GBIF - 생물종 데이터 품질의 원칙

옮긴이: 박형선, 안성수, 박재홍

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

GBIF - 생물종 데이터 품질의 원칙

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

GBIF – 생물종 데이터 품질의 원칙

초판 인쇄: 2007 년 12 월 21 일
초판 발행: 2007 년 12 월 21 일

옮긴이 | 박형선, 안성수, 박재홍
펴낸이 | 양병태

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 데이터 품질의 원칙(PRINCIPLES OF DATA QUALITY) 자료를 원저자의 허락을 받고 번역한 것입니다. 이 번역물이 국내의 생물다양성 데이터를 인터넷상에서 공유하고 활용하려고 할 때 참고 자료로 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/URL1124374433
- http://www.kbif.re.kr/Download/DIGIT/data_quality.pdf
- http://www.kbif.re.kr/Download/DIGIT/data_quality_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 전자메일로 연락 주시기 바랍니다.

한국과학기술정보연구원 박형선 (seonpark@kisti.re.kr)
한국과학기술정보연구원 안성수 (ssahn@kisti.re.kr)
한국과학기술정보연구원 박재홍 (middle75@kisti.re.kr)

표지 디자인 | 박양숙 (greenish3@kisti.re.kr)

ISBN 978-89-5884-959-9 93470

Chapman, A. D. 2005. *Principles of Data Quality*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

This paper was commissioned from 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

Amata phegea (Linnaeus 1758)

목차

목차 ..... **I**

서론 **2**

정의 **4**

데이터 품질의 원칙 **10**

분류학 및 명명 데이터 **24**

공간 데이터 **29**

수집자와 수집 데이터 **32**

서술 데이터 **33**

데이터 획득 **34**

데이터 입력과 입수 **37**

데이터의 문서화 **40**

데이터의 저장 **46**

공간 데이터 다루기 **51**

표시와 표현 **53**

결론 **58**

감사 **59**

참고문헌 **60**

색인 **65**

서론

데이터 품질 원칙은 최근에 기업(SEC 2002), 의약(Gad and Taulbee 1996), GIS(Zhang and Goodchild 2002), 원격 탐지(Lunetta and Lyon 2004)와 다른 여러 분야에서 업무의 중요한 부분이 되었지만, 박물관과 분류학 커뮤니티에서는 이제서야 널리 인식되고 있다. 분류학 및 종-발생 데이터의 교환과 가용성이 급속이 증가되고 데이터 사용자들이 이러한 정보의 품질을 점점 상세하게 요구하기 시작하여 이러한 원칙은 중요한 고려 사항이 되고 있다. 실제, 박물관 커뮤니티 밖의 일부 사람들은 박물관의 데이터 품질을 환경 보전 결정에 사용하기에 대체로 적합하지 않다고 여기는데, 이것이 실제로 데이터 품질의 결과인가 아니면 이것의 문서화 때문인가? 그러나 이러한 데이터는 매우 중요하다. 오랜 시간에 걸친 수집 때문에, 이것들은 인간이 이러한 다양성에 거대한 영향을 끼친 시간 동안 생물학적 다양성에 대한 대체할 수 없는 기준선 데이터를 제공한다(Chapman and Busby 1994). 이것들은 환경 보전을 위한 노력에 필수적인 자원으로, 농지 개척, 도시화, 기후 변화로 인해 서식지 변화를 겪었거나 다른 방식으로 변화되었을 지역에 대해 유일하게 온전히 문서화된 종 발생 레코드를 제공하기 때문이다(Chapman 1999).

이러한 것들은 필자가 아래에서 전개하고자 하는 생각들의 일부이며, 또한 박물관과 식물표본관들이 자신들의 데이터를 더 넓은 커뮤니티에 제공할 때 박물관과 식물표본관의 업무에서 핵심이 되어야 하는 데이터 품질에 관한 여러 가지 원칙을 제시하고자 한다.

환경 데이터베이스, 모델링 시스템, GIS, 의사결정 지원 시스템 등에서 데이터 품질과 데이터 내의 오류는 종종 간과되는 사항들이다. 아주 종종, 데이터에 담겨있는 오류를 고려하지 않고 데이터가 무비판적으로 사용되며, 이것은 잘못된 결과, 엉뚱한 정보, 현명치 못한 환경 관련 결정 및 예산의 증가 등을 초래할 수 있다.

> *박물관과 식물표본관에서 보유하고 있는 식물과 동물 표본 데이터는 이러한 개체들의 위치에 대한 현재 정보 뿐만 아니라 수 백 년 전으로 거슬러 올라가는 역사적 정보와 같은 광대한 정보 자원을 제공한다*(Chapman and Busby 1994).

생물 종 데이터를 다룰 때, 특히 이들 데이터에서 공간적인 측면을 다룰 때 적용되는 많은 데이터 품질 원칙들이 있다. 이러한 원칙들은 데이터 관리 과정의 모든 단계와 관련이 있다. 이러한 관리 과정 중의 어느 한 단계에서 데이터 품질 손실은 이 데이터가 적절하게 사용될 수 있는 유용성과 이용성을 감소시킨다. 이러한 것들은 다음과 같다:

- 수집 당시에 데이터의 파악과 기록,
- 디지털화 이전의 데이터 조작 (레이블 표지, 수집 노트로의 데이터 복사 등),
- 수집물 동정(표본, 관찰)과 이것의 기록,
- 데이터의 디지털화,
- 데이터의 문서화 (메타데이터를 파악하고 기록),
- 데이터 저장과 보관(archiving),
- 데이터 표현과 보급 (종이 및 전자 출판물, 웹이 가능한 데이터베이스 등),
- 데이터의 이용 (분석과 처리).

모든 이러한 것들은 최종 품질 또는 데이터 "이용에 대한 적합성"으로의 입력을 가지고 있으며 데이터의 모든 측면에 적용된다 – 데이터의 분류 또는 명명적인 부분 – "무엇을", 공간적인 부분 – "어디에서" 그리고 "누가", "언제"와 같은 다른 부분에 적용된다.

데이터 품질과 생물 종-발생 데이터에 대한 이것의 적용을 자세히 논의하기 전에, 몇 가지의 개념을 정의하고 서술할 필요가 있다 이러한 것들은 데이터 품질에 대한 용어, 종종 잘못 쓰이는 정확도와 정밀도, 그리고 1차 종 데이터와 종-발생 데이터가 무엇을 뜻하는가에 대한 것이다.

품질 개선의 단순한 특징을 과소평가해서는 안 된다. 조직의 협력작업, 교육훈련, 그리고 기강 이외에, 이것은 특별한 기술을 필요로 하지 않는다. 원한다면 누구나 훌륭한 기여자가 될 수 있다.
(Redman 2001).

정의

종-발생 데이터

여기에서 사용되는 종-발생 데이터(species-occurrence data)는 박물관과 식물표본관에 보관된 표본에 붙어있는 표본 레이블 데이터, 관찰 데이터 그리고 환경 조사 데이터를 포함한다. 일반적으로, 이 데이터는 소위 말하는 "점-기반"의 데이터이지만, 선(환경 조사에서의 횡단 데이터, 강을 따른 수집물), 다각형(국립공원과 같이 한정된 지역 내의 관찰자료, 그리고 격자 데이터(정규 격자에서의 관찰 또는 조사 레코드)도 포함된다. 일반적으로 여기에서는 지리정보를 가진 데이터 – 즉, 공간상의 특정한 장소와 연결된 지리정보를 가진 레코드 – 지리좌표가 있는 것 (예, 위도와 경도, UTM) 또는 그렇지 않은 것 (장소, 고도, 깊이를 문자로 서술) , 그리고 시간에 대한 데이터를 다루고 있다. 일반적으로 이 데이터는 또한 분류학적 이름과 연관되어 있으며, 동정되지 않은 수집물이 포함될 수도 있다. 이 용어는 종종 "1차 종 데이터(primary species data)"와 같이 혼용되어 사용되어 왔다.

1 차 종 데이터

"1차 종 데이터"는 공간적인 속성이 없는 원시 수집물 데이터를 가리킬 때 사용된다. 이것은 공간적인 특징이 없는 분류학적 그리고 명명 법적인 데이터를 포함하며, 지리 참조정보를 가지지 않는 이름, 분류군, 그리고 분류학적인 개념과 같은 것이 있다.

정확도와 정밀도

정확도(accuracy) 그리고 *정밀도(precision)*를 많은 사람들이 보통 혼동하여 사용하고 있으며 그 차이점을 일반적으로 인식하지 못하고 있다. 이 차이점은 예(figure 1)를 통하여 가장 잘 이해할 수 있다.

정확도는 figure 1에서 보는 것처럼 실제 또는 참 값 (또는 참이라고 수용되는 값 – 예를 들어, 조사 제어 지점에 대한 좌표 값)에 대한 측정된 값, 관찰 값 또는 예측 값의 근접성을 일컫는다.

*정밀도 (또는 해상도)*는 주요한 두 가지 종류로 나눌 수 있다. *통계적인 정밀도*는 반복되는 관찰 결과들이 이러한 관찰 결과들에 상응하는지에 대한 근접성이다. 이것은 참 값에 대한 이것들의 관계와 아무런 관련이 없으며, figure 1a 에 보는 것과 같이 정밀도는 높지만 정확도가 낮을 수 있다. *수치적인 정밀도*는 관찰이 기록되는 유효 숫자의 개수이고 컴퓨터가 사용되면서 더욱 분명해졌다. 예를 들어, 데이터베이스는 실제로 어떤 레코드가 최대 10-100m (3-4 개의 소수점)의 해상도를 가지고 있을 때 위도/경도 레코드를 소수점 아래 10 자리, 즉 0.01mm 까지 출력할 수도 있다. 이것은 종종 해상도와 정밀도의 대해 잘못된 정보를 전달한다.

이러한 용어들(정확도와 정밀도)은 공간적인 데이터뿐만 아니라 비-공간적인 데이터에도 또한 적용이 가능하다. 예를 들어, 하나의 수집물은 아종 단계까지 동정이 되었지만(즉, 높은 정밀도) 이것이 잘못된 분류군(낮은 정확도)일 수 있고, 또는 목(Family) 단계(높은 정확도, 낮은 정밀도)까지만 동정될 수 도 있다.

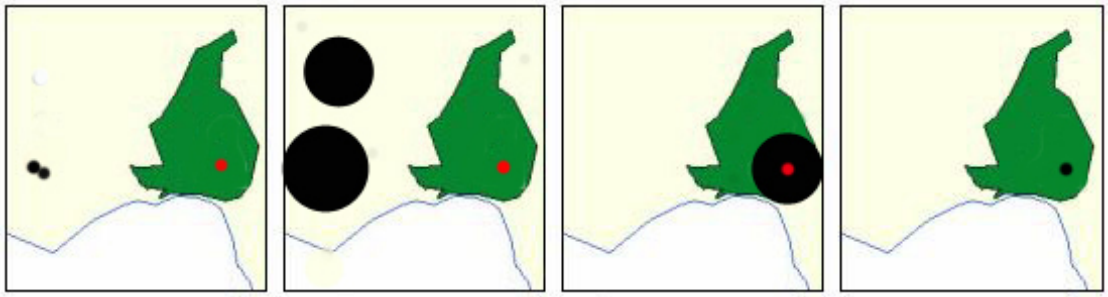

Fig. 1. *공간적인 상황에서 정밀도와 정확도의 차이를 보여주는 그림. 빨간 점은 실제 지점을 나타내고 검은 점은 수집자가 보고한 지점이다.*

a. 높은 정밀도, 낮은 정확도.
b. 낮은 정밀도, 무작위 오류를 보이는 낮은 정확도.
c. 낮은 정밀도, 높은 정확도.
d. 높은 정밀도와 높은 정확도.

품질

품질이 데이터에 적용될 때 여러 가지 정의가 있지만 지리 분야에서는 "사용에 대한 적합성" (Chrisman 1983) 또는 "잠재적인 사용성"이라는 하나의 정의로 대체적으로 수용되고 있다. 이것은 최근 대부분의 공간 데이터 전송 표준들이 채택하고 있는 정의이다 (ANZLIC 1996a, USGS 2004). 이것은 또한 경제와 기업에서와 같은 비-공간 분야에서도 점점 더 많이 사용되고 있다. 일부(예를 들어, English 1999)에서는 "사용에 대한 적합성" 정의가 다소 제한적인 면이 있다고 생각하여 미래 또는 잠재적인 사용에 대한 적합성을 또한 포함하는 정의를 주장하고 있다.

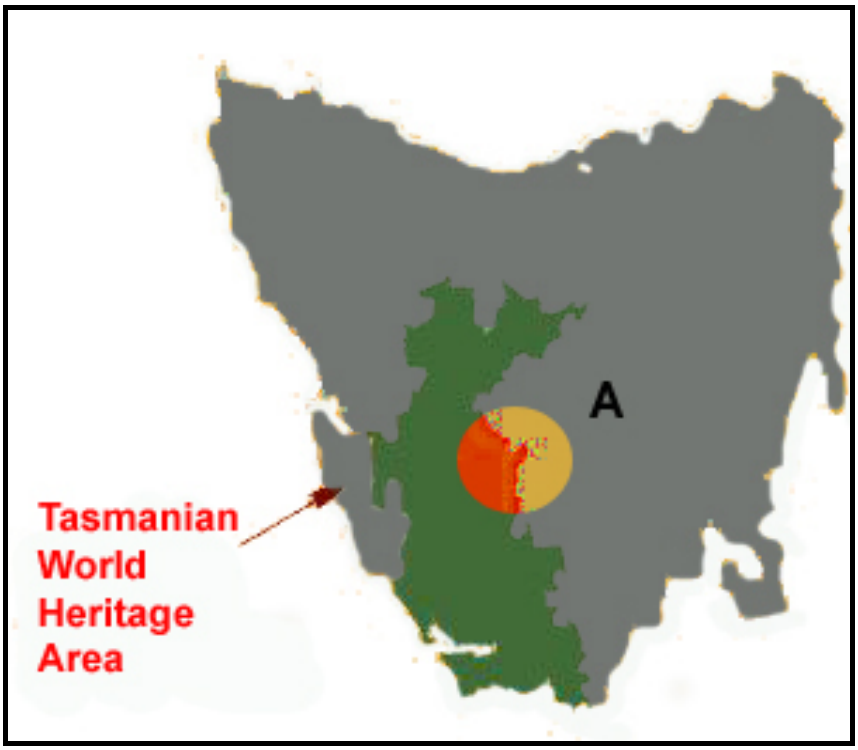

Fig. 2. *호주 태즈메이니아(Tasmania)지도에서 원으로 보여지는 곳은 0.5º (즉, 50 km)의 정확도로 수집된 레코드(A)를 나타낸다. (정밀도 값을 이용하여 판단된)수집이 예상되는 일부 지역은 태즈메이니아 세계 유산 지역과 중복된다.*

Figure 2에서 "*사용에 대한 적합성*" 개념을 사용한 예를 볼 수 있다. 특정 종 ('A'로 표시됨) 의 수집물은 위도 0.5º(즉, 50 km)의 정확도를 갖고 있다. 어떤 사람이 태즈메이니아 종의 목록을 작성하고 있고, 이 종이 태즈메이니아에서 발생하는가 알고 싶다면, 이 레코드는 이 질문에 답을 할 때 적합하다 – 이 수집물은 "사용에 적합"하고 따라서 이 목적에 높은 품질로 간주될 수 있다. 다른 한편으로, 어떤 사람이 이 종이 태즈메이니아 세계 유산 지역에 발생하는지 또는 아닌지를 알고자 한다면, 이 레코드로 이 질문에 답할 수 없다 – 그럴 수도 있고 그렇지 않을 수도 있다.이 데이터는 "이 사용에 대해 적합"하지 않고 따라서 이 목적에 낮은 품질을 가지게 된다. 데이터베이스의 위도/경도 값은 매우 정밀하고 높은 정확도를 가지고 있는 것처럼 보여지지만, 이것은 정밀도 값을 포함하지 않는 레코드를 사용하는 사용자가 잘못된 판단을 하게 할 수 있다.

유사한 경우가 비-공간적인 데이터 구성요소에서 발생하는데, 예를 들어 잘못된 동정은 데이터의 가치를 거의 없게 만들고 따라서 "목적에 대한 적합성"이 없을 수 있다. 종의 분포 (또는 이것의 생리학 또는 생태학 등)를 연구할 때 표본이나 관찰자료에 붙어있는 잘못된 이름은 그릇된 또는 거짓 결과를 산출할 수 있다.

데이터 품질은 다차원인 특징이 있고, 데이터 관리, 모델링 및 분석, 품질 관리 및 보증, 저장 그리고 표현과 관련이 있다. 크리스만(Chrisman 1991)과 스트롱(Strong *et al.* 1997)에서 각각 언급된 것처럼, 데이터 품질은 사용과 관련되어 있으며 그 사용자와 따로 분리하여 평가될 수 없다. 데이터베이스에서, 데이터는 실제적인 품질 또는 가치를 가지고 있지 않다(Dalcin 2004); 이것들은 어떤 사람이 유용한 어떤 일에 이 데이터를 사용할 때만 *구체화*되는 *잠재적*인 가치를 가지고 있다. 정보 품질은 고객을 만족시키고 고객의 요구를 충족시키는 정보의 특징과 관련되어 있다 (English 1999).

레드만(Redman, 2001)은 데이터가 사용에 적합하기 위한 다음 사항들을 제시했다: 데이터는 접근할 수 있어야 하고, 정확하고, 적시성이 있어야 하고, 완전하고, 다른 출처들과 일관되고, 관련성, 포괄성이 있어야 하며, 적절한 세부 내용을 제공하고, 읽고 해석하기 쉬워야 한다.

데이터 관리자가 고려해야 할 필요가 있는 하나의 사항은 폭 넓은 층의 고객들에게 데이터베이스의 유용성이 증가될 수 있도록 하는 것이며 (즉, 데이터베이스의 잠재적인 사용성 또는 관련성을 증가시키는 것) 그리하여 더 넓은 영역의 목적에 적합하도록 하는 것이다. 증가되는 유용성 그리고 부가적인 기능과 유용성을 추가하는데 필요로 하는 노력 간에 교환 관계(trade-off)가 발생할 수 있을 것이다. 이것은 데이터 필드들을 세분화하는 것, 지리-참조 정보의 필드를 추가하는 것 등과 같은 일을 필요로 할 수 있다.

> *데이터가 운영, 의사-결정, 계획 수립의 의도된 사용에 적합할 경우 데이터는 높은 품질을 가지고 있다고 할 수 있다*
> (Juran 1964).

품질 보증 / 품질 제어

품질 제어(quality control)와 품질 보증(quality assurance) 사이의 차이점이 항상 명확한 것은 아니다. 타울비(Taulbee 1996)는 품질 제어와 품질 보증을 구분하고 있으며 품질 목표가 충족되려면 한 쪽 없이는 다른 한 쪽도 존재 할 수 없다는 것을 강조한다. 그녀는 다음과 같이 정의한다.

- *품질 제어*는 품질을 제어하고 모니터링 하기 위해 만들어진 내부 기준, 과정, 그리고 절차에 기반하여 품질을 판단하는 것; 그리고

- *품질 보증*은 해당 과정을 외부 기준에 기반하여 품질을 판단하고 최종 제품이 이미 정의된 품질 기준에 충족되는지를 확실히 하기 위해 관련 활동과 품질 제어 과정들을 검토하는 것.

조금 더 사업-지향적인 접근 방식에서, 레드만 (Redman 2001)은 품질 보증을 다음과 같이 정의한다

> *"가능한한 최소의 비용으로 가장 중요한 고객들의 가장 중요한 요구를 충족시키면서 결함이 없는 정보가 생산될 수 있도록 설계된 그러한 활동들".*

이러한 용어들이 실제로 어떻게 적용되는지는 분명하지 않고, 대부분의 경우 이 용어들은 대체로 전반적인 데이터 품질 관리 활동을 서술하기 위해 비슷하게 사용되는 것 같다.

불확실성

불확실성은 *"완벽한 측정 장치를 이용할 수 있다면 참 값을 알 수 있는 미지의 양에 대하여 인간의 지식 또는 정보의 불완전한 정도"*로 생각할 수 있다 (Cullen and Frey 1999). 불확실성은 관찰자가 데이터를 이해하는 것에 대한 특성이며, 데이터 *그 자체* 보다는 관찰자와 더 관련이 있다. 데이터에는 항상 불확실성이 존재한다; 다른 사람들이 불확실성을 이해할 수 있도록 이것을 기록하고, 이해하고, 그리고 시각화하는데 항상 어려움이 있다. *불확실성*은 위험과 위험성 평가를 이해하는데 핵심 용어이다.

오류

오류는 데이터의 비정밀성과 이것의 부정확성 모두를 포함한다. 오류를 발생시키는 많은 인자들이 있다.

> *"오류와 불확실성에 대한 통상적인 관점은 이것들은 나쁘다는 것이다. 하지만, 이것은 항상 그렇지는 않은데, 왜냐하면 어떻게 오류와 불확실성이 발생하고, 어떻게 이것들이 관리되고 감소될 수 있을 것인가를 인지하는 것은 유용할 수 있기 때문이다... 오류와 오류 전이에 대한 올바른 이해는 적극적인 품질 제어를 도모할 수 있다"* (Burrough and McDonnell 1998).

오류는 일반적으로 무작위 또는 규칙적인 것으로 나누어진다. *무작위 오류*는 무작위한 방식으로 참 상태(true state)에서의 편차를 일컫는다. *규칙적인 오류* 또는 치우침은 값이 일정한 이동으로 인해 발생하고 지도제작 분야에서 때때로 '상대적인 정확성'을 갖는 것으로 서술된다(Chrisman 1991). '사용에 대한 적합성'을 판단할 때, 규칙적인 오류는 일부 응용 분야에는 수용될 수 있지만 다른 분야에서는 적합하지 않을 수도 있다. 하나의 사례는 서로 다른 측지학 기준점(geodetic datum)[1]를 사용하는 것이다 – 분석 전체에 사용되면 이것은 주요 문제를 일으키지 않을 수 도 있다. 그렇지만 서로 다른 출처의 데이터를 서로 다른 치우침으로 사용할 경우 문제가 발생할 것이다 – 예를 들면, 서로 다른 측지학 데이터를 사용한 데이터 출처 또는 초기 버전의 명명학적 코드를 사용하여 동정이 수행되었을 경우가 있다.

[1] 서로 다른 지리 기준점은 지구 일부 지역의 경우 실제 위치(위/경도 좌표)의 400 미터까지 규칙적인 이동을 유발시킬 수 있다.

"오류를 피할 수 없기 때문에, 이것은 데이터의 근본적인 특성으로 인지되어야 한다" (Chrisman 1991). 오류가 데이터의 형태로 포함되어 있을 경우에만 데이터의 제한점 그리고 심지어 현재 지식의 제한점에 대한 질문에 답하는 것이 가능하다. 공간, 속성, 그리고 시간의 3 차원에서의 알려진 오류를 측정하고, 계산하고, 기록하고 문서화할 필요가 있다.

검증과 정제

검증은 데이터가 부정확, 불완전, 또는 비논리적인지를 판단하는데 사용되는 과정이다. 이 과정은 형식 검사, 완전성 검사, 논리성 검사, 제한 검사, (지리, 통계, 시간 또는 환경적) 특이점 또는 다른 오류를 식별하기 위한 데이터의 검토, 그리고 주제 분야 전문가(예, 분류학 전문가)에 의한 데이터 평가를 포함할 수 있다. 이러한 과정들은 보통 의심되는 레코드를 표시하고, 문서화하고 그리고 추가적인 검사를 하는 것으로 이루어진다. 검증 검사는 또한 수용할 만한 표준, 규칙, 그리고 관례에 대한 준수 여부 검사를 포함할 수도 있다. 데이터 검증과 정제의 핵심 단계는 발견한 오류들의 근본 원인을 동정하는 것이고 그러한 오류들이 다시 발생하지 않도록 집중하는 것이다 (Redman 2001).

데이터 정제는 검증 과정 동안 동정된 데이터에서 오류를 "고치는" 과정을 일컫는다. 이 용어는 "데이터 세척(data cleansing)"과 동의어이지만 일부 사람들은 데이터 세척을 데이터 검증과 정제를 포함하는 것으로 사용한다. 데이터 정제 과정에서 데이터가 무의식중에 유실되지 않도록 하고 기존 정보에 대한 변경을 주의 깊게 수행하는 것이 중요하다. 종종 데이터베이스에 이전 (원래 데이터) 데이터와 신규 데이터 모두를 나란히 보관하는 것이 더 나은 방법인데, 만약 정제 과정에서 실수를 할 경우, 원래의 정보를 복구할 수 있기 때문이다.

최근 많은 수의 도구와 지침들이 종 데이터의 검증과 데이터 정제 과정을 지원하기 위해 개발되었다. 이러한 것은 관련 문서인 *생물종 데이터의 정제 원칙과 방법(Principles and Methods of Data Cleaning)*에서 다루어진다. 손으로 하는 데이터 정제 과정은 노동력과 시간이 많이 들며 그 자체가 오류를 발생하기 쉽다.

데이터 정제의 기본적인 틀은 다음과 같다 (after Maletic and Marcus 2000):

- 오류의 유형을 정의하고 판단
- 오류의 사례를 검색하고 동정
- 오류를 정정
- 오류 사례와 오류의 유형을 기록
- 추후에 발생 할 수 있는 오류를 줄이기 위해서 데이터 입력 절차를 수정

내용표시의 사실성

내용표시의 사실성(Truth in Labelling)은 판매 또는 제 3 자에게 이용 가능하게 할 목적으로 제품과 생산품 품질에 대한 문서화로 보통 알려져 있다. 종-발생 데이터의 경우, 이것은 보통 메타데이터로 구성되는데, 이 메타데이터는 품질, 품질 제어 절차와 방법, 그리고/또는 데이터와 관련된 측정된 품질 통계치 모두를 상세히 기록해야 한다. 내용표시의 사실성은 이것이 적합할 경우 인증과 신임에 대한 주요 기능을 한다. 대부분의 박물관과 식물표본관은 전문가의 정보 및 동정이 수행된 날짜와 관련하여 이것을 이미 수행하고 있지만 (결정 정보), 레코드 또는 관찰 및 확증되지 않은 조사 데이터의 다른 정보에 대해서는 이것은 거의 확대되어 수행되지 않고 있다.

사용자

사용자는 누구인가? 데이터의 사용자는 정보 사슬의 모든 단계에 있는 모든 사람을 포함한다 (figure 3). 1 차 종 데이터의 경우, 분류학자, 관리자, 연구자, 기술자, 수집가와 같은 생물다양성 내부 사용자 뿐만 아니라 정책 및 의사 결정자, 과학자, 농업 종사자, 산림 종사자, 그리고 원예농업 종사자, 환경 관리자, (환경과 생산) NGO 들, 의학 전문가, 약리학자, 산업 전문가, 식물원 및 동물원 관리자, (가정의 정원사를 포함하는) 일반 대중 그리고 커뮤니티 사용자들과 같은 외부의 사용자들을 포함한다. 수많은 사용자들이 종-발생 데이터를 사용하며 이런 저런 면에서 사실상 전체 커뮤니티와 관련되어 있다.

1 차 종 데이터는 광범위한 사용자 커뮤니티에 대한 고려 없이 종종 수집되었다. 전통적으로, 데이터, 특히 박물관과 식물표본관의 데이터는 분류학 또는 생물지리 연구를 위한 정보 제공이라는 주요 목적으로 수집되었다. 이것은 필수적인 과정이었고, 오늘날에 이들 기관에 연구 기금을 지원하는 제공자, 특히 정부 기관들은 자신들의 지원에 더 많은 결과를 바라고 있고, 결과적으로 데이터가 다른 추가적인 용도로 사용되어 데이터의 가치가 증가되기를 기대한다. 특히 정부 부처는 이 데이터를 환경에 관한 의사결정, 환경 관리 그리고 보전 계획에 사용하는 것을 기대하고 있으며 (Chapman and Busby 1994), 이러한 데이터의 관리자들은 이러한 사용자들 또는 이들의 요구를 무시할 수 가 없다. 우수한 피드백 절차가 갖추어진 경우, 사용자는 데이터 품질에 대한 피드백을 제공할 수 있으며, 따라서 아래에서 토론되는 것처럼 데이터 품질 사슬에서 중요한 역할을 할 수 있다.

사용자의 요구를 파악하는 것은 어렵고 힘든 일이다. 하지만 이것을 대체할 수 있는 방법은 없으며 이러한 실행에 대한 보상은 매우 크다.

데이터 품질의 원칙

> *경험에 의하면 데이터를 장기적인 자산으로 취급하고 이것을 종합적인 틀에서 관리하는 것이 상당한 비용 절감과 지속적인 가치를 산출한다* (NLWRA 2003).

데이터 품질의 원칙은 데이터 관리 절차의 모든 과정(획득, 디지털화, 저장, 분석, 표현 및 사용)에 적용될 필요가 있다. 데이터 품질 개선에 두 가지 중요한 방법이 있다 – 예방과 교정이 그것이다. 오류 예방은 데이터의 수집과 데이터베이스로의 데이터 입력, 둘 모두와 밀접한 연관이 있다. 오류 예방에 상당한 노력을 쏟을 수 있고 또는 쏟아야 하지만, 큰 데이터 집합의 경우 오류는 계속해서 존재하게 되므로(Maletic and Marcus 2000) 오류 검증과 수정을 등한시 할 수 없다.

오류 예방은 오류 탐지보다 훨씬 더 나은 것으로 여겨지는데 그 이유는 오류 탐지는 비용이 많이 들고 결코 100% 성공적이라는 것을 보장하지 않기 때문이다(Dalcin 2004). 하지만 오류 탐지는 여기에서 다루어지는 1차 종 데이터와 종-발생 데이터의 경우와 같은 오래된 수집물(Chapman and Busby 1994, English 1999, Dalcin 2004)을 다룰 때 특별히 중요한 역할을 한다.

비계획적, 비종합적, 그리고 비체계적인 "데이터 정제" 활동을 수행하기보다는 데이터 비전을 세우고, 데이터 정책을 개발하고, 그리고 데이터 전략을 실현하면서 시작하시오.

비전

기관들이 고품질 데이터에 관한 비전을 갖는 것이 중요하다. 데이터를 외부에 공개할 계획이 있는 기관들에게 이것은 특히 중요하다. 높은 데이터 품질 비전은 통상적으로 기관의 전반적인 비전을 향상시키고(Redman 2001) 기관의 운영 절차를 개선시킬 것이다. 비전을 개발할 때, 관리자들은 통합적인 관리 틀을 달성할 수 있도록 집중해야 하며, 이 틀 안에서 데이터를 관리하고 이 데이터가 고품질 정보 산출물로 바꾸어질 수 있도록 지도력, 구성원, 컴퓨터 하드웨어, 소프트웨어 애플리케이션, 품질 제어 및 데이터가 적합한 도구, 지침서, 그리고 표준과 함께 엮어져야 한다 (NLWRA 2003).

데이터 품질 비전:

- 장기적으로 기관의 데이터 및 정보의 필요성 그리고 이것과 기관의 장기적인 성공과의 관계를 고찰할 수 있도록 한다,
- 올바른 방향 – 즉, 품질쪽으로 활동을 촉진한다,
- 기관 내외부에 의사결정을 위한 견고한 기반을 제공한다,
- 데이터와 정보가 기관의 핵심 자산이라는 인식을 공식화한다,
- 기관이 가지고 있는 데이터와 정보의 사용을 최대화하고, 중복을 방지하고, 협력 관계를 촉진시키고, 공평한 접근성을 향상시킨다, 그리고
- 통합과 상호운용성을 최대화한다.

정책

비전뿐만 아니라, 기관은 비전을 실현시키기 위한 정책이 필요하다. 견고한 데이터 품질 정책의 개발은 다음과 같은 효과를 유발할 것이다:

- 기관이 품질에 대해 더욱 광범위하게 고려할 수 있도록 하고 매일 매일의 실무 활동을 재점검할 수 있도록 한다,
- 데이터 관리의 과정을 정형화한다,
- 다음 사항과 관련하여 기관의 목적을 분명히 하는데 도움을 준다
    - 비용 절감,
    - 데이터 품질 개선,
    - 고객 서비스와 관계를 개선, 그리고
    - 의사결정 과정을 개선,
- 사용자에게 기관에서 발생하는 데이터를 접근하고 사용할 때 신뢰성과 안정성을 제공한다,
- (데이터 제공자와 사용자인) 기관의 고객들과의 관계 및 의사소통 개선,
- 더 큰 커뮤니티에서 기관의 이미지 개선, 그리고
- 최선의 실행사례 표적치에 접근함에 따라 더 나은 재정 지원을 받을 수 있는 기회 증가.

전략

규모가 큰 기관들이 소장하고 있는 엄청난 양의 데이터를 고려하면 데이터를 기록하고 검사하는데 전략을 개발할 필요가 있다 (아래 *우선순위선정* 또한 참고). (데이터 입력과 품질 제어를 위해) 따라야 할 좋은 전략은 단기, 중기, 장기 목표를 세우는 것이다. 예를 들면 (Chapman and Busby 1994):

- **단기.** 6-12 개월에 걸쳐 모으고 검사할 수 있는 데이터 (이미 데이터베이스에 있는 데이터와 품질 검사의 정도가 덜 요구되는 신규 데이터가 보통 포함된다).
- **중기.** 단지 적은 자원의 투입으로 18 개월 정도의 기간에 데이터베이스에 입력될 수 있는 데이터와 간단하고 내부에서 가지고 있는 방법으로 품질에 대해 검사할 수 있는 데이터.
- **장기.** 공동 협력, 더욱 복잡한 점검 방법 등을 이용하여 조금 더 오랜 시간에 걸쳐 입력되고 검사되어야 할 데이터. 아래와 같은 방법들을 통하여 수집물에 대해 체계적으로 작업을 할 수도 있다:
    - 최근에 개정되거나 기관 내의 분류 연구가 진행 중인 분류군.
    - 중요한 수집물 (확증 표본, 특별 참고 수집물 등)
    - 핵심 분류군 (중요한 과(families), 국가적으로 중대한 분류군, 목록화된 멸종위기 분류군, 생태학/환경학적으로 중요한 분류군).
    - 핵심 지리 영역(예, 원산지 국가와 데이터 공유를 목적으로 하는 개발도상 국가 또는 해당 기관에 중요한 지리 영역)의 분류군.
    - 다른 기관들과의 공동 협약 일부를 구성하는 분류군 (예, 일련의 기관들이 동일한 분류군을 데이터베이스화 하는 협약).
    - 처음부터 끝까지 수집물을 체계적으로 이동시키는 것.
    - 축적된 수집물보다는 최근의 수집물.

전략에 포함되어야 할 좋은 데이터 관리 원칙 가운데 일부는 다음과 같다 (after NLWRA 2003):

- 정보 관리 체계를 재발명 하지 않기
- 데이터 수집과 품질 제어 절차에서 효율적인 부분을 찾기
- 될 수 있으면, 데이터, 정보, 그리고 도구를 공유하기
- 기존의 표준을 이용하거나 다른 기관과 함께 새롭고, 견고한 표준을 개발하기

- 네트워크와 협력 관계의 구축을 장려하기
- 데이터 수집과 관리의 건전한 사업 사례 제시
- 데이터 수집과 품질 제어에서 중복의 감소
- 즉각적인 사용을 넘어 고찰하고 사용자의 요구사항을 조사하기
- 올바른 문서화와 메타데이터 절차가 구현될 수 있도록 보장하기.

예방이 치료보다 더욱 낫다

수집물을 데이터베이스에 입력하는 비용은 상당할 수 있지만(Armstrong 1992) 추후에 이 데이터를 점검하고 교정하는 비용을 고려하면 이것은 단지 일부분에 지나지 않는다. 나중에 오류를 고치는 것보다 오류를 예방하는 것이 더욱 낫고(Redman 2001) 이 방법이 비용이 더 적게 든다. 과거에 처리된 데이터를 정정한다는 것은 부정확한 데이터가 정정되기 전에 이미 많은 분석에 사용되었을 수도 있다는 것을 의미하며, 이것은 품질이 낮은 데이터로 인한 비효과적인 의사결정과 그 분석을 다시 수행해야 하는데 드는 비용 등의 비생산적인 결과를 초래할 수 있다.

하지만 오류의 예방은 데이터베이스 내에 이미 존재하는 오류에 대해서는 아무 효과가 없기 때문에 데이터 검증과 정제는 데이터 품질 과정의 중요한 부분을 차지한다. 정제 과정은 이미 데이터베이스에 저장되어 있는 오류의 원인을 파악하는데 중요하며 그러한 오류가 다시 반복되지 않도록 하는 절차가 개발될 수 있도록 하여야 한다. 정제는 따로 분리되어 시행되지 않아야 한다. 분리되어 시행될 경우 이 문제는 결코 사라지지 않게 된다. 데이터 정제와 오류 예방, 이 두 가지 업무는 동시에 시행되어야 한다. 데이터를 먼저 정제하는 것을 결정하고 예방을 나중에 걱정하게 되면 보통 오류 예방은 만족할 만하게 수행되지 않으며, 이 사이 더욱 더 많은 오류들이 데이터베이스에 추가되어지는 결과를 초래한다.

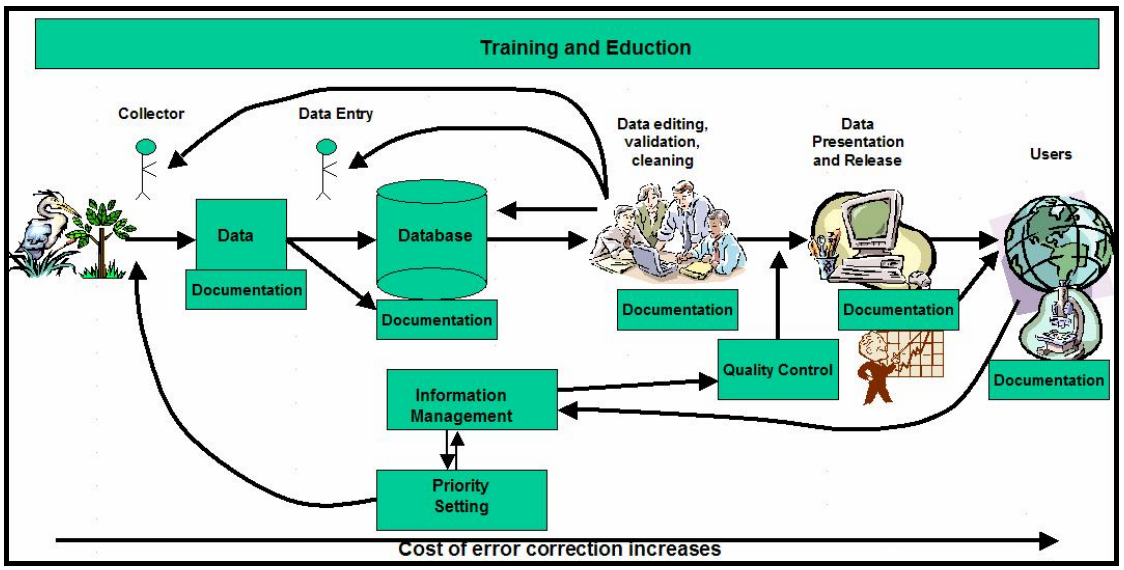

Fig. 3. *오류 수정의 비용이 사슬을 따라 진행될수록 증가한다는 것을 보여주는 정보관리사슬. 올바른 교육, 훈련, 그리고 문서화가 모든 단계에 필수적이다.*

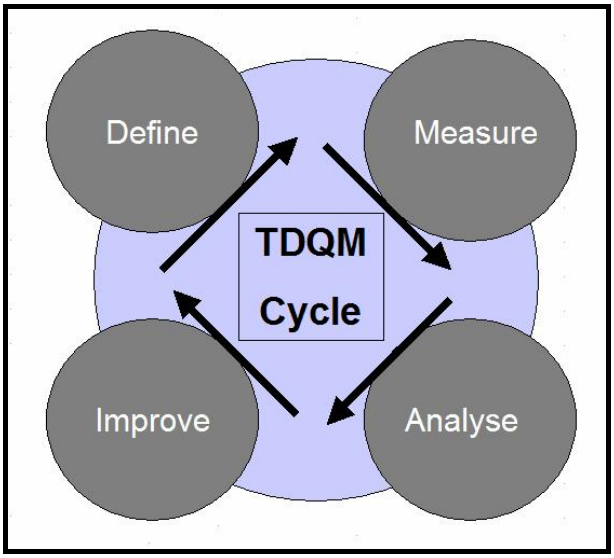

Fig. 4. *데이터 관리 과정의 순환적인 성질을 보여주는 전체 데이터 품질 관리 공정 (after Wang 1998).*

데이터의 관리자와 소유자들은 자신들이 소유하고 있는 데이터의 품질에 대한 대부분의 책임이 있다. 그럼에도 불구하고, 데이터를 공급하는 사람들과 이 데이터를 사용하는 사람들도 또한 책임이 있다.

데이터 품질에 대한 책임은 데이터를 생성한 사람에게 지워야 한다. 이것이 가능하지 않다면, 되도록이면 데이터 생성과 근접한 사람에게 책임을 지워야 한다 (Redman 2001)

1 차적인 책임은 수집자에게 있다

데이터 품질 관리의 1 차적인 책임은 데이터의 수집자에게 있다. 수집자는 다음 사항들을 명확히 하여야 한다:

- 레이블 정보가 정확하다,
- 레이블 정보가 정확하게 기록되고 문서화되었다,
- 소재지 정보는 최대로 정확하고, 정확도와 정밀도 모두가 문서화되었다,
- 수집 방법들이 자세히 기록되었다,
- 레이블과 항목 정보는 명확하고 모호하지 않다, 그리고
- 레이블 정보는 데이터 입력자가 읽을 수 있을 정도로 명확하고 분명하다.

레이블이나 수집자의 노트에 있는 정보가 명확하고 정확하지 않으면 나중에 이것을 정정하는 것은 극히 어렵다. 이것은 분류학적인 데이터일 경우 덜 중요할 수 있는데, 확증 수집물이 보관되어 있어, 보통 나중에 전문가들이 점검하고 또는 점검할 수 있는 경우에 그러하다.

장소에 대한 정보와 부가 정보가 과거에 행해졌던 것과 같이 저녁 늦게 또는 실험실에 돌아와서 기록되기보다는 수집 또는 관찰할 때 기록되는 것이 또한 중요하다.

대부분의 데이터는 "공급자"에 의해 기관으로 들어오고, 올바른 데이터 수집 실무사례를 개발하는 것이 나중에 오류를 정정하는 것보다 훨씬 쉽다.

관리자 또는 큐레이터에게 핵심 또는 장기적인 책임이 있다

(박물관, 표본관, 대학교, 보전 기관, NGO 또는 일반 개인이 소장한) 데이터의 관리자는 데이터 보유의 책임을 맡고 있는 동안 데이터의 품질을 관리하고 개선하는 장기적인 책임을 가지고 있다 (예를 들어, Olivieri *et al.* 1995 p 623 의 데이터 관리의 책임 항목들 참고). 관리 조직이 기관 내부에서 데이터 품질을 관리하기 위한 아주 중요한 책임을 맡는 것이 중요하지만, 또한 해당 기관의 모든 구성원이 그 기관이 소장하고 있는 데이터의 품질에 일부 책임이 있다는 것을 인지하는 것과 같은 데이터 품질에 대한 문화를 가지고 있는 것도 중요하다. 다음의 사항을 명확히 하는 것이 관리자의 책임이다:

- 수집가의 일지에서 데이터베이스에 데이터가 올바르고 정확하게 입력되고 있다,
- 데이터 기록 동안에 품질 제어 절차가 실행되고 적용되고 있다,
- 데이터와 데이터 품질이 적절하고 정확하게 문서화되고 있다,
- 데이터에 대한 검증 검사가 일상적으로 수행되고 있다,
- 수행된 검증 검사는 상세히 문서화되고 있다,
- 적절한 방법으로 데이터가 저장되고 보관되고 있다 (아래 저장에 대한 내용 참고),
- 이전 버전이 시스템적으로 저장되어 비교가 가능하고 "정제되지 않은" 데이터의 복구가 가능하다,
- 데이터의 일관성이 유지되고 있다,
- 사용자가 "이용에 대한 적합성"을 판단할 수 있도록 데이터가 문서와 함께 적시에 그리고 정확한 방법으로 이용 가능하다,
- 사생활, 지적재산권, 저작권, 그리고 전통/토속 소유자의 민감성에 대한 관리자의 책임성이 유지된다,
- 데이터 사용에 관한 조건과 함께 사용에 대한 임의의 제한사항 그리고 알려진 데이터의 부적합 분야를 관리하고 이용 가능하게 한다,
- 데이터에 관련된 모든 법적 사항을 존중하고 준수한다,
- 데이터 품질에 관한 사용자의 피드백이 제 때에 처리된다,
- 항상 데이터의 품질이 최고로 유지된다,
- 알려진 모든 오류에 대해서 상세히 문서화를 하고 사용자에게 알린다.

데이터 소유권과 관리권은 데이터에 대한 접근을 관리하고 제어하는 권리를 부여할 뿐만 아니라, 이것의 관리, 품질 제어 그리고 유지에 대한 책임을 또한 부여한다. 관리자는 또한 미래의 세대가 데이터를 사용할 수 있도록 하는 도덕적인 책임이 있다

사용자의 책임

데이터의 사용자 또한 데이터의 품질에 책임이 있다. 사용자는 우연히 발견하게 되는 오류, 데이터의 문서에 있는 오류, 그리고 향후 기록될 필요가 있는 추가적인 정보 등을 관리자에게 피드백 할 필요가 있다. 해당 데이터를 다른 데이터의 문맥에서 볼 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.

기관에서 데이터를 수집하는 목적에 따라서, 데이터 수집 및 검증과 관련한 향후 우선 순위 선정을 할 때 사용자는 중요한 기여와 지원을 할 수 도 있다 (Olivieri *et al.* 1995).

사용자는 또한 자신의 목적에 데이터가 적합한지를 판단해야 할 책임이 있으며, 부적합한 방법으로 데이터를 사용하지 않아야 한다.

사용자와 수집자는 관리자가 수집물에 있는 데이터의 품질을 유지하는데 지원하는 중요한 역할을 하고, 둘 모두는 데이터가 될 수 있으면 최상의 품질인 상태인 것에 관심을 가지고 있다.

협력 관계 구축

데이터 품질 유지를 위한 협력 관계 구축은 보상이 되고 비용 절약의 수단이 될 수 있다. 이것은 특히 중복 레코드가 종종 여러 개의 박물관간에 배포되는 박물관과 식물표본관에 대해 그러하다. 많은 도서관 커뮤니티는 도서관 자료의 목록화를 향상시키기 위해 협업과 협력 관계를 형성(Library of Congress 2004)하고 있으며 박물관과 식물표본관들도 비슷한 방법으로 운영될 수 있을 것이다. 이러한 협력 관계와 협업 환경은 아래 사항들과 함께 개발될 수 있을 것이다:

- 중요 데이터 수집자 (정보 유통 향상 목적으로 – 예를 들어 표준 데이터 수집물과 보고 형식, GPS 규정 등을 개발할 수 있음),
- 비슷한 데이터를 소장한 다른 기관들 (예, 중복 수집물),
- 비슷한 수준의 데이터 품질에 대한 수요를 가지면서 데이터 품질 제어 방법, 도구, 표준 및 절차를 개발하려고 하는 다른 기관들,
- 여러 데이터 제공자의 정보를 정렬하고 배포하는 역할을 하는 핵심 데이터 유통기관 (예, GBIF),
- 데이터의 사용자 (특히 분석 동안 또는 분석 전, 데이터에 대한 검증 테스트를 수행할 수 있는 사람들), 그리고
- 데이터 관리, 데이터 흐름 그리고 데이터 품질 기술에 대한 방법론을 향상시킬 수 있는 통계학자와 데이터 감리자.

데이터 품질을 취급하는 곳은 당신의 기관만 있는 것은 아니다.

우선순위선정

가능한한 최단 시간에 최대한의 사용자들에게 데이터가 최고 가치의 것이 될 수 있게 하기 위해서는, 이 데이터의 획득 그리고/또는 검증에 대해 우선순위를 선정해야 할 필요가 있을 것이다 (아래 *완전성*에 대한 내용을 또한 참고). 이것을 위해서 다음 사항이 필요할 수 있다:

- 가장 중요한 데이터에 먼저 집중,
- 개별 단위에 집중 (분류, 지리 등),
- 기준 표본과 중요한 확증자료에 대해 우선순위를 선정
- 사용되지 않거나 데이터와 품질을 보증 할 수 없는 데이터를 무시 (즉, 지리참조연산 정보가 부실한 레코드 – 그러나 지리참조연산이 부실하지만 역사성이 있는 데이터의 중요성은 고려할 것),

- 가장 광범위한 가치가 있고, 다수의 사용자에게 최대 이익을 주며, 그리고 가장 다양한 이용 가치가 있는 데이터를 고려,
- 최소의 비용으로 가장 많은 데이터가 정제될 수 있는 그러한 분야에 대해 업무를 진행.

모든 데이터가 동등하게 생성되지는 않는다, 그러므로, 가장 중요한 것에 집중하고, 데이터 정제가 요구될 경우, 이것이 ***결코 반복되지 않도록*** *신경을 써야 한다.*

완전성

기관은 데이터 (또는 우선순위선정을 통해 데이터의 개별 단위 – 예, 분류학적 범주, 지역, 등)의 완전성을 추구해야 하며, 그리하여 모든 판독이 가능한 레코드들이 데이터의 편집에 사용되어야 한다. 불완전한 데이터에 대해 수행된 분석은 포괄적이지 않으므로, 불완전한 많은 데이터를 이용할 수 있게 하는 것보다 분리된 개별 단위에 대해 데이터를 완전히 한 후 이것을 이용 가능하게 하는 것이 더욱 낫다. 데이터의 완전성을 문서화하는 정책과 함께, 누락된 데이터의 문턱 값을 정의하는 누락 데이터 정책을 수립하는 것이 또한 중요하다 (아래 *문서화* 부분 참고).

통용성(currency)과 적시성(timeliness)

데이터의 적시성 또는 통용성과 관련되어 주요한 세가지 요소가 있다:

- 데이터가 어느 기간 동안에 수집되었는가?
- 현실 세계의 변화에 맞게 데이터가 언제 갱신되었는가?
- 데이터가 얼마나 오랫동안 통용될 수 있을 것인가?

데이터 통용성은 사용자들이 흔히 제기하는 문제이다. 대부분의 데이터 관리자들은 데이터가 처음에 수집되거나 조사된 기간을 가리키는 뜻으로 통용성의 단어를 사용하는 경향이 있다. 수집과 출판 사이의 시간적 차이(생물학 데이터의 경우 이것이 매우 오랜 시간일 수 있다)때문에, 출판된 정보는 “현재”의 것이 아닌 “과거”의 것을 나타낸다. 대부분의 생물다양성 데이터 사용자들은 이러한 사실을 알고 있고, 이것이 이러한 데이터 종류의 가치 중 하나이며, 대부분의 다른 종류의 데이터와 명확히 구분되게 하는 것이다.

데이터 품질 관리 용어에서, 통용성은 데이터에 대한 “통용(use-by)” 기간(때때로 적시성이라고도 불림)의 문맥에서 자주 사용되며, 데이터의 최종 검사 또는 갱신된 때와 관련될 수 있을 것이다. 이것은 데이터에 따라붙는 이름의 경우에 특히 관련되어 있다. 이러한 이름이 언제 최종 갱신되었고 이것들이 최신의 분류 정보와 일치하는가? 현대적인 분류 명명법을 따를 경우, 종(species)이 몇 개의 더 작은 분류군으로 나뉘어질 경우, 이러한 작은 분류군 중의 하나는 광범위한 종의 이름을 보유하게 된다. 사용되는 이름이 광의 또는 협의의 개념을 가리키는 것인지를 구분하는 것이 사용자에게 중요할 수 있다. 통용성은 식료품에 대해 사용되는 것과 같은 “통용” 기간과 동등한 것으로 사용될 수 있으며, 이 기간을 넘을 경우 관리자는 레코드에 부착된 명명학적 정보를 보장하지 않게 된다.

많은 데이터집합에 대해 적시성과 통용성이 적용될 수 없거나 포함 또는 유지하는 것이 가능하지 않은 경우가 또한 있을 수 있다. 예를 들어, 이것은 규모가 큰 박물관과 식물표본관의 수집물에 적용될 수 있다. 한편으로, 이것은 확증자료가 존재하지 않거나 또는 최근의 분류학적 개정 이후 이 데이터에 갱신이 반영되지 않은 경우와 같이 관찰 또는 조사 데이터에 중요할 수 있다. 이것은 또한 수많은 외부 기증 기관에 의해 결합된 수집물을

포함하는 2 차 수집물의 경우에 또한 중요한 사항이다. 많은 개발도상 국가의 기관들이 자신의 데이터를 GBIF 포탈에 공유할 목적으로 서비스 기관에 이용 가능하게 하지만 데이터베이스로부터 실시간에 보여지지 않는 경우가 그 예이다.

갱신 빈도

데이터 집합 내에서 데이터의 갱신 빈도는 통용성 및 적시성과 관련되어 있으며 정형화되고 문서화될 필요가 있다. 이것은 정정된 데이터의 배포 빈도 뿐만 아니라 새로운 데이터의 추가도 포함한다. 이러한 것 모두는 데이터의 품질에 영향을 미치기 때문에, 사용자에게 중요한 정보가 된다. 사용자는 데이터 집합이 이제 막 갱신되고 개선되려고 할 경우 데이터집합을 애써 다운로드 하거나 획득하려고 하지 않을 것이다.

일관성

레드만(Redman, 1996)은 일관성의 두 가지 측면을 인지하였다. : *의미적 일관성(semantic consistency)* – 데이터에 대한 관점이 분명하고, 모호하지 않으며, 일관성이 있어야 한다. 그리고 *구조적 일관성(structural consistency)*: 이것 안에서, 개체 타입과 특성은 동일한 기본 구조와 형식을 가져야 한다. 의미적 일관성의 간단한 예는 데이터는 항상 같은 문항에 있어 찾기 쉬운 경우이다 – 예를 들어, 종이하 순위와 종이하 이름에 대한 분리된 필드가 있어서 종이하 이름 필드는 단지 이름 또는 종소명만을 포함하도록 하는 것이고 (Table 1 참고), 때로는 단지 이름만을 포함하고 다른 때에는 이름 뒤에 “var.” 또는 “subsp.” 접미사를 포함하는 등 섞어서 사용하지 않도록 명확히 하는 것이다 (Table 2 참고)

속(Genus)	종(species)	종이하(Infraspecies)
Eucalyptus	globulus	subsp. Bicostata
Eucalyptus	globulus	Bicostata

Table 1. *종 이하 필드에서 의미적인 비일관성을 보여주고 있다.*

속(Genus)	종(Species)	종이하 순위 (Infrasp_rank)	종이하 (Infraspecies)
Eucalyptus	globulus	subsp.	bicostata
Eucalyptus	globulus		bicostata

Table 2. *여분의 필드(“Infrasp_rank”)를 추가하여 종 이하 필드에서 의미적인 일관성을 보여주고 있다.*

관계형 데이터베이스를 올바르게 설계하면 이러한 문제들이 발생하지 않게 할 수 있지만, 수집물 기관에서 사용하는 기존의 많은 데이터베이스들은 잘 설계되어 있지 않다.

구조적 일관성은 필드 내에 일관성이 있을 경우 발생하는데, 예를 들어, “종이하 순위 (Infrasp_rank)” 필드 (Table 2)에서 아종은 항상 동일한 방식으로 기록되도록 하는 것이다 – 때로는 “sbusp.”, 다른 때는 “ssp.”, “subspecies”, “subspec.”, “sspecies” 등과 같이 기록되지 않게 하는 것이다. 이것은 잘 정의된 구조적인 속성 및 데이터베이스의 올바른 설계를 통해 피할 수 있다.

방법과 문서화 두 가지 모두에서 일관성이 중요한데, 이것으로 사용자는 어떠한 테스트가 수행되었고 어떻게, 어디에서 이 정보를 찾고, 여러 중요한 정보를 어떻게 해석해야 하는지를 알 수 있기 때문이다. 하지만 일관성은 유연성을 고려하여 균형을 맞출 필요가 있다 (Redman 2001).

유연성

데이터 관리자는 자신의 데이터 품질 제어 방법에서 유연성을 가질 필요가 있는데, 그 이유는 많은 생물학적 데이터가 본질적으로 유사하지만, 서로 다른 지역의 데이터 (예를 들어, 해당 데이터를 검사하기 위해 어떤 관련 데이터집합을 이용할 수 있는가), 서로 다른 분류학적 그룹 (수생 대 육상 개체, 등), 또는 서로 다른 데이터 획득 방법 (관찰 또는 조사 레코드 대 확증된 박물관 수집물 등)의 경우에 데이터 품질에 대한 서로 다른 접근 방식이 적합할 수 있기 때문이다.

분류학적인 견해는 실제로는 가설이며, 서로 다른 (유효한) 분류학적 견해 (가정)로 서로 다른 분류학자는 동일한 개체를 상이하게 분류할 수 있으며, 그리하여 하나 또는 그 이상의 대체 이름이 생겨날 수 있다 – 이것들 중의 각각은 동등하게 유효한 것일 수 있다 (Pullan *et al.* 2000, Knapp *et al.* 2004). 하나의 사례는 두 명의 분류학자가 상이한 속(genera) 내에서 분류군의 위치에 대해 동의하지 않는 것이다 – 예를 들어, 일부 분류학자들은 특정 종을 *Eucalyptus* 속에 위치시키지만, 다른 분류학자들은 이것들이 *Corymbia* 속에 속한다고 믿고 있다. 실제로, 특히 동물학에서는, 가장 최근의 개정자에 대한 견해에 대해 반대할 타당한 논리가 없을 경우, 이 견해가 수용된다.

유연성은 새로운 또는 상이한 요구를 수용할 수 있도록 하나의 견해가 변경될 수 있는 여지를 제공한다. TDWG(Taxonomic Databases Working Group)[2] 및 다른 단체들의 최근 활동은 이러한 대체 개념들이 표현될 수 있도록 데이터베이스의 구조에 대해 초점을 맞추고 있으며(Berendsohn 1997), 표면상으로 이러한 유연성의 성질이 품질을 저하시키는 것처럼 보이지만 실제 이것은 사용자가 사용에 대한 적합성을 판단할 때 더 큰 유연성을 가질 수 있도록 하며 따라서 그러한 경우 인지되는 품질을 증가시킬 수 있게 된다.

투명성

투명성은 데이터를 사용하는 사람들이 평가를 할 때 신뢰성을 향상시킬 수 있기 때문에 중요하다. 투명성은 오류가 숨겨지기 보다는 이것이 발견되고 보고되도록 하는 것이며, 검증과 품질 제어 절차가 문서화되어 이것을 이용할 수 있게 하는 것이며, 그리고 피드백 체계가 공개적으로 운영되며 권장되는 것을 의미한다.

투명성이 중요한 하나의 사례는 수집 방법론의 문서화가 있다 (특히 관찰 및 조사 데이터의 경우에 중요하다). 이것은 사용자가 자신의 특정 이용에 해당 데이터가 적합한지를 결정할 수 있게 도움을 준다.

성능 측정치와 목표치

성능 측정치는 품질 제어 절차에 중요한 부분이며 개별 데이터 사용자들이 데이터의 정확성이나 품질 수준에 대해서 신뢰할 수 있도록 하는 역할을 한다. 성능 측정치에는 데이터에 대한 통계적인 검사 (예를 들면, 모든 레코드의 95%는 보고된 위치에서 1,000 미터 내에 있다), 품질 제어의 수준 (예를 들어, 모든 레코드의 65%는 자격 있는 분류학자에 의해 지난 5 년 동안 검사되었다; 90%는 자격 있는 분류학자에 의해 지난 10 년 동안 검사되었다), 완전성(모든 10-분 격자가 샘플링 되었다) 등등이 포함될 수 있다.

성능 측정치는 데이터 품질을 수량화하는데 도움이 된다. 장점은 다음과 같다:

- 해당 기관은 일부 데이터가 (문서화되어) 품질이 높다는 것을 스스로 확신할 수 있다,

[2] http://www.tdwg.org/

- 이것들은 전반적인 데이터 관리와 중복 감소에 도움을 준다, 그리고
- 이것들은 데이터 품질 사슬의 다양한 측면을 조정하는데 도움이 되며, 그리하여 서로 다른 작업자가 이것들을 수행할 수 있도록 이러한 것들이 조직화될 수 있다.

데이터 품질 수준을 측정하기 전에, 첫째 사용자들이 이 결과를 어떻게 이용할지 생각하고 다음으로 이 결과가 가장 효율적으로 사용될 수 있도록 구조화하시오.

데이터 정제

데이터 정제의 원칙은 관련 문서인 *데이터 정제의 원칙과 방법(Principles and Methods of Data Cleaning)*에서 다루어진다. 말레틱과 마르쿠스(Maletic and Marcus 2000)의 자료에서 수정한 데이터 정제에 대한 일반적인 틀을 설명하는 것으로 충분할 것이다:

- 오류의 유형을 정의하고 파악한다
- 오류 사례를 찾고 파악한다
- 오류를 정정한다
- 오류 사례와 유형을 문서화한다
- 향후 비슷한 오류의 발생을 줄이기 위해 데이터 입력 절차를 정정한다.

데이터 정제 도구의 분명하면서도 단순함에 현혹되면 안 된다. 이것은 단기적으로 유용하고 도움이 되지만, 장기적으로 오류 예방을 대체할 수 있는 것은 없다.

특이점

(지리, 통계, 그리고 환경적인) 특이점 탐지는 공간 데이터에서 가능한 오류를 찾는데 가장 유용한 검사 방법 중의 하나이다. 하지만 검증 검사에서 데이터가 통계적인 특이점이라는 것 때문에 주의 없이 데이터를 삭제하지 않는 것이 중요하다. 환경 데이터는 통계적으로 특이점으로 보이지만 실제로는 유효한 레코드인 것이 많이 있다. 이것은 역사적인 진화 형태, 변화하는 기후 체계, 인간 활동의 잔재 때문일 수 있다. 특이점을 주의 없이 배제하는 것은 데이터 집합에서 유용한 레코드를 삭제하고 향후 분석을 한쪽 방향으로 치우치게 할 수 있다.

한편 사용자는 좋은 레코드에 대한 유효성 확신이 서지 않을 경우, 특이점을 분석에서 제외시킬 수 있다. 따라서 특이점 동정은 데이터 관리자가 가능한 오류를 동정하는데 도움을 줄 뿐만 아니라, 사용자에게 개별 데이터 레코드가 분석 목적으로 사용에 적합한지 아닌지를 결정할 때 도움을 준다.

특이점 탐지는 가치 있는 검증 방법일 수 있지만, 모든 특이점이 오류인 것은 아니다.

개선을 위한 목표 설정

간단하고 정량화하기 쉬운 목표를 세우는 것은 데이터 품질의 빠른 개선을 이끌 수 있다. 지리 정보가 부실하게 기록되는 신규 레코드의 비율을 2 년 동안 6 개월마다 반으로 줄이는 것과 같은 목표는 94%의 전체 오류 비율을 감소시킬 수 있다 (Redman 2001). 이와 같은 목표는 다음 사항에 초점을 두어야 한다:

- 분명하고 공격적인 기간 설정,
- 실제 품질 가치보다는 개선 비율,
- ('부실한 지리참조연산정보' 등에 대한) 명확한 정의,
- 간단하고 달성이 가능한 목표.

데이터 입력과 검증 기술을 향상시킴으로써 데이터 정제에 필요한 시간을 1 년에서 반으로 줄이는 것과 함께 조금 더 장기적인 목표를 세울 수도 있다.

성능 목표치는 한 조직이 품질 검사와 검증의 일관된 수준을 유지하는 좋은 방법이다 – 예를 들어, 모든 레코드의 95%가 수신된 후 6 개월이내에 문서화되고 검증된다.

감사(Auditability)

관리자가 어느 데이터가 언제 검사되었는지 아는 것은 중요하다. 이것은 중복을 줄이고 데이터 레코드가 손상되거나 누락되는 것을 방지하는데 도움이 된다. 이것을 수행하는 가장 좋은 방법은 문서화된 검증 감사 일지를 유지, 관리하는 것이다.

편집 제어(Edit controls)

편집 제어는 특정 필드에 허용되는 값을 결정하는 업무 규칙과 관련된다. 예를 들어, 월(month) 필드의 값은 1 과 12 사이어야 하고, 일(day)에 대한 값은 1 과 31 사이어야 하며 월에 따라 최대 값은 달라진다. 1 개의 규칙이 하나의 필드(예, 위의 월)에 적용되고, 2 개의 규칙이 2 개의 필드에 적용된다 (예, 월과 일의 조합).

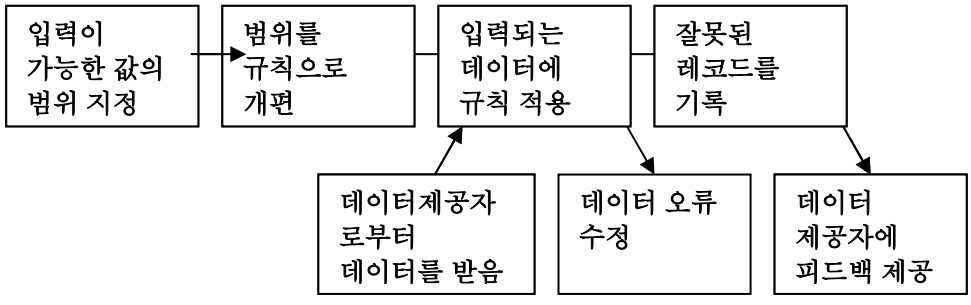

Fig. 5. *편집 제어의 사용 (Redman 2001 에서 수정).*

두 번째 예는 좌표 데이터이다. 간단한 범위 테스트(데이터가 위도와 경도 좌표인 경우)는 위도가 0 에서 90 도 사이, 분과 초가 0 에서 60 사이 등을 검사한다. 하지만 UTM 데이터일 경우, 이것은 더 복잡해진다. 아주 종종 하나의 UTM 지역(Zone)에 속하는 작은 구역의 데이터를 포함하는 데이터베이스는 이 지역(Zone)을 포함하지 않는다. 이 데이터가 다른 구역의 데이터와 결합되어 사용되지 않는다면 이것은 꽤 수용할 수 있는 것처럼 보인다. 하지만 이 데이터를 결합하게 되면, 이 데이터의 유용성은 크게 떨어진다. 따라서 편집 제어 장치는 적합한 지역이 항상 포함될 수 있도록 확실히 할 필요가 있다.

데이터의 중복과 재작업을 최소화하라

실무 경험에 의하면 정보 사슬 관리(figure 3 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.

원본 (또는 있는 그대로) 데이터의 보존

편집과 데이터 정제 과정에서 수집가가 기록하거나 또는 큐레이터가 나중에 입력한 원본 데이터가 유실되지 않도록 하는 것이 중요하다. 데이터 정제 과정 동안 데이터베이스에 행해진 변경 사항은 원본 정보가 유지되면서 추가 정보로 더해져야 한다. 한번 정보가 삭제되면, 복구하는 것이 매우 어렵거나 심지어 불가능할 수 있다. 수집자와 장소 정보의 경우 이것은 특히 중요할 수 있다. 나중에 큐레이터에게 오류인 것처럼 보이는 것이 실제 오류가 아닐 수도 있다. 하나의 장소 이름을 다른 것으로 변경하는 것 (예, 체코슬로바키아에서 체코 공화국으로 변경)은 단지 그 이름 뿐만 아니라 그 범위도 변경시킨다. “정정된” 버전이 아니라 원래 어떠한 기록이 있었는가를 아는 것이 나중에 중요할 수 있다. 아래 보관(Archiving)에 대한 내용을 또한 참고하시오.

범주화는 데이터와 품질의 손실로 이어질 수 있다

데이터의 범주화는 종종 데이터의 손실을 이끌 수 있고 따라서 전반적인 데이터의 품질 저하로 이어질 수 있다. 하나의 사례는 상세한 장소 정보(와 지리-참조연산 정보)를 가진 데이터의 수집물을 격자 셀 기반으로 이 데이터를 저장하는 것이다. 데이터를 가장 상세한 수준으로 저장한 후, 특정한 사용이 요구될 경우 그 결과에 맞게 데이터를 범주화하는 것이 대부분 더 좋다. 사용자가 10 X 10 분 격자상에 존재/부재 지도를 만들 필요가 있다면, 점으로 저장된 데이터로 이것을 만드는 것은 쉽지만, 이 데이터가 데이터베이스에 격자 셀로 저장된 경우, 더 상세한 규모로 데이터를 처리하는 것은 불가능하다. 또한 서로 다른 격자 규모 또는 원점을 사용하여 범주화된 데이터를 결합하는 것은 극히 어렵다 (아마도 불가능할 것이다). 서술 데이터의 경우도 마찬가지이다 – 만약 데이터가 키를 필요로 하는 상태(예, > 6 m = tree; <6m, = shrub)로 범주화되어 있고, 나무의 정의로 6 미터가 아닌 4 미터를 사용하고 있는 다른 출처에서 새로운 데이터가 입수되는 경우, 4 미터와 6 미터 사이의 것은 어떻게 할 것인가. 데이터를 정확한 미터로 저장하고 이것이 나무인지 또는 관목인지를 나중에 걱정하는 것이 훨씬 낫다.

지리코드의 정확도를 저장할 때 이런 경우가 자주 발생한다. 필자는 지리코드 정확도를 미터로 저장하기를 항상 권고하고 있지만, 많은 데이터베이스는 이 정보를 범주화하여

저장한다. (<10m, 10-100m, 100-1000m, 1000-10,000m). 여러분이 2km 정확도를 가지는 정보를 가지고 있는데 이것을 10km 정확도 범주에 넣을 경우 정보는 즉시 손실된다.

문서화

올바른 문서화는 데이터 관리의 핵심 원칙이다. 올바른 문서화 없는 경우, 사용자는 자신이 생각한 이용에 대해 데이터의 적합성을 판단할 수 없고 따라서 해당 목적을 위한 데이터의 품질을 판단할 수 없게 된다. 문서화에 대한 더욱 상세한 논의는 아래 *문서화*에서 다루어진다.

피드백

데이터 관리자들은 데이터 사용자들이 피드백 하는 것을 장려하고, 그 피드백을 진지하게 고려하는 것이 중요하다. *사용자 책임* 섹션에서 언급된 것처럼, 각각의 개별 데이터 관리자가 홀로 일하는 것보다 사용자는 여러 출처에서 얻은 데이터를 결합시킴으로써 특정 오류 유형을 더 많이 접하게 되는 기회를 가지게 된다.

효율적인 피드백 체계를 개발하는 것은 항상 쉬운 것만은 아니다. 질의 인터페이스 페이지에 피드백 버튼을 둘 수 있으며, 또는 데이터를 다운로드할 때 데이터 관리자에게 데이터 오류 또는 의견을 피드백 할 수 있도록 첨부 내용을 사용자에게 보낼 수 있다. 이러한 것 가운데 일부는 관련 문서인 *데이터 정제의 원칙과 방법(Principles and Methods of Data Cleaning)*에 자세하게 다루어져 있다.

사용자와 공급자의 효율적인 피드백 관계는 데이터 품질을 향상시키는 쉽고 생산적인 방법이다.

교육과 훈련

교육과 훈련은 정보 사슬의 모든 단계에서 데이터 품질을 매우 크게 향상시킬 수 있다 (Huang *et al.* 1999). 이것은 올바른 수집 절차의 사용과 데이터 사용자의 수요를 반영하기 위한 수집자들의 교육과 훈련에서부터 시작되고, 데이터 입력자와 매일 매일 데이터베이스 관리에 책임이 있는 기술 인력에 대한 훈련, 그리고 최종 사용자에게 데이터의 속성, 이것의 제한점과 잠재적인 이용에 관한 교육을 필요로 한다. 데이터 품질의 교육과 훈련에 대한 측면은 대부분 올바른 문서화에 의존한다.

데이터 품질 검사, 교육, 그리고 훈련을 통합한 예는 MaPSTeDI 지리-연산정보 프로젝트 (University of Colorado 2003)에서 볼 수 있다. 이것은 개별 지리코드 입력자의 몇몇 레코드를 검사하는 것과 관련이 있다. 관리자는 새로운 입력자의 처음 200 개 레코드를 정확도에 대해 검사한다. 이것은 데이터의 품질을 유지할 뿐만 아니라 입력자가 실수에서 배우고 향상될 수 있도록 한다. 입력자에 따라 추가적인 100 개의 레코드를 관리자가 검사할 수 있으며, 입력자가 더욱 숙련됨에 따라 레코드들 가운데 임의의 10%를 선택하여 검사하고, 최종적으로는 약 5%를 검사한다. 높은 퍼센트의 오류가 여전히 많이 발견되면, 추가적인 레코드를 검사한다.

이러한 것과 같이 잘 설계된 절차는 새로운 사용자를 교육하는데 도움을 준다. 반대로 아무런 절차가 없다면 입력자 간에 그리고 업무 간에 일관성을 유지하는 방법은 거의 없게 된다.

책임 소재

전반적인 데이터 품질에 대해 책임 소재를 정하는 것은 조직이 품질제어의 일관된 수준을 달성하도록 지원하고, 오류의 피드백에 대한 참조 지점을 제공하며, 그리고 문서화와 질의에 대한 연락처를 제공할 수 있다.

많은 데이터 품질 관리 문제의 중심에는 부족한 훈련이 있다.

분류학 및 명명 데이터

> *품질 낮은 분류학적 데이터는 관련 학문 분야를 "오염"시킬 수 있다* (Dalcin 2004).

분류학은 개체를 분류하는 이론과 실제를 말한다 (Mayr and Ashlock 1991). 우리가 여기에서 고려하고 있는 대부분의 종 데이터는 분류학적 (또는 명명학적인) 부분 (즉, 개체의 이름과 이것의 분류)을 포함한다 – 달신(Dalcin 2004)은 이것을 "분류 데이터 도메인"으로 칭하였다. 데이터의 이 부분에 대한 품질 그리고 이 품질이 어떻게 판단되는지는 데이터의 공간적인 부분과는 상당히 다르며, 그 이유는 통상적으로 이것이 더 추상적이고 수량화하기가 더욱 어렵기 때문이다.

분류학적 데이터는 다음과 같이 이루어져 있다 (모든 것이 항상 존재하는 것은 아니다):

- 이름 (학명, 일반 이름, 분류 체계, 순위)
- 명명학적 상태 (동의어, 수용 여부, 모식화)
- 참조 (저자, 출판 장소와 날짜)
- 판단 (누구에 의해 언제 해당 레코드가 동정되었는가)
- 품질 필드 (판단의 정확성, 감정자)

분류학 이름에서 오류의 주요 원천 가운데 하나는 철자 오류이다. 분류학 데이터베이스에서 철자 오류를 탐지하는 것은 과(Family)와 속(Genus) 이름과 같이 분류학 계층 구조를 나타내는 학명일 경우에는 쉬운 일이 될 수 있다. 이러한 경우 대부분의 분류학 그룹에 대한 표준 전거 파일을 일반적으로 이용할 수 있다. 또한 Species 2000(http://www.species2000.org) 그리고 GBIF의 ECat 연구 프로그램(http://www.gbif.org/prog/ecat)과 같은 프로젝트를 통해 증가되고 있는 종 이름의 광범위한 목록이 이용 가능하다. 종 이름 또는 종소명(epithets)을 관련된 속 이름 없이 홀로 전거 파일로서 이용하는 것은 거의 좋은 결과를 내지 않는데 많은 특정 종소명이 하나의 속 또는 다른 것에서 이름의 미미한 변이를 가질 수 있기 때문이다. 철자 오류 검사의 한 가지 방법은 유사성 알고리즘을 사용하여 학명에서의 오류를 탐지하고 분리시키는 것으로서, 이것은 높은 유사성을 나타내지만 정확하게 같지 않은 한 쌍의 학명을 찾기 위한 것이다 (Dalcin 2004, CRIA 2005).

지금까지 학명의 철자 오류의 가능성을 줄이는 가장 만족할만한 방법은 속(genus) 및 종(species) 이름, 과(family) 이름 등의 선택 리스트를 이용하여 데이터베이스 입력 과정에 전거 파일을 만드는 것이다. 전거 파일을 이용할 수 있는 이상적인 상황에서, 이러한 기술을 사용하는 것은 이러한 오류의 유형을 실제적으로 0 으로 줄이게 된다. 하지만 불행히도 이러한 리스트를 이용할 수 없는 세계의 여러 지역과 많은 주요 분류 그룹이 있다.

전거파일이 Catalogue of Life 또는 ECat 같은 외부 출처에서 전거 파일이 입수되는 경우, Source-Id 가 데이터베이스에 기록되어, 전거 출처의 버전간에 만들어진 변경사항이 쉽게 이 데이터베이스에 반영되고 갱신될 수 있도록 해야 한다. 기대컨대, 조만간에 이것은 GUIDs(Globally Unique Identifiers)[3]의 사용으로 더욱 쉬워질 것이다.

데이터의 분류학 품질은 이용 가능한 분류학적 전문 지식에 크게 의존한다. 분류학위기 (Taxonomic Impediment) (Environment Australia 1998) 그리고 적절하게 훈련된 분류학 연구자의 세계적인 감소는 분류학 산출물의 장기적인 품질 감소와 결과적으로 1 차 종 데이터의 품질 감소로 이어질 것이다(Stribling *et al.* 2003). GTI(Global Taxonomic Initiative)는 소위

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

분류학 위기의 해결 또는 개선을 위한 노력을 하고 있지만, 이 문제는 앞으로도 계속하여 문제가 될 것 같다. 품질은 또한 시간의 경과에 따라 저하될 수 있으며, 특히 확증된 표본이 이용될 수 없거나 보전되지 않는 경우 (예를 들어 대부분의 관찰 데이터와 다수의 조사 데이터) 또는 적절한 분류학적 전문 지식을 이용할 수 없을 경우에 그러하다.

한 기관이 고품질의 분류학 결과(문서화된 1 차 종 데이터 포함)를 산출하는 역량은 다음 요소들에 의해서 영향을 받는다 (Stribling *et al.* 2003):

- 직원의 훈련과 경험 수준,
- 기술 문헌의 접근, 참고 및 확증 수집물 그리고 분류학 전문가의 수준,
- 적절한 실험실 기자재와 시설의 소유, 그리고
- 인터넷과 이용가능한 자원의 접근성.

동정 정확성 등의 기록

전통적으로, 박물관과 식물표본관은 분류학 그룹에서 종사하는 전문가들이 때때로 표본을 검사하고 이것들의 범위 또는 동정을 결정하는 결정 시스템을 운영하여 왔다. 이것은 종종 보다 큰 개정 연구의 일부로 수행되거나, 또는 다른 기관을 방문 중인 전문가가 그곳에 머무르면서 수집물을 검사하면서 수행된다. 이것은 검증된 방법이지만 시간이 많이 걸리고 대체로 무작위로 행해졌다. 하지만 자동화된 컴퓨터 동정이 가까운 또는 심지어 장기적인 미래에도 선택사항이 될 것 같지 않기 때문에 이것 이외에 다른 방법이 있을 것 같지 않다.

하나의 옵션은 동정을 할 때 이것의 확실성에 대해 어떤 표시를 제공하는 데이터베이스의 필드를 추가하는 것이다. 결정 일자는 통상적으로 대부분의 수집물 데이터베이스에 기록된다. 이것은 코드 필드일 수 있고, 아래와 같은 맥락에서 할 수 있을 것이다 (Chapman 2004):

- 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. 필자가 이해하기로 일부 기관에서는 분명히 이런 성질의 필드를 이미 가지고 있지만, 현 단계에서 필자는 사례를 찾을 수 없었다. HISPID 표준 버전 4 (Conn 2002)는 좀 더 단순화된 버전을 포함하고 있다 – 5 개의 코드가 있는 검증 수준 플래그, 즉:

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 3. *HISPID (Conn 2000)의 검증 수준 플래그.*

많은 기관들은 또한 확실성 기록 양식을 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 확실성의 수준보다는 분류학적 개념의 변이사항을 암시한다.

또한, 이름이 분류학 전문 지식이 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

두 개 또는 그 이상의 출판물 또는 전문가의 비교를 통하여 불확실성은 보통 감소될 수 있고 품질은 향상될 수 있다. 하지만 분류학자 간의 동정에 대한 차이는 이 동정중의 하나가 꼭 오류인 것을 의미하기보다는 분류군의 배치에 대한 분류학적 견해의 차이(즉, 서로 다른 가설)를 보이는 것일 수 있다.

동정의 정밀도

스트리블링과 그의 동료들(Stribling *et al.* 2003)에 의하면, 동정 정밀도(그들은 분류학적 정밀도로 잘못 지칭함)는 두 명의 분류학자 또는 전문가가 처리한 샘플을 무작위로 선택하여 그 결과를 비교함으로써 평가될 수 있다. 서로 다른 기관이 보유한 (그리고 동정한) 중복 표본의 이름을 비교함으로써 또한 평가를 할 수 있다. 이러한 것은 꽤 추상적인 개념이고, 필자는 이 유형의 정보를 기록하는 것의 가치에 대해 확신할 수 없다.

하지만, 동정 정밀도의 두 번째 부분은 표본이 동정된 수준이다. 종 또는 아종 단계까지의 동정은 단지 과 또는 속 단계의 것보다 더욱 정밀한 동정이다. 데이터집합을 문서화할 때, 동정의 50%가 단지 속 단계(많은 동물상 그룹의 경우에 그러하다)까지라는 것을 알리는 것은 사용자에게 유용할 수 있을 것이다.

치우침(Bias)

치우침은 측정치의 균일한 이동에서 발생하는 규칙적인 오류이다 (Chrisman 1991). 이것은 일관되게 적용된 방법론에서 종종 발생하며, 이 방법론은 본질적으로 규칙적인 오류를 야기한다. 분류학적 명명에서 치우침은 동정이 정밀하지만 정확하지 않은 경우 발생할 수 있다. 이러한 치우침은 이분법 키 또는 형태적인 구조의 잘못된 해석, 유효하지 않은 명명법 또는 오래된 출판물을 이용하는 것(Stribling *et al.* 2003) 또는 부적합한 출판물을 사용하는 것에서 발생할 수 있을 것이다 (예, 다른 지역의 식물상에 대해 연구 중이지만 이 지역의 모든 관련된 분류군을 가지고 있지 않을 수 있다).

일관성

비일관성은 두 개 또는 그 이상의 이름이 "수용되는 것"으로 여겨지고 동일 분류군(예, *Eucalyptus eremaea* 과 *Corymbia eremaea*)을 나타낼 경우, 데이터베이스의 분류 도메인에서 발생할 수 있다. 이것은 분류 체계에 대한 서로 다른 의견 또는 대체 철자로 인한 오류와 관련될 수 있다 (예를 들어, *Tabernaemontana hystrix*, *Tabernaemontana histryx* 그리고 *Tabernaemontana histrix* – CRIA 2005).

완전성

모트로와 라코프(Motro and Rakov 1998, Dalcin(2004) 자료 인용)는 완전성을 "*모든 데이터가 이용 가능한 것*"으로 언급하였고, 데이터의 완전성을 파일의 완전성(누락된 레코드가 없음)과 레코드의 완전성(개별 레코드의 모든 필드가 채워져 있음)으로 구분하였다.

분류학 용어(이름 또는 분류군 데이터베이스의 경우)에서 완전성은 이름의 적용 범위를 가리킨다. 데이터베이스가 계층 구조의 모든 단계에 있는 이름(예, 아종까지 또는 단지 종)을 포함하는가? 또는 이 데이터베이스가 동물 또는 식물계의 어느 부분을 다루고 있는가? 데이터베이스에 동의어가 포함되어 있는가? 이러한 모든 것들은 사용자가 자신의 특정 목적에 대한 데이터의 적합성을 결정하는데 중요하다. 예를 들어서 Dalcin(2004)에서는 완전성을 주어진 문맥(예, 분류학적인 문맥 – 특정 분류 그룹의 모든 이름의 리스트; 또는 공간 문맥에서 – 특정 지역에 대한 모든 이름 리스트)에서 모든 가능한 이름을 포함하여 나타내는 *명명적인 완전성(nomenclatural completeness)*과 주어진 분류군에 대해 "수용된" 이름과 관련된 모든 가능한 이름(예, 모든 동의어 표)을 나타내는 *분류 완전성(classification completeness)*으로 구분한다.

표본 또는 관찰 데이터베이스의 경우, 완성도는 "모든 다윈코어(Darwin Core) 필드가 포함되었는가"와 "모든 다윈코어 필드에 데이터가 있는가"로 정의할 수 있을 것이다. 형질 데이터베이스에서는 "모든 필요한 일생-단계에 대한 형질이 존재하는가"로 정의할 수 있다. (예, 식물의 수확물, 곤충의 영(instars)).

확증 수집물

확증 수집물의 중요성은 몇 번을 강조해도 부족하지 않지만, 데이터베이스가 확증자료를 항상 포함하는 것은 가능하지 않다. 동시에 많은 관찰 데이터베이스는 확증 수집물을 만들지 않으면서 구축된다. 정치, 법, 보전 또는 다른 목적상으로 모든 경우 또는 지역에서 확증자료에 대한 샘플을 획득하는 것이 또한 항상 가능한 것은 아니다.

확증자료 구축이 가능한 경우 종-기반 프로그램의 초기 단계에 데이터 수집자와 박물관 또는 식물표본관 같은 기관간에 참조 및 확증 수집물의 기탁을 지원하는 상호 협약을 맺는 것은

종종 가치 있는 활동이 된다 (Brigham 1998). 이와 같은 협약은 처분 또는 보관 활동 이전에 최소 기간을 포함하는 적절한 보관과 처분 전략을 포함해야 한다.

공간 데이터

공간 데이터는 데이터 문서화에 대한 표준 개발에서 관련 분야를 종종 주도하였으며 (예를 들어 공간데이터전송표준(Spatial Data Transfer Standards, USGS 2004), ISPIRE(Information for Spatial Information in Europe) 프로그램[4]과 그 외 다수), 그 이후 데이터 품질 표준(예, ISO 19115 for Geographic Information – Metadata[5]) 개발의 선두에 있어 왔다. 공간 데이터의 많은 부분이 수치 특성을 가지고 있기 때문에 이것은 분류학 데이터보다 통계적 방법의 사용에 더 개방적이며, 따라서 몇몇 데이터 품질 검사 방법들의 개발이 가능하였다 (관련 문서인 *데이터 정제의 원칙과 방법 (Principles and Methods of Data Cleaning)*을 참고).

이것은 데이터 중에 공간을 다루는 모든 부분이 디지털화하기 쉽거나 정확하다는 것을 뜻하는 것은 아니다. 박물관과 식물표본관에 있는 많은 역사적인 수집물은 수집물의 장소에 관해 아주 기본적인 텍스트 서술정보만을 가지고 있고, 이러한 것을 수치적인 지리코드 또는 좌표로 변환하는 것은 큰 노력을 필요로 한다. 이러한 많은 수집물의 특성으로 인해 이것은 더욱 악화될 수 있는데, 예를 들어, 수집자가 상세한 지도를 이용할 수 없던 때에 수집된 수집물 그리고 사용되었던 장소 이름의 많은 것들이 더 이상 출판된 지명사전(gazetteers) 또는 지도에서 보이지 않는 경우에 그러하다. 지리-참조연산 정보를 과거의 레코드에 추가하는 것은, 특히 신뢰할 수 있는 과거의 지명사전이 존재하지 않을 경우, 시간이 오래 걸리고 매우 낮은 수준의 정확도를 초래할 수 있다.

사용자가 가지고 있는 데이터에 대해 지리참조연산하는 것을 지원하기 위해 온라인 도구 및 지침서를 포함하여 많은 도구들이 개발되었다. 이러한 것들은 관련 문서인 *데이터 정제의 원칙과 방법(Principles and Methods of Data Cleaning)*에서 자세히 다루어진다. 이것 뿐만 아니라 대부분의 수집자는 이제 수집할 때 지리코드를 기록하기 위해 GPS(위성항법시스템, Global Positioning Systems)를 사용하고 있다. GPS 와 연관된 정확성 토론은 "*데이터 획득(Capturing Data)*" 장을 참고하기 바란다

이미 배정된 지리참조연산값에 대한 오류 검사를 할 때 다음 사항들을 포함할 수 있다:

- 레코드 자체의 내부 또는 데이터베이스내의 레코드간의 정보를 비교하는 검사 – 예를 들어, 주, 지명 등;
- 데이터베이스를 이용하여 외부 참조정보에 대해 비교 – 레코드가 수집자의 수집 장소들과 일관되는가?
- GIS 를 이용하여 외부 참조정보에 대해 비교 – 레코드가 해상이 아닌 육상을 가리키는가?
- 지리 공간의 특이점에 대해 검사; 또는
- 환경 공간의 특이점을 검사.

이러한 방법 모두는 관련 문서인 *데이터 정제의 원칙과 방법(Principles and Methods of Data Cleaning)*에서 자세하게 다루어진다.

공간 정확성

공간 데이터의 위치 정확성은 어떻게 측정되는가?

[4] http://www.ec-gis.org/inspire/
[5] http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35

대부분의 GIS 계층(지형도 등)에서 '참(truth)'의 원천은 상대적으로 결정하기 쉬운데, 그것은 통상적으로 데이터베이스에 몇몇 특징 정보(조사 시작점(trig point), 거리와 도로의 교차로 등)의 더 높은 정확성을 가진 외부 원천이 있기 때문이다 (Chrisman 1991). 하지만 대부분의 검사 과정은 간단하지 않으며 미국 국가지도정확성표준(US National Map Accuracy Standard)과 같은 문서는 복잡하게 되어 있다. 전통적으로 공간 정확성은 몇몇의 "잘 정의된" 지점에 대한 비교와 RMSE(root-mean-square deviation)로 측정되는 명시된 수용 가능한 오류의 수준에 의해 결정된다(Chrisman 1991). 하지만 RMSE 를 개별적인 지점에 적용하는 것은 쉽지 않으며, 전체 데이터집합 또는 디지털 지도에 더욱 더 잘 적용할 수 있다. 개별적인 지점의 경우 간단한 점-반지름 방법(Wieczorek *et al.* 2004)을 이용한 참 장소에서의 거리 또는 유사한 방법들이 간단하고 사용하기 쉽다. 여기에 관련된 두 가지 요소가 있다 – '정확하게 잘 정의된 지점이 어떻게 결정될 수 있는가'에 대한 문제가 검사되고 있는 지점의 정확성을 결정하고, 검사된 지점의 측정 정확도 및 정밀도가 오류에 더해진다. 예를 들어, 교차로가 100 미터 이내로 정확하게 결정될 수 있다면, 이 지점의 정밀도가 더해지기 전에 수집물 지점의 중심점은 100 미터 원이 된다 (Wieczorek 2001 의 설명 참고).

미국연방지리데이터위원회(FGDC, US Federal Geographic Data Committee)는 1998 년에 지리공간위치정밀표준(Geospatial Positioning Accuracy Standards, GPAS)을 배포하였다. 이러한 표준은 측지선네트워크(Geodetic Networks)와 공간데이터정밀도(Spatial Data Accuracy, FGDC 1998)에 대한 섹션을 분리하여 포함하고 있다.

- *'NSSDA 는 위치의 정밀도를 측정하기 위해 평균제곱근오차(RMSE, root-mean-square error)를 사용한다. RMSE 는 동일한 지점들에 대해 더 높은 정밀도를 가진 독립적인 소스에서의 데이터집합 좌표 값과 좌표 값간의 제곱 차이값 집합의 평균 제곱근이다.'*
- *'정확성은 95% 신뢰도 수준에서 지상 거리로 보고된다. 95% 신뢰도 수준에서 보고된 정확성은 데이터집합에 있는 위치의 95%는 보고된 정밀도 값과 같거나 작은 실제 지상 위치에 따라 오류를 있을 것이다라는 것을 의미한다. 보고된 정확성 값은 측지선 제어 좌표, 편집, 그리고 산출물에 있는 지상 좌표 값의 최종 계산으로 인하여 발생되는 모든 불확실성 값을 반영하게 된다.'*

산출물과 관련하여 이와 같은 방법을 이용해서 호주에서 만든 지도 정확성 문장의 예는 다음과 같다:

- *'이 지도의 평균 정확성은 잘 정의된 상세 수평 위치에서 ±100 미터이고 고도에 대해서는 ±20 이다.'* (Division of National Mapping, Sheet SD52-14, Edition 1, 1:250,000).

종이나 디지털 지도에 기반한 임의의 수집물의 지리-연산정보를 결정할 때 이러한 정확성을 추가할 필요가 있다. 공간 데이터의 정확성에는 항상 불확실성이 있으므로, 절대적인 정확성 진술은 적용될 수 없고 알려진 정확성을 문서화하는 것이 중요하다. 오류는 정보 사슬 전반에 걸쳐 전파되며, 이것이 GIS 의 지도 결과이든 분포 모델링 소프트웨어를 사용한 종 모델이든 마지막 결과의 불확실성에 영향을 미친다 (Heuvelink 1998).

BioGeomancer 프로젝트

고돈베티무어(Gordon and Betty Moore) 재단은 1 차 종 레코드의 지리-참조연산을 향상시키고 정확성을 평가, 향상, 그리고 문서화하는 것을 지원하기 위해 최근 한

프로젝트[6]를 지원하였다. 이 프로젝트는 2006 년 중으로 개발된 도구를 공개하고 이용 가능하게 할 것이다.

거짓 정밀도와 정확도

알아두어야 할 또 다른 요소는 거짓 정밀도와 정확도에 대한 것이다. 많은 GIS 사용자들은 공간 데이터 정확도, 오류 그리고 불확실성과 관련된 문제를 알고 있지 못하며, 종종 자신들의 데이터가 절대적이라고 가정한다. 그들은 출처 데이터로 달성할 수 없는 정확도를 종종 보고한다. 많은 기관들은 지리-참조연산을 지원하고 데이터가 제공할 수 없는 수준까지 확대하는 것에 현재 GIS 를 사용하고 있다 (그리고 십진법표기도수(decimal degree)를 사용하는 것은 비현실적인 정밀도를 얻게 될 것이다). 또한, 수집 활동의 장소를 기록하기 위해 GPS 를 사용하는 경우 장소는 종종 1 또는 2 미터로 보고되는데, 실제로 많은 휴대용 GPS 기기는 아마도 약 10 미터 내외의 정확도를 가지고 있을 것이다. 이것은 특히 고도를 결정하기 위해 GPS 를 사용하는 경우에 관련이 있다 (아래 *데이터 획득(Capturing Data)*의 설명 참고).

[6] http://www.biogeomancer.org/

수집자와 수집 데이터

수집자와 수집 데이터에 관한 정보(the Collection data domain of Dalcin 2004)는 수집물 자체에 대한 정보를 포함한다 – 이러한 것에는 수집자, 수집일자 그리고 서식처, 토양, 날씨 조건, 관찰자 경험 등의 추가 정보가 있다. 이것들은 다음과 같이 범주화할 수 있다 (Conn 1996, 2000 에서 수정)

- 수집 저자(들) 그리고 수집자의 번호(들)
- 관찰자의 경험 등
- 수집 일자/기간(들)
- 수집 방법 (특히 관찰/조사 데이터)
- 연관된 데이터

이러한 사항들의 많은 부분은 수집되고 있는 데이터의 유형 – 박물관 수집물, 관찰 또는 상세 조사의 결과물에 따라 상당히 다양할 것이다. 박물관에서 사용되는 것과 같은 정적 수집물의 경우, 수집자의 이름과 번호 그리고 날짜가 핵심적인 요소이고 관련 데이터인 습관, 서식처 등 그리고 (동물의 경우) 획득 방법과 같은 것도 중요하다. 관찰 데이터의 경우, 관찰 대상의 길이, 관찰 지역, 하루의 시간 (날짜와 함께 시작 및 종료 시간), 그리고 관련 데이터인 날씨 조건, 관찰되는 동물의 성(sex), 활동 등이 중요하다. 조사 데이터의 경우에는, 관찰 데이터의 경우에 언급되었던 많은 것들과 함께, 조사 방법, 크기(격자와 전체 지역), 노력, 날씨 조건, 주기, 그리고 확증 자료와 이것의 개수가 수집되었는가 등의 정보가 중요하다.

속성 정확성

수집물 정보와 관련되어 데이터 품질에 영향을 줄 수 있는 사항은 수집자의 이름, 번호, 머리글자 등이 기록되는 방식(Koch 2003), 날씨와 시간 기록의 정확성, 그리고 수집 당시의 습관, 서식처, 토양, 식물 유형, 꽃 색깔, 성, 관련 종 등의 관련 데이터를 기록할 때의 일관성 등이 포함된다.

수집물 데이터에서 정기적으로 발생하는 문제의 예는 “수집자의 번호”이다 – 일부 수집자들은 자신의 수집물을 식별하기 위한 고유 번호를 사용하지 않는다. 이것은 품질 손실을 야기할 수 있는데 그것은 이러한 태그가 때때로 수집물의 위치, 동정, 서로 다른 기관에 있는 중복된 수집물 등의 동정을 하는데 유용하게 사용될 수 있기 때문이다.

일관성

수집물 도메인과 관련하여 사용되는 용어인 일관성은 그 사용이 매우 불규칙적이며, 특히 관련 데이터 필드가 일관적인 경우는 서로 다른 데이터집합의 경우에는 말할 것도 없고 같은 데이터집합에서도 매우 드물다.

완전성

수집물 정보의 완전성 또한 매우 다양하다. 아주 종종 많은 레코드의 경우 서식처, 수집자의 번호, 개화 시기 등이 완성되어 있지 않은 경우가 많다. 예를 들면, 이것은 단지 수집물 자체만으로 분포 지역에 대한 연구를 어렵게 만든다.

서술 데이터

서술 데이터베이스들은 데이터를 저장하고 종종 전통적인 출판물을 대신하는 출판의 방법으로서 사용되고 있으며 그 수가 증가하고 있다. 형태학적, 생리학적 그리고 생물계절학적 데이터 요소들이 이 분야의 데이터 사례들이다. 서술 데이터는 종종 클래디스틱(cladistic) 분석에 사용될 정보를 발생시키는데 사용되며, 자동으로 생성된 기재문 그리고 동정 도구에도 사용된다.

TDWG(Taxonomic Databases Working Group)는 서술 데이터베이스 분야에서 표준의 개발 및 진흥과 관련한 오랜 역사를 가지고 있다 – 처음에는 DELTA 표준(Dallwitz and Paine 1986)을 지원하였고 더욱 최근에는 "서술 데이터의 구조(Structure of Descriptive Data)" 실무 그룹의 활동(http://160.45.63.11/Projects/TDWG-SDD/)을 지원하였다.

서술 데이터의 품질은 다양할 수 있는데, 데이터 요소가 종종 측정되지만 실제로 정확성은 경우에 따라 결정될 수 있기 때문이다. 이러한 예는 데이터를 관찰할 수 없는 것(예, 과거 데이터), 관찰하기에 비실용적인 것(예, 매우 고비용이 드는 것), 그리고/또는 실제보다는 개념적인 것(예, 색깔, 풍부성 등과 같이 주관적인 평가) 등이 있다.

대부분의 경우, 서술 데이터는 표본 수준보다는 종 수준에서 저장되며, 따라서 통상적으로 평균화 또는 범위화 된다. 모스(Morse 1974)가 지적한 것처럼, 분류학 정보는 본질적으로 표본 관찰 데이터보다 신뢰도가 낮다. 이것과 상관 없이 최근에는 적어도 이러한 데이터의 일부를 (품질 향상의 결과를 염두에 두면서) 표본 수준에서 저장하려는 경향이 커지고 있다.

완전성

표본 수준에서, 서술 데이터 기록물의 완전성은 표본의 품질, 해당 년도의 시간에 종속될 수 있다. 예를 들어, 동일 표본에서 과실 또는 꽃의 특징을 기록하는 것이 가능하지 않을 수 있다. 이러한 이유로, 많은 필드가 어쩔 수 없이 공백으로 남게 된다. 다른 경우에는, 해당 특성이 객체와 관련되지 않을 수 있고 따라서 모든 속성이 기록되지는 않는다.

일관성

비일관성 문제는 관련되어 있는 데이터 항목간에 일어날 수 있다. 예를 들어, 두 종의 서술자 특징이 다음과 같이 기록될 수 있다 (Dalcin 2004):

- "HABIT=HERBACEUS" 그리고
- "USES=WOOD"

동일한 특징을 일관적이지 않게 표현하는 것은 또한 품질에 영향을 끼칠 수 있으며, 특히 부실한 특징 정의가 사용되거나 일관성있게 표준이 철저히 지켜지지 않는 경우에 그러하다. 예를 들어 (Dalcin 2004):

- "FLOWER COLOUR= CARMINE", 그리고
- "FLOWER COLOUR=CRIMSON".

표준 용어들을 사용하는 것은 오류의 정도와 잘못된 해석을 감소시키는데 상당한 도움이 될 수 있다. 표준 용어들이 여러 범위의 영역과 학문 분야에서 개발되고 있으며, 최근 연합 서술 데이터베이스의 개발 등의 활동으로 용어 사용에서 일관성이 증가되고 있다.
서술데이터의구조(Structure of Descriptive Data, SDD)에 대한 TDWG 표준 개발(TDWG 2005)만이 이 과정을 지원할 수 있을 것이다.

데이터 획득

1차 종 데이터와 종-발생 데이터를 획득하는 여러 다양한 방법이 있으며, 이것은 각각 오류와 불확실성의 출처 뿐만 아니라 자체적인 정밀도와 정확도를 가지게 된다. 이러한 것들 중의 각각은 최종 "사용에 대한 적합성", 즉 데이터의 품질에 각기 다른 영향을 주게 된다. 종 데이터와 관련되어 가장 많이 사용되는 몇몇 방법들이 여기에서 간략히 논의된다.

임의적 획득

대부분의 종-발생 데이터는 체계적이기 보다 임의적으로 수집되었다. 이러한 레코드의 많은 것들이 현재 박물관과 식물표본관에서 표본으로 저장되어 있다. 대부분의 과거 데이터는 도시의 북서쪽으로 5 km (5 km NW of a town) 등과 같은 텍스트 장소 참조정보만을 포함하였으며 수집 당시에 지리참조연산이 수행되지 않았다. 지리 참조연산정보는 보통 훗날 추가되었고, 통상적으로 수집자가 아닌 다른 사람에 의해서 이루어졌다 (Chapman and Busby 1994). 많은 관찰 데이터 또한 임의적으로 수집되었다.

이러한 데이터는 통상적으로 일괄적인 디지털 형식으로 획득되었고, 지리-참조연산은 일반적으로 물리적인 지도를 참고하여 수행되었다. 이러한 데이터의 대부분은 약 2-10 km 보다 더 정밀도가 있다고 간주될 수 없다.

현장 조사

일반적으로 현장 조사 데이터에는 종종 위도/경도 또는 UTM 참조정보의 형식으로 공간적인 참조정보가 포함되어 있다. 이 공간 참조정보는 보통 100-250 미터의 정밀도가 있다고 간주될 수 있다. 하지만 공간 참조자료가 가리키는 것이 무엇인지에 대한 주의를 기울여야 한다 – 이것이 실제 관측이 이루어진 장소가 아니고 예를 들어 횡단의 중심 점 또는 격자의 모서리(또는 중심)를 가리킬 수 있으며, 이러한 것이 항상 명확한 것은 아니다. 또한 레코드가 거의 확증표본화 되지 않았기 때문에 (즉, 물리적인 수집물을 만들고 나중에 참고하기 위해 보관), 이 분류학적인 정확성을 항상 신뢰할 수는 없다. 이것은 조사가 이루어진 시기에서 멀어지고, 분류학적 개념이 바뀔수록 특히 그러하다.

대규모 관찰

일부 생물 종 관찰 조사는 특정 경계 또는 격자 셀 내의 데이터만을 기록한다. 예를 들어, 국립 공원 내의 종 조사, 즉 10 km 격자 안에서의 행해지는 조류 관찰(예, Birds Australia 2001, 2003)이 있다. 이러한 것과 같은 레코드의 정확도는 1-10 km 또는 조금 더 큰 수치만을 나타낸다.

위성항법시스템(Global Positioning Systems, GPS)

위성항법시스템, 즉 GPS는 종 데이터 수집물의 경우에 점점 더 많이 사용되고 있다. 이러한 것은 조사 데이터 뿐만 아니라 임의적인 것과 관찰 수집물을 또한 포함한다.

GPS 기술은 삼각 측량을 이용하여 지구 표면에 있는 지점의 위치를 결정한다. 측정되는 거리는 GPS 수신기와 GPS 위성 간의 범위이다 (Van Sickle 1996). GPS 위성들은 공간상의 알려진 위치에 있기 때문에, 지구상의 위치를 계산할 수 있다. 지구 표면에 있는 지점의 위치를 결정하기 위해서는 최소 네 개의 GPS 위성이 필요하다 (McElroy *et al.* 1998, Van Sickle 1996). 오늘날 지구상 대부분의 장소에서 7개 또는 그 이상의 위성 신호를 받기 때문에

이것은 일반적으로 제한 사항이 아니지만, 과거에는 신호를 받을 수 있는 인공위성의 수가 항상 충분하지는 않았다. 2000년 5월 이전, 대부분의 민간용 GPS 기기는 "선택적인 이용 가능성(Selective Availability)"과 관련되어 있다. 이것의 제약 사항을 극복하게 되어 일반적으로 기대할 수 있는 정밀도는 크게 향상되었다 (NOAA 2002).

"선택적인 이용 가능성"의 제한점을 극복하기 전에, 대부분의 생물학자와 관찰자가 현장에서 사용했던 *휴대용(Hand-held) GPS* 수신기의 정확성은 약 100 미터 또는 이보다 더 정확하지 못했다 (McElroy *et al.* 1998, Van Sickle, 1996, Leick 1995). 하지만 그 이후 GPS 수신기의 정밀도는 증가하였고, 오늘날 4 개 또는 그 이상의 위공위성을 이용하는 경우 대부분의 휴대용 GPS 수신기 제조사는 열린 지역에서 10 미터 미만의 정확도를 보장한다. 정확도는 하나의 장소에서 여러 번 관측한 결과의 평균을 냄으로써 향상될 수 있으며 (McElroy *et al.* 1998), 평균 알고리듬을 포함하는 일부 최신 GPS 수신기들의 정확도는 약 5 미터 또는 그 미만의 결과를 낼 수 있다.

*보정 GPS(Differential GPS, DGPS)*를 이용하는 것은 정확도를 상당히 향상시킬 수 있다. DGPS 는 GPS 수신기의 위치를 측정하기 위해 알려진 위치의 GPS 기지국(통상적으로 조사 제어 지점)을 참조한다. 이것은 기지국과 휴대용 GPS 가 동시에 위성의 위치를 참조하여 작동하며, 결과적으로 기상 조건에 따른 오류를 줄이게 된다. 이러한 방식으로 휴대용 GPS 는 판단하고자 하는 위치에 적합한 정정 절차를 취한다. 사용되는 수신기의 품질에 따라, 1-5 미터의 정확도를 얻을 수 있다. 이 정확도는 기지국으로부터 수신기의 거리가 멀어질수록 감소한다. 이것 역시 평균값을 산출하는 것이 이러한 수치를 더욱 개선할 수 있다 (McElroy *et al.* 1998).

WAAS(Wide Area Augmentation System)는 GPS-기반 항해 및 착륙 시스템으로 비행기의 정밀한 유도를 위해 개발되었다. WAAS 는 지상-기반 안테나와 관련이 있으며 정밀하게 알려진 이 지상-기반 안테나의 위치와 GPS 를 같이 사용하면 더 자세한 위치의 정밀도를 제공할 수 있다. LAAS(Local Area Augmentation System)와 같은 유사 기법들이 더 세분화된 정밀도를 위해 개발되고 있다.

실시간 보정(Real-time Differential) GPS(McElroy *et al.* 1998) 또는 *정적(Static) GPS*(McElroy *et al.* 1998, Van Sickle 1996)를 이용하면 정확도를 더 많이 높일 수 있다. *정적 GPS*는 고정밀 장비와 전문가의 기술을 사용하며 보통은 측량 기사들만이 사용한다. 호주에서 이 기법을 이용하여 수행된 조사들은 정확도가 센티미터 단위로 보고되었다. 이러한 기법들은 비용과 이러한 정밀도의 요구가 일반적으로 적기 때문에 생물 레코드 수집물에 광범위하게 사용될 것 같지는 않다.

위에서 보고된 것과 같은 정확도를 얻기 위해서, GPS 수신기는 공중에 장애물이 없고 표면이 반사되지 않는 지역에 있어야 하며 그리고 수평선이 잘 보이는 위치에 있어야 한다 (예를 들면, 이것은 하늘이 보이지 않는 울창한 숲에서는 잘 작동하지 않는다). GPS 수신기는 적절한 공간 배치에 있는 최소한 네 개의 GPS 위성의 신호를 기록할 수 있어야 한다. 가장 좋은 배치는 "*바로 위에 위성이 하나 있고 나머지 세 개는 수평선을 따라 균등하게 있는 것*"이다(McElroy *et al.* 1998). GPS 수신기는 또한 해당 지역의 적합한 지점에 위치시켜야 한다.

GPS 고도. 대부분의 생물학자들은 GPS를 이용해서 측정된 고도에 대해서 알고 있지 못하다. GPS 수신기가 표시하는 고도는 지구 중심 데이터와 관련된 고도이고 평균 해수면과 관련된 고도 또는 호주 고도 데이터와 같은 표준 고도에 따른 것이 아니라는 것을 주지하는 것이 중요하다. 예를 들어, 호주에서 GPS 수신기와 평균 해수면에서 보고되는 고도의 차이는 -

35m에서 +80m까지 다양할 수 있으며 예측할 수 없을 정도로 다양하다 (McElroy *et al.* 1998, Van Sickle 1996).

데이터 입력과 입수
(데이터를 전자적으로 획득하는 것)

> *데이터 입력과 입수는 본질적으로 간단한 오류와 복잡한 오류가 쉽게 발생하는 경향이 있다.* (Maletic and Marcus 2000)

기본적인 데이터 획득

데이터 획득의 첫 번째 단계는 통상적으로 표본 레이블, 논문, 현장 일지, 수납 장부 또는 카드 목록에서 정보를 획득하는 것이다. 이것은 숙련된 또는 비숙련된 데이터 입력자 또는 스캐너를 이용하여 수행될 수 있다. 데이터 입력으로 인한 오류의 수준은 이중-입력, 스캐닝과 연관된 학습 소프트웨어의 사용, 그리고 전문가와 감독관이 샘플 자료에 대한 입력 검사를 수행함으로써 종종 감소될 수 있다 (아래 언급되는 MaPSTeDI Guidelines 참고).

사용자 인터페이스

특정한 데이터 입력을 위한 유저 인터페이스를 개발하는 것 또한 데이터 입력 오류를 줄이는 방법이다. 많은 기관들에서 미숙련 직원 또는 자원 봉사자들이 데이터 입력자로 활동하고 있으며 입력자가 쉽게 사용할 수 있는 간단한 (비-기술적인) 사용자 인터페이스를 개발하는 것은 입력의 정확성을 높일 수 있다. 이러한 인터페이스는 해당 데이터베이스와 다른 데이터베이스에 있는 전거 필드, 기존 입력 항목을 빠르게 검색할 수 있게 하고, 레이블을 읽을 때 어려움이 있거나 또는 특정 필드에 어떠한 것을 입력할 지를 결정해야 할 경우, 심지어 구글(Google)과 같은 검색 엔진을 사용하여 입력자가 올바른 철자 또는 용어인지를 판단하는 것을 지원함으로써 데이터 입력을 도울 수 있다. 일부 경우, 이것은 전거 테이블과 드롭-다운 메뉴(선택 리스트)를 통합하는 데이터베이스 설계 과정에 응용될 수 있으며, 이것은 이름, 장소 또는 서식지에 관한 결정을 해야 하는 미숙련 데이터 입력 인력에 대해서는 고려하지 않은 것이다.

지리-참조연산

지도는 정보를 전달하는 가장 효과적인 수단 중의 하나이며, 이러한 이유 하나만으로도 이것은 지리참조연산이 된 관찰 정보 획득의 증가와 함께 최근 박물관과 식물 표본관에서 표본 데이터를 데이터베이스화하고 지리-참조연산하는 활동의 증가를 정당화한다. 지도상의 향상된 데이터 처리 능력으로 오류와 불확실성을 더 잘 연구, 동정, 가시화, 문서화 그리고 정정하는 것이 가능하다 (Spear *et al.* 1996). 이것은 또한 데이터에 내재한 불확실성을 가사화하고 전달하는 강력한 방법을 제공하며, 따라서 사용자들에게 데이터의 품질 또는 사용의 적합성을 판단하는 방법을 제공할 수 있다.

데이터를 디지털 형태로 획득하고 지리코드를 추가(즉, 데이터의 지리-참조연산)하는 것은 어렵고 시간이 많이 드는 작업일 수 있다. MaPSTeDI 프로젝트 (University of Colorado 2003)의 결과에 의하면 유능한 입력자는 5 분마다 하나의 레코드를 지리참조연산할 수 있음을 나타내고 있다. 다른 연구(Armstrong 1992, Wieczorek 2002)는 지리-참조연산하는데 더 많은 시간이 소요된다는 것을 보였다 – 예를 들어, MANIS 데이터베이스에서는 미국 지역의 경우 시간당 약 9 개의 레코드, 미국을 제외한 북아메리카 지역의 경우 시간당 6 개, 그리고 북아메리카 지역이 아닌 경우 시간당 3 개의 레코드 비율을 나타내고 있다 (Wieczorek 2002).

MaNIS/HerpNet/ORNIS
Georeferencing Guidelines
http://manisnet.org/manis/GeorefGuide.html

MaPSTeDI
Georeferencing in MaPSTeDI
http://mapstedi.colorado.edu/geo-referencing.html

데이터 관리자가 지리-참조연산하는 것을 지원하기 위해 많은 우수한 방법과 지침서들이 개발되었다. 버클리 소재 척추동물박물관(Museum of Vertebrate Zoology)의 존 위에크조렉이 개발한 지리참조연산 지침서(Georeferencing Guidelines) (Wieczorek 2001)와 MapSTeDI (Mountains and Plains Spatio-Temporal Database Informatics Initiative) 지침서(University of Colorado 2003)는 이 분야에서 현재까지 수행된 가장 포괄적인 연구 가운데 두 개이며 필자는 독자들이 이것을 참조하기를 권장한다. 이 지침서들은 텍스트 형식의 장소 정보에서 도출된 지점의 정확도와 정밀도의 결정, 서로 다른 데이터의 사용으로 인해 발생하는 불확실성, 서로 다른 지도의 축척을 사용하는 효과 등을 다룬다. 이 지침서들은 관련 내용을 포괄적으로 다루고 있으며, 필자는 이 문서의 독자들이 이것들을 이 문서의 통합적인 부속 문서로 간주해 주기를 바란다.

또한 지리코드 결정을 지원할 수 있는 많은 수의 온라인 도구들이 있다 – 예를 들어, 알려진 장소로부터 주어진 거리와 장소에 대한 위치가 있다. 이러한 것은 연관 문서인 *데이터 정제의 원칙과 방법(Principles and Methods of Data Cleaning)*에서 조금 더 자세히 다루어 질 것이다.

BioGeoMancer
(Peabody Museum of Natural History)
http://www.biogeomancer.org/

geoLoc
(Reference Centre for Environmental Information)
http://splink.cria.org.br/tools/

오류

앞에서 언급된 것과 같은 도구들은 오류를 줄이고 품질을 높이는데 사용될 수 있는 강력한 것들이다. 하지만 어떠한 지리코딩 방법도 오류를 완전히 제거할 수는 없다. MaPSTeDI 지침서에 다음과 같은 내용이 있다:

> *"지리코딩이 정확한 과학은 아니고 100% 정확하게 지리코딩되는 수집물이 있을 수는 없지만, 정확하게 지리코딩되는 수집물의 비율을 품질 검사를 통해*

급격하게 개선시킬 수 있다. 모든 프로젝트는 지리코딩 연산을 계획할 때 이것을 고려해야 한다" (University of Colorado 2003).

지리참조연산 오류의 일반적인 원인 하나는 전자 지명사전을 맹목적으로 이용하는 것이다. 일부 경우 이러한 지명사전은 인쇄 목적의 지도를 출판하는 프로젝트에서 개발되었고, 지명사전에서 주어진 지점의 위치는 이름이 지도상에 쓰여진 곳의 왼쪽 하단이었으며, 이것이 가리키는 지점의 위치가 아니었다 (예, 호주토지정보그룹 (Australian Land Information Group)에서 개발한 1998 년 이전의 호주 지명사전 (The Australian Gazetteer). 대부분의 지명사전이 정정되었을 것으로 기대되지만, 이러한 값에 기반하여 지리참조연산된 값이 박물관 그리고 표본관 데이터에 이미 추가되었을 수 있다. 이와 같은 레코드의 정확도는 지명사전 또는 큰 축척의 정확한 지도에 대해 장소의 무작위 지점 검사를 통해 점검되어야 한다.

레이블 정보를 디지털화한 후 분리된 활동으로서 지리-참조연산을 수행하는 것이 종종 더 빠르고 효율적이다. 이것은 수집물을 장소, 수집자, 일자 등의 정렬 목적으로 데이터베이스가 사용될 수 있도록 하고, 지리코드 정보를 획득할 때 지도를 더욱 효율적으로 사용할 수 있도록 한다. 이것은 또한 같은 장소에서 발생한 여러 레코드의 지리코딩 연산의 중복 등을 감소시킨다.

데이터의 문서화

> *"메타데이터는 데이터에 관한 데이터이다. 이것은 특정 목적을 위해서 수집된 데이터의 특성을 서술한 것이다."* (ANZLIC 1996a).

올바른 문서화는 데이터집합과 데이터 레코드 수준 모두에서 발생한다.

메타데이터는 콘텐트, 범위, 접근성, 현재성, 완전성, 목적에 대한 적합성 그리고 사용에 대한 적합성과 같이 데이터집합에 대한 정보를 제공한다. 메타데이터가 제공될 때, 사용자는 데이터집합의 품질을 이해할 수 있으며 이것을 사용하기 전에 데이터집합에 대한 적합성을 확신할 수 있다. 좋은 메타데이터는 더 나은 데이터의 교환 및 검색을 가능하게 한다. 메타데이터는 통상적으로 전체 데이터집합을 가리키지만, 일부 사람들은 레코드-수준 메타데이터로서 (정확도의 기록과 같은) 레코드 수준의 데이터에 대한 문서화로 간주하기도 한다. 이것이 불리워지는 것에 상관없이, 올바른 문서화는 데이터집합 수준과 레코드 수준 모두에서 중요하다.

모든 데이터는 오류를 포함한다 – 이것을 피할 수는 없다! 오류가 어떠한 것이며, 그리고 이 오류가 데이터가 사용될 목적에 대해 수용 가능한 범위 내에 있는 것인가를 아는 것이 중요하다. 이것이 메타데이터가 전체 데이터집합에 대해 중요한 위치를 차지하는 경우이며, 실제로 "사용에 대한 적합성" 용어가 중요하게 부상하게 된 분야가 메타데이터 개발 부분이다. 90 년대 초반이 되어서야 사용에 대한 적합성 개념이 공간 정보와 함께 전체적으로 중요한 것으로 인식되었으며, 이러한 문맥에서 이것이 문헌에 나타나기 시작한 것은 90 년대 중반부터였다 (Agumya and Hunter 1996).

하지만, 데이터집합 수준에서만 정보를 기록하는 것이 사용자가 요구하는 정보를 항상 제공하는 것은 아니다. 레코드 수준에서 오류를 기록하는 것이, 특히 종 데이터의 경우, 이 레코드의 사용에 대한 적합성을 판단하는데 매우 중요할 수 있다. 이 정보가 제공될 때, 예를 들면 사용자는 특정 미터 값 이상의 값(예, 5,000 미터 이상)을 가진 데이터만을 요청할 수 있다. 자동화된 지리-참조연산 도구가 출력 필드에 계산된 정확도를 포함하는 것이 또한 중요하다.

데이터 사용자들이 사용에 대한 적합성 개념을 이해하는 것 또한 중요하다. 너무 자주 종 발생 데이터는 정확도에 대한 정보 없이 "record no., x, y" 형식으로 데이터베이스에서 추출된다. 이 좌표 자체는 항상 지점으로 표시되지만, 이것은, 설령 있다고 해도, 거의 참 위치를 가리키지 않는다. 일부 레코드는 임의적인 지점으로 데이터베이스에 입력되고 (예를 들어, 레이블에 "South America"로 되어있는 수집물) 정확도 필드에 5,000,000 미터의 정확도로 입력되었을 수도 있다. 일부 데이터베이스는 이렇게 데이터를 입력하고 있다! 이 레코드를 추출하여 임의의 지점을 사용하는 것은 극단적인 방향으로 결과를 초래할 수 있다. 사용자들은 정확도 필드가 존재할 경우 정확도에 대해 인지할 필요가 있으며, 이것을 어떻게 사용해야하는가에 대해 조언을 받을 필요가 있다. 데이터 제공자들이 표준 데이터 보고서를 만드는 경우, 데이터가 제공될 때 정확도 필드가 필수적으로 포함될 수 있도록 하여야 한다.

데이터는 제3 자가 데이터의 원작성자를 참고하지 않고 사용할 수 있도록 충분히 상세한 메타데이터와 함께 문서화되어야 한다.

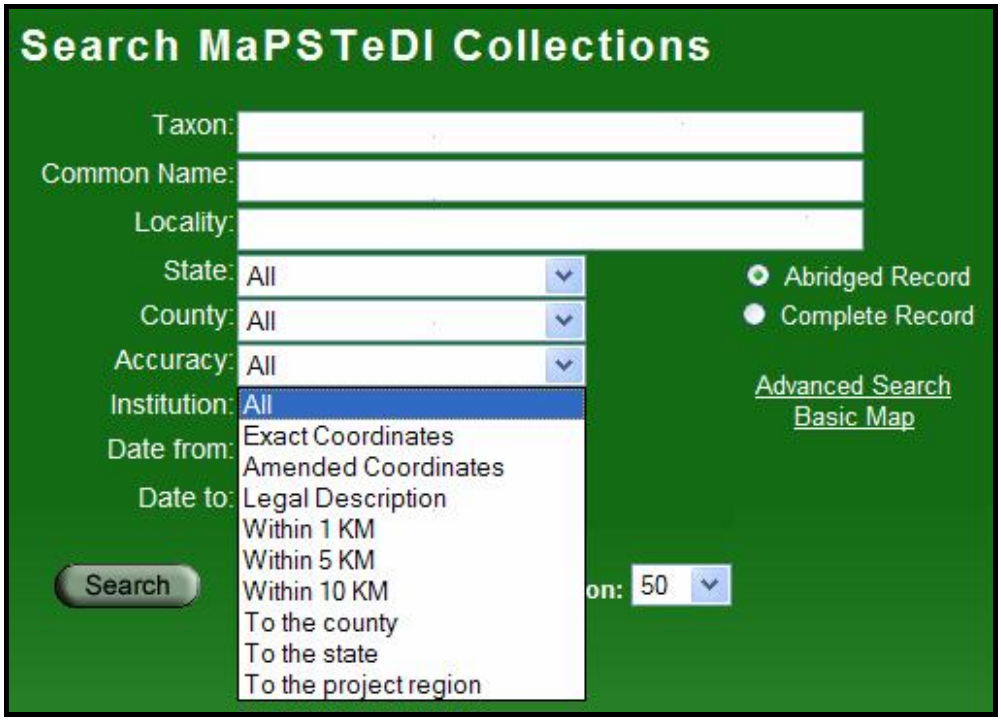

Fig. 6. *MaPSTeDI 검색 도구를 이용한 데이터 검색* http://www.geomuse.org/mapstedi/client/textSearch.html. *이 예는 레코드 수준의 문서화를 이용해 특정 정밀도를 가진 데이터의 검색이 가능하다는 것을 보여주고 있다.*

사용자가 이러한 데이터의 품질을 결정할 수 있으려면, 공간 데이터의 정확도, 정밀도, 그리고 오류를 문서화하는 것은 필수적이다. 이러한 문서화에는 다음 사항들을 (최소한) 포함해야 한다:

- 데이터집합 제목
- 데이터의 출처
- 데이터 이력 (데이터 수집 또는 파생 이후로 데이터에 수행된 작업)
- 정확도 (위치, 시간 및 특성)
- 논리적인 일관성
- 데이터의 일자 및 예상 유효 기간 (데이터의 현재성과 상황, 갱신 주기)
- 데이터 필드의 정의
- 수집 방법론
- 완전성
- 사용 조건과 제한점 (예, 저작권, 사용 제한사항 등)
- 관리사항 및 연락처 정보

모든 데이터 관리자들이 이러한 것들에 익숙한 것은 아니므로 이러한 용어의 일부를 정의하는 것이 좋다. 이러한 용어들 가운데 많은 것이 개별적인 수집 레코드 보다는 데이터베이스의 데이터 수집에 관한 것이다.

위치 정확성

위치 정확성은 좌표 서술문이 얼마나 이것의 실제 위치와 비교되어 가까운지를 뜻한다(Minnesota Planning 1999). 좌표의 위치를 판단하기 위해 사용되는 측지학 기준점(Geodetic Datum)은, 가능하고 알려진 경우, 기록되어야 한다.

데이터베이스가 각각의 개별 레코드에 대한 위치 정확도를 기록하는 필드를 포함하는 것이 또한 권고된다. 이것을 수행하는 여러 가지 방법이 있다. 일부 데이터베이스에서는 코드를 사용하지만, 간단한 미터 값을 사용하여 레코드의 정밀도를 표현하는 것이 선호된다. (Chapman and Busby 1994, Conn 1996, 2000, Wieczorek *et al.* 2004). 이것은 특정한 용도를 위해서 데이터를 추출하는 사용자들에게 중요할 수 있다 – 예를 들어, 사용자는 2000 미터 미만의 정확도를 가진 데이터만 원할 수 있다. 때때로, 레코드 수준에서 지리참조연산 정보가 어떻게 결정되었는가에 대한 필드를 포함하는 것이 중요할 수 있다. 예를 들어;

- 보정(Differential) GPS 의 사용
- 선택적 이용성(Selective Availability)에 의해 훼손된 휴대용 GPS (예, 2002 년 이전)
- 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-참조연산 소프트웨어를 통해 자동으로 획득
- 이름, 일자 및 지명사전의 버전을 포함하는 지명사전을 사용

속성 정확성

속성의 정확성은 데이터의 특징들이 실제 값과 관련하여 얼마나 정확하고 신뢰성 있게 서술되었는가에 대한 평가를 가리킨다. 이상적인 경우, 이것은 각각의 정확성에 대한 속성의 목록과 정보를 포함해야 한다. 예를 들어,

> *레코드는 경험이 많은 관찰자들이 제공한다. 추가적인 정확성은 전문적인 검증 목적으로 박물관 또는 식물표본관에 수장된 확증 표본에 대해 속성의 정확성을 테스트함으로써 얻어진다. 약 40% 의 식물 레코드, 양서류는 51%, 포유류는 12%, 파충류는 18%, 그리고 조류의 1% 가 확증 표본으로 검증된다* (SA Dept. Env. & Planning 2002).

이력 (Lineage)

이력은 데이터의 출처와 함께 데이터집합이 현재 상태가 되기까지 처리되고 취해진 방법을 가리킨다. 이것은 수집 방법(예, "10 X 10 미터 격자에서 수집된 데이터")과 데이터에 수행된 검증 테스트에 관한 정보를 포함할 수 있다. 처리 과정의 이력은 다음 사항을 포함할 수 있다:

- 데이터 획득 방법
- 모든 중간 처리 과정 또는 방법
- 최종 생산물을 만드는데 사용된 방법
- 데이터에 수행된 검증 방법.

예를 들어;

> *데이터는 20 미터 x 20 미터의 고정된 격자 안에서 수집되었다. 전체 종 개수, 구조, 그리고 다른 서식 데이터도 수집되었다. 이 데이터는 트윈스팬(Twinspan)을 이용하여 비슷한 종들로 구성된 그룹으로 분류되었다.*

논리적인 일관성

논리적인 일관성은 데이터의 항목간 논리적인 관계를 간략하게 평가할 수 있게 한다. 여기에서 수집되는 대부분의 데이터(박물관과 식물 표본관 데이터)의 경우, 이러한 항목 가운데 일부는 관련이 없을 수 있지만, 일부 관찰 데이터(국립공원 또는 생태 지역 등의 종 점검 목록)와 일부 조사 데이터에는 해당이 될 수 있다. 데이터가 디지털 형식으로 저장되는 공간 데이터의 경우, 논리적인 일관성 검사를 자동으로 수행할 수 있다. 이러한 것들은 다음과 같다

- 모든 지점, 선, 그리고 다각형에 레이블(설명)이 있고 중복된 레이블이 있는가?
- 선들은 노드에서 교차되는가 또는 뜻하지 않게 엇갈리는가?
- 다각형의 경계는 닫혀 있는가?
- 모든 지점, 선, 그리고 다각형이 위상학적으로 연관되어 있는가?

논리적인 일관성은 데이터집합내의 항목 또는 객체간에 논리적인 관계가 있는 경우에도 또한 적용될 수 있다. 이러한 경우 관계성에 대해 수행한 검사 결과의 서술문이 포함되어야 한다. 사례로는 서로 다른 필드에서 발생하는 날짜가 있을 수 있다 – 하나의 필드에 있는 날짜는 프로젝트가 'a'와 'b'년도의 사이에 수행된 것을 나타내지만 또 다른 필드에 있는 속성의 기록 날짜가 이 범위를 벗어날 경우, 이것은 논리적으로 일관적이지 않은 것이다; 또는 레코드가 지리적인 범위 밖에 있는 것이다 – 하나의 필드에는 데이터가 브라질에서 수집된 것으로 기록하고, 다른 필드에는 파라과이의 레코드에 대한 위도와 경도를 포함하고 있다면, 이것은 두 개의 필드간 논리적 비일관성을 나타내는 것이다. 수행된 검사에 대해 문서화하는 것은 메타데이터의 중요한 부분이다. 이것은 "다각형 내의 점 검사"와 같은 검사 내용이 포함될 수 있고 GIS 분야에서 비슷한 목적으로 사용된다. 관련 문서인 *데이터 정제의 원칙과 방법(Principles and Methods of Data Cleaning)*에서 상세히 설명되는 방법들을 참고하기 바란다.

완전성

완전성은 데이터의 전체 가능한 범위의 일부로서 데이터 또는 데이터집합의 시간 및 공간적인 범위를 가리키는 것이다. 완전성을 문서화하는 것은 품질을 결정하기 위한 핵심 요소이다. 다음과 같은 사례가 포함될 수 있다:

> *30º S 북쪽에 대해서는 완전하고, 30º 과 40º S 사이의 지역에 대해서는 레코드가 산재되어 있음.*
>
> *데이터집합은 1995 년 이전에 비계획적으로 수집된 레코드만을 포함하고 있으며, 주로 뉴사우스웨일즈 지역에서 수집되었으며, 다른 주에서 수집된 일부 레코드가 있다.*

사용자의 관점에서 완전성은 "필요로 하는 모든 데이터"와 관련이 있다 (English 1999). 즉, 사용자는 데이터베이스가 분석에 필요한 모든 필드를 포함하고 있는지를 알 필요가 있고 이러한 몇몇 필드의 "완전성"을 알 필요가 있다.

예를 들어, 사용자는 시간의 경과에 따른 속성을 비교하는 연구를 수행하기를 원할 수 있으며, 하지만 이 데이터베이스가 특정 년도 까지의 데이터만을 포함하고 있을 경우, 이것은 이 분석에 유용하지 않을 수 있다 (위의 두 번째 사례를 참고).

접근성

데이터가 사용자에게 가치가 있으려면, 이것에 접근할 수 있어야 한다. 모든 데이터가 온라인 상으로 접근 가능한 것은 아니고, 일부 데이터에 접근하기 위해서 사용자는 관리자에

연락을 하여 접근 허가를 요청할 필요가 있으며, 또는 필요로 하는 데이터의 복사본을 CD 로 구해야 한다. 접근 (그리고 사용) 조건의 문서화는 사용자가 데이터를 접근할 수 있도록 하기 때문에 중요하고, 따라서 데이터 품질 요소 가운데 하나이다. 접근성에 대한 문서화에 다음 사항들이 포함될 수 있다:

- 데이터의 연락자 주소
- 접근 조건
- (데이터가 전자적으로 이용 가능한 경우) 접근 방법
- 데이터 형식
- 주의사항
- 저작권 정보
- (필요한) 비용
- 사용에 대한 제약사항

시간적 정확성

시간적인 정확성은 시간과 관련된 정보의 정확성을 가리킨다. 예를 들면"*월 단위까지만 정확한 데이터*"가 있다. 데이터베이스가 '일(day)' 필드에 null 값을 넣을 수 없고 관련 정보가 이용가능하지 못할 경우, 자동으로 이 필드에 "1"을 넣는 경우에 이것은 중요할 수 있다 이것은 정밀도에 대한 잘못된 인상을 초래할 수 있다. 이것은 레코드의 년도만이 알려져 있고 데이터베이스가 자동으로 이것을 1 월 1 일로 기록하는 경우 더욱 중요할 수 있다. 예를 들어, 사용자가 식물의 개화 시기 또는 새의 이주 패턴을 연구하고 있다면, 사용자들은 이 정보를 알 필요가 있는데 그 이유는 데이터의 품질이 낮고 "사용 목적에 적합하지" 않기 때문에 이러한 레코드를 제외하기 위해서이다.

검증 절차의 문서화

데이터에 어떠한 오류가 존재하는가를 인지하는 열쇠 가운데 하나는 문서화이다. 데이터 품질 검사를 수행하고 오류가 정정되었지만 이것에 대한 문서화가 완전하게 되지 않는다면 그 소용이 거의 없게 된다. 이것은 데이터의 원제작자 이외의 다른 사람이 이러한 검사를 수행하고 있는 경우에 특히 중요하다. 인지된 오류가 전혀 오류가 아니고 변경이 오류를 추가할 가능성이 항상 있다. 검사를 반복해서 수행하지 않는 것이 또한 중요하다. 이러한 방식으로 자원을 낭비할 여력이 없다. 예를 들어, 사용자가 데이터에 대해 수행한 데이터 품질 검사는 여러 개의 의심이 되는 레코드를 식별할 수 있다. 이러한 레코드는 다음에 검사되어 올바른 레코드이고 실제 특이점인 것으로 판명될 수 있다. 이 정보가 레코드에 문서화되지 않으면, 한참 후에, 다른 어떤 사람이 와서 데이터 품질 검사를 수행하여 동일 레코드를 다시 의심되는 것으로 식별할 수 있다. 이 사람은 다음으로 이 레코드를 자신의 분석에서 제외하거나 이 정보를 다시 검사하는데 소중한 시간을 낭비할 수 있다. 이것은 기본적인 위험 관리이고, 모든 데이터 관리자와 큐레이터는 이것을 일상적으로 수행하여야 한다. 올바른 문서화의 가치와 필요성은 아무리 강조해도 지나치지 않다. 이것은 데이터가 어떠한 것이고, 품질은 어떠하며, 어떠한 목적에 이 데이터가 적합할 것인지에 관해 사용자가 알 수 있도록 돕는다. 이것은 또한 데이터 관리자와 큐레이터가 데이터와 이것의 품질을 추적하고 가정된 오류를 다시 검사하는데 자원을 낭비하지 않도록 돕는다.

문서화와 데이터베이스 설계

오류가 철저히 문서화되도록 확실히 하는 방법 중의 하나는 데이터베이스 설계와 구축의 초기 계획 단계에서 이것을 포함하는 것이다. 그 이후 추가적인 데이터 품질/정확도 필드가

반영될 수 있다. 이러한 것에는 위치 또는 지리코드 정확도, 지리참조연산 정보와 고도에 대한 정보 출처와 같은 필드, 그리고 이 정보를 추가한 사람에 대한 필드 등이 있다. 좌표 데이터는 GPS 를 사용한 수집자에 의해 추가되었는가 또는 특정 축척의 지도를 사용하여 나중에 데이터 입력자에 의해 추가 되었는가, 고도가 DEM 으로부터 자동으로 생성되었는가, 그렇다면 DEM, 날짜 그리고 축척 등의 출처는 무엇인가. 이러한 모든 정보는 이 정보가 특정 사용에 가치가 있는지 없는지를 결정할 때 나중에 가치가 있게 될 것이고, 데이터의 사용자는 이것을 결정하게 될 것이다.

> *"데이터 사용자는 적어도 일부 성능 특징에 관한 문서가 제공되지 않는 분류 데이터에 대한 생물학적 평가를 수행할 때 주의를 기울일 필요가 있다"*. (Stribling *et al.* 2003).

데이터의 저장

데이터의 저장은 여러 가지 면에서 데이터 품질에 영향을 줄 수 있다. 이러한 것 가운데 많은 것들이 명백하지 않지만 저장 용기(데이터베이스)를 설계하고 데이터 품질 사슬의 한 단위로서 고려해야 할 필요가 있다.

데이터베이스의 선택과 개발에 관한 토픽은 여기에서 다루기에는 너무나 광범위한 토픽이며, 별도의 연구 주제가 되어야 한다. GBIF 의 위탁에 의해 수행된 하나의 연구는 수집물 관리 소프트웨어를 조사하였고 필자는 이 문서를 독자들이 참고하기 바란다.

이 섹션에서는 데이터 품질 관련한 데이터 저장의 주요 원칙 중의 일부를 살펴볼 것이다.

데이터의 백업

데이터의 정기적인 백업은 일관된 품질 수준을 유지하는데 도움이 된다. 조직은 현재의 재해 복구 및 백업 절차를 운영하는 것이 중요하다. 데이터가 유실 또는 손실될 때마다 품질 손실이 동반한다.

보관(Archiving)

데이터의 보관(노후화 및 배치를 포함하는)은 더 많은 주의가 필요로 하는 데이터 관리 및 위험 관리의 분야이다. 데이터 보관은 특히 대학교, NGO 및 개인에게 데이터 관리의 우선 순위 사항이어야 한다. 대학교는 직원의 이직률이 높고 종종 연구 데이터는 분산된 방식으로 저장된다 – 통상적으로 연구자 개인 PC 또는 캐비닛에 문서화되어 보관된다. 온전하게 문서화되지 않으면, 이러한 데이터는 매우 빠르게 유용성과 접근성을 잃게 될 수 있다. 연구자가 해당 조직을 떠난 후에 때때로 이것이 버려지는 경우가 자주 있는데, 누구도 이것이 무엇인지 모르고 이것을 유지하는 노력을 기울이지 않기 때문이다. 특히 대학교가 내실 있는 문서화 및 보관 전략이 필요한 이유가 바로 이러한 것 때문이다.

주요 기관이 아닌 곳에 있는 개별 연구자들은 자신들의 데이터가 그들의 사후 또는 자신들이 이것에 대한 연구를 중지한 이후에 유지되고 보관될 수 있도록 할 필요가 있다. 비슷하게 데이터 저장에 장기적인 기금을 가질 수 없는 NGO 조직들은 장기 (보관을 포함하여) 데이터 관리 전략을 가지고 있는 적합한 조직 및 데이터에 관심이 있는 사람과 협약을 맺을 필요가 있다.

데이터 보관은 최근에 DiGIR/DarwinCore 및 BioCASE/ABCD[7] 프로토콜의 개발로 더욱 쉬워졌다. 이러한 것은 기관, 대학교의 부서 또는 개인이 이러한 형식중의 하나로 자신들의 데이터를 추출하고, 자신의 사이트 또는 서비스 기관에 전달하여, XML 형식으로 저장할 수 있게 하는 손쉬운 방법을 제공한다. 이것은 쉽게 데이터를 영구히 저장하고 GBIF 의 데이터 포탈과 같은 분산 검색 시스템을 이용하여 이용이 가능하도록 하는 방법이다.

데이터의 정리, 처분 및 보관은 웹상의 데이터에 대해서도 또한 관심사항이다. 제작자가 폐기하거나 과거의 쓰이지 않는 데이터를 담고 있는 웹 사이트는 사이버공간을 디지털 쓰레기로 (다양한 참조사이트와 함께) 더럽히게 된다. 정보 관리 사슬에 내재된 데이터 보관 전략이 조직에 필요하다. 데이터의 물리적인 보관은 여기에서 다루기에는 너무 큰 토픽이지만, CD 와 DVD 를 사용한 데이터 보관에 관한 최근 문서가

[7] http://www.tdwg.org; http://www.gbif.org/links/standards

정보도서관자원위원회(Council on Information and Library Resources) 및 미국국립표준기술원(United States National Institute of Standards and Technology)에서 출판되었다 (Byers 2003). 이것은 이 기술에 대한 가치 있는 요약 내용으로 독자들은 이 문서의 참고를 즐길 수 있을 것이다.

(법률적으로 또는 다른 이유로) 더 이상 요구되지 않는 데이터는 다른 모든 가능성(보관을 포함)을 조사하여 활용하지 않은채 폐기되거나 위험상태에 놓이지 않게 하여야 한다 (NLWRA 2003).

데이터 온전성(Data Integrity)

데이터 온전성은 데이터가 승인없이 변경되거나 폐기되지 않는 것을 가리키며, (바이러스 또는 전원 스파크 등과 같은) 사고로 또는 악의적으로 수정, 변경, 또는 폐기되지 않아야 한다.

데이터는 종종 변경된다 – 예를 들어, 재결정에 따른 레코드의 분류학적 정보가 갱신되는 경우이다 – 그러나 사용자는 컴퓨터 시스템이 데이터의 온전성을 유지하고 컴퓨터 시스템 자체는 우연히 또는 부정확하게 데이터를 변경하지 않을 것으로 기대한다. *데이터 손상(data corruption)*은 데이터 온전성이 지켜지지 않고 우연히 또는 부정확한 변경이 발생하는 것을 뜻한다.

데이터 온전성은 올바른 데이터 관리, 저장, 백업 그리고 보관을 통해 이루어진다.

오류의 패턴

다른 모든 데이터베이스처럼 분류학 및 종-발생 데이터베이스는 콘텐트 오류 패턴이 발생하기 쉽다. 잉글리시(English, 1999)는 다음과 같은 오류 패턴을 인식하였고 데이터 결함(data defects)으로 명명하였다. 달신(Dalcin, 2004)은 분류학 데이터에 사용할 목적으로 이러한 것을 채택하였다. 여기에서 보이는 값은 채프만(Chapman 1991)에서 인용된 예제와 함께 잉글리시에서 인용되었고 호주가상식물표본관[8]과 브라질의 speciesLink[9]의 데이터베이스에서 가져온 것이다:

- **도메인 값 중복** – 비표준화된 데이터 값, 또는 비슷한 값이 존재하고, 이 중에서 두 개 또는 그 이상의 값 또는 코드가 같은 의미를 갖는다. 중복성은 표준화된 용어를 사용하지 않거나 서로 다른 출처의 데이터에 대한 편집이 잘못되는 서술 데이터의 경우 매우 전형적인 것이다.
- **누락된 데이터 값** – 값이 있어야 할 데이터 필드가 없는 경우이다. 이것은 필수 필드와 데이터 획득 시에 필수 입력 필드가 아니지만 나중에 처리가 필요한 것을 모두 포함한다. 사례로는 지리-참조연산 또는 좌표 값(위도와 경도)이 있다.
- **부정확한 데이터 값** – 이러한 것은 철자 입력의 뒤바뀜, 잘못된 곳에 데이터 입력, 획득한 데이터의 잘못된 이해, 레이블의 기록을 읽을 수 없음, 또는 필수 필드에 값이

[8] http://www.cpbr.gov.au/avh/
[9] http://specieslink.cria.org.br/

요구되지만 데이터 입력자가 입력 값을 모르는 경우 등에 의해 발생한다. 부정확한 데이터 값은 가장 명백하고 일반적인 오류이며 모든 필드의 모든 값에 영향을 줄 수 있다. 학명에서 철자 오류는 분류학 및 명명학 데이터베이스에서 부정확한 데이터 값과 연관된 일반적인 패턴이고, 지리-참조연산 필드에 0을 두는 것 등이 그러하다.

- **세분화되지 않은 데이터 값** – 이것은 하나 이상의 사실이 동일 필드에 입력되는 경우 발생한다 (예, 동일 필드에 속(genus), 종(species) 그리고 저자가 있는 경우 또는 순위와 종이하 이름이 있는 경우). 이러한 종류의 오류는 통상적으로 데이터베이스 설계를 부실하게 한 결과이다. 이러한 종류의 오류 패턴은 데이터 통합을 할 때 실질적인 문제를 야기할 수 있다.

Genus	Species	Infraspecies
Eucalyptus	globulus	subsp. bicostata
Family	**Species**	
Myrtaceae	Eucalyptus globulus Labill.	

Table 4. *세분화되지 않은 데이터 값의 예.*

- **도메인 모순(schizophrenia)** – 의도되지 않은 목적으로 사용된 필드로 이것은 결과적으로 하나 이상의 속성 값을 포함하게 된다.

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 5. *도메인 모순의 예*

- **중복 발생** – 하나의 개체를 나타내는 여러 개의 레코드들. 가장 일반적인 경우는 이름에서 발생하는데 대체 철자 또는 유효 명명 대체이름이 그것이다. 이러한 것은 사용자가 이름을 검색할 때 또는 서로 다른 데이터베이스의 데이터를 결합하려고 할 때 어려운 점이 될 수 있다. 예:
    - Phaius tancarvilleae
    - Phaius tankervilliae
    - Phaius tankarvilleae
    - Phaius tankervilleae
    - Phaius tankervillae

    - Brassicaceae/Cruciferae (exact equivalents; both of which are allowed by the International Botanical Code).
- **비일관적인 데이터 값** – 데이터베이스와 관련된 데이터가 비일관적으로 또는 두 개의 데이터베이스가 서로 다른 시간에 갱신될 때 발생한다. 예를 들어, 살아있는 수집물과 식물표본관 데이터베이스, 또는 박물관 수집물 데이터베이스와 관련 사진 데이터베이스가 있다.

- **정보 품질 오염** – 이것은 정확한 데이터와 부정확한 데이터가 결합되어 발생한다. 예를 들어, 아종 수준에 있는 데이터를 종 수준까지의 데이터만 포함하는 데이터베이스에 결합하는 경우가 있다.

공간 데이터

공간 데이터의 저장은 장소 정보 (텍스트 형태의 장소 정보)와 함께 통상적으로 좌표 쌍(동방위 및 북방위)으로 주어지는 좌표 정보(지리참조연산 데이터)를 포함한다. 많은 데이터베이스가 이제 자유롭게 서술된 텍스트 형식의 장소 내용과 함께 가장 가까운 장소, 거리 그리 방향과 같은 파싱된 또는 세분화된 장소 데이터의 저장을 시작하고 있다. 이러한 세분화된 필드를 생성할 수 있도록 자연어로 된 장소 데이터에 대한 파싱을 개선하고 지리참조연산 과정을 지원하는 몇 개의 프로젝트가 현재 진행중이다. 고돈베티무어 재단에서 최근 지원한 BioGeomance 프로젝트[10]가 이러한 것중의 하나이다.

지리-참조연산 (또는 좌표) 정보는 일반적으로 위도 또는 경도 (구면 좌표 시스템) 또는 UTM (또는 관련된) 좌표(평면 좌표 시스템)로서 데이터베이스에 입력된다. 위도 및 경도와 같은 구면 좌표 시스템은 지구를 일주하며 종이 지도상에 표현되기 위해서는 투영(projections)으로 알려진 통상적이지 않은 방식으로 펼쳐져야 한다. 구면 좌표 시스템은 동등 구역을 가지고 있지 않으며 하나의 위도 1 도와 다음 위도 1 도 사이의 거리는 이것이 적도 또는 극에 가까운 정도에 따라 상당히 다를 수 있다. 평면 좌표 시스템은 동등 구역 투영에 더욱 가깝고 구역을 측정하거나 계산 목적으로 사용될 수 있다.

많은 기관들이 현재 데이터를 도(degrees), 분(minutes), 초(seconds) 또는 도(degrees)와 (많은 GPS 기기들이 보여주는 것처럼) 십진분 형태로 입력을 시작하고 있으며, 저장 목적으로 데이터베이스를 십진법 표기 도수로 변환하고 있다. GIS 에서 전송 및 사용 목적을 위해서는 데이터를 십진법 표기 도수의 형태로 저장하는 것이 일반적으로 가장 좋은데 이것이 쉬운 데이터 전송과 가장 높은 정밀도를 제공하기 때문이다.

데이터가 단지 하나의 UTM 지역(Zone)에 제한되는 경우, 이러한 기관에서 UTM 좌표로 종종 데이터를 저장한다. 이것은 위에서 논의된 것과 같은 장점이 있고 각각의 격자는 정사각형 (또는 직사각형)으로 평면 지도상에서 쉽게 표현이 가능하고 거리 또는 구역의 계산이 가능하다. 하지만 데이터를 UTM (또는 관련) 좌표 시스템으로 저장할 때 해당 지역(Zone)이 또한 저장되도록 하는 것이 중요한데, 그렇지 않으면 다른 지역 또는 기관으로부터의 데이터와 결합할 때 어려운 점이 발생한다.

십진법 표기 도수

많은 데이터베이스에서 십진법 표기 도수의 저장은 위에서 언급한 것처럼 *거짓 정밀도(False Precision)*를 이끌 수 있다. 데이터가 저장되는 곳의 정밀도를 고려해야 한다. 이 데이터베이스는 데이터베이스내의 데이터가 가지는 가장 높은 정밀도 이상의 것을 보고하지 않아야 한다. 대부분의 생물 데이터의 경우, 이것은 4 개의 자리 수 이다 (즉 , 10 미터).

기준점(Datums)

많은 가능한 측지선 기준점이 있다. 지구는 완전 구가 아닌 타원형으로 좌표 시스템을 타원의 표면에 일치시키려고 할 때 어려움이 발생한다(Chapman *et al.* 2005). 이 문제의

[10] http://www.biogeomancer.org

해결을 위해 ‘기준점(datum)’의 개념이 만들어졌다. 기준점은 구의 한 지점에서 회전하는 타원을 참조하기 위해 사용되는 일련의 점이다. 역사적으로, 지구의 서로 다른 부분에 대해 서로 다른 참조 시스템이 생성되었으며, 인공위성의 출현으로 진정한 지구 참조시스템 또는 기준점이 생성될 수 있었는데 그 이유는 인공위성이 지구의 중심을 고정하는데 사용되었기 때문이다. 서로 다른 측지선 기준점을 사용하는 지구상의 위도 및 경도에 대한 차이는 400 미터 또는 그 이상이 될 수 있다 (Wieczorek 2001).

이 차이 때문에, 데이터베이스가 사용된 기준점을 기록하는 것이 중요하며, 그렇지 않을 경우 데이터가 결합될 때 동일 장소에 대한 두 개의 기록 자료에 대한 오류는 매우 클 것이다.

공간 데이터 다루기

공간 데이터를 다루는 많은 방법이 있다. 많은 것들이 공간 데이터의 정확도에 대해 영향을 끼치지는 않지만, 일부는 그러하다. 공간 데이터의 위치 정확도에 영향을 끼치는 방법 가운데 일부 사례는 아래와 같다

하나의 형식에서 다른 것으로 데이터 변환

아마도 종 및 종-발생 데이터의 수집, 저장 그리고 사용과 관련된 사람들이 수행하는 가장 흔한 데이터 변환은 지리코드의 변환일 것이고, 이것은 도(degrees)/분(minutes)/초(seconds)를 십진법 표기 도수로의 변환(DMS 에서 DD 로), 또는 UTM 좌표를 십진법 표기 도수의 변환(UTM 에서 DD 로)이 있다. 다른 것들은 텍스트 형식의 정보 기재문에서 마일(miles)을 킬로미터로의 변환, 고도 및 깊이 기록 자료에서 피트(feet)를 미터로의 변환 등이 있다.

이러한 모든 것은 꽤 간단한 변환이지만 정밀도를 잘못 사용함으로써 정확도에 대한 잘못된 인상을 줄 수 있다. 예를 들어, 250 피트의 고도를 가진 수집물(수집자는 200 과 300 피트의 사이에 있는 값으로 기록했을 수 있다)은 미터로 변환될 경우 76.2 미터, 반올림하면 아마도 76 미터가 될 것이다. 변환된 값을 80 미터로 기록하는 것이 좋고, 아마도 (±) 20 미터를 추가할 수 있도록 하는 정밀도 필드를 포함하는 것이 더욱 나을 것이다. 정밀도를 잘못 사용하면 정밀도가 높아진 것으로 보일 수 있으나 실제로는 그 품질이 떨어진다.

기준점과 투영

하나의 측지선 기준점에서 다른 것으로 데이터를 변환하는 것은 꽤 심각한 오류를 이끌 수 있는데 이 변환이 균일하게 되지 않기 때문이다 (데이터 품질에 관한 기준점과 이것의 효과의 논의 자료 관련하여 Wieczorek 2001 참고). 많은 국가 또는 지역에서는 현재 데이터를 그들의 지역에 맞는 하나의 표준으로 데이터를 변환하고 있다 – 세계 측지선 기준점, 또는 이것에 매우 근접하게 근사화한 측지선 (호주 측지선 기준점 (AGD84), 호주에서 이것은 WGS84 와 약 10cm 차이가 있다; 유럽의 EUREF89 에서는 WGS84 와 약 20cm 차이가 있다). 예를 들어 하나의 측지선 위치에서 다른 것으로의 변환은 데이터가 약 5 또는 10 km 정도의 정밀도를 가진다면 아마도 필요하지 않을 것이다. 하지만 10-100m 정밀도를 가진 데이터를 다루고 있다면, 기준선 이동은 매우 심각하고 중요할 수 있다 (400m 까지 또는 그 이상의 일부 지역 – Wieczorek 2001).

유사하게 매핑된 데이터가 다각형 안에 있는 경우 (예, 국립공원의 수집물), 하나의 투영에서 다른 것(예, Albers 를 Geographic 으로 변환)으로 변환할 때 일어날 수 있는 오류를 인지할 필요가 있다. 이와 같은 변환을 할 때 발생하는 오류를 계산하는 표준 공식을 이용할 수 있고, 데이터와 함께 저장되는 메타데이터에 이 정보를 반영해야 한다.

격자(Grids)

데이터가 벡터 형식에서 래스터(raster) 또는 그리드 형식으로 변환될 때마다, 정확도와 정밀도는 손실된다. 이것은 벡터 데이터를 근사화할 때 사용하는 래스터 파일의 격자 셀 크기 때문에 발생한다 (Burrough and McDonnell 1998). 이 데이터를 다시 벡터 형식으로 변환해도 정밀도와 정확도를 다시 얻을 수 없다. 래스터 데이터를 사용하고 변환할 때 부딪히게 되는 문제에 관한 상세한 논의는 Chapman *et al.* (2004)를 참고하기 바란다.

데이터 온전성

지리 데이터집합이 일관성이 없을 경우 통합하기가 어렵다. 이러한 비일관성은 데이터의 공간 및 속성 특징과 관련되어 있고, 다양하면서 때때로 시간이 오래 소요되는 정정 수단들이 필요할 수 있다 (Shepherd 1991). 비일관성은 다음 사항에서 발생한다:

- 기록 또는 측정 기법의 차이(예, 관찰 데이터의 구역 크기와 경과 시간), 조사 방법(예, 그리드 크기, 횡단면의 너비) 또는 데이터 범주(예, 범주 데이터에 대해 서로 다른 범주 정의)
- 측정 수단 또는 조사 방법상의 오류 (예, 옮겨 적음, 데이터 기록, 동정상의 오류)
- 해상도의 차이 (공간, 시간 또는 속성)
- 모호하고 정밀하지 않은 정의
- 객체의 모호함 (예, 토지 또는 식물의 경계, 일부는 종(species), 다른 일부는 아종(subspecies), 그 외의 것은 단지 속(genus)에 속하는 동정)
- 용어와 명명의 해석 또는 사용에 대한 차이 (예, 서로 다른 분류가 사용됨)
- GPS 설정의 차이 (기준점, 좌표 시스템 등)

이런 통합 문제는 데이터가 아래 경우에 더욱 심하다:

- 서로 다른 유형 (예, 박물관의 표본 데이터와 조사 및 관찰 데이터의 혼합)
- 서로 다른 국가 (예, 조사 방법론이 서로 다를 수 있다)
- 다수의 출처에서 얻은 것
- 다수의 축척
- 서로 다른 데이터 타입으로 이루어진 것 (지도, 표본, 사진 등)
- 서로 다른 시간대
- 서로 다른 데이터베이스 타입, 미디어 등에 저장된 것 (예, 일부 데이터베이스 소프트웨어는 'null' 값을 허용하지 않는다)
- 다양하게 파싱되는 것 (예, 하나의 데이터집합은 전체 학명을 하나의 필드에 저장하지만, 다른 것들은 속(genus), 종(species)의 경우 분리된 필드에 나누어 저장하는 것)

데이터 통합은 데이터 관리자가 일관된 데이터 저장 표준을 따르고 사용할 때 더 높은 품질 결과를 산출한다.

표시와 표현

> *품질이 어떠하든 간에 존재하는 데이터를 가장 효율적으로 사용할 수 있도록 하는 방법이 항상 개발되어야 한다. 하지만, 데이터가 신뢰성이 있기 위해서는, 신뢰성의 수준을 나타내는 정보로 이것이 또한 검증되거나 이러한 정보와 함께 동반되어야 한다.* (Olivieri *et al.* 1995)

과학자와 기관들은 생물다양성을 이해하고, 설명하고, 수량화하고, 평가를 하는 역할에서 점점 더 정보 제공자로서 인식되고 있다. 이 인식은 의사-결정자, 관리자, 일반 대중 그리고 다른 사람들에게 신뢰할 수 있고 유용한 정보를 제공할 수 있다는 것에 기반하고 있다. 빈약하게 데이터베이스를 관리한 결과물로서 모호하고, 혼동되고, 불완전하고, 모순적이고 오류가 있는 정보는 정보 제공자와 과학적 전거기관으로서의 그들의 명성에 영향을 줄 수 있다 (Dalcin 2004).

생물학에서 디지털 데이터 처리의 핵심 목적은 정보의 사용자에게 검색과 정보를 분석하는 비용-효과적인 방법을 제공하는 것이다. 이러한 관점에서, 이것의 성공은 사용자에게 생물학 세계의 정확한 관점을 제공하는 것의 정도에 의해 결정된다. 그러나 생물학은 대단히 복잡하며, 표현되고 이해되기 위해서는 일반화, 근사화 그리고 추상화되어야 한다 (Goodchild *et al.* 1991). 이것을 하는 방법들은 지리 정보 시스템, 환경 모델링 도구 그리고 의사결정 지원 시스템을 사용하는 것이다. 하지만 이러한 도구를 사용할 때 변이를 샘플화하고 측정하고 오류와 불확실성을 서술하고 가시화하는 것이 중요하다. 최선의 실행사례로 간주되는 것에 도달하기 까지 여전히 가야 할 길이 많은 분야가 이 분야이다.

생물학은 오류 막대, 다양한 통계 방법 및 통계치를 사용하여 오류 보고의 기법을 개발한 초창기 학문 가운데 하나이다. 오류의 보고는 약점으로 여겨지지 않았는데 오류 측정치가 데이터의 정확한 해석을 위한 중대한 정보를 제공했기 때문이다. 이러한 데이터의 사용자가 이 데이터를 정확하게 해석하고 사용할 수 있도록 하기 위해, 종 데이터를 전달할 때, 유사한 오류 보고 기법이 개발되고 사용될 필요가 있다.

효과적인 데이터 품질 프로그램은 내부적으로 그리고 공공연히 조직과 개인이 난감함을 당하지 않도록 도움을 준다.

사용자의 필요를 판단하기

사용자의 필요를 판단하는 것은 간단한 과정이 아니고, 세부적인 요구사항을 만들어 이러한 요구사항을 충족시킬 수 있도록 데이터를 구조화하는 것이 쉽지 않다. 하지만 핵심 사용자를 파악하고 이 사람들과 함께 필요사항과 요구사항을 만드는 것이 중요하다. 올바른 데이터-사용자 요구사항은 더 나은 그리고 효율적인 데이터 수집, 데이터 관리 및 전반적인 데이터 품질을 이끌 수 있다.

관련성

관련성은 "품질"과 밀접한 관련이 있고 이것이 요구되는 사용에 대한 데이터의 관련성을 가리킨다. 이것은 어느 한 지역에 대한 식물군이 존재하지 않을 경우 이 지역에 대해 어느 하나의 식물군을 사용하는 것과 같은 간단한 것일 수 있으며 또는 데이터가 유용하게 되고

"관련성" 있게 만드는데 상당한 노력이 요구될 수 있는 서로 다른 투영에 있는 데이터일 수도 있다.

신뢰성

신뢰성은 사용자가 데이터를 신뢰하는 정도를 나타낸다(Dalcin 2004). 이것은 종종 사용자인지 또는 자신들의 목적상 데이터의 적합성을 평가하는 것에 대한 문제이고 과거의 경험 또는 일반적으로 수용되는 표준과의 비교에 기반할 수 있다 (Pipino *et al.* 2002). 데이터집합의 명성은 때때로 사용자가 인지하는 신뢰성 (그리고 따라서 유용성)에 따라 좌우되지만 이것은 올바른 문서화로 가끔 향상될 수 있는 부분이다.

왕과 그의 동료들(Wang *et al.* 1995)은 이러한 토픽의 많은 것을 계층적인 표현에 관계시키면서 신뢰성 및 명성 등과 같은 개체간의 관계성을 보여주는 다이어그램을 포함하였다.

공간 데이터의 불확실성

불확실성은, 특히 공간 데이터의 경우, 실제 있는 사실이지만, 종종 데이터 내의 불확실성이 종종 잘 기록되지 않고 있으며, 사용자에게 항상 명확한 것은 아니다. 사용하기 쉬운 PC 용 지도 시스템의 많은 등장으로 GIS 비전문가도 쉽게 자신들의 데이터에 대해 공간적인 관계성을 가시화하고 분석할 수 있게 되었지만 종종 부적합한 축척(Chapman *et al.* 2005)이 사용되고 있으며, 데이터 내의 본질적인 공간 오류와 불확실성을 고려하고 있지 않다 (Chapman 1999). 일부 사례에서 이것은 위험한 데이터의 오용을 야기할 수 있고 때때로 비극적인 결과를 초래할 수 있다(Redman 2001). 최근 전통적인 데스크톱 GIS 에서처럼 사용자가 공간 데이터를 보고 분석할 수 있도록 하는 간단한 온라인 지도 서비스가 증가하고 있고, 이것은 또한 서비스 제공자가 보여지는 데이터 계층과 데이터 집합의 축척을 제어 가능하도록 하고 있다. 가까운 미래에 이것은 웹 지도 서비스(Web Mapping Services, WMS)의 기능과 함께 더욱 더 확장될 것이다. 지도 제공자에 의한 데이터 계층과 축척의 제어(예, 사용자가 확대 기능을 이용할 때 자동으로 서로 다른 계층을 켜고 끄는 기능)는 범할 수 있는 일부 간단한 오류를 감소시킨다.

데이터의 불확실성을 문서화하는 것이 필수적이며, 첫째 올바른 메타데이터를 사용하고, 둘째 가시화와 표현을 통해서 문서화하여야 한다. 종 및 종-발생 데이터와 관련하여 연구될 필요가 있는 하나의 분야는 불확실성을 가시화하는 기법(예를 들어 정밀도의 궤적을 보이기)을 개발하는 것이다. 수집물 레코드를 위도 및 경도의 점으로 표기하는 대신에 레코드의 연관된 정밀도를 포함하고 따라서 궤적으로서 위치(원, 타원 등)를 표현하는 것이 필요하며 심지어 확률의 수준을 포함할 수 있을 것이다(Chapman 2002).

위치 또는 속성 정밀도와 관련하여 데이터와 데이터의 제한점을 아는 사람들이 이러한 정보를 문서화하고 이용가능하게 함으로써 사용자가 자신의 사용에 데이터의 적합성을 판단할 수 있도록 사용자를 지원하는 것이 중요하다.

오류와 불확실성의 가시화

공간 데이터의 효과적인 오류 가시화에 대한 새롭고 흥미있는 많은 방법들이 개발되고 있기는 하지만 아직 갈 길이 멀다 (예, Zhang and Goodchild 2002). 아마도 가장 쉬운 방법은 GIS 에서 부가적인 오버레이로서 오류 계층을 사용하는 것이다. 이와 같은 기법은 지도제작 분야에서 사용되었으며 하나의 계층은 지도의 서로 다른 부분에 대한 신뢰성을 보여주기 위해 서로 다른 강도(세기)의 명암을 제공한다. 다른 기법은 서로 다른 심볼(낮은 품질 또는

정밀도의 데이터를 가리키기 위한 실선에 반하여 점선, 서로 다른 크기 또는 강도의 점 등)을 사용하는 것과 관련될 수 있다. 이러한 오버레이를 사용하는 것은 종종 또한 오류의 원천에 대한 단서를 제공하며 이러한 것은 데이터의 검증과 검사에 가치 있는 도구가 될 수 있다.

통계 계산치가 가능할 경우, 행은 기대 결과값을 제공하고 열은 관찰된 결과값을 제공하는 오분류 행렬(misclassification matrix)을 사용하는 것이 유용하다. 이러한 경우 행과 관련된 오류는 누락의 오류이고 열과 관련된 오류는 행위에 대한 오류이다 (Chrisman 1991). 이와 같은 방법은 일반적으로 종-발생 데이터에 대해 사용되지 않지만, 예를 들어, 일정한 기간 동안 존재/부재 레코드가 관찰된 것과 같은 조사 데이터의 경우 가치가 있을 수 있다.

위험성 평가

정책 결정자들은 확실성을 선호한다; 하지만 자연계는 본질적으로 변이가 있으며 거의 이러한 기대에 부응하지 않는다. 위험성 평가 기법은 점점 더 정책 결정자와 환경 관리자들에게 환경에 대한 결정사항이 더 큰 확실성으로 이루어질 수 있도록 확실성과 위험성의 측정치를 제공하고 있다. 종(species)의 정확한 발생 지식이 종종 거의 없는 경우, '발생이 가능할 것 같은' 지역이 대체자로 사용될 수 있다. 하지만, '발생이 가능할 것 같은' 광범위한 지역에서, 다른 것보다 더 '가능성이 있는' 지역이 있을 수 있다(Chapman 2002).

위험성의 개념에는 일반적으로 두 가지 요소가 있다고 볼 수 있다 – 발생하는 어떤 것에 대한 가능성과 정도 그리고 이 사건이 실제 발생할 경우의 결과값(Beer and Ziolkowski 1995). 종 데이터의 문맥에서, 위험성 평가는 다른 곳에 백업절차가 구현되지 않았을 경우 데이터를 파괴하는 현장의 화재 위험성에서부터 품질 낮은 데이터의 사용으로 인한 환경적인 결정의 오류 위험성까지 확장될 수 있다. 이것의 사례는 멸종 위기 종이 해당 지역에서 발생했기 때문에 개발을 가로 막는 경비와 관련될 수 있다. 일부 환경 문제의 경우, 정부는 점점 더 중요한 환경 문제의 결정을 내릴 때 *예방적인 원칙(precautionary principle)*의 적용을 고려하고 있다.

법률 및 도덕적 책임사항

종 데이터의 품질과 표현 관련하여 법률적 및 도덕적 책임 사항이 발생할 수 있는 많은 분야가 있다. 이러한 것은 다음과 같다

- 저작권과 지적재산권;
- 사생활;
- 상품표기의 진실;
- 민감한 분류군에 대한 제한된 품질 표현;
- 토속권(Indigenous Rights);
- 책임;
- 경고 및 관련없음 선언

대부분의 경우 데이터의 *저작권과 지적재산권*은 이 데이터와 함께 기록되는 문서로 다루어질 수 있다. 이러한 것들이 레코드와 레코드간에 그 차이가 다양한 경우 이것은 레코드 수준에서 기록되어야 하고, 그렇지 않을 경우 이것은 메타데이터에서 다루어질 수 있다.

많은 국가에서 최근 *사생활* 법률을 도입하였으며 데이터 관리자들은 이러한 법률의 규정을 인지하여야 한다. 이것은 데이터가 정치적인 경계를 넘어 전송되는 경우 또는 인터넷을 통해 이용 가능하게 되는 경우에 특히 관련이 있다. 일부 국가에서는 개인에 관한 정보는

데이터베이스에 저장될 수 없고 또는 개인의 분명한 허락 없이는 이용 가능하게 할 수 없다. 이것이 종-발생 데이터에 부속된 정보에 어떠한 영향을 미칠지는 분명하지 않지만, 관리자들은 이 문제를 알고 있어야 하고 필요할 경우 이것에 대한 준비를 해야 한다.

올바른 메타데이터와 함께 올바른 품질 제어 수단은 통상적으로 "***내용표시의 사실성(truth in labelling)***" 개념을 따르도록 유도한다. 지금까지 적어도 법률에서 "내용표시의 사실성"은 대부분 식료품에 제한되어 있었다. 하지만 이것은 다음 프로젝트의 개발에 관한 논문에서 언급되고 있다: 세계공간데이터인프라구조(Global Spatial Data Infrastructure, Nebert and Lance 2001, Lance 2001), 국가공간데이터인프라구조(National Spatial Data Infrastructure for the USA, Nebert 1999) 그리고 호주뉴질랜드공간데이터인프라구조 (Australian and New Zealand Spatial Data Infrastructure, ANZLIC 1996b). 세계 SDI 논문(Lance 2001)에서, 공간데이터정보유통체계(Spatial Data Clearinghouse)는 "*'내용표시의 사실성'의 원칙하에서 소장물에 대한 보편적인 접근을 제공할 수 있는 자유로운 홍보 방법*"을 포함해야 한다는 것이 권고되고 있으며, 호주와 뉴질랜드의 아래 문서를 인용하고 있다:

> *"육지 및 지리 데이터 품질 표준은 서술적, 규범적 또는 둘 모두일 수 있다. 서술적 표준은 '내용표시의 사실성' 개념에 기반하며, 데이터 생산자가 데이터 품질로 알려진 것을 보고하도록 요구한다. 이것은 데이터 사용자가 해당 데이터의 '목적에 대한 적합성' 관련하여 판단을 할 수 있도록 가능하게 한다."*

장소 정보가 "모호"한 경우 ***민감한 종에 대한 제한된 품질 표현***이 발생할 수 있다 – 예를 들어, 이것은 멸종위기 종의 정확한 위치에 대한 지식 또는 민감한 종의 교역 등을 제한시키기 위한 것이다. 이것은 데이터의 출판된 품질의 감소로서, 실제로 발생할 경우 이것은 분명하게 문서화되어, 사용자들이 그들이 획득하고 있는 것을 알고, 그 이후 해당 데이터가 그들의 사용에 가치가 있는지 없는지를 결정할 수 있도록 해야 한다.

토속권(Indigenous rights)이 또한 데이터 품질에 영향을 줄 수 있으며, 토착민들의 민감성 때문에 일부 정보를 한정시켜야 하는 경우가 있을 수 있다. 이 경우 "토속민들의 권리에 상응할 목적으로 일부 데이터가 제한되고 있다"라는 것에 대한 문서화가 포함되어야 한다.

1998 년 엡스테인과 그의 동료들(Epstein *et al.*)은 공간 정보의 사용과 관련하여 법적 책임성 문제를 조사하였다. 그들이 제시한 몇몇 핵심 사항은 다음과 같다:

- *공간 정보내의 오류로 인하여 개인과 조직의 명성 및 일체성 모두의 소송 및 손실에 대한 **'상당한 잠재가능성'** 이 현재 있다.*
- *소송 사건에서 전통적인 포기선언은 강력한 항변이 되지 않을 수 있다.*
- *책임 소재를 한정시키기 위해, 조직은 **'최선의 능력과 지식으로'** 자신들의 생산품을 적합하고 진실되게 표기하는 높은 수준의 품질 문서를 유지하는 것이 요구될 수 있다.*

경고 및 포기선언은 데이터 품질 문서화의 중요한 부분이다. 이것들은 관리 기관에 대한 것 뿐만 아니라 사용자에게 데이터의 품질에 대해 몇몇 아이디어와 이 품질에서 기대될 수 있는 것들을 제공할 수 있는 방식으로 기록되어야 한다.

데이터를 생산하는 것과 관련된 대부분의 기관 및 그룹은 데이터와 정보를 이용가능하게 하는 편이성과 정보의 품질로 판단될 것이다. 정보를 출판, 공유, 접근, 통합 및 사용할 수 있는 사람들은 가장 큰 이익을 보게 될 것이다 (NLWRA 2003).

인증과 신임

종-발생 데이터가 인증될 수 있고 인증되어야만 하는가? 많은 기관의 데이터가 점점 더 이용가능하게 됨에 따라, 사용자는 어떠한 기관을 신뢰할 수 있고, 어느 기관이 문서화된 품질 제어 절차를 이행하고 있는지를 알기 원한다. 사용자는 단지 잘 알려진 기관을 신뢰하여야 하는가, 또는 신뢰할 수 있는 데이터를 가진 덜 알려진 기관이 또한 있는가? 잘 알려진 기관의 이용가능한 데이터 가운데 어느 것을 신뢰할 수 있고 어느 것이 그렇지 아니한가. 사용자가 자신들의 데이터에 대한 출처를 명시할 경우 *명성(reputation)* 하나만으로도 결정적인 요소가 될 수 있지만 명성은 주관적인 개념이고 행동과 판단을 지지하기에는 약한 성질이 있다 (Dalcin 2004). 이것이 우리 분야에서 우리가 원하는 것인가? 데이터 품질 절차에 대한 올바른 메타데이터와 문서화는 종종 명성과 같은 주관적인 요소를 사용자가 조금 더 과학적이고 논리적인 평가를 할 수 있도록 하는 다른 것으로 바꿀 수 있다. 아마도 우리는 최소의 데이터 품질 문서 표준과 절차를 따르는 조직에 대한 정보를 사용자에게 알리는 인증 및 신임 절차를 개발해야 할 것이다.

합의된 품질 인증의 개발은 전반적인 데이터 품질의 향상을 이끌고 데이터의 가치 관련하여 사용자에게 증가된 확실성을 제공할 수 있을 것이다. 이것은 거꾸로 인증된 기관들에 더 많은 기금을 유치할 수 있게 할 것이다. 달신(Dalcin, 2004)은 다음과 같이 제안하고 있다: *"분류군 데이터의 품질 인증은 세가지 측면에 수행할 수 있다: 1 차 데이터 출처(원천 자료), 정보 사슬(과정) 그리고 데이터베이스 (산출물)."*

데이터베이터의 동료 검토

데이터베이스에 대한 동료 검토 시스템은 종 데이터베이스에 대해 도입될 수 있다. 이러한 동료 검토 과정은 위에서 살펴본 것처럼 인증 절차의 입력이 될 수 있으며, 품질제어 절차, 문서화 그리고 메타데이터와 같은 문제들, 갱신 및 피드백 절차 등과 관련될 수 있다.

결론

> *모든 정보 전문가의 목표 하나는 불필요한 오류를 발생시키지 않게 하는 것이다. 직접적으로 오류를 인식함으로써, 오류를 수용할 수 있는 범위내로 한정시킬 수 있을 것이다. 아직도 오류를 손쉽게 또는 값싸게 피할 수는 없다.*
>
> (Chrisman 1991).

데이터 품질과 오류 검사의 중요성은 아무리 강조해도 지나치지 않다. 이 문서 전반에 걸쳐 강조한 것처럼, 성과물을 개발할 때 데이터가 실제 가치가 있도록 하는 것이 중요하며, 이것은 개선된 환경 결정 및 관리를 할 수 있도록 할 것이다. 데이터 품질은 이것이 박물관 또는 식물표본관의 수집물 데이터, 관찰 레코드, 조사 데이터 또는 종 점검 리스트 이든지 간에 모든 데이터에 중요한 사항이다. 세계의 많은 정부는 점점 더 데이터에 대해 고품질과 더 나은 문서화를 요구하고 있다. 예를 들면:

- 호주 연방, 주 그리고 지방 정부는 서비스를 개선하고 자원(데이터와 정보 자원 포함)을 더욱 효과적인 이용할 수 있도록 하는 강력한 지시를 하고 있다.
- 공공 경비를 통해 수집된 데이터는 공중이 이것을 접근할 수 있도록 하여 이것의 잠재성을 실현하고 이것과 관련된 상당한 생산 비용과 유지 보수비가 정당화될 수 있도록 적합하게 관리되어야 한다는 인식이 점점 늘어나고 있다.
- 고객이 올바른 데이터와 정보를 더욱 쉽고 더욱 빠르게 접근할 수 있도록 하고, 이에 대해 무료 또는 비용이 거의 필요 없이 제공될 수 있도록 하는 압력이 증가하고 있다.
- 데이터를 합리화하고 결합하여 효율성을 향상시키고 가치를 부가하기 위한 필요성에 대한 강조가 정부 내에서 증가하고 있다.
- 데이터가 관련성이 있어야 한다는 요구가 증가하고 있다. 이것은 새로운 수집물, 새로운 조사, 데이터 관리 그리고 출판에 적용된다.

고품질 데이터에 대한 필요는 명확하지만, 많은 데이터 관리자들은 자신들의 시스템에 저장되고 보여지는 데이터는 절대적이며 오류가 없거나 오류는 중요하지 않다고 가정한다. 하지만 오류와 불확실성은 본질적으로 모든 데이터에 존재하고, 모든 오류는 데이터가 적용되는 마지막 사용에 영향을 미친다. 데이터의 품질을 향상시키기 위한 데이터 입수와 관리 절차는 데이터 관리의 핵심 부분들이다. 종-발생 데이터에 책임을 가지고 있는 기관들은 정보 품질 사슬의 모든 부분들을 검사하고 개선할 필요가 있으며 문서화는 사용자가 데이터에 대해 알고 이해하여 “사용에 대한 적합성”, 그리하여 데이터의 품질을 판단할 수 있는 핵심 요소이다.

> *인간은 잠재적으로 공간 정보의 정확도와 신뢰도에 가장 큰 위협이 되는 요소이다. 이것은 또한 임의의 공간 데이터 집합에 내재한 약점에 대해 신뢰성과 예지력을 부여할 수 있는 하나의 요소이다.* (Bannerman 1999).

감사

세계 여러 곳에 있는 많은 동료들과 기관들이 이 논문에 직간접으로 기여하였다. 일부는 직접적으로, 일부는 30 년이 넘는 기간 동안 필자와의 토론을 통해, 일부는 간접적으로 출판된 논문을 통해 또는 단지 그들의 정보를 이 세계에 이용 가능하게 함으로써 기여를 하였다.

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 멕시코의 코나바이오, 캔사스 대학교, 호주의 CSIRO, 콜로라도 대학교, 코네티컷 피바디 박물과 그리고 버클리 캘리포니아 대학교, 또한 언급하기에 너무 많은 여러 기관들이 수행한 초기 연구로 인해 우리 분야의 종 데이터 품질 관리가 현재 상태에 이르고 있다. 필자는 이러한 사람들의 독창적인 생각과 건설적인 비판에 감사를 드린다.

덧붙여, 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 편집 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; Francisco Pando, Real Jardín Botánico, Madrid, Spain; Mervyn Mansell, USDA-Aphis, Pretoria, South Africa; 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 착수와 현재의 상태에 이르기까지 모든 과정에서 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

Agumya, A. and Hunter, G.J. 1996. Assessing Fitness for Use of Spatial Information: Information Utilisation and Decision Uncertainty. *Proceedings of the GIS/LIS '96 Conference*, Denver, Colorado, pp. 359-70

ANZLIC. 1996a. *ANZLIC Guidelines: Core Metadata Elements Version 1, Metadata for high level land and geographic data directories in Australia and New Zealand.* ANZLIC Working Group on Metadata, Australia and New Zealand Land Information Council. http://www.anzlic.org.au/metaelem.htm. [Accessed 14 Jul 2004]

ANZLIC 1996b *Spatial Data Infrastructure for Australia and New Zealand. Discussion Paper.* www.anzlic.org.au/get/2374268456. [Accessed 1 Jul 2004].

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

Bannerman, B.S., 1999. *Positional Accuracy, Error and Uncertainty in Spatial Information.* Australia: Geoinovations Pty Ltd. http://www.geoinnovations.com.au/posacc/patoc.htm [Accessed 14 Jul 2004].

Beer, T. & Ziolkowski, F. (1995). *Environmental risk assessment: an Australian perspective.* Supervising Scientist Report 102. Canberra: Commonwealth of Australia. http://www.deh.gov.au/ssd/publications/ssr/102.html [Accessed 14 Jul 2004]

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 16 Mar. 2005].

Birds Australia. 2001. *Atlas of Australian Birds. Search Methods*. Melbourne: Birds Australia. http://www.birdsaustralia.com.au/atlas/search.html [Accessed 30 Jun 2004].

Birds Australia. 2003. *Integrating Biodiversity into Regional Planning – The Wimmera Catchment Management Authority Pilot Project*. Canberra Environment Australia. http://www.deh.gov.au/biodiversity/publications/wimmera/methods.html. [Accessed 30 Jun 2004].

Brigham, A.R. 1998. Biodiversity Value of federal Collections **in** Opportunities for Federally Associated Collections. San Diego, CA, Nov 18-20, 1998.

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

Byers, F.R. 2003. *Care and Handling of CDs and DVDs. A Guide for Librarians and Archivists.* Washington, DC: National Institute of Standards and Technology and Council on Library and Information Resources. http://www.itl.nist.gov/div895/carefordisc/CDandDVDCareandHandlingGuide.pdf [Accessed 30 Jun 2004].

CBD. 2004. *Global Taxonomic Initiative Background.* Convention on Biological Diversity. http://www.biodiv.org/programmes/cross-cutting/taxonomy/default.asp [Accessed 13 Jul 2004].

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. 2002. Risk assessment and uncertainty in mapped and modelled distributions of threatened species in Australia pp 31-40 **in** Hunter, G. & Lowell, K. (eds) *Accuracy 2002 –*

Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Melbourne: Melbourne University.
Chapman, A.D. 2004. Environmental Data Quality – b. Data Cleaning Tools. Appendix I 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 57 pp. http://splink.cria.org.br/docs/appendix_i.pdf [Accessed 14 Jul. 2004]
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., Muñoz, M.E. de S. and Koch, I. 2005. Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context. *Biodiversity Informatics* **2:** 24-41.
Chrisman, N.R. 1983. The role of quality information in the long-term functioning of a GIS. *Proceedings of AUTOCART06*, 2: 303-321. Falls Church, VA: ASPRS.
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.
Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data.* Version 3. Sydney: Royal Botanic Gardens.
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 Jun. 2004].
Cullen, A.C. and Frey, H.C. 1999. *Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York: Plenum Press, 335 pages.
CRIA 2005. *species*Link. Dados e ferramentas – Data Cleaning. Campinas, Brazil: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/dc/ [Accessed 4 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. 2004].
Dallwitz, M.J. and Paine, T.A. 1986. *Users guide to the DELTA system*. CSIRO Division of Entomology Report No. 13, pp. 3-6. *TDWG Standard*. http://biodiversity.uno.edu/delta/ [Accessed 9 Jul 2004].
Davis R.E., Foote, F.S., Anderson, J.M., Mikhail, E.M. 1981. *Surveying: Theory and Practice*, Sixth Edition: McGraw-Hill.
DeMers M.N. 1997. *Fundamentals of Geographic Information Systems*. John Wiley and Sons Inc.
English, L.P. 1999. Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. New York: John Wiley & Sons, Inc. 518pp.
Environment Australia. 1998. *The Darwin Declaration*. Canberra: Australian Biological Resources Study. http://www.biodiv.org/programmes/cross-cutting/taxonomy/darwin-declaration.asp [Accessed 14 Jul 2004].
Epstein, E.F., Hunter, G.J. and Agumya, A.. 1998, Liability Insurance and the Use of Geographical Information: *International Journal of Geographical Information Science* 12(3): 203-214.
Federal Aviation Administration. 2004. Wide Area Augmentation System. http://gps.faa.gov/Programs/WAAS/waas.htm [Accessed 15 Sep. 2004].
FGDC. 1998. *Geospatial Positioning Accuracy Standards.* US Federal Geographic Data Committee. http://www.fgdc.gov/standards/status/sub1_3.html [Accessed 14 Jul. 2004].

Foote, K.E. and Huebner, D.J. 1995**.** *The Geographer's Craft Project*, Department of Geography, University of Texas. http://www.colorado.edu/geography/gcraft/contents.html [Accessed 14 Jul 2004].

Gad, S.C. and Taulbee, S.M. 1996. *Handbook of data recording, maintenance, and management for the biomedical sciences*. Boca Raton: CRC Press.

Goodchild, M.F., Rhind, D.W. and Maguire, D.J. 1991. *Introduction* pp. 3-7 In: Maguire D.J.,Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Heuvelink, G.B.M. 1998. *Error Propagation in Environmental Modeling with GIS*: Taylor and Francis.

Huang, K.-T., Yang, W.L. and Wang, R.Y. 1999. *Quality Information and Knowledge*. New Jersey: Prentice Hall.

Juran, J.M. 1964. *Managerial Breakthrough.* New York: McGraw-Hill.

Knapp, S., Lamas, G., Lughadha, E.N. and Novarino, G. 2004. Stability or stasis in the names of organisms: the evolving codes of nomenclature. Phil. Trans: Biol. Sci. 359(1444): 611-622.

Koch, I. (2003). *Coletores de plantas brasileiras.* Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 26 Jan. 2004].

Lance, K. 2001. Discussion of Pertinent Issues. pp. 5-14 in *Proceeedings USGS/EROS Data Center Kenya SCI Workshop, November 12 2001*. http://kism.iconnect.co.ke/NSDI/proceedings_kenya_NSDI.PDF [Accessed 1 Jul 2004].

Leick, A. 1995. *GPS Satellite Surveying*: John Wiley and Sons, Inc: New York.

Library of Congress. 2004. *Program for Cooperative Cataloging.* Washington, DC. US Library of Congress. http://www.loc.gov/catdir/pcc/ [Accessed 26 Jun 2004].

Lunetta, R.S. and Lyon, J.G. (eds). 2004. Remote Sensing and GIS Accuracy. Boca Raton, FL, USA:CRC Press.

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 21 November 2003].

Mayr, E. and Ashlock, P.D. 1991. *Principles of systematic zoology*. New York: McGraw-Hill.

McElroy, S., Robins, I., Jones, G. and Kinlyside, D. 1998. *Exploring GPS, A GPS Users Guide*: The Global Positioning System Consortium.

Minnesota Planning. 1999. *Positional Accuracy Handbook. Using the National Standard for Spatial data Accuracy to measure and report geographic data quality*. Minnesota Planning: Land Management Information Center. http://www.mnplan.state.mn.us/pdf/1999/lmic/nssda_o.pdf [Accessed 14 Jul. 2004]

Morse, L.E. 1974. Computer programs for specimen identification, key construction and description printing using taxonomic data matrices. *Publs. Mich. St. Univ. Mus., biol.* ser. 5, 1–128.

Motro, A. and Rakov, I. 1998. Estimating the Quality of Databases. *FQAS 1998*: 298-307

Naumann, F. 2001. *From Database to Information Systems – Information Quality Makes the Difference*. IBM Almaden Research Center. 17 pp.

Nebert, D. and Lance, K. 2001. Spatial Data Infrastructure – Concepts and Components. *Proceeedings JICA Workshop on Application of Geospatial Information and GIS. 19 March 2001, Kenya.* http://kism.iconnect.co.ke/JICAWorkshop/pdf/Ottichilo.pdf [Accessed 1 Jul 2004].

Nebert, D. 1999. *NSDI and Gazetteer Data*. Presented at the Digital Gazetteer Information Exchange Workshop, Oct 13-14, 1999. Transcribed and edited from audiotape. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/session3/nebert.htm [Accessed 1 Jul 2004].

NLWRA. 2003. *Natural Resources Information Management Toolkit*. Canberra: National Land and Water Resources Audit. http://www.nlwra.gov.au/toolkit/contents.html [Accessed 7 Jul 2004].

NOAA. 2002. Removal of GPS Selective Availability (SA). http://www.ngs.noaa.gov/FGCS/info/sans_SA/ [Accessed 15 Sep 2004].
Olivieri, S., Harrison, J. and Busby, J.R. 1995. Data and Information Management and Communication. pp. 607–670 **in** Heywood, V.H. (ed.) *Global Biodiversity Assessment*. London: Cambridge University Press. 1140pp.
Pipino, L.L., Lee, Y.W. and Wang, R.Y. 2002. Data Quality Assessment. *Communications of ACM* 45(4): 211-218.
Pullan, M.R., Watson, M.F., Kennedy, J.B., Raguenaud, C., Hyam, R. 2000. The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon* **49**: 55-75.
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.
SA Dept Env. & Planning. 2002. *Opportunistic Biological Records (OPPORTUNE).* South Australian Department of Environment and Heritage. http://www.asdd.sa.gov.au/asdd/ANZSA1022000008.html [Accessed 14 Jul. 2004].
SEC 2002. *Final Data Quality Assurance Guidelines*. United States Securities and Exchange Commission. http://www.sec.gov/about/dataqualityguide.htm [Accessed 26 Jun 2004].
Shepherd, I.D.H. 1991.Information Integration and GIS. pp. 337-360 **in:** Maguire D.J.,Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
Spear, M., J.Hall and R.Wadsworth. 1996. *Communication of Uncertainty in Spatial Data to Policy Makers* **in** Mowrer, H.T., Czaplewski, R.L. and Hamre, R.H. (eds) Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium, May 21-23, 1996. Fort Collins, Colorado. USDA Forest Service Technical Report RM-GTR-277.
Stribling, J.B., Moulton, S.R. II and Lester, G.T. 2003. Determining the quality of taxonomic data. *J. N. Amer. Benthol. Soc*. **22(4):** 621-631.
Strong, D.M., Lee, Y.W.and Wang, R.W. 1997. Data quality in context. *Communications of ACM* 40(5): 103-110.
Taulbee, S.M. 1996. *Implementing data quality systems in biomedical records* pp. 47-75 in Gad, S.C. and Taulbee, S.M. Handbook of data recording, maintenance, and management for the biomedical sciences. Boca Raton: CRC Press.
TDWG. 2005. TDWG Working Group: Structure of Descriptive Data (SDD). Taxonomic Databases Working Group (TDWG). http://160.45.63.11/Projects/TDWG-SDD/ [Accessed 4 Apr. 2005].
University of Colorado. 2003. MaPSTeDI. *Georeferencing in MaPSTeDI*. Denver, CO: University of Colorado. http://mapstedi.colorado.edu/georeferencing.html [Accessed 30 Jun. 2004].
USGS. 2004. *What is SDTS?* Washington: USGS. http://mcmcweb.er.usgs.gov/sdts/whatsdts.html [Accessed 30 Jun. 2004].
Van Sickle, J. 1996. *GPS for Land Surveyors*: Ann Arbor Press, Inc: New York.
Wang, R.Y. 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM* 41(2): 58-65.
Wang, R.Y., Storey, V.C., Firth, C.P., 1995. A frame-work for analysis of data quality research, *IEEE Transactions on Knowledge and Data Engineering* 7: 4, 623-640.
Wieczorek, J. 2001. *MaNIS: GeoreferencingGeo-referencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/GeorefGuide.html [Accessed 26 Jan. 2004].
Wieczorek, J. 2002. *Summary of the MaNIS Meeting. American Society of Mammalogists, McNeese State University, Lake Charels, LA, June 16, 2002*. Berkeley: University of California, Berkeley - MaNIS. http://manisnet.org/manis/ASM2002.html [Accessed 30 Jun. 2004].

Wieczorek, J., Guo, Q. and Hijmans, R.J. (2004). *The point-radius method for georeferencing locality descriptions and calculating associated uncertainty*. International Journal for GIS 18(8): 754-767.
Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.
Zhang, J. and Goodchild, M.F. 2002. Uncertainty in Geographic Information. London: Taylor and Francis.

색인