae/Labiatae). 후자의 경우는 데이터베이스에 대해 유효한 대체 이름중의 하나를 선택하거나 기관의 정책에 따라 링크된 동의어를 사용함으로써 처리될 수 있다. 대체할 수 있는 분류 방법이 상위 분류군 순위에 적용되었을 경우 비슷한 문제가 또한 발생할 수 있으며, 이것이 속 수준에서 적용될 경우 어떠한 분류 기준을 적용했는가에 따라 종은 유효하게 하나 이상의 속에서 발생할 수 있다 (*Eucalyptus*/*Corymbia*; 남반구의 알바트로스 종, 야생의 작은 고양이 종, 그리고 더 많은 것들).

일관성이지 않은 데이터 값 – 이것은 두 개의 관련 데이터베이스가 동일한 이름 목록을 사용하지 않을 경우 발생하고, 결합 (또는 비교)되었을 때 불일치성을 보인다. 예를 들어, 이것은 식물원의 살아있는 수집물(Living Collection)과 식물표본집(Herbarium)간에 발생할 수 있고, 두개의 전문적인 데이터베이스를 통합할 때 발생할 수 있으며, 또는 박물관의 수집물 데이터베이스와 사진 데이터베이스 사이에서 발생할 수도 있다. 이것을 올바르게 고치기 위해서는 하나의 데이터베이스를 다른 것과 검사하여 불일치성을 찾아야 한다.

Dalcin (2004)은 학명의 철자 오류를 검사하는 방법에 대해 많은 상세한 실험을 수행하였고 음성 유사성을 검사하는 일련의 도구를 개발하였다. 필자는 이러한 도구를 사용하거나 테스트하지는 않았지만, 그 세부 방법과 결과 그리고 각 방법들간의 비교 테스트는 Dalcin (2004) pp. 92-148에서 구할 수 있다. 또한 브라질의 CRIA에서도 유사한 맥락에서 이름 검사 방법을 개발하였고 (CRIA 2005) 이것들은 아래의 방법 섹션에서 언급된다.

b. 일반 이름

포르투갈어, 스페인어, 영어, 힌두어, 다양한 다른 언어, 또는 지역에 기반한 토속 이름이든간에, '일반 (common)' 또는 통속 이름에 대한 고정된 기본 규칙은 존재하지 않는다. 종종 '일반' 이름이라고 불리는 것들은 실제로 (특히 식물학에서) 구어체적 이름이고 단지 라틴어 학명의 번역으로 생성된 것일 수 있다. 일부 그룹의 경우, 예를 들어 조류(Christidis and Boles 1994 참고)와 어류(Froese and Pauly 2004), 합의된 관례와 권장 영어 이름들이 개발되었다. 많은 그룹에서 동일한 분류군은 많은 일반 이름을 가질 수 있으며, 이것들은 종종 특정 지역, 언어 또는 민족에 근거한다. 하나의 예는 *Echium plantagineum* 종으로 호주의 한 주에서는 'Paterson's Curse'로, 다른 주에서는 'Salvation Jane'으로, 그리고 다른 언어와 국가에서는 또 다른 이름(예, Viper's Bugloss, Salvation Echium)과 같이 여러 가지로 알려져 있다. 역으로, 동일한 일반 이름이 여러 분류군에 사용될 수도 있으며, 이것은 때로는 다른 지역에서 때로는 심지어 같은 지역에서도 그럴 수 있다.

하나의 언어에서, 아마도 일부 작은 그룹을 제외하고, 일반 이름을 표준화하는 것은 거의 불가능하다. 그러나 이러한 시도를 하고 그렇게 하는 것이 어떤 의미가 있는가 (Weber 1995)? 진정한 일반 이름은 오랜 시간에 걸쳐 만들어지고 발전된 이름이며, 이러한 것의 목적은 사람들이 의사 소통을 하기 위한 것이다. 필자가 여기에서 제시하고자 하는 것은 일반 이름이 표준화되어야 한다는 것이 아니고 데이터베이스에 저장될 때 표준화된 방식으로 되고 이것의 출처가 문서화되어야 한다는 것이다. 1 차 종 발생 데이터의 많은 사용자들은 일반 이름을 이용하여 데이터에 접근하기를 원하므로, 우리가 가능한한 가장 많은 사용자들에게 우리의 데이터를 최대한 이용할 수 있게 하려면 데이터베이스에 이러한 것들을 저장하는 것은 가치가 있는 일이다. 일반 이름을 기록하고 개별 이름의 출처를

문서화하는 표준 방법들을 채택함으로써, 이것이 한 개 언어 또는 여러 개의 언어로 된 하나 또는 수백 개의 종이든 간에, 우리는 검색을 할 수 있고, 따라서 더욱 효율적이고 유용하게 데이터를 추출할 수 있다.

일반 이름을 종 데이터베이스에 포함시킬 때 많은 어려운 점들이 있다. 이러한 것들은 다음과 같다:

- 데이터베이스에 저장할 때 유니코드의 사용이 요구되는 비-라틴어의 이름. 다음과 같은 문제점들이 발생할 수도 있다:
    - 데이터베이스가 단지 라틴 알파벳만을 이용하여 음성학적인 이름만을 저장하려는 경우,
    - 사람들이 화면상에서 또는 인쇄할 때 해당 문자를 올바르게 출력할 수 없는 경우,
    - "라틴" 키보드만을 가진 사용자가 검색을 수행하는 경우,
    - 라틴어와 비-라틴어 문자가 혼합된 이름을 입력하는 경우,
- 이름의 언어에 관한 정보를 저장할 필요, 특히 혼합된 언어로 구성된 이름이 포함되는 경우,
- 지역적인 요소에 대한 정보를 저장할 필요 – 이름과 관련될 수 있는 지역, 사투리 등,
- 이름의 출처에 대한 참조 문헌과 같은 정보를 저장할 필요.

오류를 줄이면서 정확하고 현저하게 효용성을 증가시키는 방식으로 일을 하는 것은 결코 쉬운 일이 아니다. 이러한 이름들을 포함하기로 결정하였다면, 아래 사항들이 어느 정도 표준화를 진행하는데 도움이 될 수 있을 것이다.

i. 데이터 입력

일반 이름을 데이터베이스화할 때, 구축 과정에서 일관된 몇몇 형태를 따르는 것이 권고된다. 아마도 각각의 개별 데이터베이스가 내부적으로 일관성이 있는 것이 가장 중요할 것이다. 가능하다면 지역적 또는 국가적 표준을 개발하는 것이 권장된다. 모든 언어와 모든 분류군에 대해 표준을 개발하기에는 너무 많은 언어와 지역적인 차이가 있어 이것을 시도할 수 없지만, 여기에서 제시되는 개념의 일부는 다루어지는 언어 이외의 다른 언어에서 표준의 기반이 될 수 있을 것이다.

영어와 스페인어에 대한 일반 이름의 경우, 필자는 환경호주(Environment Australia) (Chapman *et al.* 2002, Chapman 2004)에서 이용될 목적으로 개발된 것과 비슷한 관례를 따르기를 추천한다. 이러한 지침들은 관련 기관들이 가지고 있는 많은 데이터베이스의 일관성 유지를 지원하기 위해 개발되었다. 이러한 관례중의 하나는 이름의 각 단어를 대문자로 시작하는 것이다.

Sunset Frog

일반적이거나 그룹화된 이름의 경우 하이픈이 권고된다. 하이픈 뒤에 오는 단어는 일반적으로 대문자를 쓰지 않으며, 예외적으로 조류의 경우 이것이 더 큰 그룹의 일원일 경우 크리스티디스와 보울스(Christidis and Boles, 1994)가 권고한 것처럼 하이픈 뒤의 단어를 대문자로 시작한다.

Yellow Spider-orchid
Double-eyed Fig-Parrot ('Parrot'은 대문자로 시작하는데 이것은 Parrot 그룹의 한 구성원이기 때문이다).

포르투갈어 일반 이름은 일반적으로 모두 소문자를 사용하며, 명사의 경우 모든 단어 사이를 보통 하이픈으로 연결하고, 명사와 형용사인 경우 공백으로 분리한다. 포르투갈어 일반 이름을 쓸 때 이 관례를 따르거나 영어와 스페인어 사례를 따를 수 있도록 관례가 수정되는 것이 권고된다.

mama-de-cadela,
fruta-de-cera
cedro vermelho

아포스트로피가 일반 이름에 사용되어야 하는가에 대해 약간의 불일치와 혼란이 존재한다. 지리적인 이름의 경우, 모든 아포스트로피를 무시하는 경향이 증가하고 있고 (예: Smiths Creek, Lake Oneill), 이것은 호주에서 현재 수용되고 있는 실행사례이다 (ICSM 2001, Geographic Names Board 2003 Art. 6.1). 필자는 유사한 관례가 일반 이름에 채택되기를 추천하지만 현재 이렇게 하는 것은 요구사항이 아니다.

이름이 하나 이상의 언어로 추가되고 그리고/또는 지역, 방언, 또는 원주민간에 다를 경우, 해당 언어와 지역 정보는 이름에 덧붙여질 수 있는 방식으로 포함되어야 한다. 이것은 관계형 데이터베이스에서 추가적인 지역 및 언어 필드 등을 링킹함으로써 아주 쉽게 할 수 있다. 일부 데이터베이스에서, 단지 언어 차이만 있을 경우, 해당 언어는 모난 괄호로 이름에 종종 추가되는데, 그러나 이것은 처음에 간단한 해결책으로 보일 수 있지만, 시간이 경과함에 따라 더욱 복잡해지고, 종종 작동하지 않게 된다. 나중에 유동성을 추가하는 것보다 처음부터 이러한 문제를 해결할 수 있도록 데이터베이스를 설계하는 것이 가장 좋다.

비-라틴 알파벳으로 된 이름이 데이터베이스에 추가되는 경우, 해당 데이터베이스는 유니코드 문자 집합을 포함할 수 있도록 설계되어야 한다.

ii. 오류 검사

일반 이름은 일반적으로 학명과 연관되어 있기 때문에, 데이터베이스 내의 일관성 검사를 위해 때때로 여러 검사를 수행할 수 있다. 이것은 단조로운 절차일 수 있지만, 단지 필요할 때 불규칙적인 간격으로 수행하면 될 것이다. 중복되지 않는 모든 발생 레코드를 추출한 후 일관성이 없는 것들(예를 들어 하이픈이 빠진 것)에 대해 검사할 수 있을 것이다.

Soundex (Roughton and Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), 또는 Skeleton-Key (Pollock and Zamora 1984)와 같은 프로그램이 위의 *학명*에서 언급된 문자의 뒤바뀜과 같은 타자 오류를 검색하기 위해 다시 사용될 수 있을 것이다 (Dalcin 2004 참고).

c. 종이하 순위

종이하 순위 필드의 사용은 동물 데이터베이스보다 식물 데이터베이스에서 더욱 중요하다. 동물 분류학자들은 일반적으로 종 밑에 하나의 순위 – 아종 (subspecies)만을 사용하고 있으며 (이것의 사용은 점점 그 빈도가 줄어들고 있다), 이름은 3 항으로 표현된다:

Stipiturus malachurus parimeda.

하지만 역사적으로 일부 동물 분류학자들이 아종 외의 다른 순위를 실제로 사용했기 때문에 데이터베이스는 이러한 것을 제공할 필요가 있을 수도 있다. 이러한 경우 아래 식물 데이터베이스에 대한 설명이 동물 이름의 데이터베이스에도 또한 적용될 수 있을 것이다.

i. 데이터베이스 설계

다른 곳에서 언급되었듯이, 종이하 이름과 종이하 순위를 별도로 유지할 때 데이터 품질에 큰 장점 이익이 된다. 이것은 종이하 순위 필드를 간단하게 검사할 수 있도록 하고, 또한 종이하 이름의 검사를 더욱 쉽게 할 수 있도록 한다. 순위와 이름은 하나의 필드 내에 "콘텐트"로서 결코 포함되지 않아야 한다.

세분화된 데이터베이스에서 고려되어야 할 한 가지 문제는 일부 경우 웹 등에 표시하기 위해 이러한 필드를 연결할 필요가 있다는 것이다. 이것은 일반적으로 자동화될 수 있지만, 데이터베이스와 이것의 출력 형식을 설계할 때 이것을 어떠한 방식으로 할 것인가에 (데이터베이스에서 추가적으로 생성된 필드 또는 즉석에서 생성) 대한 고려가 있어야 한다.

ii. 데이터 입력

식물 (그리고 역사적으로 동물)의 경우, 종이하에서 사용될 수 있는 몇 가지 수준이 있다. 이러한 종이하 순위는 가장 빈번하게 아종(***subspecies***), 변종(***variety***), 아변종(***subvariety***), 품종(***forma***), 아품종(***subforma***) (식물규약(Botanical Code)은 추가적인 순위를 삽입하는 것을 배제하지 않으며, 따라서 다른 순위가 데이터집합에 존재할 가능성도 있다) 등이 있다. 아변종과 아품종은 거의 사용되지 않지만 식물 데이터베이스에서 꼭 제공할 필요가 있다. 여기에서도, 제한된 개수의 선택을 할 수 있는 선택-리스트가 설정되어야 한다. 이것이 되어 있지 않으면 오류가 조금씩 조금씩 발생하기 시작하고, 여러분은 분명히 아종이 다음과 같이 표현된 것을 보게 것이다: subspecies, subsp., ssp., subspp., 등. 이런 상황은 데이터를 찾거나 오류 검사를 수행하려고 하는 모든 사람에게 악몽이 될 수 있다. 입력을 할 때 이 선택사항을 제한하는 것이 데이터를 추출할 때 전체적인 범위를 제공하고 또는 나중에 일관성 유지를 위해 데이터 정제를 하는 것보다 더 나은 방법이다. 아래와 같이 사용되는 것이 권고된다:

subsp.	subspecies (아종)
var.	variety (변종)
subvar.	subvariety (아변종)
f.	form/forma (품종)
subf.	subform (아품종)
cv.	cultivar (재배변종, 종종 데이터베이스에서 또 다른 순위로 취급되지만 아래 설명을 참고할 것)

수집물 데이터베이스에서, 분류 체계를 포함하는 것은 하나 이상의 수준이 존재할 경우 필요하지 않은데, 왜냐하면 이것은 부가적인 혼란을 가중시키고 국제식물명명규약에 의하면 이 분류체계는 명확하게 해당 분류군을 정의할 때 필요하지 않다. 분류 체계가 포함된다면, 명확하게 분류군을 정의하는데 필요한 것만을 추출할 수 있어야 한다.

> *Leucochrysum albicans* subsp. *albicans* var. *tricolor* (= *Leucochrysum albicans* var. *tricolor*).

iii. 오류 검사

데이터베이스가 잘 설계되고 점검표의 값이 사용된다면, 심화된 오류 검사를 할 필요는 상대적으로 감소된다. 그러나 상황이 이렇지 않다면, 승인된 값들 가운데 한정된 부분집합만이 발생할 수 있도록 확실히 검사를 수행하여야 한다. 또한 수행해야 할 하나의 검사는 누락된 값에 대한 것이다.

d. 재배변종과 잡종

재배변종과 잡종은 많은 식물 데이터베이스에서 발생하지만 종종 잘 다루어지지 않는다. 재배변종은 이것들의 자체 명명규칙(Code of Nomenclature)을 따른다 (Brickell *et al.* 2004). 많은 식물 종 데이터베이스에서 이것들은 단지 또 다른 종이하 순위("cv.")로서 취급되고 일부 데이터베이스의 경우 이것은 매우 적합할 수 있다. 잡종은 대부분의 다른 그룹보다 다루기가 더욱 어렵다. 이것들은 두개의 항으로 표시할 수 있고 다음으로 동일한 순위의 (잡종 표시로 앞에 "**X**" (곱셈기호)가 나타난다) 임의의 어떤 다른 분류군으로 다루어질 수 있으며, 또는 이것들은 공식(formula) (심지어 서로 다른 순위에 있을 수 있는 두개 또는 두개 이상간의 잡종)으로 다루어질 수 있으며 여기에서 분류군 이름은 곱셈 기호로 분리되어 표시된다.

i. 데이터베이스 설계

잡종 또는 재배변종을 포함할 수도 있는 식물의 데이터베이스를 설계하고자 하는 사람에게 필자는 잡종이 레코드동정그룹의 (Record Identification Group) 부분으로 취급되는 HISPID 표준(Conn 1996, 2000)을 참고할 것을 권고한다. 이것들이 어떻게 다루어지든, 해당 이름이 재배변종 또는 잡종 등에 속하는 것을 서술하는 필드를 포함하는 것은 좋은 실행사례이다. 이런 방식으로 이것들은 분리해서 추출될 수 있고 그리고 오류 검사에서 서로 다르게 (형식화, 연결 등) 다루어질 수 있다.

ii. 오류 검사

잡종과 재배변종에 대한 오류 검사는 데이터베이스가 이러한 일을 할 수 있도록 설계되어 있지 않다면 어려운 작업이다. 이러한 검사에 대한 한가지 제안은 이것들을 그룹으로 다루는 것으로 즉, 모든 잡종을 추출하여 데이터베이스에 이것들이 저장된 방식을 고려하면서 종을 기준으로 알파벳순으로 정렬하는 것이다. 잡종 레코드를 동정할 수 있는 분리된 필드가 포함되어 있다면 이렇게 하기 더욱 쉽다. 발생할 가능성이 있는 하나의 중요 오류는 일관되지 않게 '**X**' 기호를 사용하는 것이다. 일부 데이터베이스는 곱셈 기호를 수용하지 않을 수도 있고, 때로는 이름 앞에 공백과 함께 그리고 때로는 공백 없이 일반적으로 'x' 또는 'X'로 대체된다. 이러한 종류의 일관성 사례는 쉽게 검사될 수 있다. 필자는 잡종 이름의 오류 검사를 실제적으로 잘하는 시스템을 알지 못한다.

e. 출판되지 않은 이름

i. 데이터 입력

1 차 종 데이터베이스에 저장된 모든 레코드들이 유효한 출판 분류군에 속하게 되지는 않는다. 데이터베이스에서 이러한 레코드를 검색할 수 있기 위해서는 해당 수집물에 '임시' 이름을 부여하는 것이 필요하다. 출판되지 않은 이름이 표준 형식으로 데이터베이스에 추가될 수 있다면, 이것들을 추적하고 나중에 이것들을 검색하는 것이 훨씬 쉬워질 것이다. 이것은 "nomen novum (새로운 이름 또는 대체 이름)", "nom. nov.", 그리고 "ms"와 같은 태그가 있거나 없는 출판된 이름처럼 보이는 두개 항을 가진 출판되지 않은 이름을 추가하는 것보다 더 낫고 덜 혼란이 된다. 아주 종종 "ms" 또는 "nom. nov."는 생략되고 사용자들은 출판물과 출판되지 않은 이름에 대한 참조 정보를 조사하는데 많은 시간을 소비할 수 있다. 공식을 사용함으로써 모두에게 이것은 출판되지 않은 이름이라는 것이 명백할 것이다.

1980 년대 호주에서, 식물학자들은 출판되지 않은 이름의 사용에 대한 공식에 합의하였다 (Croft 1989, Conn 1996, 2000). 이것은 "*Verticordia* sp.1", "*Verticordia* sp.2" 등과 같은 것들의 사용을 통해 발생하는 혼란을 피하기 위한 것이었다. 예를 들어 호주가상식물표본관 (Australian Virtual Herbarium) (CHAH 2002), *species*Link (CRIA 2002), 유럽생물학수집물접근서비스 (Biological Collection Access Service for Europe) (BioCASE 2003), 포유류네트워크정보시스템 (Mammal Networked Information System) (MaNIS 2001), GBIF 포탈 (GBIF 2004) 그리고 다른 많은 것들을 통하여 여러 데이터베이스가 통합되기 시작하면, 한 기관에서 "sp. 1"라고 불리는 것이 다른 기관의 "sp. 1"과 동일하다는 보장이 없기 때문에 이러한 것들과 같은 이름은 더 많은 혼란을 야기할 것이다. 이러한 데이터베이스를 정결하고 일관성 있게 유지하면서, 하나에서 다른 것으로 데이터의 매끈한 이전을 가능하게 하는 한가지 방법은 호주 식물학 커뮤니티에서 채택한 것과 유사한 공식을 사용하는 것이다.

합의된 공식은 아래 형식과 같다:

> "<속> sp. <구어체 이름 또는 서술> (<확증인>)":
>
> *Prostanthera* sp. Somersbey (B.J.Conn 4024)

나중에, 해당 분류군이 공식적으로 서술되고 명명이 되면, 이 공식-이름은 다른 임의의 동의어와 같이, 동의어로 간주될 수 있을 것이다.

이런 공식을 사용하는 것은 사용하지 않을 때 보다 데이터베이스를 더 복잡하게 만드는데, 왜냐하면 단지 한 개의 단어만을 가지는 종 필드 대신에 공식에 맞추기 위해 이제 문장을 포함해야 하기 때문이다. 종종 사용되고 있는 "sp. 1" "nom. nov." 등도 동일한 문제점이 있지만, 이 방법이 모호함의 정도가 덜하다. 이러한 것과 같은 공식을 사용하는 것은 (웹상에서 보여줄 때 등) 연결에 어려움을 발생시킬 수 있지만, 호주에서 이 방법론을 (예를 들어, 호주 환경부(DEH 2005b)의 SPRAT 데이터베이스에서 사용한 것을 참고) 사용한 경험에 의하면 이것은 잘 작동했다. 이 공식은 공백과 괄호 등이 있긴 하지만 그 외 다른 모든 면에서'종' 소명으로 취급된다.

예를 들어, 멸종위기 종에 대한 법적인 목록(예를 들어 DEH 2005a 참고)에서 출판되지 않은 이름을 사용해야 할 필요성 때문에, 비분류학적 출판물, 예를 들어 법률 문서에서 사용될 수 있도록 이러한 분류군을 명명 또는 태깅하는 일관성 있는 시스템의 존재가 필수적이다. 여기에서 제시된 것과 같은 공식을 사용함으로써, 예비학명(nomen nudum)을 실수로 우연히 출판하는 위험성은 거의 없을 것이다.

박물관과 식물표본관이 자신들의 데이터베이스에 비슷한 체계를 도입하는 것이 권고된다.

ii. 오류 검사

여기에서 제시되는 것과 같은 공식 이름에서 발생하는 가장 흔한 오류는 틀린 철자이다. 보통 이 공식은 몇 개의 단어를 포함하기 때문에, 종종 확증자의 인용 등에서 실수를 범하기가 쉽다. 이러한 이름을 검사하는 가장 쉬운 방법은 속 내에서 각각을 정렬하고 (이것들은 단지 하나 이상의 단어를 가진 종 또는 종이하 필드의 이름이어야 한다) 이것들을 유사도에 대해 검사하는 것이다. 이것은 임의의 하나의 속 내에 많은 수가 있지 않을 것이므로 그렇게 번거롭지는 않을 것이다. 위에서 언급한 Soundex 와 같은 유사 기법들이 또한 사용될 수 있을 것이다.

f. 저자 이름

종 이름의 저자가 일부 표본 데이터베이스에 포함될 수도 있지만, 대부분 이러한 것의 포함은, 포함하기 전에 철저하게 검사되지 않기 때문에, 오류를 야기하기 쉽다. 이것들은 동일한 이름이 무심코 같은 속 내의 서로 다른 분류군(homonyms)에 명명되었을 경우 또는 데이터베이스가 분류학적 개념을 포함하려고 할 경우에 실제적으로 필요하다 (Berendsohn 1997). 종 (또는 종이하) 이름 이후에 저자의 이름을 포함하면 해당된 두개의 이름 또는 개념간을 구분할 수 있다. 데이터베이스가 종 이름의 저자를 반드시 포함하는 경우, 이것들은 종의 이름 자체와 분리된 필드에 분명히 포함되어야 한다.

저자 이름과 날짜가 분리되어 저장되는 경우 데이터의 연결은 식물과 자동으로 생성된 이름(autonym)을 (아래 참고) 제외하고는 보통 주요 문제가 되지 않는다. 그러나, 식물과 동물의 혼합된 데이터베이스는 전거 기관이 약간 다르게 다루어질 경우 몇 가지 문제를 야기할 수도 있다. 이러한 것을 제공하기 위해 저자 필드가 데이터베이스에 설정되는 경우 너무 많은 문제점이 발생하지 않아야 하며 추출에 대한 규칙을 다른 계(Kingdom)에 대해 서로 다르게 할 필요가 있을 것이다.

Dalcin (2004)은 그의 명명 데이터 도메인(domain)하에서 전거 기관을 라벨 항목으로 다루고 있다.

i. 데이터 입력

동물 이름의 경우 저자 이름은 (통상적으로 모든 이름) 뒤에 항상 연도가 따라온다; 식물의 경우, 저자 이름 또는 약자가 홀로 사용된다.

동물:
Emydura signata Ahl, 1932
Macrotis lagotis (Reid, 1937)
(괄호는 Reid 가 이 종을 다른 속에 포함시켰다는 것을 나타낸다)

식물:
Melaleuca nervosa (Lindley) Cheel
synonym: *Callistemon nervosus* Lindley
(Lindley 가 처음으로 이것을 *Callistemon* 으로 서술하였다; Cheel 이 나중에 이것을 *Melaleuca* 속(genus)으로 바꾸었다).

식물의 경우, 가끔 저자 이름에서 “ex” 또는 “in” 용어가 있는 것을 볼 수 있다. “ex” 앞에 있는 저자는 이름을 지었지만 유효한 출판을 위한 여러 요구사항을 만족시키지 못했거나 또는 관계된 그룹에 대한 명명 시작일 이전에 이름을 출판한 사람이다. “in”뒤에 붙여진 저자는 이 사람의 연구, 서술 또는 분석을 다른 저자가 검토하여 출판했을 때 사용된다. ‘ex’ 앞의 저자 그리고 ‘in’ 뒤의 저자에 대한 자세한 설명과 이것의 이용에 관해서는 국제식물명명규약 (International Code of Botanical Nomenclature) (2000)의 46.2 절과 46.3 절을 참고하기 바란다. 데이터베이스 내에 저자 이름이 사용된다면 이것들은 학명과 다른 필드에 존재해야 하고 (위의 세분화 부분 설명 참고) ‘ex’ 이전 또는 ‘in’ 이후의 저자들을 인용하지 않는 것이 권고된다.

Green (1985)은 새로운 조합 *Tersonia cyathiflora* 을 "(Fenzl) A.S. George" 에서 기인한다고 하였다. Green 이 George 가 어떤 방식으로도 기여했다는 것을 어디에서도 언급을 하지 않았으므로, 이 결합된 저자는 다음과 같이 인용되어야 한다: “A.S.George ex J.W.Green” 또는 가능하다면 단지 “J.W.Green”으로 서술되어야 한다.

Tersonia cyathiflora (Fenzl) J.W.Green

W.T.Aiton 의 *Hortus Kewensis* (1813) 2 판에서, 많은 서술문에 Robert Brown 의 서명이 있으므로, Brown 이 관련 종을 서술한 것으로 가정할 수 있다. 이름의 저자는 종종 "R.Br. in Ait."으로 인용되어 있다. 하지만 저자는 단순히 "R.Br"으로 인용되는 것이 권고된다.

Acacia acuicularis R.Br.

식물 – 기준(모식) 아종 또는 변종 등 종이하 이름이 종 이름(본명)과 같은 경우, 종 이름의 저자가 사용되고 이것은 특정 종소명 뒤에 오게 된다. 이 형식은 저자 이름을 포함하는 표본 데이터베이스에서 이름을 재구성할 때 혼란을 일으키는데, 그 이유는 이것이 다른 규칙들에 대해 예외이기 때문이다.

Leucochrysum albicans (A.Cunn.) Paul G.Wilson subsp. *albicans*

식물의 경우, 저자 이름의 약자는 국제적으로 합의된 표준을 따르고 (Brummitt and Powell 1992), 이 출판물은 점검표를 설정할 때 또는 데이터 입력 또는 검증 검사에 이용될 수도 있을 것이다.

A.Cunn. = Allan Cunningham
L. = Linnaeus
L.f. = Linnaeus filius (son of-)

때때로, 머리 글자와 이름의 성(Surname) 사이에 공백을 사용하지만, 다른 것들은 그렇지 않다. 이것은 선호도의 문제이다.

ii. 오류 검사

브러밋과 파웰 (Brummit and Powell (1992))에서 사용된 저자 이름들은 식물학 데이터베이스에서 저자들을 검사할 때 사용될 수 있을 것이다. 하버드 대학교는 또한 식물학 저자들에 대한 다운로드할 수 있는 파일을 준비하였고 온라인[2]으로 이용 가능하게 하였다. 이것은 저자의 이름과 날짜를 검사하는데 매우 귀중한 파일이 될 것이다. 일부 이름 데이터베이스 또한 저자 이름을 포함한다 (예, IPNI 1999, Froese and Bisby 2002). 또한 위에서 언급한 것과 같은 Soundex 와 비슷한 기법을 이용하여 두 이름 사이의 유사점을 검사할 수 있을 것이다. 하지만 결정적인 요소는 종 이름과 저자의 조합이고 이러한 것을 검사하는 것이 항상 쉬운 것은 아니다.

저자가 사용된다면, 데이터베이스에 있는 모든 출판된 이름은 저자를 가져야 한다. 이러한 경우, 누락데이터값(Missing Data Values)에 대한 검사가 수행되어야 한다.

g. 수집자의 이름

수집자의 이름은 수집물 데이터베이스에서 일반적으로 표준화되어 있지 않지만, 식물 수집자의 이름에 대한 표준화가 브라질의 *species*Link 프로젝트(Koch 2003)의 식물 이름에 대해, 그리고 큐 식물원의 피터 서톤(Peter Sutton)에 의해 시도되고 있다.

광범위한 수집자의 이름 목록이 일부 분야에서 출판되었지만 주로 식물 수집자에 대한 것이었다 (Steenis -Kruseman 1950, Hepper and Neate 1971, Dorr 1997, Index Herbariorum 1954-1988 참고). 또한 이용 가능한 많은 온라인 자원들이 있다:

- 하버드 대학교는 최근 다운로드 할 수 있는 식물학 수집자와 수집자 팀의 파일을 준비하였고 이것을 온라인으로 이용 가능하게 하였다.

[2] http://www.huh.harvard.edu/databases/cms/download.html

http://www.huh.harvard.edu/databases/cms/download.html

- Index Collectorum – 괴팅겐 대학교
  http://www.sysbot.uni-goettingen.de/index_coll/default.htm
- 남부 아프리카의 곤충 수집자 등록부 (남부 아프리카 곤충학회)
  http://www.up.ac.za/academic/entomological-society/collectr/collectr.html
- Index bio-bibliographicus notorum hominum Nonveilleriana (크로아티아 곤충학회)
  http://www.agr.hr/hed/hrv/bibl/osobe/comentsEN.htm

또한 책자 형태의 많은 출판물이 있으며, 동물학의 여러 분야에서 이용할 수 있는 이러한 많은 색인들이 있다.

i. 데이터 입력

이름이 표준 형식으로 1 차 종 발생 데이터베이스에 포함되는 것이 권장된다. HISPID 표준 (Conn 2000)은 다음을 권장한다:

> *1 차 수집자의 가족 이름 (이름의 성) 다음에 콤마와 공백 (,) 다음에 머리글자 (모두 대문자이고 각각은 마침표로 분리).모든 머리글자와 수집자의 가족 이름의 첫 글자는 대문자로 한다. 예를 들면, Chambers, P.F.*

2 차 수집자들은 두 번째 필드에 위치시키는 것이 권장된다. 이런 경우가 아니라면, 이 수집자들은 다수의 수집자를 분리하기 위해 사용되는 콤마와 공백을 가지고 인용되는 것이 권고된다. 예를 들면:

Tan, F., Jeffreys, R.S.

혼란의 여지가 있을 때에는, 성 이외의 다른 이름이 사용되어야 한다. 다음은 Wilson, Paul G. and Wilson, Peter G.을 구분하는 예이다 (성 이외의 이름에 공백이 사용되었다; 위에서 서술한 것처럼 두 이름간에 분리자를 제외하고 구두점이 없다).

직함은 생략되어야 한다.

가족 이름(이름의 성)이 전치사와 실질명사(substantive)로 구성되면, 많은 유럽인의 이름처럼 (예, C.G.G.J. van Steenis) 전치사는 소문자로 표시하고 실질명사는 첫 글자를 대문자로 표시한다. 예를 들면:

Steenis, C.G.G.J. van

유사한 형태의 다른 이름들은 de la Salle, d'Entrecasteaux, van Royen 등이 있다. 하지만 이러한 이름들 중의 많은 것들이 영어화 되었고, 특히 미국에서, 이러한 경우 가족 이름의 두 부분이 실질명사로 취급된다. 이러한 경우, 이 이름들은 다음과 같이 변형된다:

De Nardi, J.C.

접두사가 된 O', Mac', Mc' 그리고 M' (e.g. MacDougal, McKenzie, O'Donnell)는 모두 실질명사의 일부로 취급되어야 하며 따라서 가족 이름의 일부로 변형되어야 한다. 예를 들면:

McKenzie, V.

하이픈으로 연결된 성 이외의 이름은 모두 대문자로 변형되어야 하며 처음과 마지막의 머리글자는 하이픈(공백 없이)으로 분리되고, 마지막 글자에 마침표로 종료된다. 예를 들면:

Quirico, A-L.

Peng, C-I.

레코드의 수집자가 알려져 있지 않으면, “Anonymous(익명)” 용어가 사용되어야 한다.

해석된 정보는 꺾쇠 괄호 안에 포함되어야 한다. 예,

Anonymous [? Mueller, F.]

ii. 오류 검사

위에서 언급되었듯이, 수집자의 표준 목록 없이, 수집자의 이름에 관한 오류 검사를 수행하는 것은 쉬운 일이 아니다. 이것은 표준 실행사례(위에서 언급된 것처럼 이름의 성을 첫 번째로 두는 것 등)를 따르지 않는 데이터베이스에서 특히 그러하다. 데이터베이스가 표준화되어 있으면, 데이터베이스에 있는 모든 수집자의 이름을 정렬해서 미세한 차이를 (예를 들어, 때때로 하나의 머리글자를 사용하고 다른 경우는 두개의 머리글자를 사용하는 수집자) 살피는 것은 아주 쉬운 일이다. 해당 변경이 올바른 것이라는 분명한 확신 없이 수집자의 이름을 변경함으로써 데이터베이스에 새로운 오류가 생기지 않도록 극도로 주의하여야 한다. 위의 머리글자 사례는 하나의 변경이 새로운 오류를 간단히 초래할 수 있는 경우이다. 예를 들면, 수정이 가능한 오류는 이름의 성이 잘못 표기된 철자들이다.

수집자의 목록을 만드는 한가지 방법은 분류군 이름에 대해 전거 테이블이 개발되는 것과 동일한 방식으로 데이터베이스에서 유일한 값의 목록을 만드는 것이다.

*수집일자 (Date-of-Collection)*와 같이 *수집자-이름(Collector-Name)*과 연관된 필드 또한 오류 검사에 이용될 수 있다. 역사학자들은 최근 탐험가와 수집자의 여행기, 역사적인 과학 탐험, 항해기 등을 발굴하는 것에 관한 상당한 양의 연구를 진척시켜 왔다. 종종 이러한 것은 과학자들이 아닌 역사학자들에 의해 수행되며, 우리의 과학은 이 연구에서 많은 혜택을 얻을 수 있다 (세계의 오래된 많은 박물관 및 식물표본관의 도서관에 있는 수집물 (출판물, 저널 등), SIO 디지털도서관 프로젝트의 일환으로 선박여행의 데이터를 문서화하고 갈무리(capture)한 캘리포니아 대학교, 샌디에고, 스크립스해양연구소(Scripps Institution of Oceanography), 그리고 국가과학디지털도서관의 최근 연구와 함께 위에 나열된 자원들 참고). 이러한 데이터베이스와 1 차 종 데이터베이스 간의 연결은 불일치와 오류가 발견됨에 따라 두 가지 모두의 개선을 이끌 수 있다.

공간 데이터

공간적인 위치는 많은 종-발생 레코드의 이용에 대한 적합성을 결정할 수 있다는 점에서 아주 결정적인 측면 중의 하나이다. 공간과 관련된 생물지리 연구들은 이러한 데이터를 가장 많이 이용하는 것 가운데 하나이다 – 이것들은 종 분포 모델링, 생물지리 연구, 환경 계획 및 관리, 생물-구역화 연구, 보호지 선택과 보전 계획, 그리고 환경 정책결정 지원 등과 같은 연구가 있다. 세부적인 연구에 대해서는 관련 논문인 *1 차 종-발생 데이터의 이용* (*Uses of Primary Species-Occurrence Data*) (Chapman 2005b)을 참고하기 바란다.

우리는 1 차 종 데이터를 식물 또는 동물의 발생에 대한 ***위치 (point)*** 레코드로 종종 여기지만 이것은 단지 전체중에 부분적인 것에 불과하다. 수집 장소들은 실제 위치로 간주될 만큼 정확하게 또는 정밀하게 거의 기록되지 않는다. 수집물과 연관된 정확성은 해당 지점이 실제로는 어느 구역 또는 궤적(footprint)을 나타낸다고 말할 수 있다. 예를 들어, "Campinas 에서 북쪽으로 10km (10 km north of Campinas)"로 되어 있는 텍스트 정보는, "10km" (아마도 ± 500m) 거리와 연관된 정확성이 있고, "북쪽" 방향과 연관된 정확성이 있고 (즉, 북쪽은 NW(북서)와 NE(북동) 사이의 어떤 곳이다), "Campinas"와 연관된 정확성이 있다 (이것은 시의 경계인가 (다각형(polygon)) 또는 시의 중심인가 등). 이에 대한 좀 더 자세한 논의는 Wieczorek 2001a, Wieczorek *et al.,* 2004 을 참고하기 바란다. 또한 많은 관찰 및 조사 레코드들은 한 지역에 걸쳐 (***다각형***) 기록되며, 이러한 것으로는 2 ha 지역에 걸쳐서 또는 국립공원 내에서 또는 호주 전역의 10-도 그리드 사각형에서의 관찰과 같은 정규 격자(***그리드***)에서, 또는 10m X 10m 조사 그리드에서 또는 횡단 조사 또는 도로나 강을 따라 발생하는 레코드와 같은 횡단(***선***)을 따르는 곳(규모에 따라 다르겠지만,아마도 도로나 강의 버퍼링에서 도출한 다각형으로 더욱 잘 다루어질 수 있을 것이다)에서의 조류 관찰이 있다. 더욱 자세한 논의는 아래 *오류의 시각화* 부분을 참고하기 바란다.

이전에 언급된 것처럼, 1 차 종 레코드에 포함된 지리코드의 오류를 검사하고 테스트를 지원할 수 있는 많은 프로그램이 실제 존재한다. 그 외 이용할 수 있는 다른 도구들은 (지명으로부터 거리와 방향과 같은) 장소 정보에서 데이터로 최초의 좌표 값을 배정하는 것을 지원한다.

이미 배정된 지리참조연산 정보의 오류 테스트는 아래 사항들과 관련이 있다

- 레코드 내부의 다른 정보에 대한 검사, 예를 들어, 주 또는 지명.
- 데이터베이스를 이용하여 외부 참조정보에 대한 검사 – 예, 레코드가 수집자의 수집 지역과 일치하는가;
- GIS 를 이용한 외부 참조정보에 대한 검사, "다각형내의 점(point-in-polygon)" 테스트 포함 – 예를 들어, 해당 레코드들은 바다보다는 내륙에 속하는 것이다;
- 지리 공간에서 특이점에 대한 검사; 또는
- 환경 공간에서 특이점에 대한 검사.

데이터 입력과 지리참조연산

이 문서 전반에 걸쳐 강조된 것처럼, 오류 예방은 오류 탐지보다 더 선호되며, 지리참조연산 또는 레코드에 대한 지리코드 입력은 종-발생 데이터를 데이터베이스화할 때 가장 많은 오류 원천중의 하나이다. 1 차 종 데이터에 좌표(특히 위도와 경도)를 추가하는 과정을 지원하는 새로운 많은 도구들이 현재 개발되고 있다. 하지만 이것은 쉽지 않은 과정으로 특히, 과거 데이터(과거 300 년 또는 400 년에 걸쳐 수집된 박물관과 표본식물관의 초기 수집물들)의 많은 것들이 어느 장소에서 수집되었는가에 대한 일반적인 서술 이외에 지리 정보를 거의

가지고 있지 않기 때문이다 (Chapman and Milne 1998). 이러한 수집물은 때때로 근대적인 정착 도구와 방법들이 확립되고 세워지기 전에 그리고 도로가 건설되기 전에 만들어졌다. 많은 것들이 마지막 정착 시대에 말 위에서 또는 배에서 수집되었고 참조 지점들을 종종 결정하기 어려웠다. 참조 지점들 가운데 많은 장소들은 더 이상 현대 지도에 나타나지 않고, 많은 경우 이것들이 실제 있더라도 모호한 것이 되어버렸다. 지리코드가 있는 경우, 이것들은 종종 아주 정확하지 않고 (Chapman 1999) 일반적으로 나중에 (*retrospective georeferencing* – Blum 2001) 수집자가 아닌 다른 사람들에 의해 추가되었다 (Chapman 1992).

i. 정의:

계속하기 전에, 이 문서에 사용되는 몇몇 용어에 대한 정의가 필요하다. 이러한 용어들 중의 일부는 다른 곳에서는 다르게 쓰이고, 다른 분야에서는 유사한 과정을 정의할 때 다른 용어를 사용한다.

지리코드 (Geocode): 이 문서에서 사용될 때, 지리코드(geocode)는 표준 지리 정보 시스템에 맞게 레코드의 공간 위치를 기록하는 코드이다 (보통 x, y 좌표). 여기에서 이것은 지구 표면상의 위치를 기록할 때 사용된다. 종-발생 데이터의 경우, 지리코드는 전지구횡단메르카토르(Universal Transverse Mercator)와 같은 몇몇 표준 지리 참조 시스템 중의 하나에 따라 주어지고, 위도와 경도는 가장 흔하게 사용되고, 여러 방법 (미터; 십도; 도, 분, 초; 도, 십진분, 등) 중의 하나로 기록될 수 있다. 지리코드 용어에 대한 정의는 다양하고 폭이 넓다. 많은 GIS 애플리케이션에서 이것은 주소와 우편 번호를 가리키고, 마케팅 용어로 이것은 지역의 인구 특징을 가리키며, 그리고 일부 경우에 (Clarke 2002) 이것은 "*컴퓨터가 읽을 수 있는 형식*"의 위치만을 가리킨다. 때때로 이것은 또한 ***지리참조정보(georeference)*** 또는 ***좌표(coordinate)***라고 불리기도 한다.

지리참조연산 (Georeferencing): 이 문서에서 지리참조연산은 레코드에 지리좌표를 할당하는 과정을 서술하는 것에 사용되며 이것은 지구상의 지리적인 위치에 이 레코드를 연결하는 것이다. 이것은 또한 때때로 ***지리코딩(geocoding)***으로 불리기도 한다.

ii. 데이터베이스 설계

1 차 종-발생 데이터에 대한 데이터베이스를 설계할 때 장소 필드(서식지와 서식 정보 그리고 지리적인 사항에 대한 요약과 같은 데이터)에 가끔 잘못되어 저장되는 정보가 올바르게 처리될 수 있는 필드를 두어야 한다. 분포 필드에 여러 가지 정보(Fishbase[3]에 있는 *Perca fluviatilis* 에 대해)가 섞여 있는 예는 다음과 같다:

> "*Throughout Europe and Siberia to Kolyma River, but not in Spain, Italy or Greece; widely introduced. Several countries report adverse ecological impact after introduction (스페인, 이탈리아 또는 그리스는 제외하고 유럽과 시베리아에서 콜리마 강까지 광범위하게 도입되었다; 몇몇 국가들은 도입 후 심각한 생태계 영향을 보고하였다)*".

이렇게 혼합된 필드들은 파싱 알고리듬을 이용하여 자동화된 방식으로 다루는 것이 매우 어려우며 이러한 정보를 메모(Memo) 필드들에 저장할 수 있는 관계형 데이터베이스의 철학과 설계에 맞지 않는다.

데이터 정제 지원 목적으로 종-발생 데이터베이스에 추가될 수 있는 몇몇 추가적인 필드들이 있고 이것은 데이터 품질 문서화를 향상시킬 수 있다. 이러한 것들은 다음과 같다:

[3] http://www.fishbase.org/

- *공간적인 정확성* – 레코드의 위치가 판단될 때의 정확성을 기록하는 필드 (미터가 선호되지만, 때로는 미리 정해진 코드 형식).
- *지명, 거리 그리고 방향* – 일부 데이터베이스는 일반 텍스트 형식의 장소 필드 이외에 "가장 가까운 지명", "거리", 그리고 "방향" 필드를 포함한다. 이러한 필드를 포함하는 것은 오류 검사뿐만 아니라 지리코드 결정에 도움을 줄 수 있다.
- *지리코드 방법* – 지리코드가 어떻게 결정되었는지를 기록하는 필드 – 다음을 포함할 수 있다 (Chapman 2005a)
  - 보정(Differential) GPS 사용;
  - 선택적 이용성(Selective Availability)에 의해 훼손된 휴대용 GPS (즉, 2000 년 5 월 1 일 전에 사용된 것);
  - 1:100,000 의 축척에서 손쉽게 식별할 수 있는 특징을 이용하여 삼각 측정법으로 구한 지도 참조;
  - 추측 항법(dead reckoning)을 사용하는 지도 참조;
  - 원거리에서 획득한 지도 참조 (예, 헬리콥터);
  - 점-반지름 방법을 사용하는 지리-참조 소프트웨어를 통해 자동으로 획득;
  - 과거 지리참조연산된 위치를 이용하여 데이터베이스에서 획득.
- *지리코드 유형* – 지리코드를 결정하는데 사용되었던 장소 서술 정보의 유형을 기록한다.

  장소 서술정보를 지리참조연산하는 점-반지름 방식에 관한 논문에서, 위에크조렉(Wieczorek 2004)과 그의 동료들은 자연사 수집물에서 발견되는 9 가지 유형의 장소 서술문에 대한 테이블을 제공하고 있다. 이러한 것들 중에 앞부분의 세 가지는 지리참조연산이 수행되지 않아야 된다고 이들은 권고하고 있으며, 그렇지만 왜 이것들이 지리참조연산 되지 않았는지에 대한 설명은 제공되어야 한다. 일부 데이터베이스는 매우 큰 정확성 수(예, 100,000,000 미터)를 가진 중심점을 사용한다. 이것은 연관된 정확성 필드를 사용하지 않고 단지 지리코드만을 사용하여 데이터를 추출하는 사용자에게는 단점이 되며, 결국 이것과 연관된 거대한 반지름을 사용하지 않고 하나의 점으로만 인식하게 된다. 위에크조렉(Wieczorek) 방법은 이러한 방식을 야기하는 지리코드를 제공하지 않음으로써 이 단점을 극복한다. Wieczorek *et al.* (2004) 에서 나열된 9 가지 범주는 다음과 같다 (예는 수정되었음):

  1. *모호함* (예, 'Isla Boca Brava?')
  2. *위치를 파악할 수 없음* (예, 'Mexico', 'locality not recorded')
  3. *명백하게 부정확함* (예, 상반되는 문장들을 포함)
  4. *좌표* (예, 위도 또는 경도, UTM 좌표가 있음)
  5. *지명* (예, 'Alice Springs')
  6. *상대적 거리* (예, '5 km outside Calgary')
  7. *경로에서의 상대적 거리* (예, '24 km N of Toowoomba along Darling Downs Hwy')
  8. *교차 방향에서의 상대적 거리* (예, '6 km N and 4 km W of Welna')
  9. *진행방향에 대한 상대적 거리* (예, 50 km NE of 'Mombasa')

해당 논문에서 논의되는 것처럼, 이러한 것들의 각각은 서로 다른 정확성 계산 방법을 필요로 할 것이다 (Wieczorek *et al.* 2004).

iii. 지리참조연산 지침서

데이터 관리자들이 지리참조연산을 할 수 있도록 지원하는 두개의 훌륭한 지침서가 개발되었다. 버클리(Berkeley)에 있는 척추동물박물관(Museum of Vertebrate Zoology)의 존 위에크조렉(John Wieczorek)이 개발한 지리참조연산 지침서(Georeferencing Guidelines)와 콜로라도 대학교(University of Colorado)에서 개발한 MaPSTeDI (Mountains and Plains Spatio-Temporal Database Informatics Initiative) 지침서(University of Colorado 2003)는 가장 포괄적인 두가지이다. 멕시코의 코나바이오(Conabio)에서 개발된 지침서(CONABIO 2005)들이 또한 있으며, 이것들은 현재 영어로 번역되고 있어서 조만간 스페인어와 영어로 이용 가능할 것이다.

iv. 편집 컨트롤

편집 컨트롤(edit controls)은 특정 필드에 허용되는 값을 결정하는 실무 규칙과 관련이 있다. 공간 데이터베이스에서 가장 빈번한 오류중의 하나는 남반구 또는 동반구의 레코드에서 '–'(마이너스)기호를 우연히 빠뜨리는 것이다. 만약 해당 데이터베이스가 전부 남반구 레코드(예를 들어 호주 레코드에 대한 데이터베이스)만을 저장하는 데이터베이스라면, 모든 레코드는 자동적으로 "마이너스(negative)"위도를 가져야 한다. 물론, 혼합된 레코드를 가진 데이터베이스는 처리하기 더욱 어렵지만, 위도와 경도에 대한 검사에 국가와 주 필드가 사용될 수 있을 것이다.

모든 데이터베이스가 초기에 제대로 설정되지는 않으며, 이것 때문에 결코 발생하지 않아야 할 오류가 발생하게 된다. 예를 들어, 90º 보다 크거나 –90º도 작은 위도 그리고 180º보다 크거나 –180º보다 작은 경도. 만약 이러한 것들이 데이터베이스에서 허용된다면, 해당 데이터베이스는 수정될 필요가 있고, 또는 이러한 오류가 확실히 발생되지 않고 고쳐질 수 있도록 정기적인 검사가 수행될 필요가 있다.

v. 지리코드 결정에 기존의 데이터베이스 레코드를 사용하기

데이터베이스에 이미 포함된 정보를 이용하여 추가되는 새로운 레코드에 지리참조연산 정보를 할당할 수 있다. 간단한 보고 절차를 추가하여, 같은 장소의 표본이 이미 데이터베이스되었고 지리코드가 할당되었는지를 확인할 수 있는 검색을 가능하게 할 수 있다.

예를 들어, 사용자가 지금부터 "Campinas 에서 북서쪽으로 10 km (10 km NW of Campinas)"의 위치 정보가 있지만 지리참조연산 정보가 없는 수집물을 데이터베이스화하려고 한다고 가정하자. 사용자는 "Campinas"에 대해 데이터베이스 검색을 수행할 수 있고 "Campinas 에서 북서쪽으로 10km"의 정보를 가진 또 다른 수집물에 지리코드가 이미 할당되어 있는지를 알아보기 위해 이미 데이터베이스화된 수집물을 조사할 수 있다. 이 과정은 데이터베이스 구조가 전통적으로 자유롭게 서술되는 텍스트 장소 필드 이외에 "가장 가까운 지명 (Nearest Named Place)", "거리 (Distance)", 그리고 "방향 (Direction)" 또는 유사한 것에 대한 필드를 포함하면 더욱 더 간단해질 수 있다.

이 방법론은 만약 첫 번째 지리코드가 오류를 가진 채 할당되면, 이 오류가 데이터베이스에서 계속해서 존재하게 되는 단점을 가지고 있다. 하지만 지금까지 데이터베이스화된 수집물 중에 이런 오류가 어느 하나에서 발견되면 전체적인 수정이 가능하다. 지리코드 결정에 이러한 방법이 사용되면 *지리코드 방법* 필드에 관련 방법이 문서화되어야 한다 (위 참고).

호주가상식물표본관(Australian Virtual Herbarium) (CHAH 2002), *species*Link (CRIA 2002), 또는 GBIF 포탈 (GBIF 2004)과 같은 링크를 가진 데이터베이스에서, 유사한 방식으로 협업적인 지리코딩 이력이 개발되고 사용될 수 있도록 하는 온라인 작업절차가 설정될 수 있을 것이다. 이와 같은 협업은 웹서비시스(Web Services) (Beaman *et al.* 2004, Hobern and Saarenmaa 2005)의 사용을 통해 수행될 수 있다. 물론 이것의 단점 하나는 여러분의 데이터베이스에 대한 어느 정도 관리권한 손실이 있으며, 다른 데이터베이스의 오류가 우연히 여러분의 데이터베이스에 복사될 수 있다는 것이다. 이렇게 될 경우, 추후 갱신과 수정이 반영될 수 있도록 소스 식별자(source-id)가 레코드에 추가되어야 한다. 첫째 오류가 무심결에 계속해서 남아있지 않도록 하기 위해서, 둘째 탐지된 오류 정보가 처음 유래한 데이터베이스뿐만 아니라 다른 의존적인 데이터베이스에서 피드백 될 수 있도록, 기관간에 올바른 피드백 메커니즘이 개발될 필요가 있을 것이다.

많은 식물 수집물은 다른 수집 기관들에 '복사물(duplicates)'로서 배포된다. 전통적으로 이것은 지리참조연산을 수행하기 전에 행해져 왔고, 서로 다른 많은 기관에서 정확하게 동일한 수집물이 각각 다른 지리참조연산 정보를 가지고 있는 것을 종종 발견할 수 있다. 이러한 불일치를 피하기 위해, 배포 전에 지리코드가 추가될 필요가 있고, 또는 기관간에 공동의 협력관계를 구축할 필요가 있다. 앞에 설명된 것처럼, 지리코드를 추가하는 것은 많은 시간과 경비가 소요되고, 여러 기관들이 독립적으로 동일한 수집물의 지리참조연산에 시간과 자원을 소비하는 것은 극도로 소모적인 활동이다. 이러한 소모적인 활동은 서로 다른 지리코드가 각기 다른 기관에 있는 동일한 수집물에 할당되면 더욱 더 복잡해진다.

vi. 자동화된 지리코드 할당

자동화된 지리참조연산 도구들은 알려진 위치에서의 거리와 방향을 이용하여 텍스트 장소 정보로부터 위도와 경도를 결정하는 것을 기반으로 한다. 이상적으로 데이터베이스는 최소한 "가장 가까운 지명", "거리", 그리고 "방향", 또는 더욱 낫다면 "지명 1", "거리 1", "방향 1", "지명 2", "거리 2", "방향 2"의 필드를 포함해야 한다. 따라서, "5 km E of Smithtown, 20 km NNW of Jonestown (Smithtown 에서 동쪽으로 5km, Jonestown 에서 북북서쪽으로 20km)"은 위에서 언급된 6 개의 필드에 적절하게 전달될 수 있을 것이다.

대부분의 데이터베이스가 이렇게 구성되어 있지 않으므로, 자유로운 텍스트 형식으로 된 장소 서술문을 기본적인 "가장 가까운 지명", "거리" 그리고 "방향" 필드에 맞게 파싱하는 자동화된 파싱 소프트웨어를 개발하려는 시도가 이루어지고 있으며, 적합한 지명사전(gazetteers)과 함께 이러한 필드를 이용하여 지리참조정보를 결정하고 있다 (아래 *BioGeomancer* 참고). 지리코드가 이러한 이런 방식으로 결정되는 것과 동시에, 지리코드 정확성(Geocode Accuracy)이 추가 필드에 기록되어야 하고, 기대치 않은 오류를 피할 수 있도록 가능하다면 원본에 대해 전문가가 검사한 결과가 기록되어야 한다. 어떠한 경우든, 이러한 파싱이 어떠한 방식으로도 원래의 "장소" 데이터(필드)를 함부로 수정하지 않아야 하고, 추가적인 정보가 더해져야 한다. 이것은 파싱의 정확성 검사에 항상 사용될 수 있다.

이 방법론의 단점들은 지명사전(대부분의 공개적으로 구할 수 있는 지명사전은 상당한 수의 오류를 가지고 있다. 예를 들어, figure 15 참고)에 있을 수 있는 오류를 포함하고, 가장 가까운 지명 위치는 꽤 넓은 지역(아래 정확성 할당에 대한 설명 참고)을 가리킬 수 있고, 많은 위치 필드는 위에서 인용된 것처럼 명확하지 않고, 때때로 역사적인 지명이 사용되며, 그리고 수집물 라벨상에 있는 많은 거리는, 비록 이것이 라벨 자체에 거의 서술되어 있지 않지만, 직선 거리보다는 "길에 따른" 거리 간격이 사용된다는 것이다. 정확성 필드는 이러한 문제와 함께 벡터 거리에 내재된 오류 – "남서"가 "남쪽"과 "서쪽" 사이인지 또는 남남서와

서남서의 사이인지를 의미하는 것인가를 고려할 필요가 있다. 시작 지점에서 이러한 거리가 늘어날수록, 이러한 것들에 있는 본질적인 오류도 또한 빠르게 증가할 것이다 (Wieczorek *et al.* 2004 의 논의 참고). 간단한 GIS 와 함께 이 방법을 사용하면 조작자에게 지도상의 레코드를 본 후, 더욱 적합한 장소에 대한 지점(예를 들어 가장 가까운 도로)으로 "끌어서 옮기기(grab and drag)"하는 기회를 제공할 수 있을 것이다.

vii. 지리코딩 소프트웨어

이용자들이 자신들의 수집물에 지리참조연산 수행을 지원하는 많은 수의 온라인, 독립형 프로그램이 개발되었다. 여기에서 다섯 개를 소개한다 – 두 개는 '온라인'이고 세 개는 '독립형'이다.

BioGeoMancer

BioGeoMancer 는 자연사 수집물에 대한 자동화된 지리참조연산 시스템이다 (Wieczorek and Beaman 2002). BioGeoMancer 의 현재 상태는 프로토타입 시스템으로, 아래 설명은 사용자 친숙성을 개선할 것으로 확실히 기대되는 이것의 계획된 여러 가지 향상된 기능에 대해서는 고려하지 않은 것이다. BioGeoMancer 는 영어로 된 지명 서술문을 파싱해서 이 서술문과 연관된 일련의 위도와 경도 좌표를 제공할 수 있다. 일반 텍스트, 즉 영어로 된 장소 데이터에 대한 파싱은 가장 가까운 지명, 거리 그리고 방향의 형식으로 결과물을 제공한다 (Wieczorek 2001b):

- 2.4 km WNW of Pandemonium
- Springfield, 22 miles E
- Springfield, 0.5 mi. E of Pandemonium

몇몇 다른 프로그램(예, Diva-GIS, eGaz)처럼, 이것은 파싱된 정보를 가지고, 적합한 지명사전과 함께, 일련의 위도 및 경도 좌표를 계산한다. BioGeoMancer 는 텍스트 파싱 기능을 제공한다는 점에서 다른 지리코딩 프로그램에 비해 장점이 있다. 이것은 인터넷을 통해서 일반 대중과 연구자들이 첫 번째로 이용할 수 있는 종류의 지리-파싱 프로그램이다.

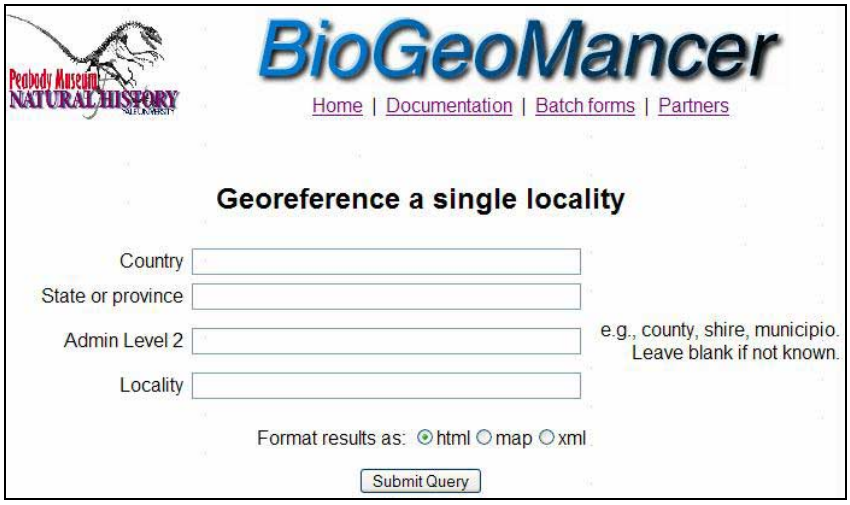

Fig. 2. *하나의 장소에 대한 BioGeoMancer 입력 형태 http://biogeomancer.org/ (Peabody Museum n.dat.).*

BioGeoMancer 프로그램은 두 가지 형태로 존재한다. 첫 번째는 하나의 표본에 대한 웹 검색을 할 수 있는 형태로 사용자가 장소 필드에 입력을 하면 지리참조연산된 결과가 반환된다 (figure 2). 두 번째 형태인 일괄 처리는 콤마로 구분된 버전 (figure 3) 또는 SOAP/XML 버전의 HTTP/CGI 인터페이스를 통해 데이터를 받고 지리참조연산된 레코드를 구분자가 있는 형식, html, 테이블 (figure 4), 또는 xml 형식으로 된 반환 파일을 제공한다. 이 프로젝트는 최근 사업비 측면에서 상당한 지원을 받았으며 새롭고 향상된 지리참조연산 도구를 개발을 추진하는 세계적인 협업 프로젝트로 확장되었다.

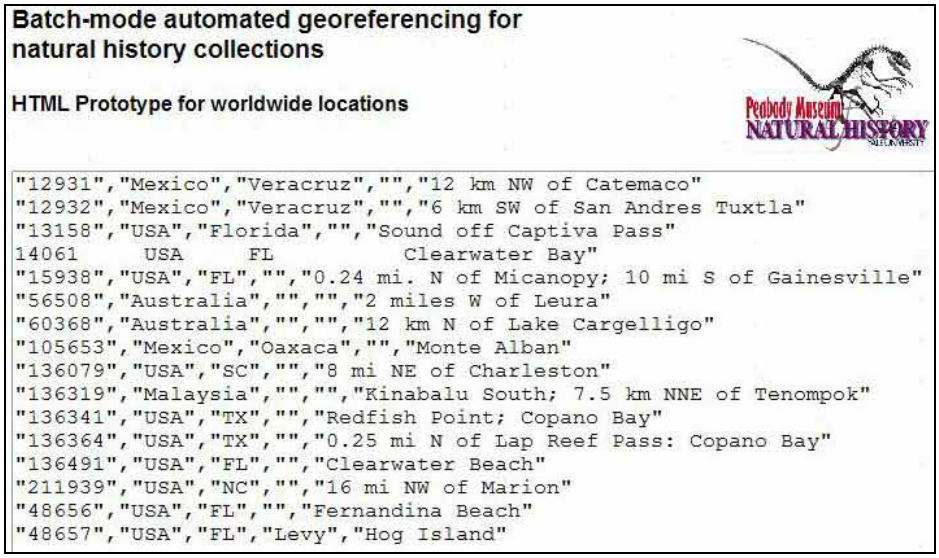

Fig. 3. *자연사 수집물에 대한 자동화된 지리참조연산 도구인 BioGeoMancer 웹-기반 일괄-모드에 대한 입력 형태* http://georef.peabody.yale.edu/yu/bgm-forms/batch-int02.html *(Peabody Museum n.dat.).*

Biogeomancer Results

Summary

Query Id	Query Country	Query Adm1	Query Adm2	Query Locality	Number of records matched	Centroid Latitude	Centroid Longitude	Error radius (km)	Multipoint match	Bounding box
12931	Mexico	Veracruz		12 km NW of Catemaco	1	18.49331	-95.19701	0.0	MULTIPOINT(-95.19701 18.49331)	BOX(-95.19701 18.49331, -95.19701 18.49331)
12932	Mexico	Veracruz		6 km SW of San Andres Tuxtla	1	18.41167	-95.25682	0.0	MULTIPOINT(-95.25682 18.41167)	BOX(-95.25682 18.41167, -95.25682 18.41167)
13158	USA	Florida		Sound off Captiva Pass	1	26.60917	-82.22222	0.0	MULTIPOINT(-82.22222 26.60917)	BOX(-82.22222 26.60917, -82.22222 26.60917)
14061	USA	FL		Clearwater Bay	1	27.97222	-82.82083	0.0	MULTIPOINT(-82.82083 27.97222)	BOX(-82.82083 27.97222, -82.82083 27.97222)
15938	USA	FL		0.24 mi. N of Micanopy; 10 mi S of Gainesville	1	29.50614	-82.325	0.0	MULTIPOINT(-82.32500 29.50614)	BOX(-82.32500 29.50614, -82.32500 29.50614)
56508	Australia			2 miles W of Leura	2	-28.449995	149.925235	587.4	MULTIPOINT(149.55188 -23.18333, 150.29859 -33.71666)	BOX(149.55188 -33.71666, 150.29859 -23.18333)

Fig. 4. *자연사 수집물에 대한 자동화된 지리참조연산 도구인 BioGeoMancer 웹-기반 일괄-모드에서 산출된 테이블 형식의 샘플 일부 (Peabody Museum n.dat.).*

하나 이상의 옵션이 가능할 경우, 모든 것은 해당 ID 하에 출력된다. 달리 다른 옵션이 없으면, 해당 레코드는 반환되지 않는다. *경계 박스 (Bounding Box)* 열은 계산된 정확성을 제공한다. 이 시스템은 많은 데이터에 대한 잘 작동하지만, 위의 지명, 거리 그리고 방향으로 쉽게 파싱되지 않는 텍스트 데이터의 경우는 처리에 실제 어려움이 있다. 현재 버전에서 아래와 같은 주의할 문제들이 있다. (이것들을 감소시킬 차후 개선 기능들이 계획되고 있다):

- 영어 서술문에 제한된다.
- 현재 버전에서는 정확도가 경계 박스로만 출력되고, 이것은 개선될 수 있을 것이다. John Wieczorek (2001b)이 이미 개발한 관련 프로그램(Georeferencing Calculator)은 이 정보(http://mainset.org/manis/gc.html)를 제공할 수 있고 이것은 추후에 BioGeoMancer 에 링크될 가능성이 있다. 이미 지리참조-연산을 하고 연관된 불확실성을 계산하기 위한 "점-반지름 방법"이라고 알려진 것을 통해 정확성을 자동으로 할당하는 방법에 대한 연구가 이미 시작되었다 (Wieczirek *et al.* 2004)
- 두개의 지명은 (예, "10 km W of Toowoomba toward Dalby") null 결과를 산출한다.

또 다른 파싱 프로그램인 RapidMap Geocoder (NMNH 1993)는 1993 년 미국 자연사국립박물관(US National Museum of Natural History)과 하와이 버니스 비숍 박물관 (Bernice P. Bishop Museum in Hawaii)에서 개발되었고 이것의 목적은 하와이의 지명에만 사용하는 것이었다. 그렇지만 이것은 성공적으로 여겨지지 않아 중단되었다. 그러나 파싱 방법론에 사용된 일부 유용한 정보는 인터넷에서 구할 수 있다: http://users.ca.astound.net/specht/rm/tr_place.htm.

GeoLoc-CRIA

GeoLoc 은 공보화된 장소로부터 브라질의 장소, 알려진 거리 그리고 방향을 검색하는 간단한 웹 기반 프로그램이다. 이것은 CRIA(Marion *et al.* in prep.)에서 개발되었다. GeoLoc 은 eGaz 프로그램(아래 참고)과 유사한 방식으로 작동하고 http://splink.cria.org.br/tools/ (CRIA 2004a)에서 관련 정보를 볼 수 있다. 이것의 프로토타입은 많은 지명사전을 포함하고 한 지역에 대한 지명사전이 하나 이상 있을 경우, 그 중 하나를 선택하는 기능을 제공하며, 계산된 오류 값을 또한 제공한다.

한가지 예를 figure 5 에서 볼 수 있으며, 브라질 상파울루(São Paulo)에 있는 Campinas 에서 북동쪽으로 25km 떨어진 지점의 위도와 경도를 검색하고 있다. 우선 여러 개의 지명사전 중의 하나 또는 *species*Link 레코드를 (주로 상파울루 주에 있는 일련의 데이터베이스들에 대해 분산 검색을 통해서 얻은 레코드) 이용하여 Campinas 의 위치를 검색한다. 다음으로 여기에 "25km"과 "북동쪽(NE)"(동그라미로 표시됨)을 추가한 후 관련된 'Campinas' (화살표)를 클릭하면, 해당 결과가 지도상에 (figure 6) 표시된다. 지리코드는 -46.9244, -22.7455, 오차는 9.754km 를 (동그라미로 표시됨) 보여주고 있다. 이 정보(위도, 경도, 그리고 오류)는 이미 마이크로소프트 버퍼에 저장되어서 마이크로소프트 호환 파일인 워드(Word), 엑셀(Excel), 액세스(Access)에 붙여넣기 할 수 있다. 이 지도는 "Campinas"의 위치를 세가지 소스로부터 또한 보여준다 – 빨간색은 선택된 것으로 "Campinas 에서 북동쪽으로 25km" 지점을 가리킨다. 이 지도는 확대와 축소를 할 수 있고, 다양한 환경 계층 정보를 덧붙이거나 제거할 수 있다.

이 프로그램은 장소 정보를 포함하는 엑셀 워크시트에 링크할 수 있고 추후 검색을 위해 결과 값을 html 테이블 또는 엑셀 워크시트 형태로 만들 수 있다. 이 프로그램의 주요 단점은

브라질의 장소에 대해서만 이용 가능하다는 점이다. 이 알고리듬은 더 큰 프로젝트인 Biogeomancer 에 현재 통합되고 있다.

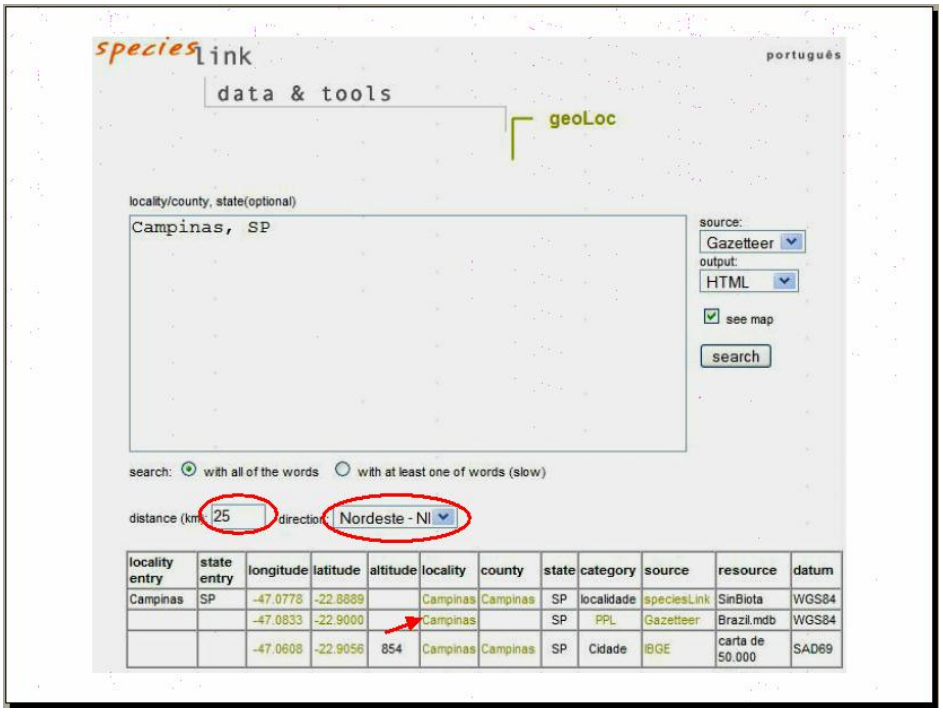

Fig. 5. *CRIA 의 geoLoc 프로그램을 이용하여 상파울루(SP), Campinas 의 25km 북동쪽에 있는 장소의 지리코드를 찾기*

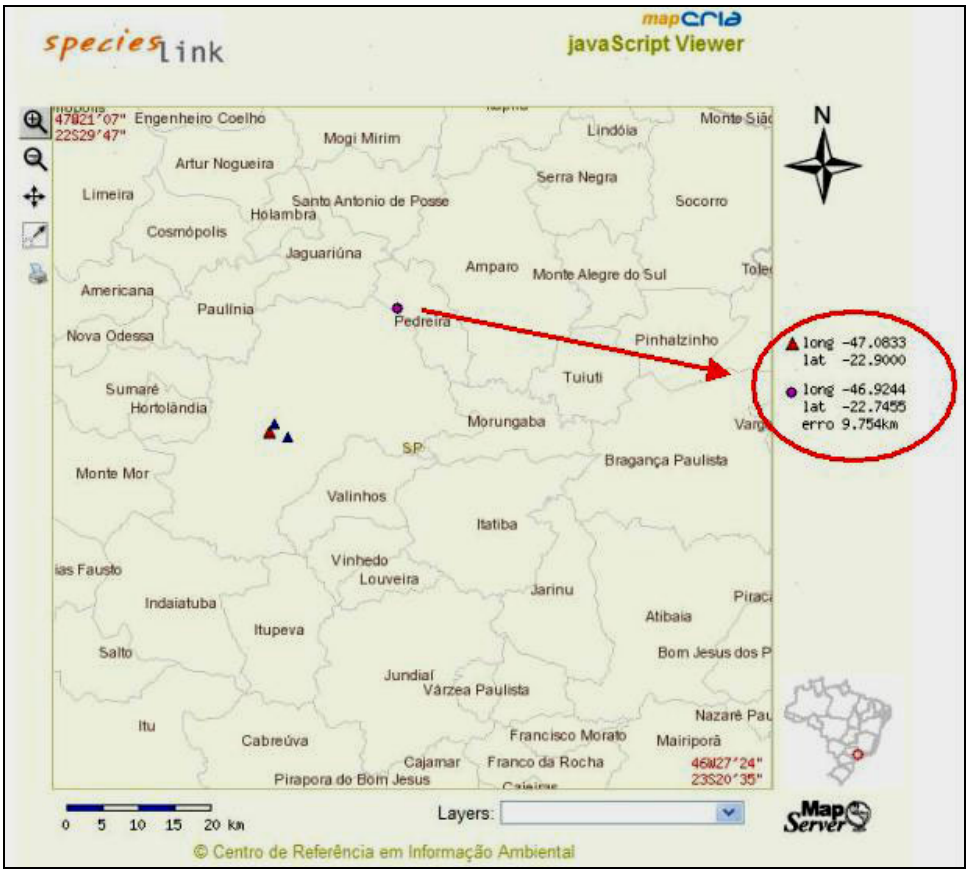

Fig. 6. *관련된 지리코드 정보와 오류를 표시하면서 (다양한 소스로부터) "Campinas" 위치와 Campinas 에서 북동쪽으로 25km 지점을 보여주는 위 선택의 결과 (동그라미로 표시됨)*

GEOLocate

GEOLocate (Rios and Bart *n.dat*)은 툴레인(Tulane) 대학교의 자연사박물관에서 개발된 지리참조연산 프로그램으로 자연사 수집물과 연관된 장소 데이터에 지리 좌표를 할당하는 일을 수월하게 하기 위해 설계되었다. GEOLocate 의 주 목적은 다음과 같다:

- 텍스트 형식의 자연사 데이터를 북미의 위도와 경도로 변환하는 알고리듬 개발;
- 생성된 좌표의 가시화와 상세 조정을 위한 인터페이스 제공;
- 사용자들이 자신의 데이터를 가져오고 지리참조연산할 수 있도록 간단한 해결책 제공;
- 자동-갱신 기능 제공.

이 알고리듬은 처음에 장소에 대한 문자열을 일반 용어로 표준화한 후에, 거리, 방향, 그리고 지명과 같은 주요 지리 식별자를 파싱한다. 다음으로 이 정보는 지리 좌표를 결정하기 위해 지명사전과 (지명, 강의 마일, 토지이용 그리고 도로/강 교차 데이터) 함께 사용된다. 이 프로그램은 또한 사용자가 수계(water-body)에 가장 가까운 장소를 검색 가능하도록 한다.

이 프로그램은 툴레인(Tulane) 대학교로부터 이용 가능하고, 온라인 데모는 다음 사이트에서 볼 수 있다: http://www.museum.tulane.edu/geolocate/demo.aspx (figure 7).

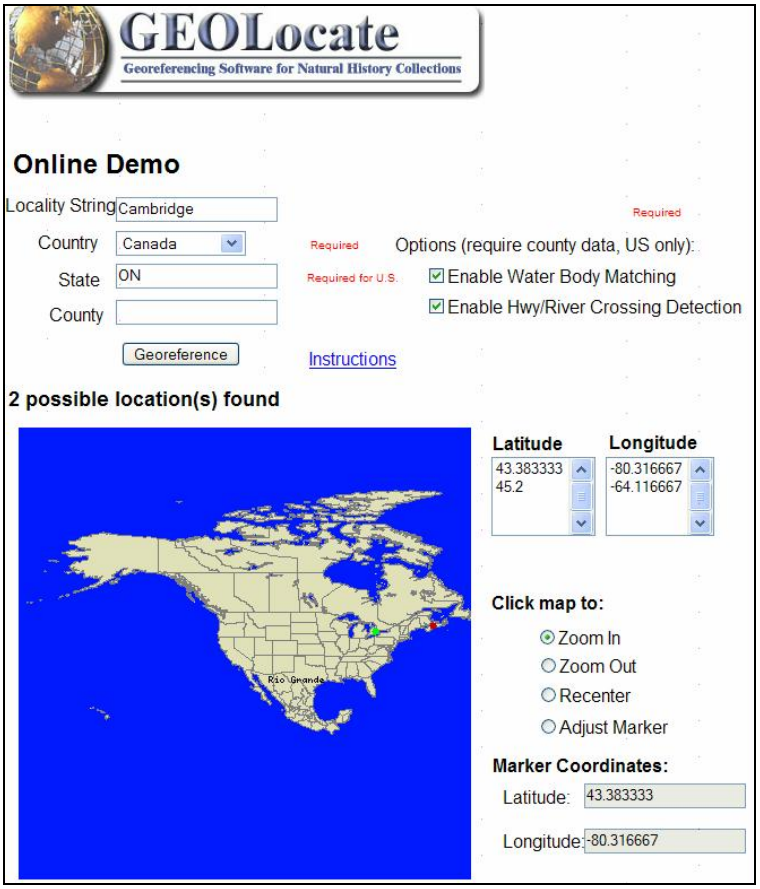

Fig. 7. *온타리오, 캠브리지에 대한 지리좌표를 식별하기 위해 온라인 데모 버전을 사용하는 GEOLocate 인터페이스의 예.*

이 프로그램은 북미 (멕시코, 미국, 그리고 캐나다) 지역에 대해서만 작동하지만, 관련 개발자들은 전 세계를 포함할 수 있도록 현재 이것을 확장하는 작업을 하고 있다. 그외 다른 개발 기능은 DiGIR 호환성, 다국어 지원, 그리고 고급 검증 기법을 포함할 것이다 (N.Rios pers.com. 2004).

eGaz

eGaz (Shattuck 1997)는 박물관과 식물표본관이 자신들의 표본 레코드에 대해 동정하고 지리코드 추가를 지원할 목적으로 CSIRO 의 호주국립곤충수집물(Australian National Insect Collection)에서 개발된 프로그램이다. 데이터 입력 및 표본 관리 소프트웨어인 BioLink (Shattuck and Fitzsimmons 2000)의 개발로, 이것은 BioLink 패키지로 통합되었다. eGaz 는 BioLink 패키지의 일부로 이용 가능하다 (아래 참고).

eGaz 는 도시, 마을, 산, 호수 그리고 그 외 다른 지명에 대한 위도와 경도 측정에 종이로 된 지도와 자를 필요하지 않게 하고 있다. eGaz 는 또한 어느 한 지명으로부터 알려진 거리와 방향을 가진 장소에 대한 위도와 경도를 계산할 수 있다. 이 프로그램은 임의의 어느 지역에 대한 지명사전을 쉽게 추가할 수 있게 하였으며, 세계 많은 지역의 지명사전을 CSIRO 사이트에서 (http://www.biolink.csiro.au/gazfiles.html) 다운로드할 수 있다.

eGaz 는 두개의 창, 즉 지명사전 창과 지도 창을 제공하는 마이크로소프트 윈도우 기반 제품이다 (figure 8). 이것은 "지명", "거리", 그리고 "방향"의 형태로 된 장소의 정보에 대해 사용자가 지리코드를 구하고 이것을 파일로 전송할 수 있도록 한다.

Figure 8 에 나온 예시는 호주, 퀸스랜드(Queensland)의 "80 km SSW of Toowoomba (Toowoomba 에서 남남서쪽으로 80km)" 위치의 위도와 경도를 구하고 있다. 첫 번째 단계는 적합한 지명사전을 로딩하고 이것에서 "Toowoomba"를 선택하는 것이다 **(A)**. 이 때 많은 옵션이 있고, 필자는 Toowoomba City 를 선택했다 (거주지라는 뜻에서 POPL(Populated Place)이라고 분류되어 있다). Toowoomba 의 위치는 지도에서 빨간색으로 나타난다 **(B)**. 거리 필드에 거리 "80"을 입력하고 "km"와 "SSW"을 선택하기 위해 풀다운 메뉴가 사용되었다 **(C)**. 선택된 위치는 지도상에서 파란색 점으로 나타난다 **(D)**. 위도, 경도와 함께 해당 위치가 지명사전 창의 밑에 또한 나타난다 **(E)**. 이 영역에 대해 오른쪽 마우스 클릭하고 "Copy"를 선택하면 해당 정보는 임의의 마이크로소프트 호환 파일 (워드, 엑셀, 액세스)에 복사되어 붙여질 수 있다. 위도와 경도(1 분 정확도)도 또한 나타나고 **(F)**, 이것은 같은 방식으로 파일에 복사될 수 있다. 다른 방식으로, Edit 메뉴에서 "Copy Lat/Long"을 선택하면 이 지리코드는 1 초 정확도로 복사될 수 있다.

사용자는 지도 자체로 이동하여 해당 지점을 확대할 수 있다. 도로 네트워크(ESRI Shape 파일 형식)와 같은 다른 계층 정보를 로딩하여 더욱 정확하게 해당 위치를 알 수 있다 – 즉, 수집활동이 운송수단에서 행해졌다면 가장 가까운 도로로 이것을 옮길 수 있을 것이다. 다음으로 선택 도구를 사용하여 해당 지점을 클릭하면 1 초 해상도로 지리코드를 얻을 수 있다. 다시 마우스로 오른쪽 클릭을 하거나 Edit/Copy Lat/Long 을 이용하면 해당 정보를 적합한 파일에 복사 할 수 있다.

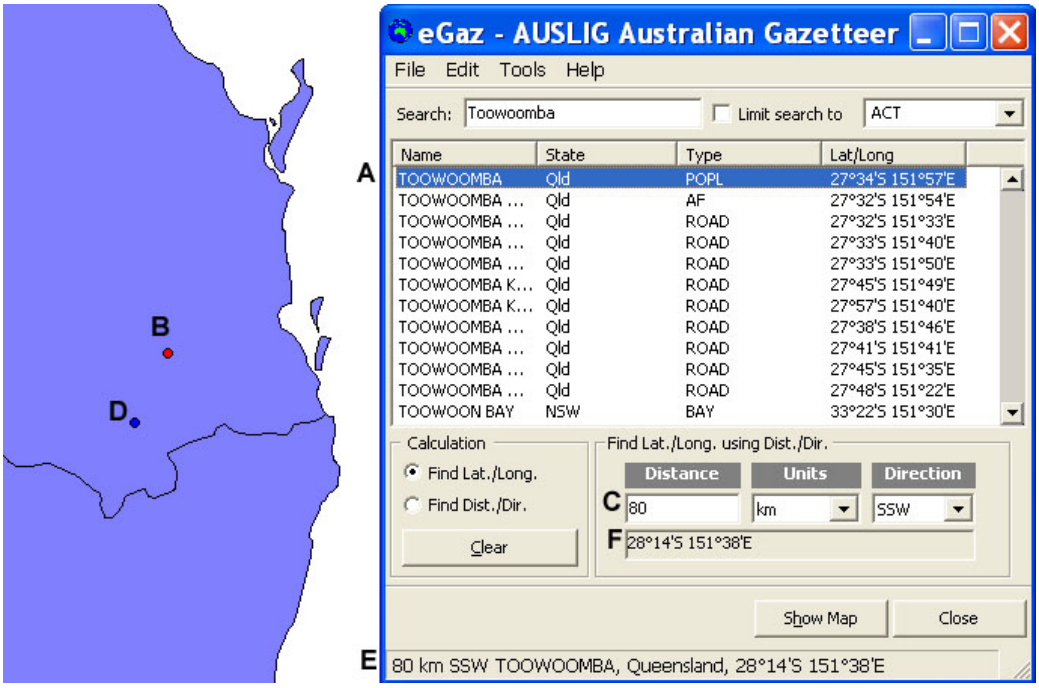

Fig. 8. *호주, Toowoomba 에서 남남서 방향으로 80km 떨어진 지점의 위도와 경도를 측정하는 것을 보여주는 eGaz 의 샘플 출력 화면.* ***A.*** *지명사전에서 Toowoomba 의 정보.* ***B.*** *Toowoomba 의 지도상 위치.* ***C.*** *하이라이트된 위치에 대해 80km SSW 입력.* ***D.*** *Toowoomba 에서 남남서 방향으로 80km 떨어진 지점의 지도상 위치.* ***E.*** *지점 세부정보.* ***F.*** *새로운 위치의 위도와 경도.*

Diva-GIS

Diva-GIS 는 박물관과 식물표본관에서 사용할 목적으로 개발된 무료 GIS 프로그램이다. 이것은 좌표가 없는 표본 데이터에 좌표 할당을 지원하는 알고리듬을 포함하고 있다. 이 프로그램이 수용할 수 있는 형식으로 데이터를 구조화하기 위한 일부 전처리 작업이 필요하지만, 많은 데이터베이스가 이미 이러한 방식으로 데이터를 구조화하고 있다. 입력 파일에 포함된 텍스트 형식의 위치 데이터는 몇 개의 특수 필드에 파싱된다. 이러한 것은 "지명 1", "거리 1", "방향 1", 그리고 "지명 2", "방향 2", 그리고 "거리 2"가 있다. 예를 들어 아래의 장소 레코드는:

"*growing at a local place called Ulta, 25.2 km E of Chilla*"

다음과 같이 파싱될 것이다:

지명 1:	Ulta
거리 1:	
방향 1:	
지명 2:	Chilla
거리 2:	25.2km
방향 2:	E

그리고

"14 km ESE of Sucre on road to Zudanez"

은 다음과 같이 파싱될 것이다:

지명 1:	Sucre
거리 1:	14 km
방향 1:	ESE
지명 2:	Zudanez
거리 2:	
방향 2:	

단지 한 쌍의 "지명", "거리", 그리고 "방향"에 대한 정보로 많은 레코드들에 대해 지리코드를 제공할 수 있을 것이며, 이것은 대부분의 기관이 가지게 될 모든 정보이다. Diva-GIS 의 개발자들(Hijmans *et al.* 2005)은 데이터 내의 부정확성을 보상하고, 직선이 아닌 실제로는 길을 따라 25 km 라는 것을 의미하는 어느 한 장소에서 북쪽으로 25km 의 경우를 처리하기 위해 거리를 정수로 내림할 것을 권장한다. 필자는 오히려 반대의 경우를 권장하고 싶은데, 필자는 주어진 가장 정확한 숫자를 기록하고, "정확성" 필드에 미터 단위로 정확성 숫자를 기록할 것이다.

입력 파일이 선택되고, 출력 파일의 이름이 정해지고, 그리고 풀다운 리스트에서 적절한 필드 이름이 선택되었다면, 알고리듬이 실행되고 출력 파일을 생성한다 (figure 9). 이 알고리듬은 좌표를 할당하기 위해 적합한 지명사전을 사용한다.

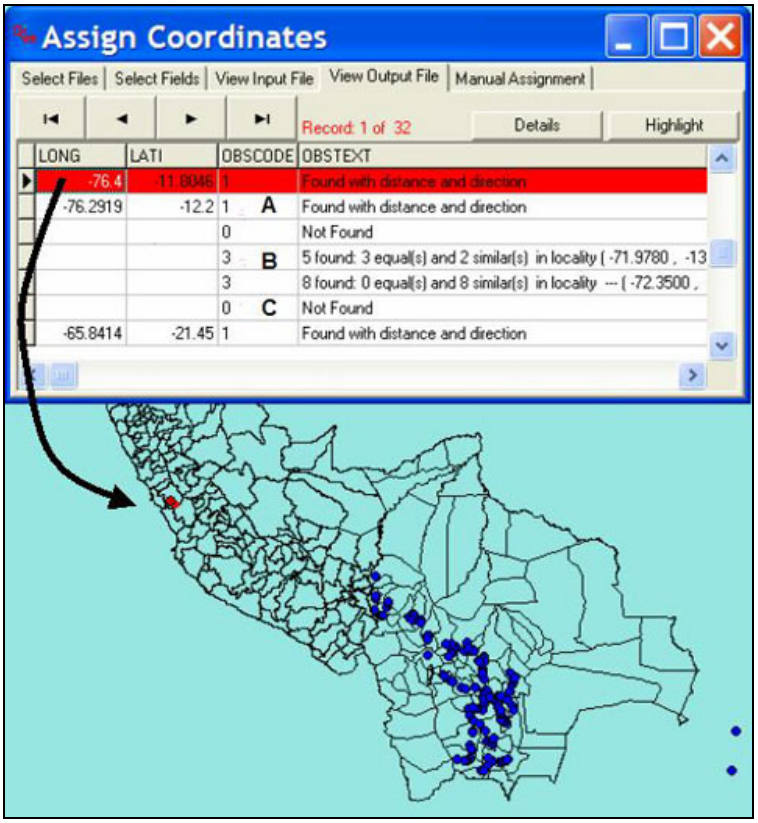

Fig. 9. *지리코드가 자동으로 할당된 점 레코드를 보여주는 Diva-GIS 의 결과 화면.* ***A.*** *프로그램이 명확한 지리코드를 발견하고 할당.* ***B.*** *모호한 지리코드 발견.* ***C.*** *적합한 지리코드가 발견되지 않음.*

이 예에서 보는 것과 같이 (figure 9), 이 프로그램은 입력파일의 "지명" 필드를 이용하여 지명사전에서 많은 수의 레코드에 대해 명확한 짝을 찾아서 이러한 레코드에 적합하게 계산된 지리코드를 할당했다 **(A).** 일단 출력 파일이 로딩되고 형태 파일이 만들어지면, 이러한 레코드의 각각은 지도상에 플래싱 점(flashing point)으로 나타내기 위해 하이라이트 될 수 있다. 다른 많은 경우, 이 프로그램은 "지명"에 대해 지명사전에서 몇 가지 가능한 짝을 찾아 적절하게 이것을 보고하였다 **(B).** 그러나 그외 다른 경우**(C)**에는 프로그램은 지명사전에서 짝을 찾지 못했다.

몇 가지 가능한 짝이 발견된 레코드의 경우, 사용자는 **(B)** 레코드중의 하나를 더블 클릭하고 또 다른 출력 파일을 생성하는 다음 단계로 이동할 수 있다 (figure 10).

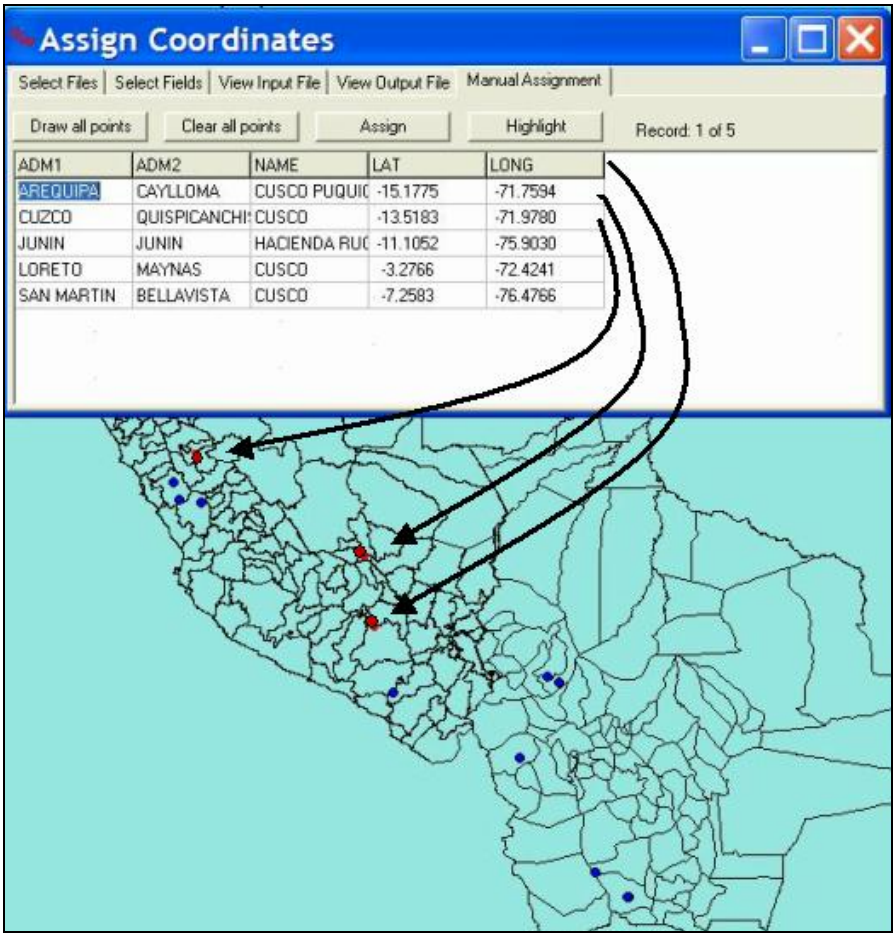

Fig. 10. *지명사전에서 신뢰할 수 있는 여러 개의 지명이 추출될 경우 대체 가능한 지리 코드들을 보여주는 Diva-GIS 의 결과화면.*

Figure 10 에서 보여진 레코드의 경우, 이 프로그램은 지명사전에서 5 개의 가능한 대체 위치를 식별하였고 사용자가 선택할 수 있도록 GIS 상에 이러한 대체 정보를 제시하였다. 하나를 선택한 후, 이것이 출력파일에 저장될 수 있도록 하기 위해서는 간단히 "Assign" 버튼을 클릭하면 된다. 다른 방식으로, 사용자는 전혀 다른 위치를 선택한 후 지리코드를 추가하거나 할당된 것 중의 하나를 수정하기 위해 "수동할당(Manual Assignment)"을 사용할 수 있다.

지리코드 검사 및 검증

일단 데이터베이스화된 표본 레코드의 지리코드를 검사하고 검증하는데 사용될 수 있는 네 가지 주요 방법이 있다. 이것들은 내부적인 불일치 검사를 위한 데이터베이스의 이용, 지리정보시스템의 이용, 특이점 검사에 대한 환경 공간 이용, 그리고 지리 또는 환경공간에서 특이점 검사에 통계를 이용하는 것이다.

i. 데이터베이스를 이용하기

a. 내부 검사

대부분의 종 및 종-관련 데이터베이스는 어느 정도 중복된 정보를 포함하고 있다. 예를 들면, 수집 활동이 이루어졌던 주(State) 필드와 텍스트 형식의 위치 정보에 대한 필드가 그것이다. 일부 데이터베이스는 또한 "가장 가까운 장소"를 포함하고 있는데 이것은 장소 필드내의 정보와 또한 중복될 수 있다. 다음으로 한 필드에 있는 인용된 도시 또는 가장 가까운 장소가 올바른 주 또는 구역, 또는 심지어 다른 필드에서 인용된 국가 내에 위치하는지에 대한 검사를 진행할 수 있다.

데이터베이스 내부에 있는 유사한 레코드간의 정보를 검사하는 것 또한 가능하며, 예를 들어, 제공되는 위도와 경도에 대해 모든 지명을 검사할 수 있다. 어떤 사람이 한 장소에서 수집된 5 개 수집물에 대한 데이터베이스를 가지고 있을 수 있다 – 예를 들어, "SP, 캠피나스에서 북쪽으로 10 km (10 km N of Campinas, SP)". 이것들이 모두 같은 위도와 경도를 가지는가 또는 하나 또는 그 이상의 것들이 다른 것들과 서로 다른가? 아래 *순서연관규칙(Ordinal Association Rules)*에 관한 논의를 참고하기 바란다.

Data Cleaning (speciesLink)

CRIA 의 speciesLink 분산정보시스템(Distributed Information System) (CRIA 2002)에 있는 데이터 정제 모듈은 수집물 관리자들이 자신들의 데이터를 처리할 때 도움이 되도록 잠재적인 오류를 동정하는 많은 기능을 포함하고 있다. 현재 이것은 포르투갈어 버전으로만 되어 있지만 영어 버전이 추진되고 있다. 이 도구의 일부분은 이름내에서 오류를 동정하는 것이 있다. 이것은 아래와 같은 여러 기능을 포함하고 있다:

- 접근되는 데이터베이스의 발생 횟수와 함께 모든 이름(과, 속, 종, 아종)을 나열. 하나의 예 (figure 11)를 잠깐 살펴보면 명백히 많은 문제가 있음을 알 수 있다. 첫 번째 줄에서 보면 과 하위 어느 수준에서도 동정되지 않은 데이터베이스의 레코드가 101 개 발생하고 있음을 알 수 있다. 두 번째 줄에서는 "4606euphorbiaceae"라는 과 이름이 1 개 있으며, 세 번째 줄에서는 과 수준에서만 동정된 Acanthaceae 에 대한 5 개의 레코드를 보여주고 있다.

family	genus	species	subspecies	ocor_col
[]	[]	[]	[]	101
[4606euphorbiaceae]	sp [Julocroton]	[humilis]	[var. subpannosus]	1
[Acanthaceae]	[]	[]	[]	5

Fig. 11. *몇몇 잠재적인 오류를 보여주는 CRIA 데이터정제 모듈의 일부*

- 속 이름에 있는 잠재적인 오류 검사. 이것은 과명이 같고, 종명도 같고, 속 이름도 유사하지만 (Soundex 와 유사한 알고리듬을 이용하여 동정됨) 속 이름의 철자가 다른

경우이다. 이 출력물은 조사중인 데이터베이스에 있는 각각의 발생 빈도와 speciesLink 를 통해 접근한 데이터베이스의 총 발생 빈도를 보여준다. 이 예(figure 12)는 *Hieronyma* 속의 (*alchornioides* 철자와 함께 있지만, 이것들은 다른 루틴에서 식별된다) 두 개의 서로 다른 철자를 각각의 발생 빈도와 함께 보여주고 있다. 사용자가 "sp" 버튼을 클릭하면 해당 기관의 내부 뿐만 아니라 외부를 포함하는 일련의 데이터베이스를 검색할 수 있고, 이것은 국제식물이름색인(International Plant Name Index, IPNI), species 2000 등의 자원을 포함하며, 이것 모두는 사용자가 어느 것이 올바른 철자인지를 판단할 때 도움을 줄 수 있다.

family	genus	species	subspecies	ocor_col	ocor_total
[Euphorbiaceae]	sp [Hyeronima]	[alchorrneoides]	[]	3	3
[Euphorbiaceae]	sp [Hieronyma]	[alchorrneoides]	[]	1	1
[Euphorbiaceae]	sp [Hyeronima]	[alchornioides]	[]	9	9
[Euphorbiaceae]	sp [Hieronyma]	[alchornioides]	[]	17	17

Fig. 12. *몇몇 잠재적인 오류를 보여주는 CRIA 데이터정제 모듈의 일부.*

- 종 이름 또는 종소명에서 잠재적인 오류를 조사하기. 속 이름과 같이, 이것은 속 이름이 동일하고, 종소명에 대한 발음은 같지만, 종소명의 철자가 다른 것을 검사한다. 이전처럼 이 출력결과는 조사중인 데이터베이스에서 전체 발생 빈도수, 접근되는 모든 데이터베이스의 전체 발생 빈도수, species2000[4]에서 해당 이름의 상태를 보여준다. 아래 예(figure 13)는 대체 이름을 가진 몇몇 종 이름을 보여주고 있다. 각 이름의 발생 빈도는 species2000 에서의 상태 정보가 이용 가능하다면 이것과 함께, 여러 철자 중에 어느 것이 오류인지를 가리킬 수 있다.

genus	species	subspecies	ocor_col	ocor_total	status_sp2000
sp [Acacia]	[polyphylla]	[]	83	217	accepted name
sp [Acacia]	[polyphyllla]	[]	1	1	
sp [Banisteriopsis]	[argyrophylla]	[]	25	164	
sp [Banisteriopsis]	[argirophylla]	[]	2	2	
sp [Bauhinia]	[cuyabensis]	[]	19	19	provisionally accepted name
sp [Bauhinia]	[cuiabensis]	[]	1	2	
sp [Bignonia]	[unguis-cati]	[]	1	5	unambiguous synonym
sp [Bignonia]	[unguiscati]	[]	2	2	

Fig. 13. *몇몇 잠재적인 오류를 보여주는 CRIA 데이터정제 모듈의 일부.*

- 저자명에서 차이점과 잠재적인 오류를 조사하기. Figure 14 는 단지 하나의 종 이름에 대해 여러 개의 가능한 이름을 보여주고 있다. 이전과 마찬가지로 여기에서 "sp" 버튼을 클릭하면 각기 다른 여러 데이터베이스에 대해 검색을 수행할 수 있어 어느 것이 사용하기에 가장 적합한 것인지를 결정할 때 도움이 된다.

[4] http://www.species2000.org

genus	species	subspecies	author	ocor_col	ocor_total
sp [Actinostemon]	[concolor]	[]	[Müll.Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spreng.) M.Arg.]	0	13
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Muell.Arg.]	0	17
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	1	2
sp [Actinostemon]	[concolor]	[]	[(spreng.) Müll. Arg.]	0	2
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll. Arg.]	0	52
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll.Arg.]	88	139
sp [Actinostemon]	[concolor]	[]	[(Spreng) Müll. Arg.]	0	6
sp [Actinostemon]	[concolor]	[]	[(spr.) Muell. Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spr.) Muell.Arg.]	0	2

Fig. 14. *몇몇 잠재적 오류를 보여주는 CRIA 데이터정제 모듈의 일부.*

- 과 이름과 아종 이름의 차이점 조사도 비슷한 방식으로 작동한다.

다른 기능들은 데이터집합에서 잠재적인 지리 오류를 동정하기 위해 사용되며, 이것들은 아래 *공간 데이터*에서 다루어진다. CRIA 는 데이터 소유기관이 아니며, 데이터에 어떤 변경을 하지 않지만 데이터 소유기관들에게 자신들의 데이터베이스에서 잠재적인 오류를 동정하는데 도움이 되도록 서비스를 제공한다. 어느 것이 올바른 형식이고 어느 레코드가 수정되어야 하고 그렇지 않은지를 결정하는 것은 데이터 소유기관에서 결정한다.

b. 외부 데이터베이스

외부 데이터베이스에 연결됨으로써, 종-발생 데이터의 다양한 측면에서 오류가 동정될 수 있다. 이러한 데이터베이스는 수치표고모델 (Digital Elevation Models), 공간 지형 데이터베이스, 지명사전, 그리고 수집자들의 여행 일정기를 포함할 수 있다.

조금 더 정교한 데이터베이스들이 고도 필드의 정확성을 검사에 사용될 수 있으며, 이것은 인용된 고도와 데이터베이스화된 수치표고모델 (Digital Elevation Model, DEM)의 고도와 비교함으로써 가능하다. 이 DEM 이 적절한 축척으로 사용되는 것이 중요하며, 대부분 표본 데이터의 각기 다른 정확성 때문에 주의 깊게 사용되지 않으면 거짓 또는 잘못된 오류를 발생시킬 수 있다. 이러한 기술은 호주의 ERIN (Environmental Resources Information Network)에서 10 년 넘게 성공적으로 사용되었다 (Chapman unpublished). 이 과정은 오라클 (ORACLE) 데이터베이스를 이용한 일괄 처리를 사용하며 분당 3000 개 이상의 고도 레코드를 검사 (또는 할당) 할 수 있다.

아주 최근 ESRI 의 공간데이터베이스엔진 (Spatial Database Engine) (ArcSDE®) (ESRI 2003)과 PostGIS 와 같은 정교한 공간 데이터베이스가 개발되었으며, 이것들은 지리코드 자체를 이용하여 더욱 더 복잡한 데이터베이스 검색을 가능하게 한다. 하지만 이러한 종류의 소프트웨어는 매우 값이 비싸며, 아주 극소수의 박물관 또는 식물표본관만이 이러한 것을 감당하거나 이러한 것에 대한 필요가 있을 것으로 예상되고, 이러한 이유로, 이 방법들은 더 이상 이 논문에서 다루어지지 않을 것이다.

지명사전은 전세계 대부분의 지역에 대해 이런 저런 형태로 존재하고, 종종 이것들은 데이터베이스 형식으로 다운로드할 수 있다. 이것들은 표본 데이터베이스 내부의 적절한 필드에 대해 정확성을 검사하는데 사용될 수 있다. 종종 이러한 데이터베이스 자체가 오류를 포함하고 있기 때문에 (예를 들어 figure 15 참고) 이러한 많은 데이터베이스를 사용할 때 주의가 필요하고, 해당 지역의 올바른 지명사전과 적절한 축척을 사용하는 것이 중요하다. 또한 많은 지명이 모호할 수도 있고 (예, 호주에는 수백개의 “Sandy Creek”이 있다) (Chapman and Busby 1994) 또는 현대의 지명사전에 발생하지 않는 역사적인 지명이 있을 수 있다. 지명이 실제로 어떤 것을 뜻하는지에 대한 문제도 또한 있을 수 있다 (Wieczorek 2001a). 새로운 BioGeoMancer 프로젝트(다른 곳의 설명 참고)의 특징 중 하나는 웹 서비시스 기술을

이용하여 생물학적 데이터베이스에 지명사전을 통합하는 것이다. 일반 대중의 참여를 통하여 지명사전을 개선하는 것과 특히 역사적인 수집물 장소를 포함하는 일을 시작하는 것이 또한 기대되고 있다.

거의 사용되지는 않지만 매우 큰 잠재력을 가진 한 방법은 수집자들의 장소에 관한 여러 데이터베이스에 대해 교차 검색을 하는 것이다. 오늘날까지 이러한 형태의 데이터베이스는 거의 없고 (그러나 앞에서 참조된 하버드 대학교 데이터베이스는 식물학 관련하여 좋은 시작점이 될 수 있을 것이다[5]), 그 외 다른 것들은 서서히 개발되고 있다. Peterson *et al.* (2003)은 최근 사례로서 멕시코의 조류를 이용한 새로운 통계적인 방법을 제시하였다. 이들은 특정 수집자의 수집물을 시간 순으로 정렬하였으며 매일 (또는 며칠) 마다 예상되는 최대 이동 반지름을 계산했다. 엑셀에서 공식-기반 접근방식을 사용하여, 이들은 계산된 범위 밖으로 분류되는 표본들의 잠재적인 오류를 동정하였다. 이것과 유사한 방법이 데이터베이스 자체에서 수행될 수 있을 것이다 – 아래 *순서연관규칙* 논의를 참고하기 바란다. 그렇지만, 이러한 방법은 수집자의 데이터베이스화된 수집물이 이러한 여행 일정기를 만들 수 있을 만큼 대용량일 경우에만 작동할 것이다.

ii. GIS 검사

지리정보시스템(Geographic Information Systems, GIS)은 최근 더욱 더 사용자에게 친숙해진 매우 강력한 도구이다. GIS 는 고비용, 고기능 시스템부터 무료이면서, 제한된 기능을 가진 일반 상품까지 다양하다. 그렇지만 많은 무료 GIS 시스템은 식물표본관 또는 박물관에서 필요로 하는 많은 기능을 제공할 만큼 강력하며, 일련의 데이터 검사와 데이터 정제 기능을 제공할 수 있도록 쉽게 고칠 수 있다.

	Points	Lines	Polygons
Points	▪ is a neighbour of ▪ is allocated to	▪ is near to ▪ lies on	▪ is a centroid of ▪ is within
Lines		▪ crosses ▪ joins	▪ intersects ▪ is a boundary of
Polygons			▪ is overlain by ▪ is adjacent to

Table 3: *객체의 클래스간의 관계 (Gatrell 1991 자료)*

GIS 는 또한 데이터베이스 내부의 논리적인 일관성을 검사하는데 사용될 수 있다. 지형적인 인코딩에서 과잉(redundancy)은 누락된 데이터 또는 라벨이 되어있지 않은 다각형과 같은 데이터 구조에서 결점을 탐지하는데 사용될 수 있다 (Chrisman 1991). GIS 는 공간 계층 상호관계를 통해 오류를 탐지할 수 있도록 하며 이것은 가시화와 함께 주요 장점이다.

다각형(지역, 주, 국가, 경작지)에 대해 점(표본 레코드)의 위치를 결정하는데 간단한 GIS 를 사용하는 것은 (지리 또는 고도) 데이터의 부적합성을 간파하는데 도움을 줄 수 있다. 이것은 GIS 시스템에 이용되는 일반적인 테스트이며 "다각형내의 점(point-in-polygon)" 방법으로 알려져 있다 – 이것은 해양의 부표가 내륙상에서 발생하지 않도록 하고, 강이 범람 지역 밖에서 발생하지 않도록 하는 것 등을 보장하기 위해 GIS 에서 사용된다. GIS 가 1 차 종 데이터에 대해서 수행할 수 있는 가장 중요한 테스트 가운데 하나는 내륙에 있어야 할 데이터가 실제로 내륙에 있고 해양에 있어야 할 데이터가 해양에 있다는 것을 검사하는 것이다. 대량의 데이터 집합을 처음 GIS 에 로딩할 때 많은 레코드가 명백히 잘못된 위치에 있다는 것을 단지 이 간단한 검사로 알 수 있다. 위치가 잘못된 레코드에 대해 GIS 를 이용한

[5] http://www.huh.harvard.edu/databases/cms/download.html

검사는 간단한 가시적인 조사에서부터 좀 더 상세한 자동화 검사까지 나눌 수 있다. 예를 들어, 가시적인 조사는 레코드가 관련된 올바른 국가에 속하는지를 결정할 때 또한 가치가 있을 수 있다. 여러분이 브라질 레코드에 대한 데이터베이스를 가지고 있다면, 여러분은 GIS 를 이용하여 이런 저런 방식으로 브라질 외부에 존재하는 것과 같은 위치가 잘못된 레코드를 빠르게 식별할 수 있을 것이다. 예를 들어, figure 15 에서, 공개적으로 이용할 수 있는 브라질 지명에 관한 지명사전의 레코드는 일부 명백한 오류가 있다. 표본 레코드의 오류도 이 방법론을 이용하여 유사하게 식별될 수 있다.

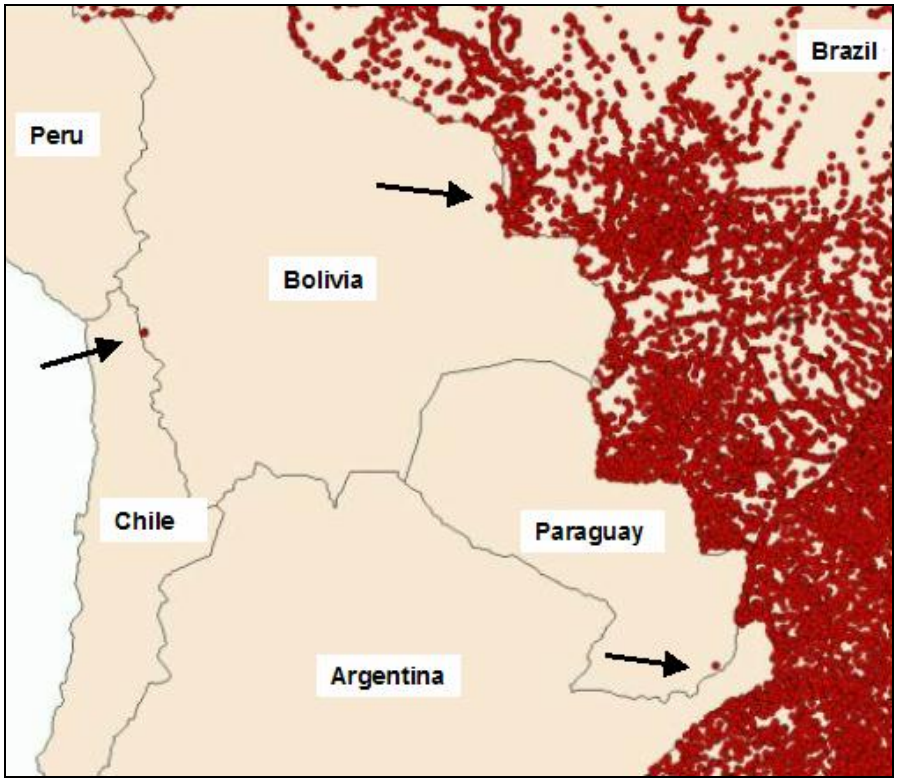

Fig. 15. *몇몇 오류(화살표)를 보여주는 브라질 지명에 관한 지명사전의 레코드. 분명한 오류 하나는 칠레-볼리비아 국경에 있고 다른 하나는 파라과이 남부에 있다.*

많은 관련 도구들, 예를 들어 Diva-GIS (Hijmans *et al.* 2005)와 CRIA 데이터 정제 도구 (CRIA 2005)는 이러한 오류 동정을 지원하는 기능을 가지고 있다.

GIS 는 특정 식물의 종류, 토양의 종류 또는 지질 등에서 벗어나는 레코드의 검색에 또한 사용될 수 있다. 일부 종은 특정한 지질적 유형과 높은 관련성이 있다 – 예를 들면, 석회석, 사암, 그리고 사문석(figure 16)이 있다. 여러분이 이러한 것들의 경계선을 알고 있다면, 이 경계 밖에 있는 임의의 레코드는 잠재적인 특이점으로 간주될 수 있고 추후 검사를 위해 표시해 둘 수 있을 것이다 (Chapman *et al.* 2001). Figure 16 에서, 고광물질의 사문석 토양에서만 발생하는 종이 맵핑되었고 두개의 레코드('a'와 'b'로 표시됨)는 오류일 가능성을 보여준다. 검사에서 레코드 'a'는 사문석이 표출된 장소에서 가장 가까운 마을인 'Goomeri' 장소만을 가지고 있고, 이 마을의 위도와 경도로 지리코드 되어 있다. 레코드 'b'는 이 표출된 장소에 아주 가까이 있지만 주어진 지리코드의 정밀성 때문에서 위치가 잘못 지정되었을 가능성이 있다.

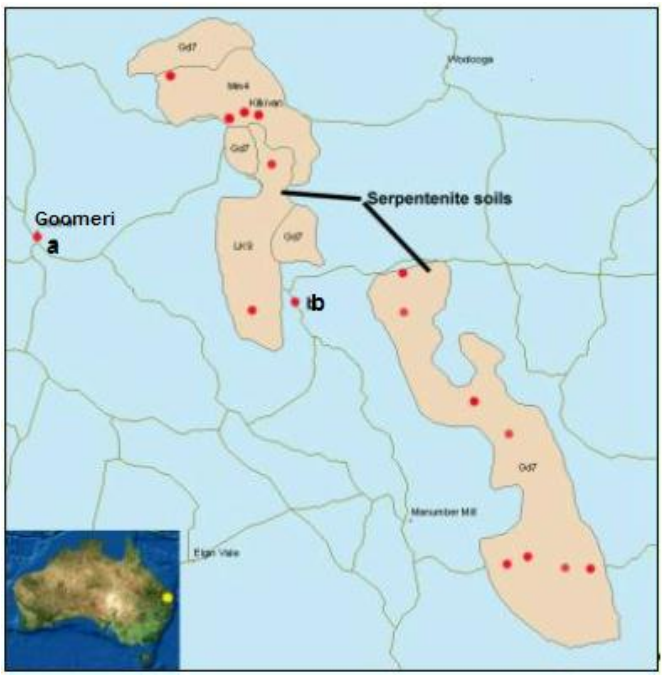

Fig. 16. *고광물질인 사문석 토양에서만 발견되는 종의 레코드(빨간색). 'a'와 'b' 표시된 레코드는 지리코딩에서 오류가 포함될 수 있을 가능성이 있다.*

수집가의 일정기를 파악하는 것(Chapman 1988, Peterson *et al.* 2003)은 잠재적인 오류를 검사할 수 있도록 하며, 예를 들어, 수집 일자가 해당 수집가의 특정 패턴에 맞지 않는 것이 있다. 이것은 특히 헬리콥터, 비행기 또는 자동차를 이용하여 하루동안 광활한 범위를 조사할 수 있었던 수집자들보다 이전 시대인 18 세기, 그리고 19 세기의 수집자들에 대해 유용할 수 있을 것이다. Figure 17 의 예에서, 2 월 22 일과 25 일 사이, 그리고 3 월 초에 수집된 수집물은 Pentland-Lolworth 지역(동그라미)에 있어야 하고, 그 지역 밖에 있으면, 수집 일자 또는 지리코드에 오류가 포함되어 있을 가능성이 있다 (Chapman 1988). 이전과 마찬가지로, GIS 를 이용해서 여행 일정기와 종의 레코드를 맵핑하는 것은 여기에서도 매우 유용할 수 있다. 다른 예는 네팔에서 애니메이션 기능이 있는 GIS 을 이용한 것으로 강을 따라서 수집자의 경로를 추적하였다 (Lampe and Reide 2002).

GIS 의 다른 이용은, 예를 들면, 그럴듯한 장소들의 버퍼링을 포함하는데 그 예로는 어류와 수생식물에 대한 하천, 연해 종에 대한 해안, 고산 종에 대한 고산 산맥 또는 독특하게 고산 지대에만 서식하는 것들이 있다. 이러한 방식으로, 이 버퍼 밖에 있는 임의의 것은 검사될 필요가 있을 것이다. 여기에서 주의를 기울일 필요가 있는데, 예를 들어, 어류의 경우 버퍼 영역 밖에 있는 레코드들이 전혀 오류가 아니고, 이 종은 작은 하천에 존재하지만, 그 개체 수를 지도에 맵핑하기에는 너무 작을 수도 있는 것이 있을 수 있기 때문이다. 이러한 검사는 일반적으로 단지 의심되는 레코드를 표시할 수만 있으며, 레코드 내에서 어느 것이 진짜 오류이고, 어느 것이 실제 특이점인지는 개별 검사를 해야 한다.

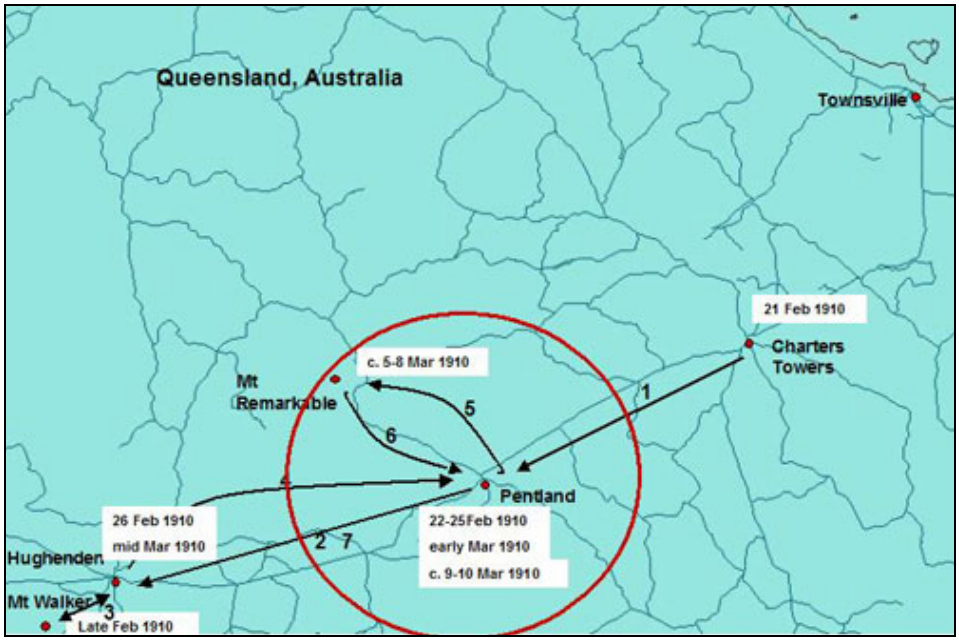

Fig. 17. *1910 년 Queensland 에서 Karl Domin 의 수집지역 (Chapman 1988).그는 Townsville 에서 Hughenden 으로 기차로 여행하였으며,도중에 Charters Towers 과 Pentland 에 멈추었다. 기차로 Hughenden 으로 돌아가기 전에, 그는 다시 돌아와서 말을 타고 Pentland, Mount Remarkable, Lolworth 지역에 10 일을 보냈다. 날짜는 단지 대략적인 것이다.*

iii. 지리 및 환경 공간의 특이점

데이터의 특이점을 탐지하는 많은 방법이 있고, 아래에서 이것들이 설명된다. 자연사 데이터는 매우 다양하고 일반적으로 표준 통계 분포를 따르지 않으며, 이에 따라 말레틱과 마르쿠스(Maletic and Marcus 2000)에서 제안된 것처럼, 대부분의 특이점을 갈무리하기 위해서는 하나 이상의 방법들이 종종 필요하다.

지리적인 특이점 탐지

브라질의 CRIA 에서 만든 프로그램(spOutlier)은 사용자가 인터넷 상의 상자에 레코드를 입력 또는 복사해서 붙여넣거나, 파일에 대한 링크를 하거나, 또는 표본 레코드에 대한 XML 파일을 전송하여 지리 특이점에 대한 정보를 받을 수 있도록 하고 있다. 이 레코드들은 "id, 위도, 경도, 고도" 의 형태로 송신되고, 프로그램은 텍스트 형태와 지도상의 인터페이스 두가지로 오류일 가능성이 있는 정보를 반환한다 (Marino *et al.* in prep). 이것은 또한 사용자가 자신들의 데이터 집합을 해변-위 (지상) 또는 해변-바깥(해상)의 것으로 동정할 수 있도록 하고 프로그램은 일치되지 않은 쌍의 리스트를 반환한다. 이것은 독창적인 프로그램으로 생물학자들에게 매우 유용할 것이다. 또한 사용자가 온라인으로 문서를 전송하여 잠재적인 오류에 대해 주석이 있는 정보를 받는 것이 가능하다. 온라인 버전은 http://splink.cria.org.br/tools/ 에서 볼 수 있다 (CRIA 2004b).

Figure 18 에서, 장소 리스트는 오류일 것 같은 네 개의 레코드를 반환하였고, 위도에서 1 개, 경도에서 3 개, 고도에서 1 개의 오류 가능성이 있다. 그 다음 이 점들은 연관된 지도상에 빨간색으로 표시되고 이것은 오류 가능성이 있는 레코드를 나타낸다.

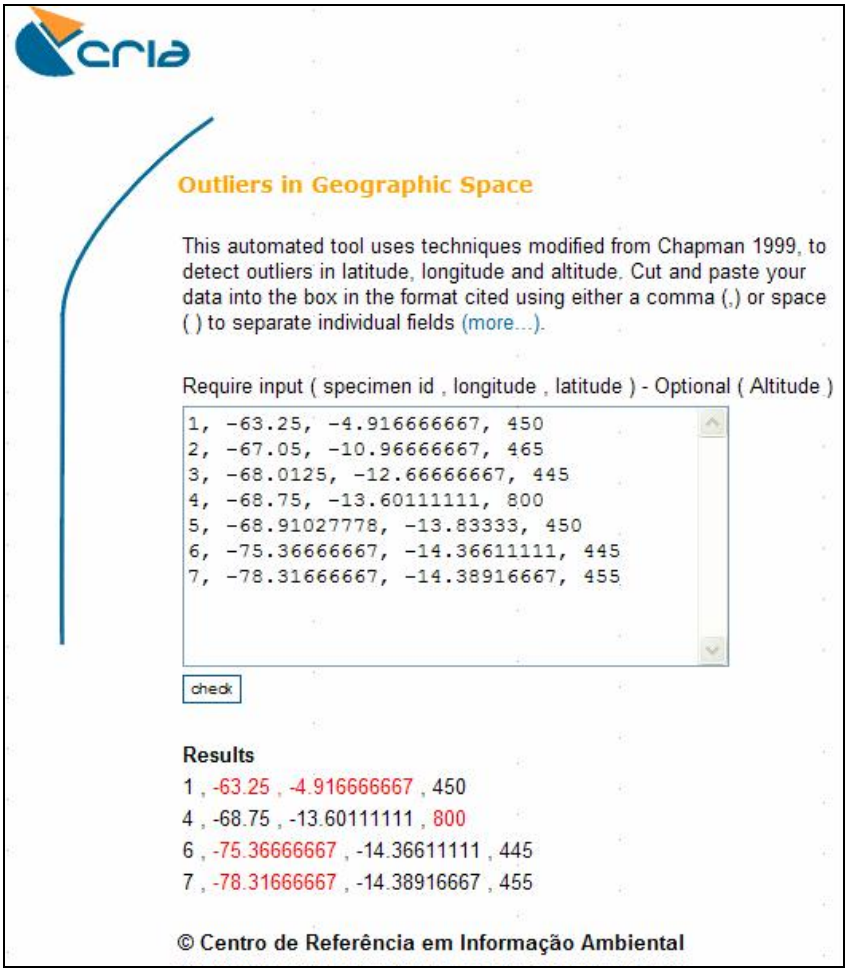

Fig. 18. *CRIA 의 지리공간상의 특이점 (Outliers in Geographic Space) 시스템 프로토타입 - 지리코딩에서 오류가 있을 것으로 예상되는 레코드 1, 4, 6, 그리고 7 를 식별하고 있다.*

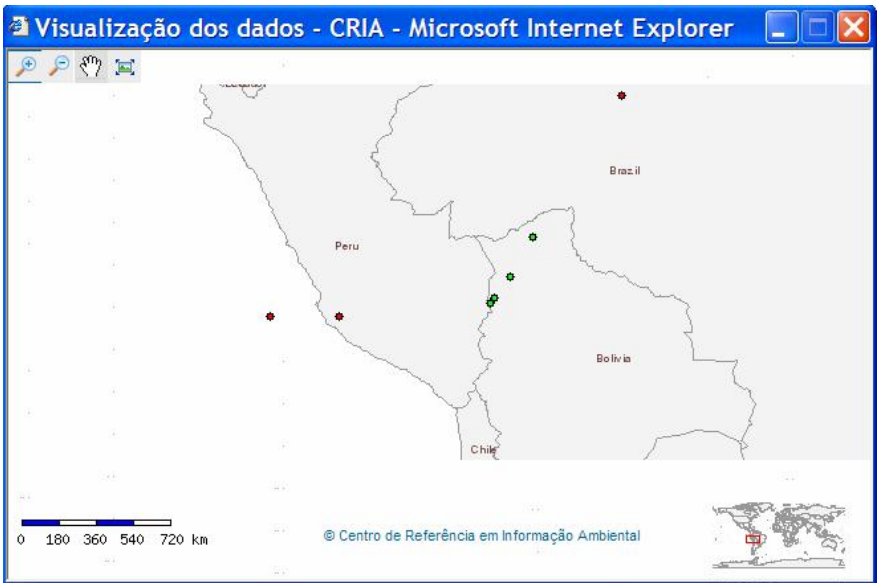

Fig. 19. *figure 18 에서 의심되는 레코드(빨간색)로 식별된 것을 보여주는 관련 지도 출력물*

이 방법을 이용하는 공개 프로그램:

- **spOutlier-CRIA** (CRIA 2004b, Marino *et al*. in prep).
- **Data Cleaning-CRIA** (CRIA 2005).
- **Diva-GIS** (Hijmans *et al.* 2005)

누적 빈도 곡선

BIOCLIM (Nix 1986, Busby 1991) 프로그램의 초기 버전은 해당 분류군에 대해 기후정보 데이터 중 어느 한 항목이 90% 범위를 벗어나는 레코드를 제외시키거나 또는 퍼센트 숫자를 변경할 수 있는 누적 빈도 곡선을 이용함으로써 (Busby 1991, Lindemeyer *et al.* 1991) 잠재적인 특이점을 탐지하는데 사용되었다. 비록 이러한 기법들이 여전히 사용되고 사용하기 쉽지만 (Houlder *et al.* 2000, Hijmans *et al.* 2005), 이것들은 오류가 전혀 없거나 많은 오류를 포함하는 분류군에 대해서는 작동하지 않는다. 이것들은 또한 매우 작은 건수의 샘플의 경우 의심스러운 결과를 보였다 (Chapman and Busby 1994, Chapman 1999).

Diva-GIS 소프트웨어의 최근 수정판(Hijmans *et al.* 2005)은 아래 논의되는 역잭나이핑 (Reverse Jackknifing) 방법론(Chapman 1999)을 포함하였고, 이것은 역잭나이프 (Reverse Jackknife) 방법하에서 동정된 레코드와 함께 누적빈도곡선에 연계되었으며, 이 레코드들은 각각의 인자에 대해 누적빈도곡선상에서 강조된다.

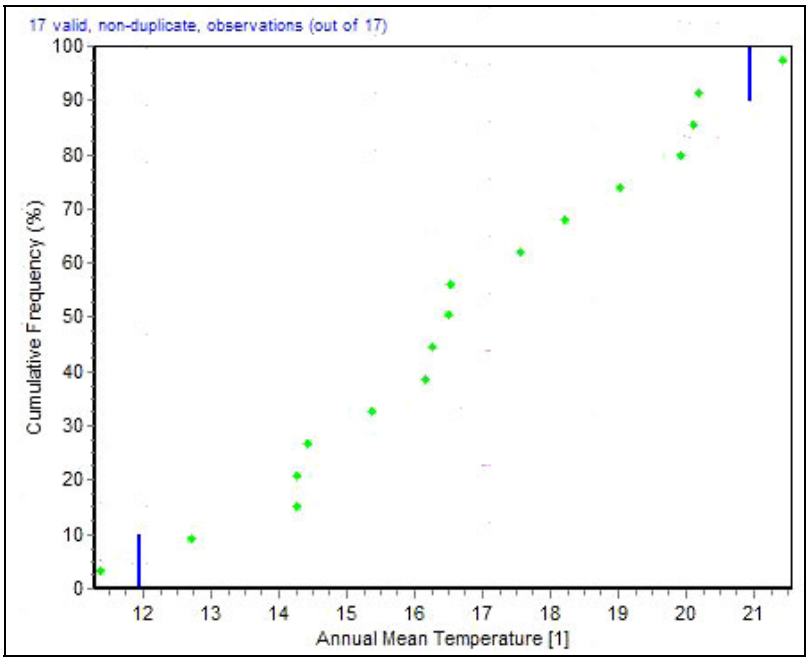

Fig. 20. *연평균 기온을 사용하는 기후 공간에서 특이점을 탐지하는데 이용된 누적빈도곡선. 파란선은 97.5%를 나타내고 왼쪽 하단의 점 (또는 왼쪽 하단의 두개의 점)은 잠재적인 특이점으로 간주될 수 있어 지리코드 오류에 대하여 검사할 가치가 있을 것이다.*

이 방법을 사용하는 공개 프로그램들:

- **Diva-GIS** (Hijmans *et al.* 2005)
- **ANUCLIM** (Houlder *et al.* 2000).

중요 구성요소 분석

한 기후 계층에 대한 다른 기후 계층의 중요구성요소분석(Principal Components Analysis)에서 점들의 흩어짐을 이용함으로써 사용자는 잠재적인 특이점을 동정할 수 있고 따라서 지리코딩상의 잠재적인 오류를 동정할 수 있다. 이것은 꽤 강력한 데이터 검증 방법이지만, 어떤 식으로든 다중 특이점 레코드를 동정하는 과정이 자동화되어 있지 않으면, 이 방법은 아주 단조로운 작업이 될 수 있는데, 그 이유는 사용자가 사용되고 있는 기후 구성요소의 많은 조합을 조작하고 다루어야 하기 때문이다.

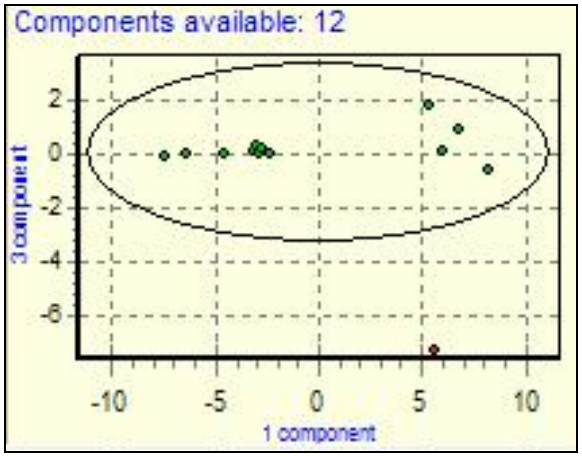

Fig. 21. *특이점으로 동정된 (따라서 잠재적인 오류일 수 있는) 한 점(빨간색)을 보여주고 있는 주요구성요소분석 (FloraMap, Jones and Gladkov 2001).*

이 방법을 사용하는 공개 프로그램:

- **FloraMap** (Jones and Gladkov 2001)
- **PATN** vers. 3.01 (Belbin 2004)

클러스터 분석

유클리드 또는 다른 거리 측정법에 기반한 클러스터링을 이용하여 특이점을 동정하는 것은 필드 수준에서의 방법들로 식별되지 않는 특이점을 때때로 동정할 수 있다 (Johnson and Wichern 1998, Maletic and Marcus 2000). 클러스터 분석은 유사한 군집의 다중 그룹을 식별하는 것을 돕는데 사용될 수 있고 (기후 공간 또는 다른 어떤 기준을 사용), 다른 클러스터와 현저한 거리로 분리되어 있어 유니케이트(unicates) 또는 소그룹으로 격리된 클러스터를 파악하는데도 또한 사용될 수 있다. 이것은 매우 유용하고 견고한 방법론처럼 보이지만, 사용되는 클러스터 방법에 매우 크게 의존적이고 계산적으로 복잡할 수 있다.

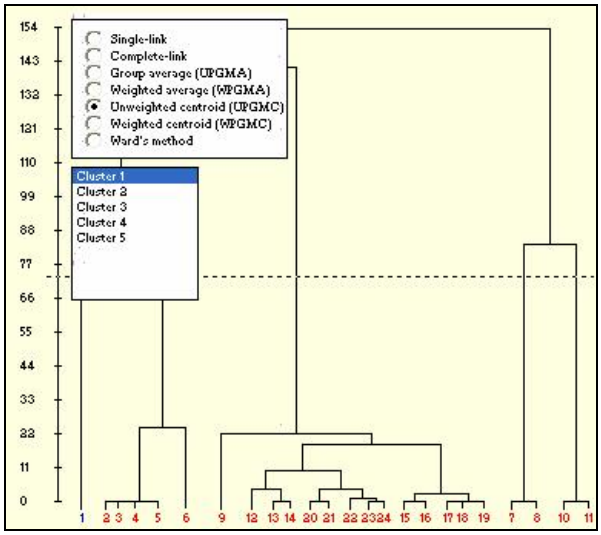

Fig. 22. *특이점으로 간주될 수 있는 유니케이트 클러스터(unicate cluster)를 보여주는 클러스터 분석 (FloraMap 자료, Jones and Gladkov 2001).*

이 방법을 이용하는 공개 프로그램:

- **FloraMap** (Jones and Gladkov 2001)
- **PATN** Vers. 3.01 (Belbin 2004)

Climatic Envelope

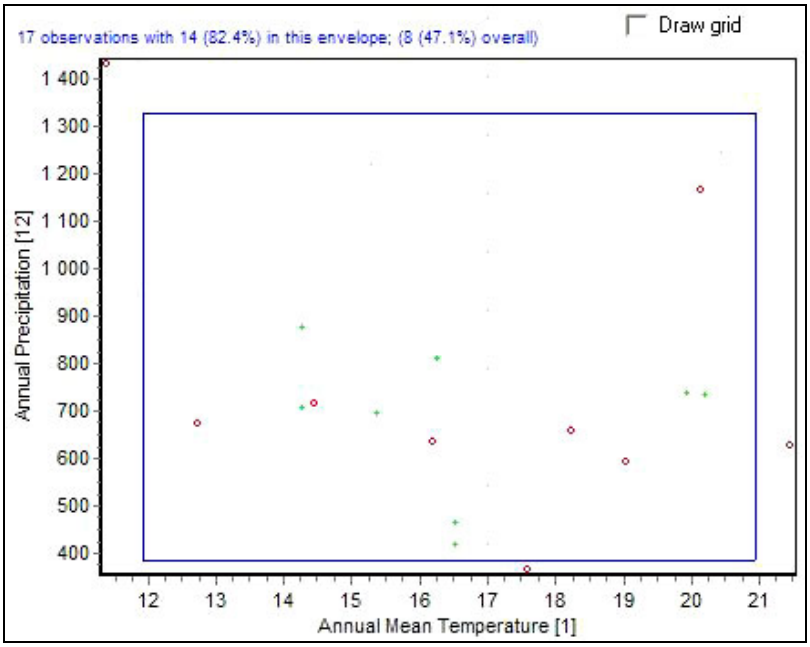

Fig. 23. *연평균 기온과 연평균 강수량에 대해 97.5% 봉투를 사용하는 BIOCLIM 의 기후 봉투 (climatic envelope). 빨간색으로 표시된 레코드는 64 개의 가능한 봉투중의 어느 하나에도 속하지 않는 것들이다.*

기후봉투(Climatic Envelope) 방식은 위에서 언급된 누적빈도곡선 방법론을 확장한 것이지만, 주요구성요소분석과 유사하게 한번에 이차원씩 조사될 수 있도록 기후 계층의 각각을

다차원 박스 또는 봉투로 그룹화한다. 기후 계층의 전체성을 구성하는 임의의 누적빈도곡선에 있는 특이점들은 이 방식으로 식별될 수 있다.

이 방법을 이용하는 공개 프로그램:

- **Diva-GIS** (Hijmans *et al.* 2005)

Reverse Jackknife

이 기법은 많은 기후 정보자료(profiles)중의 임의의 하나에서 한 배열의 점들중 양끝에 있는 특이점을 추출하기 위해 수정된 역잭나이프 기법을 사용한다. 이 방법은 기후 공간에서 자동으로 특이점을 탐지하기 위해 1992 년 호주 ERIN 에서 개발되었고 (Chapman 1992, 1999, Chapman and Busby 1994), 당시에 모델링되고 있었던 수 천 개의 종 가운데 의심이 가는 데이터를 탐지하였다. 이 방법은 의심되는 레코드를 자동적으로 식별하는 측면에서 신뢰성이 매우 높은 것으로 증명되었으며, 의심되는 것으로 파악된 것 가운데 많은 것(약 90%)이 실제 오류로 증명되었다.

$$x < \bar{x}$$

if

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(\bar{x} - x_{(i)}\right)$$

else

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(x_{(i+1)} - \bar{x}\right)$$

then

$$C = \frac{y_{(i)}}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_{(i)} - \bar{y}\right)^2}{n-1}}}$$

Fig. 24. *특이점 탐지 알고리듬에서 임계값 (Critical Value, C) 결정을 위한 공식들, 여기에서 C 는 임계값 (Chapman 1999 자료). 이 공식은 1992 년부터 호주에서 환경(기후) 공간에서 특이점을 탐지하는데 이용되었다. 이 공식은 최근 (2005) C 의 값을 'x'의 범위로 나누는 것으로 수정되었고 Diva-GIS 버전 5.0 에 적용되었다 (Hijmans et al. 2005). 이것은 강수량, 고도 등과 같이 큰 값을 기준으로 하는 것에 사용될 수 있도록 신뢰도가 향상되었다.*

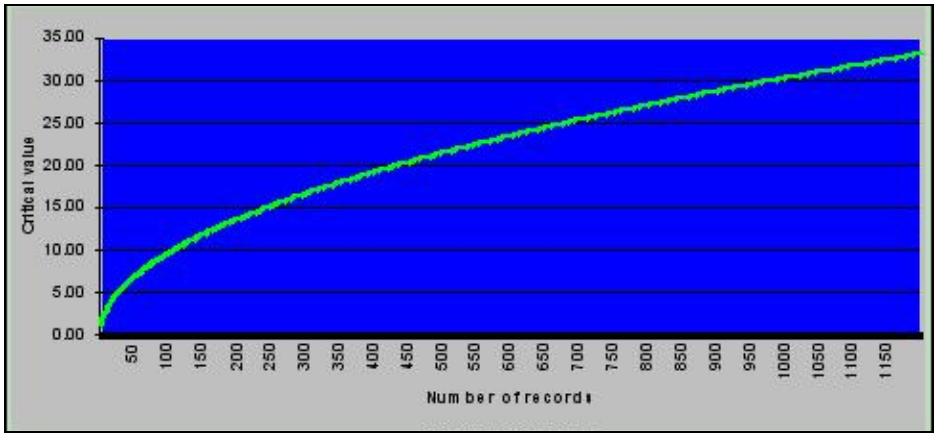

Fig. 25. *임계값 곡선 (T=0.95(√n)+0.2, 여기에서 n 은 레코드의 개수). 곡선보다 높은 값은 '의심되는 것'으로 간주되고, 곡선보다 낮은 값은 '유효한 것'으로 간주된다 (Chapman 1999 의 그림).*

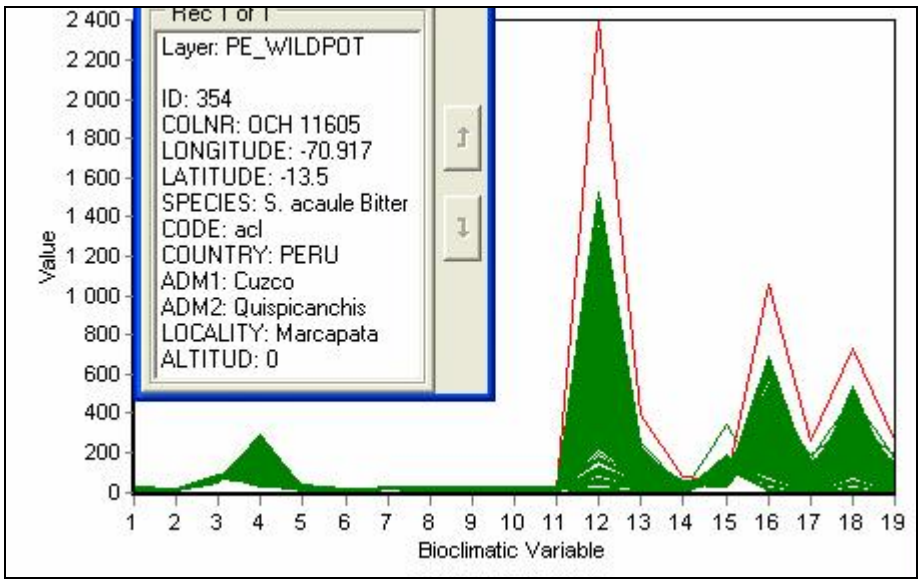

Fig. 26. *Diva-GIS 에서 역잭나이프 기법을 사용하는 특이점탐지 알고리듬. 이 프로그램은 잠재적인 하나의 특이점을 식별하였다 (19 개의 가능한 기준 가운데 적어도 6 개 이상의 기준에서 특이점인 레코드만을 보이기 위해 선택 옵션을 사용).*

이 방법을 이용하는 공개 프로그램:

- **Diva-GIS** (Hijmans *et al.* 2005)
- 또한 2006 년 중반에 이용 가능하게 될 새로운 BioGeomancer 툴킷 안으로 프로그램 되고 있음

인자 극단법

인자극단법(Parameter Extremes)은 기후봉투 방법과 비슷한 방법으로 각 누적빈도곡선의 극단에 있는 레코드들을 식별해서 이것들을 출력 로그 파일에 묶어 출력한다. 이 방식으로 사용자는 하나 이상의 기후 인자에서 극단에 있는 특정 레코드를 동정할 수 있다.

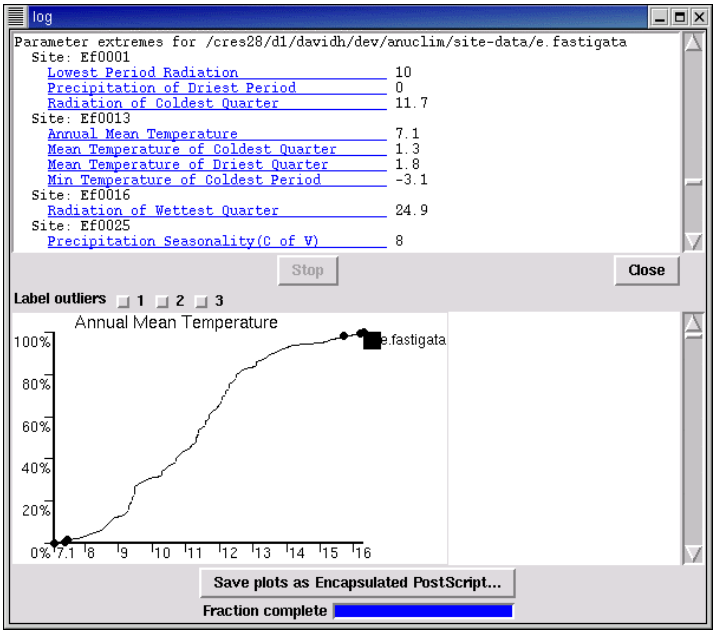

Fig 27. *인자 극단(위) 법과 이것과 관련된 종 누적 곡선(아래)을 보여주는 ANUCLIM 버전 5.1 (Houlder et al. 2000)의 Eucalyptus fastigata 로그 파일.*

이 방법을 이용하는 공개 프로그램:

- **ANUCLIM** (Houlder *et al.* 2000).

그 외 다른 방법들

아래 나열되는 많은 방법론은 간단하며 많은 표준 통계 패키지에서 이용 가능하다. 이것 중의 일부는 생물학적 데이터의 오류를 탐지하는데 이용되지 않았던 것처럼 보이지만, 비슷한 종류의 데이터를 이용한 사례들로 볼 때 이것들은 시도해 볼만할 것이다. 이러한 방법 및 다른 많은 방법들이 레젠드레와 레젠드레(Legendre and Legendre (1998))에 자세히 설명되어 있다. 시도해 볼만한 다른 많은 탐지 방법들을 바넷과 루이스(Barnett and Lewis (1994))의 출판물에서 찾을 수 있다.

i. 평균으로부터의 표준 편차

아마도 이러한 다른 방법들 중에서 가장 가능성이 있는 것은 체비셰프(Chebyshev) 이론에 기반한 평균으로부터 다양한 많은 표준 편차를 조사하는 것일 것이다 (Barnett and Lewis 1994). 말레틱과 마르쿠스(Maletic and Marcus) (2000)는 같은 종류(날짜)의 78 개 필드를 가진 해군 인사 레코드 5000 개를 이용하여 평균으로부터 많은 수의 편차를 테스트하였고, 표준 편차의 5 배에 해당되는 값을 이용하는 것이 가장 좋은 결과를 산출한다는 것을 발견하였다. 많은 수집물 데이터집합에 대해 테스트를 수행하는 것이 필요할 것이고, 특히 말레틱과 마르쿠스가 사용했던 것보다 훨씬 적은 수의 레코드로 테스트하는 것이 필요할 것이다. 필자 자신이 고도를 사용하여 적은 수의 레코드로 실험한 예비 테스트는 지금까지 좋은 결과를 보이지 않았다.

ii. 중앙값으로부터의 편차

매개 변수를 이용하지 않는 또 다른 통계 테스트 그룹은 평균보다는 중앙값과의 관계를 이용한다. 두개의 가능한 방법은 맨-휘트니(Mann-Whitney) U 테스트와 쿠스콜-왈리스 (Kuskall-Wallis) 테스트가 있으며 이것은 두 개 (Mann-Whitney), 또는 세 개 또는 그 이상 (Kuskall-Wallis)의 집단이 단지 중앙값만 다르다는 교호 가설(alternate hypothesis)을 조사한다 (Barnett and Lewis 1994, Lowry 2005). 필자는 종-발생 데이터의 특이점 탐지에 적용된 이러한 사례를 본 적이 없지만, 이것은 테스트해 볼 가치가 있을 것이다.

iii. 모델링된 분포 이용

GARP (Stockwell and Peters 1999, Pereira 2002) 또는 Lifemapper(University of Kansas 2003b)를 이용하여 생성된 것들과 같은 종 분포 모델링에 기반한 분포 모델은 예측된 분포 밖에 있는 새로운 레코드를 식별하는데 이용될 수 있을 것이다. 이 방법은 쓰기 쉬운 방법이긴 하지만 예측된 분포의 품질에 따라 제한된다. 종에 대한 모든 레코드가 이 모델을 개발하는데 사용되지 않았다면, 이 모델에는 결함이 있을 수 있다. 또한, 단지 해당 분포의 외부 경계수치만을 이용하는 것은 지리 분포의 광범위한 전체성 내에서 적합한 공간을 식별하는 좋은 모델들의 흩어진 성질을 고려하지 않는 것이다.

iv. 패턴 분석

패턴 분석은 데이터 중에서 기존 패턴에 맞지 않는 레코드를 식별하는데 이용될 수 있다. 다양한 방식이 패턴의 분석에 이용될 수 있으며, 이것은 연관 (Association), 분할 (Partitioning), 분류 (Classification), 군집 (Clustering), 순서화 (Ordination), 그리고 최소스패팅트리 (minimum spanning tree)처럼 네트워크(Networks)을 이용하는 방법이 있다 (Belbin 2004). 이러한 방법들 중 일부는 위에서 상세하게 논의되었다. 일반적으로 패턴은 비슷한 특징을 갖는 레코드의 집단으로 정의될 수 있지만 (Maletic and Marcus 2000), “알맞은 참조 패턴을 선택하는 것”은, 만약 이러한 것이 존재한다면, 그 결과에 영향을 미칠 수 있다 (Weiher and Keddy 1999).

이 방법을 이용하는 공개 프로그램:

- **PATN** (Belbin 2004)

v. 순서연관규칙 (Oridinal Assoication Rules)

연관 규칙은 많은 비율의 레코드에서 적용되는 경향을 보이는 순서적인 관계를 찾으려고 시도한다 (Marcus *et al.* 2001). 이것들은 범주 데이터와 수량 데이터 둘 모두에 사용될 수 있다. 간단히 설명하면, 이것들은 대부분의 경우 A<B 인데, 한 레코드에서 A>B 이면, 오류가 있을 것이라는 패턴을 조사한다. 수량 데이터의 경우, 이 규칙은 특이점 탐지 목적으로 평균, 중앙값, 표준 편차 그리고 백분위수 범위를 사용하는 다른 통계적인 방법과 함께 사용될 수 있다. 이 방법에서 레코드의 수가 많으면 많을수록 그 결과의 신뢰성이 높아질 수 있고, 대개의 경우 단지 하나의 종보다는 데이터베이스 전체에 대해서 적용될 수 있을 것이다. 예를 들어, 종 A 가 대부분의 경우 기준(모식) 식물 B 에서 발생한다면, 기준(모식) 식물 C 에서 발생하는 정보를 가진 레코드는 오류일 가능성이 있다. 다른 예로 어느 한 수집자가 수집한 모든 레코드의 수집날짜는 수집자의 출생일에서 15 년 이내 또는 100 년 이상이 되지 않아야 하며, 또는 사후가 되지 않아야 한다. 이와 같은 규칙은 수집자의 예상 활동범위와 함께 또한 사용될 수 있을 것이다 (위 참고). 예를 들어, 수집물이 1900 년 이전의 것이라면, 같은 날에 수집된 두 개의 수집물은 x 킬로미터보다 먼 거리를 떨어져 있지 않아야 한다.

이 방법을 쓰는 공개 프로그램:

- **PATN** (Belbin 2004).

서술 데이터

꽤 다양한 성질의 것들이 이러한 데이터베이스에 포함될 수 있기 때문에 서술 데이터 (Descriptive Data)에 대한 오류 검사를 여기에서 다루기가 더욱 어렵다. 하지만 이러한 데이터베이스의 구조화된 성질로 인해 데이터베이스가 설정될 때 여러 가지 규칙을 설정하는 것이 가능하다.

i. 데이터베이스 설계

서술 데이터베이스와 관련하여 높은 데이터 품질을 유지하는 비결은 올바른 설계 절차를 따르는 것이고, 가능하다면 DELTA (Dallwitz *et al.* 1993) 또는 분류학데이터베이스연구그룹 (Taxonomic Databases Working Group, TDWG)에서 개발되고 있는 새로운 SDD (Structure of Descriptive Data) (http://160.45.63.11/Projects/TDWG-SDD/) 표준과 같은 것들에 맞추어 데이터베이스를 설계하는 것이다.

ii. 편집 컨트롤

서술 데이터베이스의 구조화된 성질로 인해, 편집 컨트롤을 사용해야 한다. 예를 들어, 대부분의 서술 데이터 필드는 다양한 내부 제약사항을 가지고 있고, 종종 잘 개발된 문자 집합들이 있어 이것에서 입력 사항을 선택한다. 하지만 오류는 여전히 발생할 수 있고, 특히 단위가 혼동될 수 있는 연속데이터의 경우에는 그러하다 (예, 밀리미터와 센티미터). 사용되는 단위는 기록되어야 하고, SDD 표준에서 권고하는 것처럼 분리된 필드에 저장되는 것이 좋다. 또한 하나의 데이터베이스 내에서 단위의 표준화가 가능할 경우 수행되어야 한다 – 즉, 작업 전반에 걸쳐 mm 또는 cm 등의 사용을 합의하여야 하며, 그렇지 않고 이것을 섞어서 사용하면 특히 다수의 입력자가 데이터 입력을 할 경우 오류를 발생시킬 수 있다. 이러한 필드에 대해 평균 또는 중앙값 등에서 표준 편차를 사용하여 특이점을 조사하면서 극단값을 조사(예, 위의 *공간 데이터* 에서 서술했던 누적빈도곡선을 사용함으로써)하는 테스트를 수행할 수 있다. 종종, 이 결과를 그래프에 나타냄으로써 사용자는 오류일 것으로 예상되는 레코드를 또한 찾을 수 있다. 오류를 찾기 위해 이용될 수 있는 일부 다른 오류의 종류는 다음과 같다 (English 1999 자료):

- ***입력되지 않은 데이터***
  값이 있어야 하지만 비어있는 필드를 검색. 어느 한 필드에 "null" 또는 값이 입력되지 않아야 할 필요가 있을 경우, 분리된 필드에 null 값에 대한 근거를 기록하는 것이 좋은 실행사례이다 – 예를 들어, "관련 없음, 측정 되지 않음 또는 알려지지 않음".
- ***부정확한 데이터 값***
  이것은 인쇄상의 오류, 자판 입력시 글자의 뒤바뀜, 잘못된 위치에 입력된 데이터 (즉, 숫자 필드에 입력된 알파벳과 숫자 문자) 그리고 값이 요구되는 필드에 강제로 넣은 데이터 값(데이터 입력자가 관련 값을 몰라서 임의의 값을 추가한 경우)을 찾는 것과 관련 있다. 더미 값(dummy value)은 통계적인 방법에서 비어있는 필드 또는 0 값이 허락되지 않는 경우 이것을 "속이기" 위해 필드에 때때로 추가된다. 이것은 주의해서 사용되어야 한다.
- ***세분화되지 않은 데이터 값***
  하나 이상의 사실이 입력된 필드를 검색.
- ***도메인 불일치***
  의도되지 않았던 목적으로 사용되는 필드를 검색.
- ***중복 발생***

동일한 실제 값을 가리킬 수도 있는 값을 검색. 이것은 서로 다른 용어를 사용했던 두개의 데이터베이스를 통합할 때 꽤 자주 발생한다.

- *일관성이 없는 데이터 값*
  두개의 관련된 데이터베이스가 동일한 목록의 값을 사용하지 않을 경우 발생하고 통합될 때 일관성이 없음을 나타낸다. 이것은 위에서 언급한 SDD 표준과 같은 변환 표준이 역할을 할 수 있는 부분이다.

오류의 문서화

관련 문서인 *데이터 품질의 원칙*(Chapman 2005a)에서 언급되었듯이, 오류의 문서화와 오류 검사는 데이터 품질을 유지하고 오류 검사의 중복을 예방하는데 필수적이다. 올바른 문서 없이, 사용자는 이용에 대한 데이터의 적합성을 결정할 수 없을 것이다.

데이터 품질 검사가 수행되고, 수정사항이 반영되더라도, 이러한 것이 온전히 문서화되지 않으면, 어느 누구에게도 그 소용이 거의 없게 된다 (Chapman 2005a). 인지된 오류가 전혀 오류가 아닐 수도 있고, 반영된 변경사항이 새로운 오류를 더할 수 있는 가능성이 항상 있으므로, 데이터 수정에 대한 감시 추적(audit trail)이 유지될 필요가 있다. 감시추적이 없다면, 이러한 "수정사항"을 다시 되돌리는 것은 가능하지 않을 것이다. 이것은 해당 데이터를 처음 수집한 사람 이외의 다른 사람이 이러한 오류를 수정하는 경우에 특히 중요하다 (Chapman 2005a).

감시 추적을 개발하는 몇 가지 방법이 있다 (즉, 시간의 경과에 따라 데이터베이스에 반영된 수정사항을 기록하고 어떠한 데이터 품질 관리 검사가 수행되었고 언제 되었는지를 기록하는 것). 감시 추적은 중요하며 그 이유는 오류가 복구될 수 있도록 하고, 큐레이터와 데이터 관리자가 이미 수행되었던 검사를 수행하지 않도록 하며, 그리고 데이터의 변경사항 및 추가사항이 법률적 또는 다른 목적(예를 들어, 데이터를 사용했었을 사용자에게 데이터를 마지막으로 접근한 이후로 데이터의 변경된 사항들을 알 수 있도록 공지하여 주는 것)으로 문서화될 수 있도록 하기 때문이다. 감시추적 절차를 만드는 한가지 방법은 연속된 타임스탬프를 추가하는 시간 데이터베이스 응용 프로그램을 통해서 가능하며, 예를 들어, 트랜잭션 타임스탬프 기간 동안 하나의 사실이 데이터베이스에 저장되어야 한다 (Wikepedia[6]). 또 다른 방법은 변경된 레코드의 데이터 또는 변경이 일어난 데이터의 일부를 정기적으로 XML 데이터로 추출하는 것이다.

Chapman (2005a)에서 언급되었듯이:

> *"이용자가 데이터에 대해 데이터 품질 검사를 수행하면 의심되는 많은 레코드를 발견할 수 있다. 이러한 레코드들은 검사 후에 아주 좋은 레코드로 판명되거나 완전한 특이점으로 판명될 수 있다. 이 정보가 해당 레코드에 문서화되지 않으면, 시간이 지난 후에, 다른 사람이 동일한 레코드를 의심되는 것으로 다시 식별하는 데이터 품질 검사를 수행할 수도 있다. "*

관련 문서인 *데이터 품질의 원칙*(Chapman 2005a)에서 또한 언급되었듯이:

> *오류가 충분히 문서화되도록 확실히 하는 방법중의 하나는 데이터베이스 설계 및 구축의 초기 계획 단계에서 이것을 포함하는 것이다. 이 때 추가적인 데이터 품질/정확성 필드가 포함될 수 있다. 지리코드의 정확성, 지리코드와 고도에 대한 정보의 출처와 같은 필드, 이 정보를 추가한 사람에 대한 필드 – 지리코드는 GPS 를 사용한 수집자에 의해 추가되었는가 또는 나중에 특정 축척의 지도를 사용한 데이터 입력자에 의해 추가되었는가, 고도는 DEM 에서 자동으로 생성되었는가, 그렇다면 이 DEM, 이것의 날짜와 축척의 원천은 무엇인가, 등. 이러한 모든 정보는 나중에 이 정보가 특정한 사용에 가치가 있는지 또는 없는지를 결정할 때 중요할 것이고, 그 다음으로 이 데이터의 사용자가 이러한 것을 결정할 것이다.*

[6] http://en.wikipedia.org/wiki/Temporal_database

또한 데이터 검증에 관한 필드(수행된 검증 검사와 관련된 “누가, 언제, 어떻게 그리고 무엇을”)는 데이터베이스에 대해 수행된 검증, 오류 검사 그리고 데이터 정제를 추적하고 감시하기 위해 데이터베이스에 추가되어야 한다. 이상적인 경우, 이러한 것은 위에서 제안된 것처럼 레코드 수준에서 또한 추가되어야 할 것이다.

오류의 가시화

1 차 종 데이터에 대해 오류를 가시화하는 좋은 방법들을 개발하기 위해 해야 할 일이 많이 있다. 가시화에 요구되는 두 가지 사항은 다음과 같다

- 오류 검사와 정제에 대한 가시화;
- 표현(presentation)을 위한 가시화.

이 중 두 번째 것 – 표현을 위한 가시화는 관련 문서인 *데이터 품질의 원칙* (2005a)에서 다루어졌다 (Chapman 2005a).

GIS 는 공간 오류를 가시화하는 목적으로 사용되는 가장 흔한 검사 방법이다. 단지 1 차 종 데이터를 맵핑하고 이것을 지형적 계층에 겹치게 하는 것 만으로도 오류를 탐지하는데 도움이 될 수 있다. GIS 시스템은 온라인 맵핑과 정보 표현을 위해 가장 많이 사용되는 간단한 온라인 시스템부터, 간단한 기능에서 시작하여 매우 복잡한 기능이 있는 독립형 시스템까지 다양하다.

많은 기관들이 맵핑에 이미 GIS 를 사용하고 있으며, 이러한 것은 오류 검사에 사용될 수 있도록 쉽게 바꿀 수 있다. 하지만 다른 기관들은 일상적으로 GIS 를 사용하지 않으면서 자신들의 예산을 초과하는 GIS 시스템 구입을 고려하고 있으나, 배우기 쉽고 사용하기 간단한 무료 GIS 프로그램들을 구할 수 있으며, 이것들은 소규모 수집물 기관의 대부분의 요구사항을 적합하게 수행할 수 있다. 적어도 이러한 것들 가운데 하나인 – Diva-GIS (Hijmans *et al.* 2005)는 소규모 박물관과 식물표본관에서 사용할 수 있도록 특수하게 설계되었고 모델링 및 가시화 알고리듬과 함께 이 문서에서 서술한 몇 가지 오류 탐지 방법들을 포함하고 있다.

비-공간 데이터의 경우, 스프레드시트와 그래프를 이용해서 오류를 가장 잘 가시화 할 수 있다. 간단한 값의 그래프는 패턴에 맞지 않는 레코드를 빠르게 식별할 것이다. 간단한 그래프는 설정하기 쉬우며 데이터베이스에서 가져온 데이터를 표준 오류 검사 방법으로 쉽게 처리할 수 있다.

공간 커뮤니티 분야에서 예상되는 오류의 범위와 중요성에 대한 예상치를 만들기 위해 몬테카를로분석(Monte Carlo Analysis) 같은 기법을 이용하는 경향이 증가하고 있다 (Flowerdew 1991). 몬테카를로 분석은 가시화에 적합하며 사용자에게 오류를 전달하는데 좋은 방법이다. 몬테카를로 방법을 포함하는 일부 소프트웨어는 매우 비싸졌지만 (예, 윈도우용 Canoco 4.5[7]와 S-Plus[8]), 무료 대체 프로그램들, 예를 들어 마이크로소프트 엑셀에 추가할 수 있는 PopTools (Hood 2005)가 실제 존재한다.

정확성을 가시화하기

위의 *지리참조연산(Georeferencing)*에서 언급된 것처럼, 1 차 표본 레코드의 점 레코드는 실제적으로는 지점이 아니고, 이것관 연관한 오류 숫자를 포함하고 있다. 해당 지점을 연관된 정확성과 함께 맵핑함으로써, 해당 수집물이 실제적으로 의미하는 것과 실제 세계에서 이것의 관계에 대한 올바른 이해, 즉 *"궤적(footprint)"*이 가시화될 수 있다.

이것은 1 차 종 데이터와 관련하여 시급하게 연구가 필요한 분야 중의 하나이다 – 불확실성을 가시화하고 정확성의 궤적을 가시화하는 기법의 개발. 위도와 경도의 점으로서 표현되는 수집물 레코드 대신에, 레코드와 연관된 정확성을 포함할 필요가 있으며, 이후

[7] http://www.microcomputerpower.com/
[8] http://www.insightful.com/products/splus/default.asp

해당 위치를 이것의 궤적(원, 타원, 다각형 또는 심지어 그리드)으로 표현할 수 있을 것이다. 버퍼링 같은 GIS 기법은 강 또는 도로를 따르는 것과 같은 궤적을 개발하는데 훌륭한 도구를 제공한다. Biogeomancer 프로그램은 이러한 일부 측면을 조사하고 있지만, 가까운 시일 내에 완전한 기능이 작동되는 시스템을 개발하기는 어려울 것 같다.

인용된 도구들

1. 소프트웨어 자원

ANUCLIM

설명: 최신 버전의 BIOCLIM 을 포함하고 있고, 여러 가지 프로그램으로 구성된 생물기후 모델링 패키지. 이 프로그램에는 입력되는 표본 데이터의 오류를 찾을 수 있는 많은 방법들이 포함되어 있다.
버전: 5.1 (2004).
소유기관: 호주, 캔버라, 호주국립대학교, 자원환경연구센터(Centre for Resource and Environmental Studies, CRES).
가격: $AUD1000.
참고문헌: Houlder et al. 2000.
다운로드: http://cres.anu.edu.au/outputs/software.php

BioLink

설명: 명명법, 분포, 분류, 생태, 형태, 삽화, 멀티미디어, 그리고 문헌과 같은 분류군-기반 정보를 관리하기 위해 설계된 소프트웨어 패키지.
버전: 2.1 (2005).
소유기관 호주, 캔버라, CSIRO, 호주국립곤충수집물(Australian National Insect Collection).
가격: 무료.
참고문헌: Shattuck and Fitzsimmons 2000.
다운로드: http://www.biolink.csiro.au/.

BIOTA

설명: 생물다양성과 수집물 데이터에 대한 생물다양성데이터관리 시스템. 이것의 사용하기 쉬운 그래픽 인터페이스로 관계형 데이터베이스의 장점을 편리하게 이용할 수 있다.
버전: 2.03 (2004).
소유자: 미국, 코네티컷, Robert K. Colwell.
가격: 데모 버전 무료: 정식 버전: $US200-600.
참고문헌: Collwell 2002.
다운로드: http://viceroy.eeb.uconn.edu/Biota2Pages/biota2_download.html

Biótica

설명: 큐레이션, 명명, 지리, 서지, 그리고 생태학적 데이터를 처리하고 갈무리와 갱신을 지원하기 위해 설계됨.
버전: 4.0 (2003).
소유기관 멕시코, 멕시코 시, 코나바이오(CONABIO).
가격: $US290.
참고문헌: Conabio 2002.
다운로드: http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/distribu_v4.0.html.

BRAHMS

설명: 식물학 연구와 수집물 관리를 위한 데이터베이스 소프트웨어. 이것은 이름 관리, 수집물 큐레이션 그리고 분류학 연구를 지원한다.

버전: 5.58 (2005).
소유기관 영국, 옥스포드, 옥스포드 대학교.
가격: 무료.
참고문헌: University of Oxford 2004.
다운로드: http://storage.plants.ox.ac.uk/brahms/defaultNS.html.

Desktop GARP

설명: 야생 종 분포의 예측 및 분석을 위한 소프트웨어 패키지
버전: 1.1.3 (2004)
소유기관 미국, 캔사스, 로렌스, 캔사스 대학교 그리고 브라질, 캠피나스, CRIA (Centro de Referência em Informação Ambiental).
가격: 무료.
참고문헌: Pereira 2002.
다운로드: http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1.

Diva-GIS

설명: 생물다양성 데이터의 분석을 위해 개발된 지리 정보 시스템. 이것은 몇가지 간단한 모델링 도구와 많은 데이터 품질 검사 알고리듬을 포함한다.
버전: 5.0 (2005).
소유자: 버클리, 캘리포니아 대학교, R.J. Hijmans *et al.*
가격: 무료
참고문헌: Hijmans *et al.* 2005
다운로드: http://www.diva-gis.org

eGaz

설명: 박물관과 식물표본관이 자신들의 표본 레코드를 동정하고 지리코드의 추가를 지원하기 위해 개발된 프로그램.
소유기관 호주, 캔바라, CSIRO, 호주국립곤충수집물.
가격: 무료
참고문헌: Shattuck 1997.
다운로드: http://www.biolink.csiro.au/egaz.html

FloraMap

설명: 야생에 존재하는 식물과 다른 개체의 분포를 예측하기 위한 소프트웨어 도구.
버전: 1.02 (2003).
소유기관 콜롬비아, Centro Internacional de Agricultura Tropical (CIAT).
가격: $US100.
참고문헌: Jones and Gladkov 2001.
다운로드: http://www.floramap-ciat.org/ing/floramap101.htm.

GeoLocate

설명: 자연사 수집물과 연관된 장소 데이터에 지리 좌표 할당을 쉽게 할 수 있도록 하는 지리참조연산 프로그램.
버전: 2.0 (2003).
소유기관 미국, LA, 벨 차세, 툴레인 자연사박물관.
가격: 무료

참고문헌: Rios and Bart *n.dat*.
주문: http://www.museum.tulane.edu/geolocate/order.aspx.

PATN
설명: 다변수가(multivariate) 데이터에서 패턴을 추출하고 보여주는 포괄적이면서 다기능이 있는 소프트웨어 패키지.
버전: 3.01 (2004).
소유기관 Blatant Fabrications Pty Ltd (Lee Belbin)
가격: $US299.
참고문헌: Belbin 2004.
다운로드: http://www.patn.com.au/.

PopTools
설명: PopTools 는 마이크로소프트 엑셀 PC 버전에 대한 다용도 기능을 가진 추가 프로그램(add-in)으로 행렬 군집 모델의 분석과 시늉 그리고 추측 통계학적인 과정을 쉽게 할 수 있도록 한다.
버전: 2.6.6 (2005).
소유자: Greg Hood, Albany, W.A., Australia.
가격: 무료
참고문헌: Hood 2005
다운로드: http://www.cse.csiro.au/poptools/.

Specify
설명: 자연사 박물관과 식물표본관을 위한 수집물 관리 시스템.
버전: 4.6 (2004).
소유기관 미국, 캔사스, 로렌스, 캔사스 대학교, 생물다양성연구센터.
가격: 무료
참고문헌: University of Kansas 2003a
다운로드: http://www.specifysoftware.org/Specify/specify/download.

2. 온라인 자원들

BioGeoMancer
설명: 자연사 표본의 수집자, 큐레이터, 그리고 사용자들을 위한 지리참조연산 서비스.
소유기관 미국, 코네티컷, 피바디 자연사박물관.
참고문헌: Peabody Museum *n.dat.*
위치: http://www.biogeomancer.org
참고: BioGeomancer 프로젝트는 최근 (2005) 지리참조연산과 데이터 품질 검사에 대한 도구를 개선할 목적으로 자연사 박물관들이 참여하는 세계적인 협업 프로젝트로 확대되었다. 이러한 도구들은 2006 년 중반에 웹 서비스 뿐만 아니라 독립형 제품으로 범용적인 이용이 가능할 것이다.

Data Cleaning (CRIA)
설명: 큐레이터들이 speciesLink 분산 정보 시스템을 통해 이용 가능하게 된 데이터집합에 대해 자신들의 데이터베이스에 있는 잠재적 오류의 동정을 도울

수 있도록 CRIA 에서 개발된 온라인 데이터 검사 및 오류 동정 도구. 명명 및 지리에 관련된 오류를 수정하는 것이 가능하다.

소유기관 Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.

위치: http://splink.cria.org.br/dc

참고: 이 도구에서 개발된 일부 알고리듬(특히 지리 도구)은 세계적인 협업 프로젝트의 일부로서 BioGeomancer 툴킷에 통합되고 있으며 이 프로젝트의 종료일은 2006 년 중반이다.

geoLoc

설명: 생물학적 수집물 데이터의 지리참조연산을 지원하는 도구.

소유기관 Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.

위치: http://splink.cria.org.br/geoloc?&setlang=en

Georeferencing Calculator

설명: 박물관-기반 자연사 수집물에서 발견되는 것과 같은 서술적인 장소의 지리참조연산을 지원하기 위해 개발된 자바 애플릿.

소유기관 미국, 캘리포니아, 버클리, 캘리포니아 대학교.

위치: http://manisnet.org/manis/gc.html

Lifemapper

설명: 화면보호기 소프트웨어로 인터넷을 사용하여 자연사 박물관의 식물과 동물의 레코드를 추출하고 분포를 예측하기 위해 모델링 알고리듬을 사용한다.

소유기관 미국, 캔사스, 로렌스, 캔사스 대학교, 생물다양성연구센터.

위치: http://www.lifemapper.org/

spOutlier

설명: 위도, 경도, 그리고 고도에서 특이점을 탐지하고 자연사 수집물 데이터에서 정도를 벗어난 해상 또는 비해상 레코드를 식별하기 위해 사용되는 자동화된 도구.

소유기관 Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.

위치: http://splink.cria.org.br/outlier?&setlang=en

3. 표준과 지침서

DELTA

설명: DELTA (DEscription Language for TAxonomy) 형식은 컴퓨터 처리를 위한 분류학적 서술자료의 인코딩 방법으로 융통성이 있다.

표준: TDWG 에서 데이터 교환 표준으로 채택됨

참고문헌: Dallwitz *et al.* 1993.

위치: http://biodiversity.uno.edu/delta/

HISPID

설명: 데이터 교환을 위한 식물표본관 정보 표준과 프로토콜 (Herbarium Information Standards and Protocols for Interchange of Data)

소유기관 호주 식물표본관장 위원회. TDWG 표준으로 채택됨.

참고문헌: Conn 1996, 2000.

위치: http://plantnet.rbgsyd.nsw.gov.au/Hispid4/

MaNIS Georeferencing Guidelines

설명: 지리 좌표 할당과 장소 서술자료 관련한 좌표의 최대 오차 거리 정보를 포함하고 있다.

소유기관 미국, 캘리포니아, 버클리, 캘리포니아 대학교.

위치: http://manisnet.org/manis/GeorefGuide.html.

Manual de Procedimentos para Georreferenciar

설명: 자연사 수집물의 지리참조연산을 위한 지침서로서 멕시코의 CONABIO 에서 개발된 매뉴얼. 영어 초록을 포함하는 스페인어 매뉴얼이 준비 중에 있다.

참고문헌: CONABIO 2005.

위치: 아직 온라인으로 이용할 수 없음.

MaPSTeDI Georeferencing Guidelines

설명: MaPSTeDI 프로젝트에서 표본의 지리참조연산 과정을 위한 안내서.

소유기관 미국, 콜로라도, 덴버, 콜로라도 리젠트 대학교.

위치: http://mapstedi.colorado.edu/geocoding.html

Plant Names in Botanical Databases

설명: 이 표준의 목적은 어떻게 식물의 학명이 식물학 데이터베이스에서 구조화 될 수 있는지를 명시하는 것이다.

소유기관 분류학데이터베이스연구그룹 (Taxonomic Databases Working Group, TDWG).

위치: http://www.tdwg.org/plants.html

SDD

설명: TDWG 의 소그룹인 SDD 는 개체에 대한 서술 데이터를 갈무리하고 관리하는 것에 대한 국제적인 XML 기반 표준을 개발하기 위해 만들어졌다.

소유기관 분류학데이터베이스연구그룹 (Taxonomic Databases Working Group, TDWG).

위치: http://160.45.63.11/Projects/TDWG-SDD/index.html

TDWG Standards

설명: 분류학데이터베이스연구그룹(Taxonomic Databases Working Group, TDWG) 은 오랜 기간 생물다양성 데이터에 사용될 수 있는 표준을 개발해 왔다. 저장, 문서화, 및 종의 분포 그리고 종-발생 데이터와 관련된 일련의 쟁점사항에 대한 표준이 개발되었고 개발되고 있다.

소유기관 분류학데이터베이스연구그룹 (Taxonomic Databases Working Group, TDWG).

위치: http://www.tdwg.org/standrds.html
http://www.tdwg.org/subgrops.html

결론

Errores ad sua principia referre, est refellere
오류의 출처를 명시하는 것은 이것들을 반박하는 것이다.
(Ref. 3 Co. Inst. 15)

정보화 시대의 도래로 수집물 기관들은 환경 정책 수립 과정에서 빠뜨릴 수 없는 부분이 되었고 정치인들은 이러한 기관에 투자한 자원에 대해 적절성과 가치를 점점 더 찾고 있다. 그러므로 수집물 기관들이 관련 사업비를 제공하는 이들에게 가치를 더하는 자원으로 계속해서 보이려면 수집물 기관들이 고품질의 결과를 산출하는 것은 수집물 기관 자신들의 최대 이익이 된다.

조사 및 관찰 정보를 유지하고 있는 박물관, 식물표본관 및 여러 기관의 데이터베이스화된 정보에 관한 최선의 실행사례는 이러한 데이터를 가능한한 정확하게 만들고, 데이터가 최상의 상태로 유지될 수 있도록 가장 적합한 기법과 방법론을 사용하는 것이다. 이 목적을 달성하기 위해서, 데이터 입력 오류를 최소로 유지해야 하며 계속되는 데이터 정제와 검증 작업이 매일 매일의 데이터 및 정보 관리 프로토콜에 반드시 통합되어야 한다.

좋은 품질의 데이터 또는 나쁜 품질의 데이터와 같은 것은 없다 (Chapman 2005a). 데이터는 데이터일 뿐이고, 이러한 것의 이용이 데이터의 품질을 결정하는 것이다. 그럼에도 불구하고, 데이터 제공자는 데이터가 만들어질 때 데이터의 오류가 없도록 확실히 해야 할 필요가 있다. 어떤 한 가지 검사만으로도 결코 데이터 집합의 모든 오류를 동정하기에 충분하지 않을 것이고, 따라서 데이터를 사용하는 기관의 상황과 해당 기관에 있는 데이터의 환경에 가장 적합한 여러 가지 방법을 조합하여 사용하는 것이 중요하다. 또한 개별 수집물 기관 내부의 데이터 품질과 함께 데이터의 조합이 발생함에 따라 관련 수집물의 전체적인 데이터 품질을 높이기 위해, 해당 데이터의 이용자 뿐만 아니라 기관, 데이터 제공자, 과학자 그리고 IT 전문가간의 협업이 필요하다.

아마도 가장 중요한 데이터 관리 실행사례는 올바른 문서화일 것이다. 데이터에 어떤 검사가 행해지든, 이것들은 철저하게 문서화 되어야 한다. 이 방식으로만 해당 데이터의 사용자는 진정한 데이터의 성질과 정확성을 알 수 있다.

데이터와 정보 교환이 증가되고 있는 이 때, 수집물 기관의 명성은 과거처럼 소속 과학자들의 자질에 따라 좌우되기보다는 해당 기관에 있는 정보의 품질과 가용성에 따라 좌우될 것이다 (Redman 1996, Dalcin 2004). 이것은 현실 세계의 사실이고, 이 두개는 더 이상 분리될 수 없다. 올바른 데이터 및 정보 관리는 올바른 과학과 병행되어야 하고, 이것들은 함께 올바른 데이터와 정보로 이어져야 한다.

감사의 글

세계 여러 곳에 있는 많은 동료들과 기관들이 이 논문에 직간접으로 기여하였다. 여기에 설명된 개념들은 비슷한 생각을 가진 사람들과 토론이 없었다면 진전되지 못했을 것이고, 이것은 지난 30 년의 기간에도 그랬다. 다른 사람들이 우리의 지식을 이용할 수 있도록 하고, 그리하여 그들이 이 지식을 바탕으로 새로운 것을 만들고 한 단계 앞으로 전진할 수 있게 됨으로써, 우리의 과학은 진보할 것이다. 최근 몇 년 동안 출판 전에 정보를 기꺼이 공유하고, 사회적 정치적 경계를 넘어 협업 프로젝트를 추진하는 활동이 크게 증가되고 있다. 인터넷의 도래는 이것을 더욱 쉽게 하였고, 이 방법으로만 우리의 분야는 요구되는 데이터와 정보 서비스에서 큰 진전을 이루어 낼 수 있을 것이다.

특별히, 필자는 브라질, 캠피나스의 CRIA 그리고 호주 캔버라에 있는 ERIN 의 전현직 직원들에게 특별한 감사를 표시하고 싶은데, 이들은 여러 생각과, 도구, 이론에 대해 기여하였고 훌륭한 위원회를 통해 필자가 생각을 명확하게 정리할 수 있게 하도록 도움을 주었다. 수년에 걸쳐 이 사람들과 함께 한 환경 정보에 관한 오류와 정확성에 대한 토론 그리고 여러 기관들, 특히 코네티컷 피바디 박물과 그리고 버클리 캘리포니아 대학교에서 수행한 초기 연구로 인해 이 일이 가능하였다. 또한, 멕시코의 코나바이오, 캔사스 대학교, 호주의 CSIRO 그리고 콜로라도 대학교와 같은 기관들 그리고 언급하기에 너무 많은 여러 기관들이 1 차 종 데이터 품질 관련하여 오늘날 우리가 이용할 수 있는 현재의 상태까지 발전시켰다. 필자는 이러한 사람들의 독창적인 생각과 건설적인 비판에 감사를 드린다.

캔사스 대학교의 타운 피터슨(Town Peterson)과 다른 사람들, 코네티컷 웨슬리안 대학교의 바리 체르노프(barry Chernoff), 예일 대학교의 리드 비이만(Read Beaman), 버클리 캘리포니아 대학교의 존 위에크조렉(John Wieczorek)과 로버트 히즈만(Robert Hijmans), 암스테르담 ETI 의 피터 숄크(Peter Shalk)와 다른 사람들, 캘리포니아 과학원의 스탠 블럼(Stan Blum), 코펜하겐의 GBIF 직원들은 필자에게 여러 생각과 도전 과제를 제시하였으며, 논문 중의 일부로 이러한 것들이 표현되었다. 하지만 임의의 오류, 누락 또는 논쟁거리는 필자의 몫이다.

필자는 이 문서의 편집과 검토 기간 동안 비판, 의견 그리고 제안을 제공했던 사람들, 그리고 특히 GBIF DIGIT (Digitisation of Natural History Collection Data) 소위원회의 다음 위원들에게 또한 감사를 드린다: Anton Güntsch, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany; Mervyn Mansell, USDA-Aphis, Pretoria, South Africa; Francisco Pando, Real Jardín Botánico, Madrid, Spain; A. Townsend Peterson, University of Kansas, USA; Tuuli Toivonen, University of Turku, Finland; Anna Wietzman, Smithsonian Institution, USA as well as Patricia Mergen, Belgian Biodiversity Information Facility, Belgium.

GBIF 의 래리 스피어즈(Larry Speers)는 이 보고서의 착수, 그리고 검토 및 편집에 큰 도움이 되었다.

끝으로 필자가 2003-2004 년 브라질에 머무는 동안 데이터 품질 관리에 관한 필자의 생각을 확장할 수 있도록 기회와 지원을 제공한 브라질 FAPESP/Biota 프로젝트와 이 논문을 작성할 수 있도록 지원하고 격려를 해 준 GBIF 에 감사를 드린다.

참고문헌

Armstrong, J.A. 1992. The funding base for Australian biological collections. *Australian Biologist* **5(1):** 80-88.

ABRS. 2004. *Australian Faunal Directory*. Canberra: Australian Biological Resources Study. http://www.deh.gov.au/biodiversity/abrs/online-resources/abif/fauna/afd/index.html [Accessed 12 Apr. 2005].

ANBG. 2003. *Australian Plant Name Index*. Canberra: Australian National Botanic Gardens. http://www.anbg.gov.au/apni/index.html [Accessed 12 Apr. 2005].

Barnett, V. and Lewis, T. 1994. *Outliers in Statistical Data*. Chichester, UK: Wiley and Sons.

Beaman, R.S. 2002. Automated georeferencing web services for natural history collections **in** *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002* http://www.cria.org.br/eventos/tdbi/flora/reed [Accessed 12 Apr. 2005]

Beaman, R., Wieczorek, J. and Blum, S. 2004. Determining Space from Place for Natural History Collections in a Distributed Library Environment. *D-Lib Magazine* Vol. 10(5). http://www.dlib.org/dlib/may04/beaman/05beaman.html [Accessed 12 Apr. 2005].

Belbin, L. 2004. *PATN vers. 3.01*. Blatant Fabrications http://www.patn.com.au [Accessed 12 Apr. 2005].

Berendsohn, W.G. 1997. A taxonomic information model for botanical databases: the IOPI model. *Taxon* 46: 283-309.

Berendsohn, W., Güntsch, A. and Röpert, D. (2003). *Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections*. Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Members/irc/gbif/digit/library?l=/digitization_collections/contract_2003_report/ [Accessed 13 Apr. 2005].

BioCASE. 2003. *Biological Collection Access Service for Europe*. http://www.biocase.org [Accessed 12 Apr. 2005].

Birds Australia. 2004. *Birds Australia Rarities Committee (BARC)*. http://users.bigpond.net.au/palliser/barc/barc-home.html [Accessed 12 Apr. 2005].

Bisby, F.A. 1994. *Plant Names in Botanical databases.* TDWG Standard. http://www.tdwg.org/plants.html [Accessed 12 Apr. 2005].

Bisby, F.A., Zarucchi, J.L., Schrire, B.L., Roskov, Y.R., Heald, J. and White, R.J. (eds). 2002. *ILDIS World Database of Legumes* ver. 6.05. http://www.ildis.org/ [Accessed 12 Apr. 2005].

Blakers, M., Davies, S.J.J.F. and Reilly, P.N. 1984. *The Atlas of Australian Birds*. Melbourne: Melbourne University Press.

Blum, S. 2001. *Georeferencing Natural History Collection Localities at the California Academy of Sciences*. http://www.calacademy.org/research/informatics/GeoRef/index.html [Accessed 12 Apr. 2005].

Brickell, C.D., Baum, B.R., Hetterscheid, W.L.A., Leslie, A.C., McNeill, J., Trehane, P., Vrugtman, F. and Wiersema, J.H. (eds) 2004. *International Code for Cultivated Plants* ed. 7. Edinburgh, U.K.: ISHS. http://www.actahort.org/books/647/ [Accessed 11 Apr. 2005].

Brummitt, R.K. and Powell, C.E. (eds). 1992 *Authors of Plant Names.* Kew: Royal Botanic Gardens, Kew. http://www.ipni.org/index.html [Accessed 12 Apr. 2005]

Burbidge, A.A. 1991. Cost Constraints on Surveys for Nature Conservation **in** Margules, C.R. and Austin, M.P. (eds). *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. Canberra: CSIRO.

Burrough, P.A. and McDonnell, R.A. 1998. *Principals of Geographical Information Systems*. Oxford, UK: Oxford University Press.

Busby, J.R. 1991. BIOCLIM – a bioclimatic analysis and prediction system. Pp. 4-68 **in** Margules, C.R. and Austin, M.P. (eds) *Nature Conservation: Cost Effective Biological Surveys and data Analysis*. Melbourne: CSIRO.

Chapman, A.D. 1988. Karl Domin in Australia **in** *Botanical History Symposium. Development of Systematic Botany in Australasia. Ormond College, University of Melbourne. May 25-27, 1988*. Melbourne: Australian Systematic Botany Society, Inc.

Chapman, A.D. 1991. Australian Plant Name Index pp. 1-3053. *Australian Flora and Fauna Series* Nos 12-15. Canberra: AGPS.

Chapman, A.D. 1992. Quality Control and Validation of Environmental Resource Data **in** *Data Quality and Standards: Proceedings of a Seminar Organised by the Commonwealth Land Information Forum, Canberra, 5 December 1991*. Canberra: Commonwealth land Information Forum.

Chapman, A.D. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 **in** Lowell, K. and Jaton, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.

Chapman, A.D. *et al*. 2002. *Guidelines on Biological Nomenclature*. Canberra: Environment Australia. http://www.deh.gov.au/erin/documentation/nomenclature.html [Accessed 12 Apr. 2005].

Chapman, A.D. 2004. Guidelines on Biological Nomenclature. Brazil edition. Appendix J to *Sistema de Informação Distribuído para Coleções Biológicas: A Integração do Species Analyst e SinBiota. FAPESP/Biota process no. 2001/02175-5 March 2003 – March 2004.* Campinas, Brazil: CRIA 11 pp. http://splink.cria.org.br/docs/appendix_j.pdf [Accessed 12 Apr. 2005].

Chapman, A.D. 2005a. *Principles of Data Quality*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/data_quality [Accessed 1 Aug. 2005].

Chapman, A.D. 2005b. *Uses of Primary Species-Occurrence Data*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/uses_of_data [Accessed 1 Aug. 2005].

Chapman, A.D. and Busby, J.R. 1994. Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 **in** Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.

Chapman, A.D. and Milne, D.J. 1998. *The Impact of Global Warming on the Distribution of Selected Australian Plant and Animal Species in relation to Soils and Vegetation*. Canberra: Environment Australia

Chapman, A.D., Bennett, S., Bossard, K., Rosling, T., Tranter, J. and Kaye, P. 2001. Environment Protection and Biodiversity Conservation Act, 1999 – Information System. *Proceedings of the 17th Annual Meeting of the Taxonomic Databases Working Group, Sydney, Australia 9-11 November 2001*. Powerpoint: http://www.tdwg.org/2001meet/ArthurChapman_files/frame.htm [Accessed 12 Apr. 2005].

CHAH 2002. *AVH - Australian's Virtual Herbarium*. Australia: Council of Heads of Australian Herbaria. http://www.chah.gov.au/avh/avh.html [Accessed 12 Apr. 2005].

Chrisman, N.R., 1991. The Error Component in Spatial Data. pp. 165-174 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Christidis, L. & Boles, W.E. 1994. *Taxonomy and Species of Birds of Australia and its Territories*. Royal Australasian Ornithologists Union, Melbourne. 112 pp.

Clarke, K.C. 2002. *Getting Started with Geographic Information Systems*, 4th edn. Upper Saddle River, NJ, USA: Prentice Hall. 352 pp.

Colwell, R.K. 2002. *Biota: The Biodiversity Database Manager*. Connecticut, USA: University of Connecticut http://viceroy.eeb.uconn.edu/Biota [Accessed 12 Apr. 2005].

CONABIO. 2002. *The Biótica Information system*. Mexico City: Comisión national para el conocimiento y uso de la biodiversidad. http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/acerca_biotica.html [Accessed 12 Apr. 2005
CONABIO. 2005. *Manual de Procedimentos para Georreferenciar.* Mexico: Comisión para el Conocimiento y Uso de la Biodiversidad México (CONABIO).
Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data.* Version 3 (Draft 1.4). Sydney: Royal Botanic Gardens. http://www.bgbm.org/TDWG/acc/hispid30draft.doc [Accessed 10 Apr 2005]
Conn, B.J. (ed.) 2000. *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data.* Version 4 – Internet only version. Sydney: Royal Botanic Gardens. http://plantnet.rbgsyd.nsw.gov.au/Hispid4/ [Accessed 30 Jul. 2003].
Croft, J.R. (ed.) 1989. *HISPID – Herbarium Information Standards and Protocols for Interchange of Data*. Canberra: Australian National Botanic Gardens.
CRIA. 2002. *speciesLink*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/ [accessed 12 Apr. 2005]
CRIA. 2004a. *GeoLoc-CRIA*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [Accessed 12 Apr. 2005].
CRIA. 2004b. *spOutlier-CRIA*. Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [accessed 1 Mar. 2005]
CRIA (2005), *speciesLink. Dados e ferramentos. Data Cleaning*. Centro de Referência em Informação Ambiental. http://splink.cria.org.br/dc/ [Accessed 12 Apr. 2005].
Dalcin, E.C. 2004. *Data Quality Concepts and Techniques Applied to Taxonomic Databases.* Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf [Accessed 7 Jan. 2005].
Dallwitz, M.J., Paine, T.A. and Zurcher, E.J. (1993). *User's guide to the DELTA System: a general system for processing taxonomic descriptions*. 4th edn. http://delta-intkey.com/ [Accessed 12 Apr. 2005].
DEH. 2005a. *Threatened Species*. Canberra: Department of Environment and Heritage. http://www.deh.gov.au/biodiversity/threatened/species/index.html [Accessed 12 Apr. 2005].
DEH. 2005b. *Species Profile and Threats Database*. Canberra : Department of Environment and Heritage. http://www.deh.gov.au/cgi-bin/sprat/public/sprat.pl [Accessed 7 Apr. 2005].
Dorr, LJ. 1997. *Plant Collectors in Madagascar and the Comoro Islands*. Kew, UK: Royal Botanic Gardens, Kew.
English, L.P. 1999. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc., New York.
ESRI. 2003. *ArcSDE: The GIS Gateway to Relational Databases*. http://www.esri.com/software/arcgis/arcinfo/arcsde/overview.html [Accessed 12 Apr. 2005
Farr, E. and Zijlstra, G. (eds). *n.dat. Index Nominum Genericorum (Plantarum).* On-line version. http://ravenel.si.edu/botany/ing/ [Accessed 21 Jul. 2004].
Flowerdew, R., 1991. Spatial Data Integration. pp. 375-387 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
Froese, R. and Bisby, F.A. (eds). 2004. *Catalogue of Life 2004*. Los Baños, Philippines: Species 2000. http://www.sp2000.org/AnnualChecklist.html [Accessed 7 Apr. 2005].
Froese, R. and Pauly, D. 2004. *Fishbase.* Ver. 05/2004. The Philippines: World Fish Center. http://www.fishbase.org/ [Accessed 10 Apr. 2005].

Fundación Biodiversidad 2005. *Proyeto Anthos – Sistema de información sobre los plantas de España*. http://www.programanthos.org/ [Accessed 8 Apr. 2005].
Gatrell, A.C. 1991. Concepts of Space and Geographical Data. pp. 119-134 **in:** Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
GBIF. 2003a. *What is GBIF?* http://www.gbif.org/GBIF_org/what_is_gbif [Accessed 12 Apr. 2005].
GBIF. 2003b. *GBIF Work Program 2004*. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/GBIF_org/wp/wp2004/GB7_20WP2004-v1.0-approved.pdf [Accessed 12 Apr. 2005].
GBIF. 2004. *Data Portal*. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.net/portal/index.jsp. [Accessed 12 Apr. 2005].
Geographic Names Board. 2003. *Guidelines for Naming of Roads*. Sydney: Geographic Names Board of New South Wales. http://www.gnb.nsw.gov.au/newsroom/road_naming_guideline.pdf [Accessed 21 Jul. 2004].
Greuter, W. *et al.* 1984-1989. *Med-Checklist: a critical inventory of vascular plants of the circum-mediterranean countries*. 4 Vols. Botanical Garden and Botanical Museum Berlin-Dahlem.
Hepper, F.N. and Neate, F. 1971. *Plant Collectors in West Africa.* Utrecht, The Netherlands: Oosthoeks' Uitgeversmaatschappij.
Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrentes, I. and Rojas, E. 2005 *DIVA-GIS Version 5. A geographic information system for the analysis of biodiversity data*. http://www.diva-gis.org [Accessed 30 Jul. 2004].
Hobern, D. and Saarenmaa, H. 2005. *GBIF Data Portal Strategy*. Draft Version 0.14. Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/portal_strategy_1/ [Accessed 7 Apr. 2005].
Hood, G.M. 2005. *PopTools version 2.6.6*. Canberra: CSIRO Sustainable Ecosystems. http://www.cse.csiro.au/poptools [Accessed 13 Apr. 2005].
Houlder, D. Hutchinson, M.J., Nix, H.A. and McMahaon, J. 2000. *ANUCLIM 5.1 Users Guide*. Canberra: Cres, ANU. http://cres.anu.edu.au/outputs/anuclim.php [Accessed 12 Apr. 2005].
IAPT. 1997. *Names in Current Use for Extant Plant Genera ver. 1.0*. on-line version. International Association for Plant Taxonomy. http://www.bgbm.org/iapt/ncu/genera/Default.htm [Accessed 12 Apr. 2005].
ICSM. 2001. *Guidelines for the Consistent Use of Place Names*. Intergovernmental Committee on Survey and Mapping: Committee for Geographic Names in Australia. http://www.icsm.gov.au/icsm/cgna/consistent_pnames.pdf. [Accessed 12 Apr. 2005].
Index Herbariorum. (1954-1988) *Index Herbariorum Part 2: Collectors*. Various compilers. Utrecht/Antwerp, The Hague/Boston
Part 2(1): Collectors A-D (1954). *Regnum Vegetabile* vol. 2 (A-D),
Part 2(2): Collectors E-H (1957). *Regnum Vegetabile* vol. 9 (E-H),
Part 2(3): Collectors I-L (1972). *Regnum Vegetabile* vol. 86 (I-L),
Part 2(4): Collectors M (1976). *Regnum Vegetabile* vol. 93 (M),
Part 2(5): Collectors N-R (1983). *Regnum Vegetabile* vol. 189 (N-R),
Part 2(6): Collectors S (1986). *Regnum Vegetabile* vol. 114 (S),
Part 2(7): Collectors T-Z (1988). *Regnum Vegetabile* vol. 117 (T-Z).
IPNI. 1999. *International Plant Names Index*. http://www.ipni.org/index.html [accessed 12 Apr. 2005].
IOPI. 2003. *Global Plant Checklist*. International Organization for Plant Information (IOPI). http://www.bgbm.fu-berlin.de/IOPI/GPC/ [Accessed 12 Apr. 2005].

Johnson, R.A. and Wichern, D.W. 1998. *Applied Multivariate Statistical Analysis*. 4th edn. New York, NY: Prentice Hall.
Jones P.G. and Gladkov, A. 2001. *Floramap Version 1.01*. Cali, Colombia: CIAT. http://www.floramap-ciat.org/ing/floramap101.htm [Accessed 12 Apr. 2005].
Koch, I. 2003. *Coletores de plantas brasileiras*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 12 Apr. 2005].
Lampe, K.-H. and Riede, K. 2002. *Mapping the collectors: the georeferencing bottleneck*. Poster given to TDWG meeting, Indaiatuba, Brazil. http://www.cria.org.br/eventos/tdbi/bis/Poster-200dpi.html [accessed 12 Apr. 2005].
Legendre P. and Legendre L. (1998): Numerical Ecology. *Developments in Environmental Modeling* 20, Second English Edition, Elsevier, Amsterdam, 853p. http://www.bio.umontreal.ca/legendre/numecol.html [Accesssed 12 Apr. 2005].
Lindemeyer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F. and Tanton, M.T. 1991. The Conservation of Leadbeater's Possum, *Gymnobelidus leadbeateri* (McCoy): A Case Study of the Use of Bioclimatic Modelling. *J. Biogeog.* **18:** 371-383.
Lowry, R. 2005. Concepts and Applications of Inferential Statistics. *VassarStats: Web Site for Statistical Computation*. http://faculty.vassar.edu/lowry/webtext.html [Accessed 8 Apr. 2005]
Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond Integrity Analysis pp. 200-209 in *Proceedings of the Conference on Information Quality (IQ2000)*. Boston: Massachusetts Institute of Technology. http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf [Accessed 12 Apr. 2005].
MaNIS. 2001. *The Mammal Networked Information System*. http://manisnet.org/manis [Accessed 12 Apr. 2005].
Marcus, A., Maletic, J.I. and Lin, K.-I. 2001. Ordinal Association Rules for Error Identification in Data Sets pp. 589-591 in *Proceedings of the 10th ACM Conference on Information and Knowledge Management (ACM CIKM 2001). Atlanta, GA.* http://www.cs.wayne.edu/~amarcus/papers/cikm01.pdf [Accessed 12 Apr. 2005].
Margules, C.R. and Redhead, T.D. 1995. *BioRap. Guidelines for using the BioRap Methodology and Tools*. Canberra: CSIRO. 70pp.
Marino, A., Pavarin, F., de Souza, S. and Chapman, A.D. in prep. *Simple on line tools for geocoding and validating biological data*. To be submitted.
Neldner, V.J., Crossley, D.C. and Cofinas, M. 1995. Using Geographic Information Systems (GIS) to Determine the Adequacy of Sampling in Vegetation Surveys. *Biological Conservation* 73: 1-17.
Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes in Longmore, R.C. (ed). Atlas of Australian elapid snakes. *Australian Flora and Fauna Series* No. **7:** 4-15. Canberra: Australian Government Publishing Service.
NMNH. 1993. *RapidMap. Geocoding locality descriptions associated with herbarium specimens*. U.S. National Musuem of Natural History and Bernice P. Bishop Museum, Honolulu. http://users.ca.astound.net/specht/rm/ [Accessed 12 Apr. 2005].
Peabody Museum. *n.dat*. *BioGeoMancer*. http://www.biogeomancer.org [Accessed 12 Apr. 2005].
Peterson, A.T., Navarro-Siguenza, A.G. and Benitez-Diaz, H. 1998. The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* **140:** 288-294.
Peterson, A.T., Stockwell, D.R.B. and Kluza, D.A. 2002. Distributional Prediction Based on Ecological Niche Modelling of Primary Occurrence Data pp. 617-623 **in** Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale.* Washington: Island Press.
Peterson, A.T., Navarro-Siguenza, A.G. and Pereira, R.S. 2003. Detecting errors in biodiversity data based on collectors' itineraries. *Bulletin of the British Ornithologists' Club* 124: 143-151. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PNP_BBOC_2004.pdf. [Accessed 12 Apr. 2005].

Pfeiffer, U., Poersch, T. and Fuhr, N. 1996. Retrieval effectiveness of proper name search methods. *Information Processing and Management* **32(6):** 667-679.
Platnick, N.I. 2004. *The World Spider Catalog*. New York: The American Museum of Natural History. http://research.amnh.org/entomology/spiders/catalog81-87/INTRO3.html [Accessed 12 Apr. 2005].
Podolsky, R. 1996. *Software Tools for the Management and Visualization of Biodiversity Data*. NY, USA: United Nations Development Project. http://www3.undp.org/biod/bio.html [Accessed 13 Apr. 2005].
Pollock, J.J. and Zamora, A. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of ACM* **27(4):** 358-368.
Redman, T.C. 1996. *Data Quality for the Information Age*. Artech House Inc.
Redman, T.C. 2001. *Data Quality: The Field Guide*. Boston, MA: Digital Press.
Rios, N.E. and Bart, H.L. Jr. *n.dat. GEOLocate. Georeferencing Software. User's Manual*. Belle Chasse, LA, USA: Tulane Museum of Natural History. http://www.museum.tulane.edu/geolocate/support/manual_ver2_0.pdf [Accessed 12 Apr. 2005].
Roughton, K.G. and Tyckoson, D.A. 1985. Brousing with sound: Sound-based codes and automated authority control. *Information Technology and Libraries* **4(2):**130-136.
Ruggiero, M. (ed.) 2001. *Integrated Taxonomic Information System*. http://www.itis.usda.gov/ [Accessed 12 Apr. 2005].
Pereira, R.S. 2002. *Desktop Garp*. Lawrence, Kansas: University of Kansas Center for Research. http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1 [Accessed 13 Apr. 2005].
Shattuck, S.O. 1997. eGaz, The Electronic Gazetteer. *ANIC News* **11:** 9 http://www.ento.csiro.au/biolink/egaz.html [Accessed 12 Apr. 2005].
Shattuck, S.O. and Fitzsimmons, N. 2000. *BioLink, The Biodiversity Information Management System*. Melbourne, Australia: CSIRO Publishing. http://www.ento.csiro.au/biolink/software.html [Accessed 12 Apr. 2005].
Steenis-Kruseman, M.J. van 1950. Malaysian Plant Collectors and Collections. *Flora Malesiana* Vol. 1. Leiden, The Netherlands.
Stockwell, D. and Peters, D. 1999. "The GARP modelling system: problems and solutions to automated spatial prediction." *International Journal of Geographical Information Science* **13**(2): 143-158.
University of Colorado Regents. 2003a. *mapstedi. Geocoding*. Denver: University of Colorado MaPSTeDI project. http://mapstedi.colorado.edu/geocoding.html [Accessed 12 Apr. 2005].
University of Colorado Regents. 2003b. *GeoMuse*. Denver: University of Colorado MaPSTeDI project. http://www.geomuse.org/mapstedi/client/start.jsp [Accessed 12 Apr. 2005].
University of Kansas. 2003a. *Specify*. Biological Collections Management. Lawrence, Kansas: University of Kansas http://www.specifysoftware.org/Specify/ [Accessed 12 Apr. 2005].
University of Kansas. 2003b. *LifeMapper*. Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. http://www.lifemapper.org/ [Accessed 12 Apr. 2005
University of Oxford. 2004. *BRAHMS. Botanical Research and Herbarium Management System*. Oxford, UK: University of Oxford http://storage.plants.ox.ac.uk/brahms/defaultNS.html [Accessed 27 Jul 2004].
Weber, W.A. 1995. Vernacular Names: Why Oh Why?. *Botanical Electrical News* No. 109. http://www.ou.edu/cas/botany-micro/ben/ben109.html [Accessed 7 Apr. 2005].
Weiher, E. and Keddy, P. (eds). 1999. *Ecological Assembly Rules: Perspectives, Advances, Retreats*. Cambridge, UK: Cambridge University Press. 418 pp.
Wieczorek, J. 2001a. *MaNIS: Georeferencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/GeorefGuide.html [Accessed 12 Apr. 2005].
Wieczorek, J. 2001b. *MaNIS: Georeferencing Calculator*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/gc.html [Accessed 12 Apr. 2005].

Wieczorek, J. and Beaman, R.S. 2002. Georeferencing: Collaboration and Automation in *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002* http://www.cria.org.br/eventos/tdbi/bis/georeferencing [Accessed 12 Apr. 2005].

Wieczorek, J., Guo, Q. and Hijmans, R.J. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18(8): 745-767.

Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.

Williams, P.H., Marguiles, C.R. and Hilbert, D.W. 2002. Data requirements and data. sources for biodiversity priority area selection. *J. Biosc.* **27(4):** 327-338.

색인