생물종 데이터 품질의 원칙

옮긴이: 박형선, 안성수, 박재홍

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

생물종 데이터 품질의 원칙

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

생물종 데이터 품질의 원칙

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

옮긴이 | 박형선, 안성수, 박재홍
펴낸이 | 조영화

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

이 책은 Arthur D. Chapman 이 GBIF DIGIT 연구 프로그램의 산출물로 작성한 생물종 데이터 품질의 원칙(PRINCIPLES OF DATA QUALITY) 자료를 원저자의 허락을 받고 번역한 것입니다. 이 번역물이 국내의 생물다양성데이터를 인터넷상에서 공유하고 활용하려고 할 때 참고자료로 사용되고 도움이 될 수 있기를 바랍니다. 단, 이 책을 참조할 경우 참조한 사실을 반드시 인용해야 합니다.

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

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

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

ISBN 89-5884-641-0 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

서론

데이터 품질 관리는 기업(SEC 2002), 의료 업계(Gad and Taulbee 1996), GIS(Zhang AND Goodchild 2002), 원격 탐지(Lunetta and Lyon 2004) 그리고 여러 분야에서 일상적인 것이 되었지만 박물관과 분류학에서는 이제서야 도입되고 있다. 분류와 생물 종 발생 데이터가 교환되고 이용이 많아지면서 이를 이용하는 사용자들이 데이터의 품질에 대한 정보를 요구하게 되어 데이터 품질 관리가 중요하게 되었다. 박물관에서 일하지 않는 몇몇 사람들은 박물관이 가지고 있는 데이터가 환경 정책 수립 과정에서 쓰지 못할 정도로 품질이 나쁘다고 하는데 이는 데이터의 품질 때문인가 아니면 문서화의 품질 때문인가? 하지만 이 데이터는 매우 중요하다. 대규모로 멸종이 일어나기 전부터 오랜 시간에 걸쳐서 모은 데이터이기 때문에 생물 종 다양성에 대한 중요한 정보를 갖고 있다(Chapman and Busby 1994). 인간에 의해서 생태계의 변화가 일어난 지역에 대한 유일한 기록이기 때문에 이들 정보는 환경 보존에 매우 중요하다(Chapman 1999).

이 논문에서는 데이터를 공개하는 박물관과 식물원들이 중요하게 여겨야 할 데이터 품질 관리의 원리를 설명하면서 이에 대한 저자의 생각을 추가 할 것이다.

환경 정보 데이터베이스, 모델링 시스템, GIS, 정책 보조 시스템 등등에서 데이터 품질과 데이터 오류 문제는 중요하지 않은 문제로 치부 되는 경우가 많다. 충분한 검증 없이 데이터가 쓰이는 바람에 결과가 다르게 나타나거나 비용 증가 또는 잘못된 환경 보존 정책이 나오게 된다.

> *박물관과 식물원에 있는 표본 데이터는 현재에 대한 정보뿐만 아니라 수 백 년의 역사에 걸친 정보도 제공한다*
>
> (Chapman and Busby 1994)

생물 종 데이터를 다룰 때, 특히 이들 데이터에서 지리에 관련 데이터를 다룰 때 적용이 가능한 데이터 품질 관리의 원리들이 많다. 이 원리들은 데이터 관리의 전 과정에 쓰인다.

관리 과정 중에 품질 저하는 데이터의 유용성을 떨어뜨린다. 여기에 포함 되는 것에:

- 수집 당시에 데이터의 파악과 기록,
- 디지털화 이전에 데이터 처리(표지, 데이터 복사, 등),
- 수집물 파악(표본, 관찰) 그리고 기록,
- 데이터의 디지털화,
- 데이터의 문서화(데이터에 대한 정보를 파악하고 기록),
- 데이터 저장과 보관,
- 데이터 프레젠테이션과 보급(출판물, 온라인 데이터베이스),
- 데이터의 이용(분석과 처리).

이 중에서 모든 항목에 데이터의 최종 품질에 영향을 미치며 데이터의 모든 면 –분류와 명명 부분, “무엇을”, “어디에”, “누가” 그리고 “언제”-에 대해서 적용됩니다.

데이터 품질과 생물 종 발생 데이터에 대한 논의를 하기 전에 정의해야 할 개념들이 몇 개 있다. 데이터 품질, 정확성, 정밀도는 잘못 쓰이는 경우가 많고 생물 종 데이터와 종 발생 데이터는 정의하고 넘어가야 한다.

품질 개선의 간단함을 무시해서는 안된다. 팀웍크, 훈련, 그리고 기강만 필요하고 특별한 기술이 핑요없다. 원한다면 모두가 효율적인 기여를 할 수 있다.

(Redman 2001).

정의

종 발생 데이터

이 논문에서 종 발생 데이터는 박물관과 식물원의 표본 데이터, 관찰 데이터 그리고 환경 조사 데이터를 포함한다. 이 데이터는 소위 말하는 "지점 중심"의 데이터인데 여기에 줄(환경 조사의 횡단 조사, 강을 따른 조사), 폴리곤(일정한 구역 내의 조사), 그리고 격자(일정한 지역을 격자로 나누어서 조사) 데이터도 포함된다. 흔히 다루게 되는 데이터는 지리적인 정보가 포함되어 있는 정보인데 여기에 지리 좌표, 소재지 서술 그리고 시간 데이터가 포함 되어 있다. 보통은 분류학에서 쓰는 이름과 연결이 되어 있는 경우가 많지만 미감정 표본들도 포함 될 수 있다. 종 발생 데이터는 일차적 종 데이터와 혼용해서 쓰기도 한다.

일차적 종 데이터

"일차적 종 데이터"는 가공 되지 않은 수집 데이터를 가리킬 때 쓰는 말이며 지리적인 정보가 전혀 업다. 지리적인 정보가 없는 분류와 명명 데이터가 포함되어 있다.

Accuracy and Precision

정확도와 정밀도는 혼동되는 경우가 많으며 많은 사람들이 그 차이점을 이해하지 못한다. 그림 1에서 이 두 개념의 차이점이 설명되어 있다.

정확도는 측정 된 값, 관찰, 그리고 예측한 결과와 실제 결과와의 근접성을 일컫는다(또는 참이라고 여겨지는 값).

정밀도(또는 해상도)는 두 가지 개념으로 나눌 수 있다. 통계적인 정밀도는 반복되는 관찰 결과가 일정한 값을 나타내는 정도를 말한다. 그림 1a 에 나오는 것처럼 실제 값과의 관계성은 상관이 없으며 정밀도는 높지만 정확도가 낮을 수 있다. 수치적인 정밀도는 기록의 유효 숫자의 개수를 가리키며 컴퓨터가 보급이 되면서 이의 중요성이 높아졌다. 예를 들어서 데이터베이스는 위도/경도를 소수점 아래 10 자리까지 계산하는데 원본 기록은 소수점 아래 4 자리 밖에 없어서 해상도와 정확성에 대한 잘못된 인상을 심을 수 있다.

정확도와 정밀도는 공간적인 정보가 없는 데이터에도 적용이 가능하다. 예를 들어서 어떤 수집 기록에서 아종 단계까지 감정이 제대로 되었지만(높은 정밀도) 과가 틀리거나(낮은 정확도) 목 단계까지만 감정(높은 정확도, 낮은 정밀도)이 되었을 수 있다.

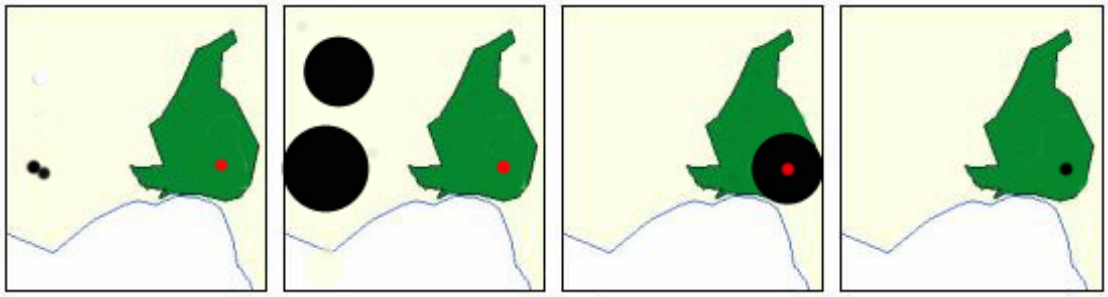

그림. 1. *공간 정보에서 정밀도와 정확도의 차이. 빨간 점은 실제 지점, 검은 점은 보고된 지점이다.*

a. 높은 정밀도, 낮은 정확성.
b. 낮은 정밀도, 높은 정확성. 무작위 오류들이 많음.

c. 낮은 정밀도, 높은 정확성.
d. 높은 정밀도와 정확성.

품질

품질이라는 개념이 데이터에 적용 될 때 정의가 여러 가지 있지만 지리 학계에서는 "사용 적합성"(Chrisman 1983) 또는 "잠재적인 사용 적합성"이라는 의미로 쓴다. 현대적인 공간 데이터 전송 표준의 대부분은 이 정의를 따르고 있다(ANZLIC 1996a, USGS 2004). 지리적인 정보를 다루지 않는 경제와 기업에서도 이 정의를 쓰기 시작했다. 몇몇 사람들은(English 1999, 등)은 "사용 적합성" 정의가 제한적인 면이 있어서 미래나 잠재적인 사용 적합성을 포함하는 새로운 정의를 만들어야 한다고 주장하고 있다.

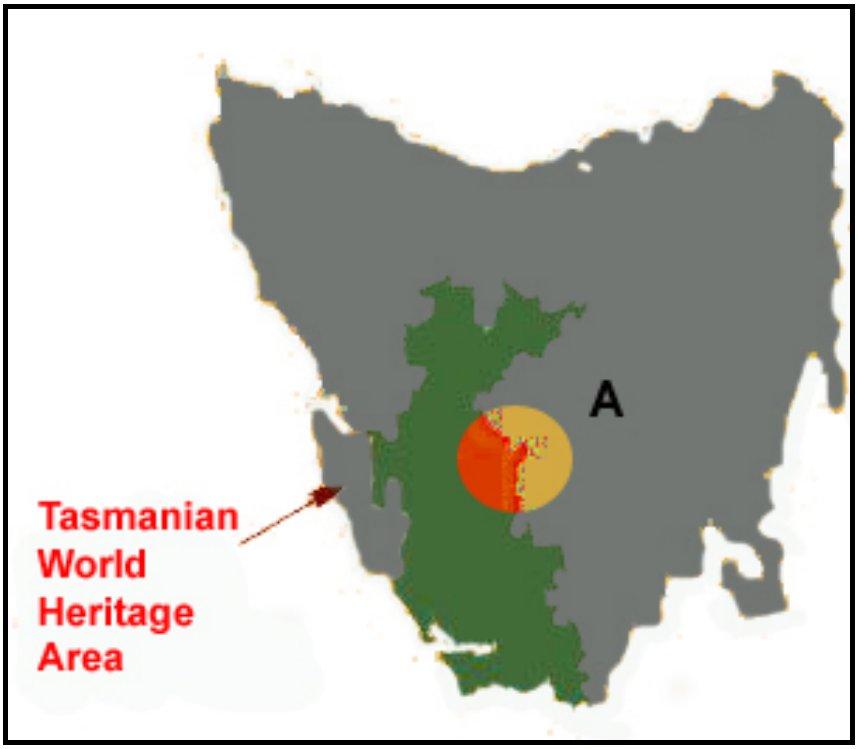

Fig. 2. *오스트레일리아의 Tasmania 지도에서 50km의 정확도로 A 지역에서 이루어진 기록이 표시 되어 있다. A 지역의 일부분은 Tasmania의 세계 유산 지역과 겹친다.*

그림 2에서 "사용 적합성" 개념의 예를 볼 수 있다. A 지역에서 수집된 기록은 약 50km의 정확도를 갖고 있다. Tasmania에 사는 종의 목록을 작성하고 있고 해당 종이 Tasmania에 살고 있는가를 알고 싶으면 해당 기록은 그 질문에 답할 수 있고 이에 따라서 적합성이 높다고 할 수 있다. 반면에 해당 종이 Tasmania 세계 유산 지역에 분포하는지 알고 싶을 때 해당 기록에서 답을 얻을 수 없다. 따라서 데이터는 그 질문에 답할 수 없고 낮은 품질의 데이터가 된다. 데이터베이스의 위도/경도 수치는 정밀도가 높아서 정확도까지 높다는 잘못된 인상을 주어서 사용자에게 잘못된 인상을 줄 수 있다.

지리적인 정보가 아니더라 하더라도 감정을 잘못하면 데이터의 가치를 떨어뜨리고 용도에 적합하지 않은 것이 될 수 있다. 종의 분포를 연구하고 있다면 표본이나 관찰에 잘못된 이름을 기재하는 것은 예상과 다른 결과를 낼 수 있다.

데이터 품질에는 여러 면이 있으며 데이터 관리, 모델링, 분석, 품질 관리와 보장, 그리고 보관 및 표시가 있다. Chrisman(1991)와 Strong(1997)이 말한 것처럼 데이터의 품질은 용도에

의해서 결정되며 사용자와 따로 떼어놓아서 데이터를 평가하는 것이 불가능하다. 데이터베이스 내에서는 데이터는 품질과 가치는 없고(Dalcin 2004) 실제로 사용 될 때 가치가 생기는 잠재적인 가치만이 있을 뿐이다. 정보의 품질은 사용자의 요구 사항을 맞춰줄 능력이 직접적인 관계가 있다.

Redman(2001)에 의하면 데이터가 용도에 적합하게 되는 조건은 다음과 같다: 정확하다, 접근성이 좋다, 제 시간에 전달이 된다, 완성도가 높다, 다른 기록과 일관성이 있다, 용도에 들어 맞는다, 적절한 세부 내용, 그리고 읽고 해석하는 것이 쉬워야 한다.

데이터 관리자 고려해야 하는 문제 중에 데이터베이스가 폭 넓은 층의 사람들에게 쓸모가 있으려면 무엇을 해야 하는가에 대한 문제가 있다. 이렇게 하려면 다양한 목적에 맞는 데이터가 들어 있어야 하는데 커진 접근성과 추가 기능과 사용 적합성을 늘리려는 노력 사이에 반비례 관계가 형성된다. 사용 적합성을 늘리려는 노력 중에 데이터 문항을 쪼개거나 지리 좌표 정보 추가 등이 있다.

> ***데이터는 정책 결정, 계획 수립 및 조사에서 이용되면 고품질의 데이터이다.*** (Juran 1964).

품질 보증/관리

품질 관리와 보증 사이의 차이점은 분명하지 않은 경우가 많다. Taulbee(1996)에서 품질 보증과 품질 관리의 차이점에 대해서 설명하고 있으며 한 쪽 없이는 다른 쪽도 존재 할 수 없다는 사실을 강조하고 있다. 내려진 정의는 다음과 같다:

- 품질 관리는 내부 기준, 과정, 그리고 절차에 따라서 품질을 판단해서 이를 조절하고 관리하는 것이고
- 품질 보증은 외부의 기준에 의해서 품질을 평가해서 품질 관리 절차들을 재평가 하여 이들 기준에 결과물이 외부의 품질 기준을 맞추도록 하는 것이다.

Redman(2001)에서는 품질 보증을 다음과 같이 기업의 시선에서 다음과 같이 정의 한다:

> ”가장 중요한 고객들에게 결함이 없는 정보를 최소의 비용에 제공하기 위한 활동들”.

실제 상황에서 이들 개념이 적용되는 것은 분명하지 않으며 품질 관리 조절을 가리키는 동의어로 쓰이는 경우가 많다.

불확정성

불확정성은 “측정이 가능하다면 모든 성질을 파악할 수 있는 알지 못하는 대상에 대해서 지식이나 정보의 불완전성의 정도”로 정의하면 된다(Cullen and Frey 1999). 불확정성은 관찰자의 데이터에 대한 이해의 일부분이며 데이터의 성질이라기 보다는 관찰자의 성질에 가깝다. 데이터에는 항상 불확정성이 존재하고 남이 이해 할 수 있도록 이 불확정성을 기록, 이해, 그리고 시각화 하는 과정이 어려운 것이다. 위험성을 이해하고 평가하는데 있어서 불확정성은 중요한 개념이다.

오류

오류는 정확성과 정밀도를 포함하는 개념이다. 많은 요인들이 오류에 기여한다.

"보통은 오류와 정확하지 않은 사항에 대해서는 나쁜 것으로 간주한다. 하지만 꼭 이럴 필요성은 없고 이들의 원인을 파악하면 이들을 조절하고 줄일 수 있다. 오류와 오류의 발생에 대해서 제대로 이해하게 되면 품질 관리가 활성화 된다" (Burrough and McDonnell 1998).

오류에는 무작위 오류와 체계적인 오류가 있다. 무작위 오류는 무작위 한 방식으로 참 값에서 벗어나는 것을 일컫는다. 체계적인 오류는 측정 값이 일정하게 변하는 것을 일컫는 개념이며 지리 학계에서 '상대적인 정확성'을 갖는 것으로 여겨지는 경우가 있다(Chrisman 1991). 사용 적합성을 판단하는데 있어서 체계적인 오류는 몇몇 용도에 대해서 적합 할 수 있다. 예를 들어서 측지선 자료는 계속 쓰인다면 문제를 일으키지 않는다. 하지만 출처가 다른 데이터를 분석에 쓴다면 문제는 일어날 것이다.

"오류는 필연적으로 일어나므로 데이터의 근본적인 특성으로 여겨야 한다"(Chrisman 1991). 오류가 데이터의 일부로 인정이 되어야 데이터와 현재 학계의 지식에서의 제한에 대한 질문에 대답 할 수 있을 것이다. 시공간 오류는 측정, 계산, 기록 그리고 문서화 되어야 한다.

검증과 청소

검증은 데이터의 정확성, 완성도, 그리고 논리성을 따지는 과정이다. 여기에 형식 검사, 완성도 검사, 논리성 검사, 제한 검사, outlier 검사 그리고 전문가에 의한 평가가 포함 되어 있다. 검사를 하면 의심이 가는 기록에 대해서 표시, 문서화 그리고 추가 검사를 하는 것이 보통이다. 검증에는 표준, 지침, 그리고 관행 준수 여부 검사도 포함된다. 발견한 오류들의 원인을 찾고 이 오류들이 재발을 막는 것은 데이터 검증과 청소의 중요한 단계 중 하나이다.

데이터 청소는 검증 과정에서 발견된 오류를 고치는 과정을 일컫는다. "데이터 세척"과 동의어이지만 일부에서는 데이터 세척에 데이터 검증과 청소를 포함 시키는 개념으로 정의한다. 데이터 청소 과정에서 데이터가 유실 되지 않도록 유의해야 하며 기존의 정보를 수정 할 때 조심을 기울여야 한다. 이런 경우에는 기존의 데이터와 새롭게 수정 된 데이터 양쪽을 보관해서 청소 과정에 잘못이 생기면 원래 정보가 복구 되도록 해야 한다.

최근에 데이터 검증과 청소를 도와 주는 도구와 지침들이 개발되었다. 이에 대한 논의는 Principles and Methods of Data Cleaning 에서 다루고 있다. 데이터 청소를 수동으로 하는 것은 노동력과 시간이 많이 들며 청소 자체가 오류에 노출되어 있다(Maletic and Marcus 2000).

데이터 청소를 하는데 있어서 기본적인 틀은 다음과 같다(Maletic and Marcus 2000):

- 오류 종류를 정의하고 파악
- 오류를 찾고 명시
- 오류 수정
- 오류 발생 지점과 종류를 기록
- 추후에 발생 할 수 있는 오류를 줄이기 위해서 데이터 입력 과정 수정

레이블의 진실성

Truth in Labelling 은 제품과 생산품의 품질에 대한 문서화로 간주 되는 경우가 많다. 생물 종 발생 데이터의 경우에 이는 보통 품질, 품질 관리 절차 아니면 품질을 수치적으로 표현한 metadata 으로 이루어져 있는 경우가 많다. Truth in labelling 은 필요한 경우에 보증으로 이어진다. 대부분의 박물관과 식물원들은 검증 전문가와 표본 감정 날짜에 대해서 이런 절차를 행하고 있지만 다른 정보에 대해서 이러한 절차가 행해지는 경우는 그리 많지 않다.

사용자

사용자들은 누구인가? 데이터의 사용자는 정보 사슬의 모든 단계에 있는 모든 사람을 포함한다(그림 3). 생물 종 데이터의 경우에 기관의 표본 학자, 관리자, 연구원, 기술자, 수집가 아니면 외부의 정책 결정자, 과학자, 농업 학자, 산림 학자, 환경 관리자, NGO, 의료 학자, 약학자, 산업계 인사, 식물원 및 동물원 관리자, 그리고 일반인이 포함 되어 있다. 생물 종 발생 데이터는 수많은 사용자들이 있으며 사회의 거의 전 구성원이 포함된다.

생물 종 데이터는 비전문가 사용자 집단의 수요에 대한 고려 없이 수집된 경우가 많다. 전통적으로 데이터는 분류학 또는 생물 지리 연구를 위해서 수집이 되었다. 이 것은 중요한 작업이지만 오늘날에 이들 기관에 재정적인 지원을 해주는 기관들은 데이터가 다른 용도로도 쓰일 수 있도록 요구를 하고 있다. 또한 정부들은 이 데이터를 환경 정책 결정 과정에서 이 데이터를 쓰려고 하고 있으며 데이터 관리자들은 이러한 요구를 무시하지 못한다.(Chapman and Busby 1994). 우수한 피드백 절차가 있다면 사용자들은 데이터 품질에 대한 피드백을 제공 할 수 있으며 데이터 품질 사슬에서 중요한 역할을 차지 할 수 있다.

사용자의 요구를 파악하는 것은 어려운 일이다. 하지만 이를 대신 할 수 있는 방법은 없으며 요구를 파악하는 것의 대가는 가치가 있다.

데이터 품질의 원리

> *경험에 의하면 데이터를 장기적인 자산으로 고려해서 이를 체계적으로 관리하는 것은 많은 비용을 아낄 수 있다*(NLWRA 2003).

데이터 품질의 원리는 데이터 관리 절차의 모든 과정에 적용이 되어야 한다(기록, 디지털화, 보관, 분석, 제시, 및 사용). 데이터 품질을 개선 하는데 있어서 방지와 수정이 중요하다. 오류 방지는 데이터 수집과 입력에 밀접한 연관을 갖고 있다. 오류를 수정하는데 많은 노력을 기울여야 하지만 데이터가 많은 경우에 오류는 여전히 존재 할 것이며(Maletic and Marcus 2000) 오류 검증과 수정은 무시하지 못하는 부분이다.

오류 방지는 오류 검사가 100% 정확하지 못하기 때문에 이보다 효율적이다(Dalcin 2004). 하지만 오래된 데이터를 다룰 때, 특히 생물 종 데이터와 종 분포 데이터의 경우에 오류 검사를 할 때 중요한 역할을 차지한다(Chapman and Busby 1994, English 1999, Dalcin 2004).

데이터 청소를 무계획적으로 하는 것보다 데이터에 대한 비전을 설정한 다음에 데이터에 대한 정책을 만들어서 청소를 체계적으로 하는 것부터 시작해야 한다.

비전

기관들이 고품질 데이터에 관한 비전을 갖는 것이 중요하다. 특히 데이터를 외부에 공개할 경우에 이 사실이 중요하다. 제대로 세워진 데이터 품질 비전은 기관의 전체적인 비전을 향상 시키고(Redman 2001) 전반적인 절차를 개선 할 수 있을 것이다. 비전을 설정하는데 있어서 관리자들은 리더쉽, 구성원, 하드웨어, 프로그램, 품질 관리 및 데이터가 적절한 도구, 지침, 그리고 표준으로 묶여져서 데이터의 품질을 관리 하는 것에 초점을 맞추어야 한다(NLWRA 2003).

데이터 품질 비전:

- 기관으로 하여금 보유하고 있는 장기 보관 데이터, 정보 수요 그리고 장기적인 목적에 대해서 생각하도록 유도한다,
- 올바른 방향으로 행동을 촉진한다,
- 기관 내부와 밖에서 각종 결정에 사용될 수 있도록 하는 견고한 기반을 제공한다,
- 데이터와 정보가 기관의 핵심 자산이라는 사실을 공고히 한다,
- 기관이 가지고 있는 정보와 데이터의 사용을 최대화 하고, 중복을 방지하며, 협력 관계 구축에 도움을 주고 정보와 데이터에 대한 접근을 용이하게 한다,
- 통합과 협력을 최대화 할 수 있다.

정책

비전뿐만 아니라 기관은 비전을 실현 시키기 위한 정책이 필요하다. 견고한 데이터 품질 비전을 개발 함으로써 얻을 수 있는 효과는 다음과 같다:

- 기관으로 하여금 품질에 대한 고려를 하고 일상적인 관행을 점검 하도록 유도한다,
- 데이터 관리 절차를 분명히 하게 된다,
- 다음 사항에 관련된 목적을 분명히 하는데 도움을 준다,
    - 비용 절감,

    - 데이터 품질 개선,
    - 고객 서비스와 관계를 개선,
    - 그리고 결정 과정을 개선 할 수 있다.
- 기관이 제공하는 데이터를 이용하는 사용자들에게 안정과 자신감 제공한다,
- 기관의 고객들과의 관계와 의사 소통 개선(데이터 제공자와 사용자),
- 기관의 이미지 개선, 그리고
- 재정 지원을 얻을 확률을 증가 시킬 수 있다.

전략

규모가 큰 기관들이 소유 하고 있는 데이터의 양을 고려하면 데이터를 기록하고 검사하는데 전략을 개발 할 필요가 있다(우선 순위 항목도 참고). 데이터 입력과 품질 관리에서 따를 수 있는 기간에 따른 전략을 세우는 것이 좋은 전략이다. 예를 들어서(Chapman and Busby 1994):

- **초기**. 6-12 개월 동안 정리되고 검사가 가능한 데이터는 이미 데이터베이스에 있는 데이터와 품질 관리의 필요성이 떨어지는 새로운 데이터가 있다.
- **중반**. 적은 자원으로 18 개월 정도의 기간에 데이터베이스에 입력 될 수 있는 데이터와 간단한 방법으로 검사 할 수 있는 데이터.
- **장기**. 협력 관계를 이용하여 입력 될 수 있는 데이터와 복잡한 검사를 필요로 하는 데이터가 이에 속한다. 전체 데이터를 다음 그룹으로 분류해서 우선적으로 처리 할 수 있다:
    - 최근에 개정된 분류군이나 기관에서 연구 중인 분류.
    - 중요한 표본 모임.
    - 중요한 분류군(국가적으로 상징적인 분류군, 위협 받고 있는 분류군, 생태/환경적으로 중요한 분류군).
    - 중요한 지역의 분류군(기관에 중요한 지역, 자료를 공유하는 개발 도상국).
    - 다른 기관과 협력 하에 데이터를 같이 데이터베이스화 하는 경우.
    - 처음부터 끝까지 데이터를 체계적으로 처리하는 경우.
    - 최근에 얻은 데이터를 밀려 있는 데이터 보다 우선적으로 처리하는 경우.

전략에 포함 되어야 할 데이터 관리 원리에는 다음 사항들이 포함 되어 있다 (NLWRA 2003):

- 정보 관리 체계를 재 발명 하지 말 것
- 데이터 수집과 품질 관리 절차에서 효율적인 부분 찾기
- 가능하다면 데이터, 정보, 그리고 도구를 공유
- 기존의 표준을 이용하거나 다른 기관과 같이 새로운 표준 이용
- 네트워크와 협력 관계의 구축 장려
- 데이터 수집과 관리의 사업성 제시
- 데이터 수집과 품질 관리에서 중복 방지
- 눈 앞의 사용만 따지는 것이 아니고 사용자들의 필요성을 점검
- 문서화를 철저히 되고 metadata 에 대한 절차를 준수.

치료 보다는 방지가 효율적이다

표본 수집을 데이터베이스에 입력 하는 비용은 클 수 있지만(Armstrong 1992) 추후에 데이터를 점검하고 수정 하는 비용의 몇 분의 일 밖에 되지 않는다. 나중에 수정 하는 것보다 오류를 방지하는 것이 효율적이며(Redman 2001) 이 방법이 비용이 적게 든다. 나중에 수정을

하게 되면 수정되는 데이터가 이미 분석에 쓰였을 수 있기 때문에 이에 부적절한 데이터로 이루어진 결정으로 인한 비용이나 다시 분석함으로써 생기는 비용이 추가된다.

하지만 데이터베이스 내에 이미 존재하는 오류에 대해서는 오류 방지는 효과를 갖지 못하기 때문에 데이터 검증과 청소는 데이터 품질 관리 체계의 중요한 부분을 차지한다. 청소 과정에서 오류의 원인을 발견하고 이에 대한 대비 책을 세워야 한다. 하지만 이 과정은 공개되어야 오류들이 다시 발생하지 않는다. 데이터 청소와 오류 방지는 항상 같이 시행되어야 한다. 데이터를 먼저 청소하고 방지를 나중에 하는 것은 오류 방지의 효율적인 시행을 막고 데이터베이스에 오류들이 더해지게 된다.

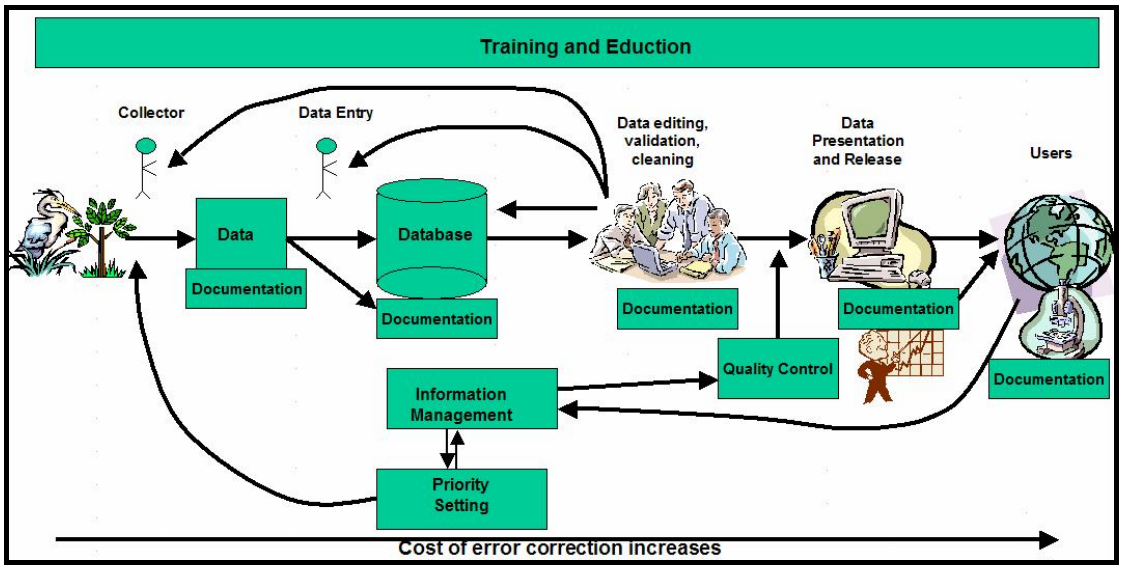

그림. 3. 사슬이 진행될수록 오류 수정의 비용이 증가한다. 문서화를 철저히 하는 것과, 교육, 그리고 훈련이 모든 단계에서 중요한 역할을 한다.

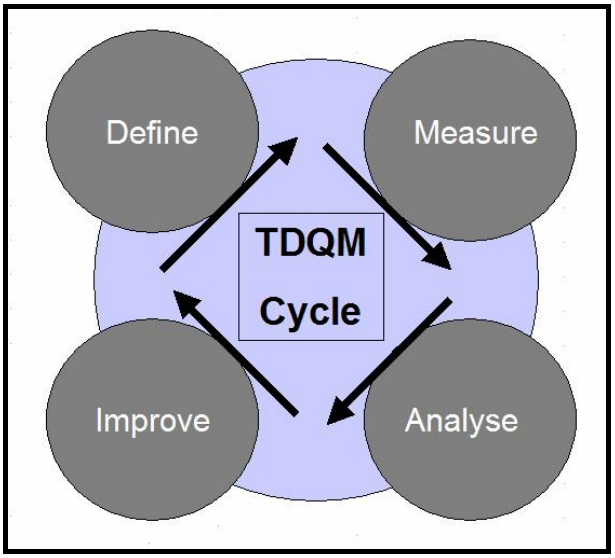

그림 4. 데이터 관리 과정의 순환을 보여주는 데이터 품질 관리 싸이클(Wang 1998).

데이터의 관리자와 소유자들은 소유하고 있는 데이터의 품질에 대한 책임이 있다. 하지만 책임은 데이터 제공자와 사용자에게도 있다.

> 데이터를 만든 사람에게 품질에 대한 책임을 지워야 한다. 이 것이 가능하지 않다면 가장 근접한 사람에게 책임을 지워야 한다(Redman 2001).

일차적인 책임은 수집가에게 있다

데이터 품질 관리의 일차적인 책임은 데이터 수집가에게 있다. 다음 사항들은 수집가들의 책임이다:

- 레이블 정보가 올바르다,
- 레이블 정보가 정확하게 기록되고 문서화 되어야 한다,
- 소재지 정보는 최대로 정확하고, 정확도와 정밀도가 기록되어야 한다,
- 수집 방식이 자세히 기록되어야 한다,
- 레이블과 수집 일지가 명확하고 불분명하게 작성 되어있지 않다, 그리고
- 데이터를 입력하는 사람들이 쉽게 읽을 수 있도록 레이블이 명확하고 분명해야 한다.

레이블이나 수집 일지의 정보가 명확하고 정확하지 않으면 나중에 이를 수정하는 것이 상당히 어려워진다. 데이터 중에서 분류군에 관한 정보는 나중에 전문가들이 점검하는 것이 가능하므로 상대적으로 명확성이나 정확성의 중요성이 떨어진다.

또한 소재지나 장소에 연관된 정보는 나중에 기록 하는 것이 아니고 수집 당시에 기록하는 것이 중요하다.

> *대부분의 데이터는 기관에 공급되며 데이터 수집 관행을 개선하는 것이 나중에 오류를 수정하는 것 보다 쉽다.*

관리자는 핵심, 또는 장기적인 책임을 가지고 있다.

데이터의 관리자는 그 책임을 맡고 있는 이상 데이터의 품질을 관리하고 개선하는 장기적인 책임을 가지고 있다(Olivieri et al 1995 p 623 참고). 관리 기관이 기관 내부에서 데이터 품질을 책임지는 사람을 임명하는 것도 중요하지만 기관의 모든 구성원이 데이터의 품질에 대한 책임의 일부를 가지고 있다는 사실을 인식 시키는 것도 중요하다. 관리자의 책임은 다음과 같다:

- 수집가의 일지에서 데이터베이스에 데이터가 정확하게 입력되고 있다,
- 데이터 정리 중에 품질 관리 절차가 적용되고 있다,
- 데이터와 데이터 품질이 정확하게 문서화 되고 있다,
- 데이터에 대해서 수시로 검증을 하고 있다,
- 검증 검사에 대한 문서화를 철저히 하고 있다,
- 적절한 방법으로 데이터가 저장되고 보관되고 있다,
- 이전 버전도 저장해서 청소가 안된 데이터로 회귀 가능성을 열어 둔다,
- 데이터의 일관성이 유지 되고 있다,
- 사용자가 용도 적합성을 판단 할 수 있도록 문서화와 함께 데이터를 제공한다,
- 사생활, 지적재산권, 저작권, 그리고 전통을 보호한다,
- 데이터 사용에 관한 조건이 지켜지고 제한 조건이나 기타 사항을 알려준다,
- 데이터에 관련된 모든 법적 사항을 지킨다,
- 데이터 품질에 관한 사용자의 피드백은 제 때 처리한다,
- 언제나 데이터의 품질을 최고로 유지한다,
- 현재까지 알려진 모든 오류에 대해서 문서화를 하고 이를 사용자들에게 알린다.

데이터에 대한 소유권과 관리는 데이터에 대한 관리와 조정 권한을 부여해주는 동시에 이에 대한 관리, 품질에 대한 책임을 지운다. 관리자들은 후손들이 쓸 수 있도록 데이터를 관리할 윤리적인 책임이 있다.

사용자의 책임

데이터의 사용자들에게도 데이터의 품질에 대한 의무가 있다. 사용자들은 데이터 오류, 문서화 오류, 그리고 추가적으로 필요한 정보에 대해서 관리자에게 피드백을 해줄 의무가 있다. 데이터의 오류나 outlier 을 찾는 것은 보통 사용자들이다. 한 박물관은 한 지역에 대한 데이터의 일부만 가지고 있을 수 있으며 다른 지역의 데이터와 합쳐져야 데이터의 오류가 드러나는 경우가 있다.

기관에서 데이터를 수집하는 목적에 따라서 사용자는 데이터 수집과 검증에서 미래에 가져야 하는 우선 순위를 정하는데 역할을 할 수 있다(Olivieri et al 1995).

또한 사용자는 데이터의 적합성을 판단하고 이를 부적절하게 쓰지 말아야 할 의무가 있다.

사용자와 수집가들은 데이터 품질을 유지하는데 있어서 관리자를 돕는 역할을 하며 양쪽 모두 데이터가 최고의 품질이어야 한다는 사실에 동의 할 것이다.

협력 관계 구축

데이터 품질 유지를 위해서 협력 관계를 구축하는 것은 비용을 크게 절감 할 수 잇는 방법이다. 중복되는 기록을 가지고 있는 경우가 많은 박물관과 식물원의 경우에 비용 절감 효과가 더 크다. 도서관 사이에 협력 관계를 구축해서 소장 도서의 목록을 개선하며 박물관과 식물원들도 비슷한 방법으로 협력 할 수 있다. 협력 관계가 구축 될 수 있는 경우는 다음과 같다:

- 주요 데이터 수집가(정보 소통을 원활하게 하기 위해),
- 비슷한 데이터를 갖고 있는 기관들,
- 데이터 품질에 관해서 수준이 비슷한 기관들과 품질 관리 방법, 도구, 표준 및 절차를 개발하고 있는 기관들,
- GBIF 과 같은 주요 데이터 공급 기관,
- 데이터 사용자, 그리고
- 데이터 관리 방법, 흐름 그리고 품질 관리를 개선 할 수 있는 통계학자와 데이터 수정 작업자.

데이터 품질을 고려하는 것은 당신의 기관 만이 아니다.

우선 순위

최단의 시간 안에 최대한의 이용자들이 최고 가치의 데이터를 이용 할 수 있도록 하기 위해서는 데이터 정리/검증을 우선시 해야 할 필요가 있을 가능성이 있다. 이를 위해서는:

- 가장 중요한 데이터에 집중,
- 개별 단위에 집중(분류, 지리),
- 대표적인 종과 중요한 것으로 알려진 종에 집중
- 사용 되지 않은 데이터와 품질을 보증 할 수 없는 데이터를 무시한다. 하지만 여기에서 오래된 데이터는 제외한다.
- 가장 많은 사용자에게 혜택을 줄 수 있는 데이터와 가장 많은 용도에 사용할 수 있는 데이터를 선택한다.
- 최소의 비용으로 가장 많은 데이터를 청소 할 수 있는 부분에 집중한다.

데이터는 평등하지 않기 때문에 가장 중요한 데이터에 집중을 해야 하며 청소를 할 경우에 이를 중복하지 않도록 해야 한다.

완성도

기관들은 판독이 가능한 기록이 데이터 취합에 사용 되어서 데이터의 완성도를 높이도록 노력 해야 한다. 정보가 없는 불완전한 데이터를 공개하는 것 보다 데이터를 하나의 주체로 완성해서 이를 공개하는 것이 좋다. 또한 부족한 데이터에 관한 정책을 수립해서 부족한 데이터의 기준을 마련해서 이에 대한 대응책을 마련하고 이에 대한 완성도를 문서화 하는 것도 중요하다.

가치와 시대성

데이터의 가치와 시대성에 관련된 요소가 세가지 존재한다:

- 데이터는 언제 수집 되었는가?

- 현실 세계의 변화에 맞도록 데이터가 언제 업데이트 되었는가?
- 데이터가 얼마 동안 들어 맞는가?

데이터의 가치는 사용자들이 흔히 제기하는 문제점 중의 하나이다. 데이터 관리자들은 데이터가 처음에 수집되거나 조사된 기간과 관련해서 가치라는 개념을 쓴다. 수집과 공개 사이의 기간을 고려하면 공개된 정보는 "지나간 사실"의 모음이다. 생물 다양성 데이터 사용자들은 이러한 사실을 잘 알고 있으며 이 사실은 이러한 데이터가 가지는 가치의 일부분을 형성한다.

데이터 품질 관리에 쓰이는 개념에서는 가치는 "유효 기간"이라는 의미로 쓰이는 경우가 많으며 데이터가 마지막으로 검증된 시기와 관련 되어 있다. 데이터에 이름이 포함된 경우라면 더욱 그렇다. 이들 이름이 언제 업데이트 되고 현재 쓰이고 있는 분류와 들어 맞는가? 현대적인 분류 명명법을 따른다면 종이 몇 개의 군으로 나뉘어질 경우에 나뉜 군 중의 하나는 원래 종의 이름을 따른다. 사용자의 입장에서는 그 이름이 어느 쪽을 가리키는 것인가를 아는 것이 중요할 수 있다. 또한 가치는 "유효 기간"에 해당되는 개념으로 쓸 수 있는데 이 경우에 "유효 기간"이 지나면 데이터의 품질에 대한 보증을 해 줄 수 없는 것으로 간주 할 수 있다.

하지만 다른 경우에는 시대성과 가치는 의미가 없거나 포함 시키거나 유지 하는 것이 불가능 할 수 있다. 박물관과 식물원의 데이터가 그럴 수 있다. 한편으로는 보증인이 없는 경우에는 시대성과 가치가 중요 할 수 있고 명명법 개정 후에 업데이트가 되지 않은 경우에도 그렇다. 여러 기관에서 표본을 모아서 수집한 경우에는 더더욱 그렇다. 예를 들어서 개발 도상국의 기관 몇 개가 모여서 자신들의 데이터를 GBIF 포탈에 입력을 할 수 있도록 제공하는 경우가 그렇다.

업데이트 빈도

데이터가 업데이트 되는 빈도는 가치와 시대성과 관련이 있으며 표준화 되어서 문서화 되어야 한다. 여기에 새로운 데이터의 추가와 수정된 데이터가 공개되는 빈도가 포함 되어야 한다. 이 두 요소는 데이터의 품질에 영향을 미치며 사용자들에게 중요한 정보가 된다. 사용자들은 조만간 업데이트 되고 개선될 데이터를 다운 받는 수고를 덜 수 있게 될 것이다.

일관성

Redman(1996)은 일관성에 두 가지 요소가 있는 것으로 간주하고 있다: 의미 일관성 – 데이터의 관점이 분명하고 일관성을 가지고 있으며; 그리고 구조적인 일관성: 각 문항과 특성이 같은 구조를 가지고 있어야 한다. 데이터는 항상 같은 문항에 있어서 찾는 것이 쉽다는 것이 의미 일관성의 요지인데 예를 들자면, 종 하위 단계의 서열과 이름에 대해 문항이 따로 존재해서 이름 문항에 이름 또는 접미사만이 존재하며 모든 데이터에 대해 같은 형식을 보인다(see Table 2).

속	종	종 이하
Eucalyptus	globulus	subsp. bicostata
Eucalyptus	globulus	bicostata

Table 1. 종 이하 단계에서의 의미 일관성을 해치는 예.

속	종	종 하위 서열	종 하위 단계
Eucalyptus	globulus	subsp.	bicostata
Eucalyptus	globulus		bicostata

표 2. *종 하위 단계에서 서열 문항을 포함해서 의미 일관성 유지*

하지만 relational 데이터베이스를 제대로 설계한다면 이러한 문제점들이 일어나지 않겠지만 현존하는 데이터베이스들은 그렇지 못한 경우가 많다.

구조적인 일관성은 문항 내에 일관성이 있을 때 생긴다. 예를 들어서 종 하위 단계 문항은 아종을 항상 같게 표현 했을 것이다. 이 것은 데이터베이스를 제대로 설계 함으로써 피할 수 있는 문제이다.

방법과 문서화에서의 일관성은 사용자들에게 검사가 언제, 어떻게 시행 되었고 정보는 어디서 찾을 수 있으며 그리고 정보를 해석하는 방법에 대해서 알려 줄 수 있다. 하지만 일관성은 신축성을 고려해서 조정 될 필요가 있다(Redman 2001).

신축성

생물학적 데이터는 유사점이 많은 반면에 지역, 분류, 아니면 기록 방법에 따라서 다른 방식을 쓸 수 있어야 하기 때문에 데이터 관리자들은 데이터 품질 관리 방식에서 신축성을 가지고 있어야 하지만

분류에 대한 의견은 가설일 뿐이며 이를 바라보는 방식에 따라서 다른 전문가들이 같은 생물을 다르게 분류해서 똑같이 유효한 다른 이름을 붙일 수 있는 가능성이 충분히 존재한다. 예를 들어서, 두 분류학자는 같은 종을 다른 과에 넣을 수 있다. 현실에서는 특별한 반대 의견이 존재하지 않으면 가장 최근에 개정을 행한 사람의 의견이 받아들여진다.

신축성은 새로운 요구에 부응 할 수 있는 능력을 부여해준다. TDWG(Taxonomic Databases Working Group)과 다른 기관들이 최근에 데이터베이스의 구조에 대해서 내놓은 결과물들을 보면 신축성에 중요성을 두는 데이터베이스 구조를 많이 개발하고 있으며 겉보기 품질이 떨어지는 반면에 사용자들이 용도 적합성을 판단하는데 더 많은 신축성을 부여해서 실제로는 품질을 향상 시킨다.

투명성

투명성은 데이터 사용자의 데이터에 대한 자신감을 높이기 때문에 중요하다. 투명성을 높이려면 오류를 숨기지 않고 이를 찾아서 보고하고, 검증과 품질 관리 절차가 문서화 되어서 이를 공개하며 피드백 과정이 열려 있어야 한다.

투명성이 중요한 예로 수집 방식의 문서화가 있다. 문서화를 통해서 사용자는 자신의 목적에 데이터가 적합한가의 여부를 판단 할 수 있다.

성능 측정과 목표치

성능 측정은 품질 관리 절차에 중요한 부분이며 사용자들이 데이터의 정확성이나 품질에 대해서 알 수 있게 해준다. 데이터에 대한 통계적인 검사, 품질 관리의 수준, 완성도, 등등이 성능 측정에 포함 될 수 있다.

성능 측정은 데이터 품질을 구체화 시키는 것을 돕는다. 구체화의 장점은 다음과 같다:

- 해당 기관은 일부 데이터가 품질이 높다는 것을 스스로에게 확신 시켜 줄 수 있다,
- 데이터 관리에 도움을 주고 중복을 줄인다, 그리고
- 데이터 품질 사슬 관리에 도움을 주어서 다른 사람들이 각자의 역할을 맡을 수 있도록 한다.

데이터 품질을 측정하기 전에 결과의 사용자들이 이를 어떻게 이용할지 생각하고 이를 효율적으로 하기 위해서 결과를 조정하라.

데이터 청소

데이터 청소의 원리는 *Principles and Methods of Data Cleaning*에서 다룰 것이다. 데이터 청소 절차의 윤곽은 다음과 같다(Maletic and Marcus 2000):

- 오류의 종류를 정의하고 파악한다
- 오류를 찾고 파악한다
- 오류를 수정한다
- 오류 발생과 종류를 문서화 한다
- 비슷한 오류가 발생하는 것을 막기 위해서 데이터 입력 절차를 개선한다.

데이터 청소 도구의 간단함에 현혹되면 안 된다. 단기적으로는 도움이 되지만 장기적으로 보았을 때 오류 방지를 대체 할 수 있는 것은 없다.

Outliers

Outlier 의 탐색은 지리적인 데이터에서 오류를 찾는데 가장 유용한 검사이다. 하지만 검사에서 찾은 outlier 를 무조건 지워서는 안 된다. 환경 데이터에 보기에는 outlier 이지만 사실은 정상적인 데이터인 경우가 있다. 이 것은 진화 경로, 기후 변화, 인간 활동의 잔재 때문일 수 있다. Outlier 을 무조건 제거하는 것은 귀중한 데이터를 지워서 추후에 이루어지는 분석의 결과에 영향을 미칠 수 있다.

한편에 사용자들은 분석에서 그 것이 올바른 기록인지 모르는 채로 outlier 을 지울 수 있다. 따라서 outlier 을 찾는 것은 오류일 가능성이 있는 기록을 찾는 의미도 있고 사용자들이 데이터의 적합성에 대해서 알려주는 행위이다.

Outlier 을 찾는 것은 중요한 검증 방법이지만 모든 outlier 는 오류가 아니다.

개선 목표 설정

간단하고 구체적인 목표를 설정하는 것은 데이터 품질을 빠르게 개선 시킨다. 지리적인 정보가 부실한 기록의 비율을 2 년 동안 6 개월마다 반으로 줄이는 것은 오류 발생률을 최대 94% 줄일 수 있다(Redman 2001). 목표를 설정 할 때 염두에 두어야 하는 점은 다음과 같다:

- 분명하고 공격적인 스케쥴
- 품질 가치 보다는 개선 속도,
- 명확한 개념 정의,
- 간단하고 성취가 가능한 목표.

반 년 마다 데이터 입력과 검증 방식을 개선해서 데이터 청소에 필요한 시간을 줄이는 장기적인 목표를 세우는 것도 가능하다.

Performance targets are a good way for an organisation to maintain a consistent level of quality checking and validation – for example 95% of all records are documented and validated within 6 months of receipt.

감사

관리자들이 어느 데이터가 언제 검사 되었는지 아는 것은 중요하다. 이렇게 하는 것은 중복을 줄이고 검사에서 빠트린 데이터가 없게 하기 위한 것이다. 이를 위해서 문서화 된 감사 일지 같은 것이 필요하다.

수정 권한

수정 권한은 특정 문항에 들어 갈 수 있는 입력 값을 제한한다. 예를 들어서 수집 월 문항에는 1 에서 12 까지만 입력 가능하고 수집일 문항에는 월에 따라서 1 에서 31 까지만 입력 가능하다.

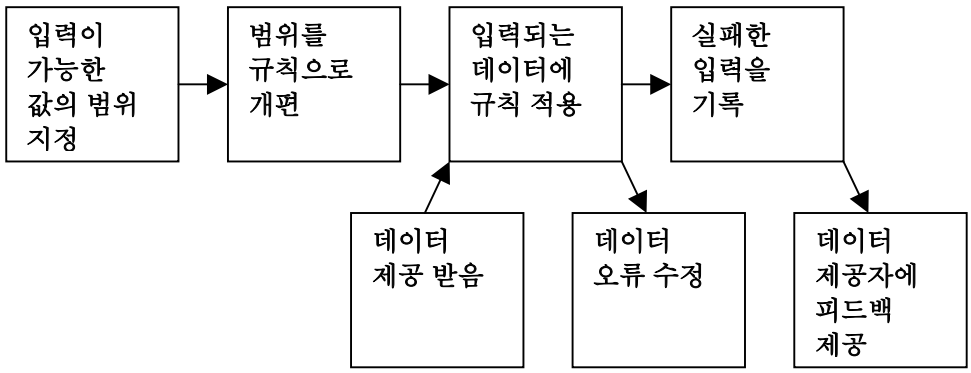

그림 5. *수정 권한(Redman 2001).*

두 번째 예는 좌표 데이터이다. 이 것은 단순한 범위 조사로 검사 할 수 있다. 하지만 시간대 데이터는 틀리다. 좁은 지역에서 수집된 데이터는 데이터가 어느 시간대에서 수집이 되었는지에 대한 기록이 없는 경우가 많다. 다른 지역에서 수집 된 데이터와 함께 쓰이지 않는다면 이 것은 문제를 일으키지 않는다. 하지만 데이터를 합치려고 한다면 데이터의 적합성은 크게 떨어질 것이다. 따라서 입력 제한을 설정해서 시간대를 입력 하도록 해야 한다.

중복과 데이터 재입력 최소화

기업계에서는 정보 관리 사슬(그림 3)의 사용은 중복과 데이터의 재정리를 50%까지 줄이는데 효과를 보였으며 관련 비용은 67% 감소했다(Redman 2001). 이 것은 데이터 관리와 품질 관리에서 책임 소재를 분명히 하고 병목 현상과 대기 시간을 줄이며 품질 관리의 순환을 통한 중복의 최소화, 그리고 작업 방식의 개선을 통해서 이루어졌다.

원본 데이터 보존

데이터 정리 작업을 하면서 수집가가 기록하거나 관리자가 입력한 원본 기록이 유실 되지 않도록 하는 것이 중요하다. 데이터 정리 작업 중에 데이터베이스의 변경 사항은 추가 정보로 포함이 되어야 하며 원본 기록도 보관을 해야 한다. 한번 지워진 기록은 복구하는 것이 매우 어렵거나 불가능 할 수 있다. 수집가와 소재지 정보를 다룰 때 원본 기록을 보관하는 것이 특히 중요한데 나중에 관리를 맞게 되는 사람이 보기에 틀린 정보가 실제로는 맞는 경우가 있다. 지역 명의 변경은 이름만 바꾸는 것이 아니라 상황도 변화 시킬 수 있다. 그러므로 원본 기록의 내용을 아는 것이 중요하다.

조직화는 데이터의 유실과 품질 저하로 이어질 수 있다.

데이터를 조직화 하는 것은 데이터 유실과 데이터 품질 저하로 이어질 수 있다. 예를 들면 서술로 이어진 소재지 정보를 격자 형식 정보로 변환하면 데이터가 많이 변질 될 수 있다. 그래서 데이터 해상도를 최고로 해놓은 상태에서 우선 저장하고 나중에 필요에 따라 조직화 하는 것이 좋다. 사용자가 격자에 분포 지도를 만들고 싶다면 지점 데이터로부터 지도를 만드는 것은 쉽지만 격자 좌표 형식으로 저장 되어 있다면 더 높은 해상도로 데이터를 표시하는 것은 불가능하다. 또한 다른 축적으로 조직화 된 데이터를 통합하는 것은 어려운 일이 될 수 있다. 서술 데이터의 경우도 마찬가지이다. 기록 내에서 사용된 기준이 다르다면 통합이 어려울 것이다. 그러므로 데이터를 처음부터 자세하게 기록하는 것이 좋다.

지리 좌표의 정확성을 저장 할 때 이런 경우가 자주 발생한다. 미터법을 쓰는 것이 가장 좋지만 많은 데이터베이스들은 이 정보를 몇 개의 간격 정보로 저장한다(<10m, 10-100m, 100-1000m, 1000-10,000m). 2km 해상도를 가지는 정보가 있다면 그 정보는 10km 간격으로 들어가 정보가 유실된다.

문서화

철저한 문서화는 데이터 관리에 필수적이다. 철저한 문서화 없이는 사용자는 데이터의 적합성을 판단 할 수 없다.

피드백

관리자들은 데이터 사용자들이 피드백을 권장해야 하며 받은 피드백을 진지하게 받아들여야 한다. *사용자의 책임* 문항에 언급 되어 있듯이, 단독으로 일하는 데이터 관리자 보다는 사용자가 여러 출처에서 얻은 데이터를 결합 시킴으로써 오류를 찾을 가능성이 높다.

효율적인 피드백 절차를 개발하는 것은 쉬운 일이 아니다. 쿼리 페이지에 피드백 기능을 추가 할 수 있거나 데이터를 다운 받는 사용자들에게 피드백 방법에 대한 공지를 띄울 수 있다. *Principles and Methods of Data Cleaning* 에 피드백에 대한 내용이 자세하게 다루어져 있다.

효율적인 피드백 관계는 데이터 품질을 높이는 쉽고 생산적인 방법이다.

교육과 훈련

정보 사슬의 모든 단계에 있는 구성원들을 교육 시키는 것은 데이터 품질을 크게 향상 시킨다(Huang et al 1999). 수집가들이 받는 기록 절차와 사용자 수요 충족 방법부터 시작해서 입력 작업자와 기술 인력을 데이터베이스의 일상적인 관리, 그리고 데이터의 성질, 제한, 그리고 잠재적인 사용에 대한 사용자 교육과 훈련까지 포함 되어 있다. 이 과정은 철저한 문서화가 많은 도움을 준다.

데이터 품질 관리, 교육, 및 훈련의 통합은 MaPSTeDI 지리 좌표 지정 프로젝트에서 볼 수 있다. 입력 작업 담당이 바뀔 때마다 관리자가 일정한 수의 입력을 직접 검토한다. 이렇게 함으로써 데이터의 질이 유지 될 뿐만 아니라 입력 담당이 자신의 실수에서 배울 수 있다. 입력 담당에 따라서 관리자가 검토하는 기록의 수가 늘어 날 수 있고 담당이 익숙해지면서

검토되는 기록의 수가 줄어들게 된다. 오류가 여전히 많이 발견되면 검토되는 기록의 수가 늘어난다.

설계 잘 된 절차들은 새로운 사용자를 교육 시키는데 큰 역할을 한다. 반대로 이러한 절차가 없다면 입력 담당자 간에 일관성을 유지할 방법이 없다.

책임 소재

책임 소재를 정해서 데이터 품질 관리의 수준을 일정하게 할 수 있으며, 피드백을 도우며 문서화와 쿼리 사이에 접점을 마련해준다.

질적으로 낮은 교육과 훈련은 데이터 품질 관리에 많은 문제점을 유발한다.

분류와 명명 데이터

질이 낮은 분류 데이터는 다른 영역에 나쁜 영향을 끼친다 (Dalcin 2004).

분류학은 생물의 특성을 파악하고 이를 분류하는 학문이다. 여기에서 다루게 될 종 데이터는 분류에 관한 데이터가 포함 되어 있으며 공간 정보와 달리 품질을 파악하는 것이 쉽지 않다.

분류 데이터는 다음과 같이 이루어져 있다:

- 이름(학명, 일반 이름, 서열)
- 분류 상태(동의어, 채택 여부)
- 출처(저자, 논문 출판 저널 및 일자)
- 기록 감정 책임자 및 일시
- 품질 문항(감정의 정확성)

분류학 명명에서 오류가 가장 많이 나는 것은 철자이다. 속과 과 이름의 경우에는 권한 파일을 이용하면 철자 오류를 찾는 것은 쉽다. 또한 Species 2000 (http://www.species2000.org)나 GBIF의 ECat 프로그램 (http://www.gbif.org/prog/ecat)을 통해 종 이름의 목록을 받을 수 있다. 권한 파일에 종 이름이나 접미사와 속 이름을 넣어서 이용하는 것은 접미사가 속 간에 차이가 많이 나기 때문에 권장되는 방법이 아니다. 철자 오류를 찾는 다른 방법으로 유사성을 찾는 알고리즘을 이용해 유사하지만 다른 점이 존재하는 학명을 찾는 것이다 (Dalcin 2004, CRIA 2005).

현재로서 학명의 철자 오류를 막는 가장 좋은 방법은 권한 파일을 이용해 종, 속, 과 이름을 선택 하도록 하는 것이다. 권한 파일을 이용할 수 있는 이상적인 상황이라면 철자 오류를 거의 완벽하게 없앨 수 있다. 하지만 권한 파일이 존재 하지 않는 지역이나 분류군들이 존재한다.

Catalogue of Life 나 ECat 같은 외부 출처의 권한 파일을 이용한다면 이들의 버전을 기록해서 권한 파일 소스 버전 변경이 데이터베이스에 편입이 될 수 있도록 해야 한다. 미래에는 GUID 이용을 통해서 이러한 작업이 쉬워질 것이다.

데이터 중 분류학 데이터의 질은 이를 검증하는데 동원할 수 있는 분류학 전문가에 달려있다. 소위 말하는 “Taxonomic Impediment”과 세계적으로 분류학자의 감소는 분류학과 분류 데이터의 품질 저하로 이어질 것이다(Stribling el at 2003). GTI(Global Taxonomic Initiative)는 이러한 문제를 해결 하려고 노력하고 있지만(CBD 2004) 완전한 해결은 요원한 상태이다. 감정된 표본이 유지가 안되거나 없는 경우, 그리고 해당 분야에 대한 전문가의 도움이 존재하지 않으면 시간이 흐르면서 품질이 저하 될 수 있다.

기관이 고품질의 분류학 생산물(문서화된 생물 종 데이터 포함)을 내놓을 수 있는 능력은 다음 요소들에 의해서 영향을 받는다(Stribling *et al.* 2003):

- 직원의 훈련과 경험,
- 기술 문헌, 참고 문헌, 감정된 표본 모음, 그리고 분류학 전문가에 대한 근접성,
- 적절한 연구 기자재와 시설의 존재 여부, 그리고
- 인터넷 이용 여부.

감정 신뢰성의 기록

전통적으로 박물관과 식물원들은 분류를 할 때 분류군 마다 전문가 여러 명이 수집물들을 검사하면서 분류를 재확인한다. 이 작업은 분류 연구의 일환으로 진행되거나 방문 중인

연구원에 의해서 이루어진다. 이 방식은 검증이 된 방식이지만 시간이 많이 걸리고 실수가 발생할 여지가 많다. 하지만 컴퓨터를 이용한 자동 분류가 이루어질 것으로 보이지 않기 때문에 이 방식이 유일한 방식이다.

데이터베이스에서 데이터의 정확성에 대한 항목을 만드는 것이 가능하다. 이런 항목을 데이터베이스에 포함 시키는 방법에 몇 가지가 있으며 표준적인 방법은 아직 존재하지 않는다. 코드 항목으로 하는 것이 좋으며 다음과 같이 할 수 있다(Chapman 2004):

- 세계적인 해당 분류군의 전문가에 의해 높은 신뢰도로 분류
- 세계적인 해당 분류군의 전문가에 의해 보통 신뢰도로 분류
- 세계적인 해당 분류군의 전문가에 의해 낮은 신뢰도로 분류
- 지역적인 해당 분류군의 전문가에 의해 높은 신뢰도로 분류
- 지역적인 해당 분류군의 전문가에 의해 보통 신뢰도로 분류
- 지역적인 해당 분류군의 전문가에 의해 낮은 신뢰도로 분류
- 해당 분류군의 비전문가에 의해 높은 신뢰도로 분류
- 해당 분류군의 비전문가에 의해 보통 신뢰도로 분류
- 해당 분류군의 비전문가에 의해 낮은 신뢰도로 분류
- 수집가에 의해 높은 신뢰도로 분류
- 수집가에 의해 보통 신뢰도로 분류
- 수집가에 의해 낮은 신뢰도로 분류.

이 분류를 어떻게 등급화 하는 문제는 논의해야 할 문제이고 이런 분류가 최선의 분류인가도 논의해야 할 문제이다. 이미 몇몇 기관에서는 이런 형식의 항목을 데이터베이스에 포함시킨다. HISPID 표준 버전 4 에서는 이를 단순화 시켜서 포함시킨다(Verification Level Flag with five codes – 표 1):

0	기록의 이름이 검사 된 적 없음
1	이름을 갖고 있는 식물과 비교해서 기록의 이름 판명
2	기록의 이름이 분류학자나 식물원/도서관/기록을 이용한 능력이 있는 사람에 의해서 판명됨
3	분류군에 대해서 체계적으로 개정 작업을 하는 분류학자에 의해서 기록 이름이 판명됨
4	다른 기록에서 같은 기록이 발견됨

Table 3. HISPID(Conn 2000)에서 Verification Level Flag

많은 기관들은 “aff.”, “cf.”, “*s. lat*.”, “*s. str*.”, “?”와 같은 약자를 이용하여 신뢰성을 표시한다. “aff.”, “cf.” 처럼 의미가 분명한 것도 있지만 개인에 따라서 이용이 다를 수 있다. 또한 *sensu stricto* 와 *senso lato* 의 이용은 분류에서의 변화를 의미한다.

다른 방법으로는 이름의 출처를 적는 방법도 있다(after Wiley 1981):

- 새 분류군의 묘사
- 분류군의 개정
- 분류 예시
- 분류 기준
- 동물과 식물 연구
- 전집
- 일람표

- 체크 리스트
- 핸드북
- 명명법의 법칙
- 계통 연구

불확정성을 줄이는 것이 가능하며 두 개 이상의 논문이나 전문가의 비교를 통하여 품질을 개선 할 수 있다. 하지만 분류학자 간의 분류에 대한 이견은 오류가 있다는 것을 의미하는 것이 아니고 단순히 분류에 대한 의견 차이 일 수 있다.

감정의 정확성

Stribling(2003)에 의하면 감정의 정확성은(논문에서는 분류의 정확성으로 잘못 기재 되어 있다) 두 명의 전문가에 의해서 무작위로 고른 표본 감저으이 차이를 살펴보면 된다. 또한 다른 기관에 있는 두 표본의 이름을 비교하는 방법으로도 할 수 있다. 하지만 이는 이론적인 생각일 뿐이기 때문에 이러한 정보가 의미가 있을지는 의문이다.

감정의 정확성은 표본 감정이 어느 단계까지 이루어졌는가에 의해서도 영향을 받는다. 종 또는 아종 단계까지 감정이 된 것은 속이나 과 단계까지만 감정이 된 것에 비해서 정확하게 감정이 된 것이다. 데이터베이스에 대한 정보를 문서화 할 때 50%의 표본에 대해 속 단계까지만 감정이 되었다고 알리는 것은 사용자에게 중요할 수 있다.

치우침

모든 측정치가 한꺼번에 같은 방향으로 쏠리면서 생기는 오류가 치우침이다(Chrisman 1991). 보통은 오류를 유발하는 처리 방식을 계속 이용함으로써 나타나게 된다. 분류에서 치우침은 감정이 정밀하게 이루어지지만 정확하지 않을 때 나타난다. 이러한 치우침은 이분 기준이나 형상의 잘못된 해석, 무효한 명명법이나 문헌의 사용, 아니면 부적절한 문헌의 사용으로 인한 것일 수 있다.

일관성

데이터베이스에서 두 개 이상의 이름이 같은 종을 나타내는 경우(*Eucalyptus eremaea* and *Corymbia eremaea*)에 일관성을 해칠 수 있다. 분류에 대한 의견이 다르거나 다른 철자에 의한 것일 수 있다(예를 들어서, *Tabernaemontana hystrix, Tabernaemontana histryx* and *Tabernaemontana histrix* – CRIA 2005).

완성도

Dalcin(2004)에서 Motro와 Rakov(1998)은 완성도를 데이터의 존재 여부로 정의하고 데이터의 완성도를 기록의 존재 여부와 기록의 완성도 여부로 나누었다.

분류학에서 완성도는 알고 있는 이름의 적용 범위를 가리킨다. 데이터베이스 내에 분류학의 모든 단계에 대한 이름이 모두 포함이 되어 있는가? 또는 동물과 식물계에서 어느 부분을 데이터베이스가 다루고 있는가? 데이터베이스에 동의어가 포함 되어 있는가? 이 질문들은 사용자가 데이터가 용도에 적합한가를 판단하게 해준다. 예를 들어서 Dalcin(2004)에서는 완성도를 존재 가능한 이름 포함 여부를 나타내는 명명 완성도(어느 분류군의 모든 이름과 일정한 지역에 관련된 모든 종 이름)와 어느 분류군에 대해서 가능한 모든 이름을 나타내는 분류 완성도로 나누었다.

표본이나 관찰 데이터베이스의 경우에 완성도는 “Darwin Core 문항이 모두 포함이 되었는가”와 “모든 Darwin Core 문항에 데이터가 있는가”로 정의 할 수 있을 것이다. 형질 데이터베이스에서는 “일생 주기의 모든 단계가 모두 포함 되어 있는가”로 정의 할 수 있다.

인증서 목록

인증서 목록의 중요성은 몇 번을 강조해도 부족하지 않지만 모든 데이터베이스에 인증서를 포함 할 수 없다. 관찰 기록 데이터베이스들은 만들어지면서 동시에 인증서 목록을 만들지 않는 경우가 많다. 또한 정치, 법률, 보존을 포함한 다른 이유로 인증을 목적으로 샘플을 채취하는 것도 불가능하다.

인증이 가능하다면 생물 종 데이터 수집 작업 초기에 데이터 수집가들과 기관 사이의 협력 관계를 구축해서 참고 문헌과 인증서 목록을 만드는 것이 중요하다(Brigham 1998). 협력 관계를 구축한다면 보관과 처리 전략을 세우는 것도 중요하다.

공간 데이터

공간 데이터는 데이터 문서화 표준에 관해서는 가장 앞서 있으며(Spatial Data Transfer Standards 개발, ISPIRE(Information for Spatial Information in Europe) 개발), 데이터 품질 표준 개발(ISO 19115 for Geographic Information – Metadata)에 대해서도 가장 앞서 있다. 공간 데이터는 수치로 표현 되기 때문에 분류 데이터 보다는 통계적인 처리에 알맞으며 이를 이용해 데이터 품질 검사 몇 가지가 개발 되었다.

이 것은 데이터 중에 공간을 다루는 부분이 모두 디지털화 하는 것이 쉽거나 정확하다는 것은 아니다. 박물관과 식물원에 있는 오래된 표본 수집 기록에 기초적인 내용만 적혀 있는 서술로 된 소재지 정보를 갖고 있으며 이를 수치화 하는 것은 노력이 많이 드는 일이다. 표본 수집이 정확한 지도가 없던 시절에 이루어졌거나 지명이 많이 변한 경우에 수치화가 더욱 어려워진다. 수집 당시의 상황을 보여주는 안내서가 있지 않은 한 오래된 기록에 공간 정보를 추가 하는 것은 시간이 오래 걸리고 정확도가 낮을 것이다.

사용자들이 가지고 있는 데이터에 지리 좌표를 배정하는 작업을 도와 주는 도구들이 몇 개 개발 되었으며 이 중에 온라인 도구와 지침서도 포함 되어 있다. 이들은 *Principles and Methods of Data Cleaning* 에서 다룰 것이다. 최근에는 수집가들은 GPS 를 이용하여 수집된 지리 좌표 정보를 기록하고 있다. GPS 를 이용한 기록의 정확성에 관한 것은 "데이터 기록" 단원을 참고 하시기 바랍니다.

이미 지정된 지리 좌표에 대한 오류 검사는 다음 사항들을 포함 할 수 있다:

- 기록 내의 다른 정보와 비교하거나 데이터베이스의 다른 기록과의 비교 – 예를 들어서, 주, 지명 등;
- 데이터베이스를 이용하여 외부 정보와의 비교 – 예를 들어서 수집가가 보통 수집 활동하는 장소와 일치하는가?
- GIS 를 이용하여 외부 정보와의 비교 – 기록은 바다를 가리키는가 아니면 육지를 가리키는가?
- 지리 공간에서의 outlier 점검; 그리고
- 환경적인 outlier 검사.

Principles and Methods of Data Cleaning 에서 자세하게 다룰 것입니다.

지리적인 정확성

지리적인 데이터의 위치 정확성은 어떻게 측정 되는가?

대부분의 GIS 지도에서는 몇 가지 사항에 대해서는 매우 높은 정확도를 갖는 요소(길, 교차로, 측량 지점)가 몇 가지 있기 때문에 기록의 정확성을 측정하는 것이 비교적 쉽다(Chrisman 1991). 하지만 대부분의 경우에 검사 과정이 간단하지 않으며 문서화도 복잡하게 되어 있다(예: US National Map Accuracy Standard). 전통적으로 지리적인 정확성은 정의가 잘 된 몇 개의 점과의 비교에 의해서 이루어지며 RMSE(root-mean-square deviation)로 측정되는 오차 한계로 표시된다(Chrisman 1991). RMSE 를 개별적인 점에 적용하는 것은 쉽지 않고 디지털화 된 지도나 데이터 세트에 적용 되는 것이 보통이다. 개별적인 지점의 경우에 지점-반경 방법이나(Wieczorek *et al.* 2004) 이와 비슷한 방법을 쓸 수 있다. 여기에 두 가지 요소가 작용한다 – 정의된 지점이 파악 될 수 있는 정확성이 지점의 정확성을 가름하고

지점에 대한 측정이 오차를 더할 것이다. 예를 들어서 교차로가 100 미터까지만 위치가 파악되면 수집 지점에도 같은 오차가 추가 된다.

FGDC(US Federal Geographic Data Committee)는 1998 년에 GPAS(Geospatial Positioning Accuracy Standards)을 내놓았다. 이 표준은 Geodetic Networks 와 Spatial Data Accuracy (FGDC 1998)을 포함 하고 있다.

- *'NSSDA 는 RMSE 를 이용해서 위치의 정확성을 측정한다. RMSE 는 독립된 정보 소스와 데이터 좌표 수치 간의 제곱된 차이의 평균의 제곱근이다. '*
- *'정확성은 95% 신뢰도 구간으로 지상 거리로 보고 된다. 95% 신뢰도 구간은 데이터의 각 지점 중 95%가 보고된 정확성과 같거나 작은 오차를 갖고 있다는 뜻이다. 보고된 정확성 수치는 측지선 조정 좌표, 기록, 그리고 지상 좌표 수치 계산 등에 의해서 영향을 받는다.'*

위와 같은 방법을 이용한 지도의 정확성에 대한 정확성에 대한 예 입니다:

- *'이 지도의 평균적인 정확성은 평면으로는 ±100 미터 이며 고도에 대해서는 ±20 입니다.'* (Division of National Mapping, Sheet SD52-14, Edition 1, 1:250,000).

이러한 정확성에 관련된 수치들은 종이나 디지털 지도 위의 수집 목록을 지리 좌표 작업 도중에 포함이 되어야 한다. 지리 데이터의 정확성에 항상 불확정적인 요소가 있기 때문에 절대적인 정확성에 대한 수치는 적용이 불가능하고 알려진 정확성은 문서화 하는 것이 중요하다. 오류를 정보 사슬을 통해서 전파되며 최종 결과에 불확정적인 요소를 추가한다.

BioGeomancer 프로젝트

생물 종 데이터의 지리 좌표 배정 작업을 개선하고 정확도를 파악, 개선, 그리고 문서화 하기 위한 프로젝트가 Gordon and Betty Moore 재단의 지원을 받기 시작했다. 이 프로젝트는 2006 년 중으로 개발된 도구들을 공개 할 수 있을 것이다.

잘못된 정확도와 정밀도

사용자들은 잘못된 정확도와 정밀도에 대해서 인식하고 있어야 한다. 많은 GIS 사용자들은 지리 데이터의 정확도, 오류, 그리고 불확정성에 대해서 모르고 있으며 데이터가 절대적으로 정확하다고 보는 경우가 많다. 출처 데이터로 얻을 수 없는 정확도를 보고하는 경우가 많다. 많은 기관들은 GIS 을 이용하여 지리 좌표 배정을 돕고 데이터가 지원하지 않는 정도로 확대를 하는 것은 비현실적인 정밀도를 만들어내게 된다. GPS 를 사용하는 경우에 사용한 GPS 기기로 나올 수 없는 정확도를 보고하는 경우도 있다. 이 문제는 GPS 를 이용하여 고도를 산정 하는 경우에 문제가 된다. 데이터 기록 단원을 참고하시기 바랍니다.

수집가와 수집 목록 데이터

수집가와 수집 목록에 대한 데이터는 목록 자체에 대한 데이터가 포함되어 있다. 데이터는 다음과 같이 규정 될 수 있다(Conn 1996, 2000):

- 수집가와 번호
- 관찰자의 경험, 등등,
- 수집 일자와 기간,
- 수집 방법,
- 관련 데이터.

이 문제들은 수집되는 데이터의 종류에 따라서 많이 달라질 것이다. 박물관을 위한 정적인 수집의 경우에 수집가의 이름, 번호, 그리고 날짜가 핵심적인 요소이며 습관, 분포 지역, 그리고 포획 방법도 포함 될 수 있다. 관찰 데이터의 경우에 관찰 기간, 지역, 시간, 날씨, 동물의 성별, 동물의 활동이 포함 될 수 있다. 조사 데이터의 경우에 조사 방법, 크기, 노력, 날씨, 조사 주기, 그리고 인증 방법의 존재 여부가 포함 될 수 있을 것이다.

부여된 정확성

수집가의 이름, 번호, 약자, 기록 시간의 정확성, 기록의 일관성과 같은 데이터 품질에 영향을 끼칠 수 있는 요소들은 기록된다(Koch 2003).

예를 들어서 몇몇 수집가들은 수집의 번호를 고유하게 배정하지 않을 경우에 문제가 발생 할 수 있다. 이들 태그가 수집의 장소, 감정, 그리고 중복 여부를 판정하는데 쓰이기 때문에 품질의 저하로 이어질 수 있다.

일관성

수집에 일관성이라는 단어는 사용이 불규칙적이며 서로 관련된 데이터 문항이 데이터 내에서 일관성을 유지하는 일은 거의 없다.

완성도

수집 정보의 완성도도 불규칙하다. 분포 지역, 수집 번호, 개화 시기 같은 정보는 완성이 되어 있지 않은 경우가 많다. 이런 경우에 분포 지역 연구가 수집 데이터 만으로 어렵게 된다.

서술 데이터

서술 데이터베이스들은 데이터를 저장하는 역할만이 아니라 출판의 매개체로도 쓰이고 있습니다. 데이터의 형태, 주기, 그리고 생리적인 요소들이 서술 데이터의 예시라 할 수 있다. 서술 데이터는 보통 진화 계통 연구, 그리고 자동적으로 생성되는 정보와 감정 도구를 위해서 쓰이고 있다.

TDWG(Taxonomic Databases Working Group)는 서술 데이터베이스 표준 개발과 보급에 관해서는 오랜 역사를 가지고 있다. 처음에는 DELTA 표준을 지원했으며(Dallwitz and Paine 1986) 최근에는 "Structure of Descriptive Data" 위원단을 구성했다 (http://160.45.63.11/Projects/TDWG-SDD/).

서술 데이터의 품질은 고르지 못하며 데이터의 각 요소가 측정이 되는 경우가 많지만 현실적으로는 관찰이 불가능하거나, 현실적이지 못하거나 아니면 실제로 존재하지 않는 요소에 의해서 정확성이 정해진다.

서술 데이터는 보통 표본 단위 보다는 종 단위에서 저장되며 평준화 되어 있는 경우가 많다. Morse(1974)에 의하면 분류 정보는 표본 관찰 데이터 보다 신뢰도가 떨어진다. 하지만 이와 상관 없이 최근에는 서술 데이터의 일부를 표본 단위로 저장해서 품질을 높이는 경우가 있다.

완성도

표본 단위에서는 서술 데이터의 완성도는 표본의 품질과 수집 기간의 영향을 받을 수 있다. 예를 들어서 같은 표본이라 하더라도 과실이나 꽃에 대한 정보를 수집 할 수 없는 경우가 있을 것이다. 이런 경우에 많은 수의 문항이 빈 칸으로 남겨져 있을 것이다. 다른 경우에는 해당 특성이 얻으려는 데이터와 상관이 없어서 그 특성을 측정하지 않을 수 있다.

일관성

서로 관련되어 있는 데이터 항목에 대해서 일관성 문제가 일어날 수 있다. 예를 들어서 두 종의 형상 특징이 다음과 같이 매겨질 수 있다(Dalcin 2004):

- "HABIT=HERBACEUS"
- "USES=WOOD"

같은 특징에 대한 일관성 없는 표현은 형질 정의가 철저하지 않거나 표준에 일관성이 없는 경우에 품질에 영향을 끼칠 수 있다. 예를 들어서(Dalcin 2004):

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

단어의 표준화는 오류와 잘못된 해석을 많이 줄일 수 있다. 몇 개의 분야에서 표준 단어들이 개발되고 있으며 연합된 서술 데이터베이스의 개발이 단어 사용에 일관성을 증가 시켜 주었다. TDWG 의 SDD 표준 개발은 이 과정을 도울 것이다.

데이터 기록

생물 종 데이터와 종 분포 데이터를 기록하는데 여러 가지 방법이 존재하며 정확도와 정밀도에 차이가 존재하고 오류가 생기는 원인도 다르다. 적합성에 모두 영향을 끼치게 된다. 생물 종 데이터를 기록 할 때 많이 쓰이는 기록 방법에 대해서 논의를 할 것입니다.

기회가 생기는 대로 기록

대부분의 종 발생 데이터는 기회가 생기는 대로 기록이 되었다. 이 기록의 대부분은 박물관과 식물원에서 표본으로 저장 되어 있다. 예전에 기록된 데이터는 소재지 정보를 단순히 어느 지점에서의 방향과 거리만 기록하고 지정된 정확한 지리 좌표가 없다. 지리 좌표를 차후에 지정된 경우가 많으며 수집가 이외의 사람에 의해서 이루어지는 경우가 많다(Chapman and Busby 1994). 관찰 데이터의 대부분은 이런 형식으로 수집 되었다.

이러한 데이터는 batch 형식으로 디지털화 되며 지리 좌표 지정은 지도를 참고해서 이루어진다. 보통은 정확도와 정밀도가 낮으며 2-10km 보다 정확도가 높다고 보는 것이 힘들다.

현지 조사

현지 조사 데이터에 위도/경도나 시간대 정보의 형태로 소재지 정보가 포함 되어 있다. 보통은 100-250 미터 오차가 존재한다. 하지만 소재지 정보가 가리키는 대상을 분명히 해야 한다. 실제로 관측이 이루어진 지점이 아니고 사각형의 중심이나 모서리이거나 교차 지점의 중간 일 수 있다. 또한 기록이 인증 되는 경우가 적기 때문에 분류의 정확성이 항상 신뢰성 있는 것이 아니다. 조사가 이루어진 시대에서 벌어지고 분류 체계가 변하면서 이러한 문제점이 심해진다.

대규모 관찰

몇몇 생물 종 관찰 조사는 어떤 경계선의 안쪽이나 격자선 안에서만 데이터를 기록한다. 예를 들어서 국립 공원 내의 종 조사나 10km 격자 속의 새 관찰 기록이 이에 해당된다. 이런 기록의 오차율은 최소한 1-10km 단위이다.

GPS

GPS는 최근에 생물 종 데이터의 수집에 많이 쓰이고 있다. 여기에 관찰 데이터 뿐만 아니라 즉흥적이거나 관찰 조사도 포함된다.

GPS는 삼각 측량을 이용하여 지구상의 위치를 계산한다. 측정되는 거리는 GPS 기기와 위성 간의 거리이다. GPS 위성들은 궤도상의 알려진 위치에 있기 때문에 지구상의 위치를 계산 할 수 있다. 어느 한 지점의 위치를 파악하기 위해서는 최소한 네 개의 GPS 위성이 필요하다(McElroy *et al.* 1998, Van Sickle 1996). 오늘날 지구상 어느 지점이라 하더라도 7개의 위성을 이용 할 수 있어서 문제가 되지 않지만 예전에는 이용 가능한 위성의 숫자가 충분하지 않았다. 2000년 5월 이전에는 민간용 GPS 기기는 이용 가능한 위성의 수에 제한을 받아서 정확도가 떨어졌지만 지금은 그러한 문제가 존재하지 않는다(NOAA 2002).

GPS 위성의 이용 제한이 풀리기 전에 휴대용 GPS 기기의 정확성은 100m 보다 작지 못했다(McElroy *et al*. 1998, Van Sickle, 1996, Leick 1995). 하지만 최근에 상황이 좋아져서

4 개 이상의 위성을 이용하는 경우에 휴대용 GPS 기기의 오차가 10m 미만이 되었다. 같은 장소에서 관측을 여러 번 한 결과의 평균을 사용 하는 것은 정확도를 높일 수 있고(McElroy *et al.* 1998) 몇몇 최신 GPS 기기들은 정확도가 5m 미만이다.

위상차 GPS 의 이용은 정확도를 높일 수 있다. DGPS 는 GPS 지상 중계국과 GPS 기기가 동시에 위치를 계산해서 두 결과를 비교하는 방식으로 대기로 인한 오차를 줄인다. 두 결과를 비교 함으로써 휴대용 GPS 기기는 계산 결과가 적절한 오차 수정을 할 수 있다. 기기의 품질에 따라서 1-5m 의 정확도를 얻을 수 있다. 이 정확도는 지상 중계국과의 거리가 멀어짐에 따라서 감소한다. 여러 번 측정한 결과를 평균 내는 것은 정확도를 높일 수 있다(McElroy *et al.* 1998).

WAAS(Wide Area Augmentation System)은 GPS 를 이용하는 비행기의 정확한 유도를 위한 항해 및 착륙 시스템이다. WAAS 는 위치가 정확하게 파악 되어 있는 지상 안테나를 이용하여 GPS 의 정확도를 높인다. LAAS(Local Area Augmentation System)도 정확도를 높이는 시스템이다.

실시간 위상차 GPS(McElroy *et al.* 1998)나 정적 GPS(McElroy *et al.* 1998, Van Sickle 1996)를 이용하면 이용 하면 정확도를 더 높일 수 있다. 정적 GPS는 정밀도가 높은 기구와 기법을 이용하며 보통은 측량 기사들에 의해서 이용된다. 오스트레일리아에서 이 기법을 이용한 조사들은 정확도가 센티미터 단위로 보고 되었다. 하지만 비용이나 필요성 문제 때문에 생물종 데이터 수집에서 이러한 기법들이 쓰일 가능성은 적다.

위에서 언급된 정확도를 얻기 위해서는 GPS 기기는 하늘에 노출된 지역에 있어야 하며 최소한 네 개의 GPS 위성의 신호를 받을 수 있어야 한다. 최선의 조건은 바로 위에 위성이 하나 있고 나머지 세 개는 골고루 분포하는 것이다(McElroy *et al.* 1998). GPS 기기의 지역 설정이 올바르게 되어 있어야 한다.

GPS 고도 설정. 대부분의 생물학자들은 GPS를 이용해서 측정된 고도에 대해서 알고 있을 것이다. GPS에 의해서 측정되는 고도는 Earth Centric Datum 기준이며 해수면이나 다른 표준 고도에 의한 것이 아니다. 오스트레일리아에서는 GPS의 고도와 해수면에 의한 고도의 차이가 -35m에서 +80m까지 날 수 있으며 차이는 일정하지 않다(McElroy *et al.* 1998, Van Sickle 1996).

데이터의 입력과 입수

. 데이터 입력과 입수는 간단한 오류와 복잡한 오류에 취약하다.
(Maletic and Marcus 2000)

기본적인 데이터 기록

데이터 기록의 첫 단계는 표본 레이블, 일기, 현장 일지 아니면 카드 목록의 정보를 뽑아내는 것이다. 이 작업은 데이터 입력 인력 아니면 스캐너를 이용할 수 있다. 데이터 입력으로 인한 오류는 이중으로 목록 작성, 스캐너와 관련된 학습 소프트웨어, 그리고 데이터를 무작위로 검사 할 전문가나 관리자를 이용 함으로써 줄일 수 있다(see the MaPSTeDI Guidelines mentioned below).

유저 인터페이스

데이터 입력을 위한 유저 인터페이스를 개발 하는 것도 오류를 줄이는 방법이다. 많은 기관들은 미숙련자를 데이터 입력을 하는데 쓰고 있으며 입력을 하는 사람들을 편하게 하는 간단한 유저 인터페이스의 개발은 입력의 정확성을 높일 수 있다. 이러한 인터페이스는 권한 문항, 기존의 데이터, 관련 데이터베이스, 또는 외부 검색 엔진을 빠르게 검색해서 알맞은 철자법을 찾거나 문항에 적절한 데이터를 찾을 수 있다. 어떤 경우에는 데이터베이스 설계 과정에서 권한 파일이나 목록을 포함해서 이런 결정을 쉽게 할 수 있다.

지리 좌표 지정

정보를 전달하는데 있어서 지도가 가장 효율적인 수단 중의 하나이며 이 때문에 박물관과 식물 표본관의 표본 데이터의 데이터베이스화와 지리 좌표 지정 그리고 지리 좌표가 지정된 관찰 정보 기록의 증가가 정당화 된다. 지도는 오류와 확실하지 않은 부분을 연구, 감정, 시각화 그리고 문서화 하는 것을 도와준다(Spear et al 1996). 또한 데이터의 확실하지 않은 부분을 시각화하고 전달하는데 좋은 도구이며 사용자들에게 데이터의 품질과 적합성에 대한 정보를 전달 할 수 있다.

데이터를 전자 매체에 기록해서 지리 좌표를 지정하는 것은 시간이 많이 드는 어려운 작업일 수 있다. MaPSTeDI 프로젝트 (University of Colorado 2003)의 결과에 의하면 기록 하나에 지리 좌표를 배정하는 것은 평균 5 분이 걸린다. 다른 조사에 따르면 지리 좌표를 배정하는 것은 시간이 MaPSTeDI 프로젝트 보다 훨씬 오래 걸릴 수 있다 (Armstrong 1992, Wieczorek 2002). MANIS 데이터베이스는 미국의 경우에 시간당 9 개, 미국을 제외한 북미 대륙은 시간당 6 개, 그리고 북미 밖 지역의 기록은 시간당 3 개 걸린다(Wieczorek 2002).

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

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

지리 좌표 지정 과정을 돕는 몇 개의 우수한 기법과 지침들이 개발되었다. Museum of Vertebrate Zoology in Berkeley (Wieczorek 2001)의 John Wieczorek 이 개발한 Georeferencing Guidelines 과 University of Colorado 에서 개발한 MapSTeDI(Mountains and Plains Spatio-Temporal Database Informatics Initiative) 지침들이 가장 구체적이며 이들을 독자들에게 권장한다. 이들 지침들은 서술된 소재지 정보를 바탕으로 추리해 낸 지점의 정확성과 정밀도를 파악하는 방법, 다른 데이터 형식을 써서 생기는 불확정성, 다른 축적으로 사용해서 나타나는 문제 등을 다룬다. 지침들은 이런 문제들을 상세하게 다루며 독자들이 이 지침들을 이 문서의 부속 문서로 대우해주기 바란다.

There are also a number of on-line tools that can assist with the determination of geocodes – for example for places at a given distance and direction from a known locality. These will be covered in more detail in the associated document on *Principles and Methods of Data Cleaning*. 알려진 지점에서 거리와 방향을 지정하면 지리 좌표를 지정하는데 도움을 주는 온라인 도구들이 몇 개 존재한다. 이들은 *Principles and Methods of Data Cleaning* 에서 다룰 것이다.

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

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

오류

이전에 언급된 도구들은 오류를 줄이고 품질을 높이는데 쓸 수 있지만 오류를 완전히 제거하는 것은 불가능하다. MaPSTeDI 지침에 다음과 같은 내용이 있다:

> "지리 좌표를 지정하는 것은 정확하지 ㅇ낳고 모든 수집품들이 100% 정확하게 지리 좌표를 지정할 수 없지만 품질 관리에 신경 쓰는 것은 수집품 중에서 지리 좌표가 올바르게 지정될 확률을 높일 수 있을 것이다. 모든 프로젝트는 이러한 점을 염두에 두어야 할 것이다"(University of Colorado 2003).

지리 좌표를 지정하는 과정에서 오류가 많이 나는 원인으로 안내서를 검토 없이 이용하는 것이다. 몇몇 경우에 이들 안내서들은 일반적인 지도를 출판하는 과정 중에 만든 것이며 이름이 표시 되어야 할 위치에 해당 지점이 존재하기도 한다 (e.g. The Australian Gazetteer prior to 1998 developed by the Australian Land Information Group). 대부분의 안내서들은 수정이 되겠지만 박물관과 식물 표본관 데이터에 잘못된 안내서를 이용해서 지리 좌표가 지정된 데이터가 있을 수 있다. 정확한 지도를 이용하여 무작위로 이들 기록의 정확성을 검사 해야 할 것이다.

표지의 디지털화 이후에 지리 좌표를 따로 배정하는 것이 빠르고 효율적이다. 이렇게 함으로써 데이터베이스를 소재지, 수집가, 그리고 일자 등으로 수집품들을 정렬해서 지리 좌표 정보를 얻는데 더 효율적으로 쓸 수 있다. 또한 같은 지역의 기록을 여러 번 지리 좌표 배정하는 일을 막을 수 있다.

데이터의 문서화

> *"Metadata 는 데이터에 관한 데이터이다. 특정한 목적을 위해서 수집 된 데이터 특성의 표현이다"* (ANZLIC 1996a).

철저한 문서화는 데이터와 데이터 기록할 때 이루어진다.

Metadata 는 데이터의 내용, 범위, 접근성, 가치, 완성도, 그리고 적합성에 대한 정보를 제공한다. 제공이 되는 경우에 사용자는 데이터의 품질과 적합성에 대한 이해를 얻을 수 있다. 좋은 metadata 는 데이터의 교환, 검색, 그리고 회수를 용이하게 해준다. Metadata 는 보통 데이터 전체를 가리키지만 기록의 단계에서 데이터에 대한 문서화를 기록 수준의 metadata 로 간주하기도 한다. 명칭이 어떠하던 간에 전 데이터와 개별적인 기록 수준에서의 철저한 문서화는 중요하다.

모든 데이터에 필연적으로 오류가 포함 되어 있다. 오류의 정체를 밝히고 데이터의 용도에 허용 될 수 있는 오류인지 파악하는 것이 중요하다. 이런 경우에 metadata 가 중요해지며 용도 적합성이라는 개념이 metadata 와 관련이 되어 있을 때 중요하다. 용도 적합성이라는 개념은 90 년대 초반 전까지는 지리 정보에서 중요하게 여기지 않았으며 90 년대 중반이 되어서야 논문에 등장하기 시작했다(Agumya and Hunter 1996).

전체적인 데이터의 수준에서만 정보를 기록하는 것은 사용자에게 필요한 정보를 제공 할 것이라는 보장은 없다. 기록의 단위에서 오류를 기록하는 것은 데이터의 용도 적합성을 판단하는데 있어서 중요한 판단 기준이 될 수 있다. 이 정보가 제공 된다면 사용자는 기준을 설정해서 그 기준에 맞는 데이터만 요구 할 수 있다. 지리 좌표 지정을 자동적으로 하는 도구들이 계산된 정확성을 결과에 포함 시키는 것도 중요하다.

사용자들이 용도 적합성의 개념을 이해하는 것도 중요하다. 데이터베이스에서 종 발생 데이터를 추출 할 때 정확성에 대한 정보 없이 "record no., x, y" 형식으로만 나오는 경우가 많다. 좌표는 항상 지점으로 표시 되지만 실제로 한 지점을 가리키는 경우가 많지 않다. 몇몇 기록은 수집 지점이 임의적인 지점과 큰 오차 범위로 지정 되어 있다. 그래서 이 기록을 그대로 쓰는 것은 바람직하지 않다. 사용자들은 정확도 문항이 존재 한다는 사실을 인식하고 이를 이용하는 방법을 알아야 한다. 데이터 제공자들이 표준 데이터 보고서를 만든다면 정확도 문항이 필수적으로 포함 되어야 할 것이다.

The data must be documented with sufficient detailed metadata to enable its use by third parties without reference to the originator of the data.

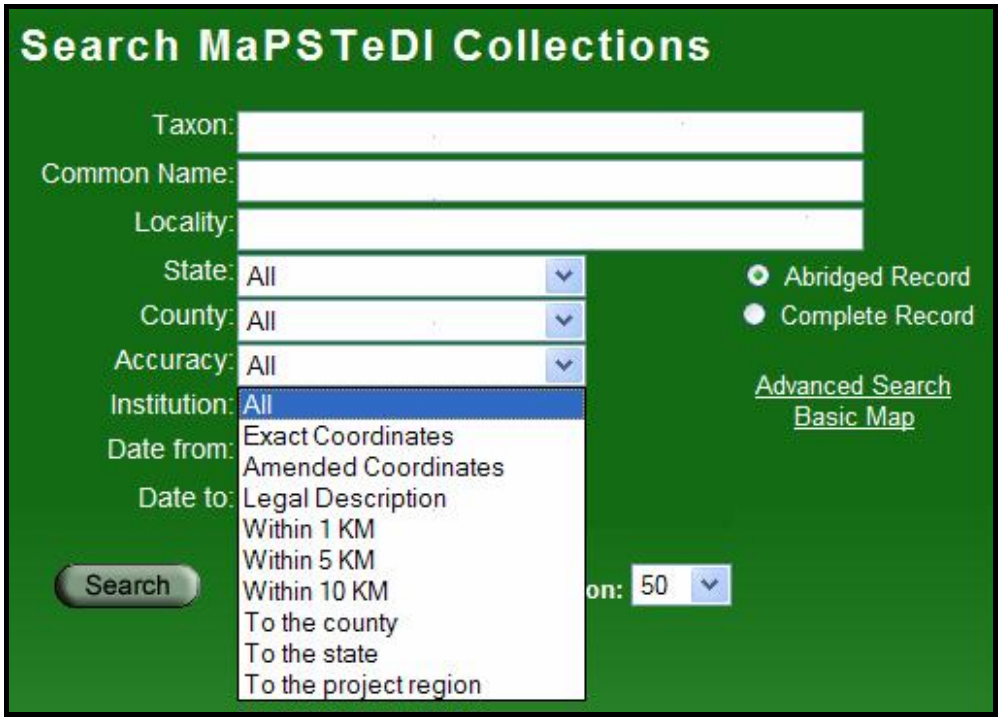

Fig. 6. *MaPSTeDI 검색 도구를 이용한 검색 예시* http://www.geomuse.org/mapstedi/client/textSearch.html. *예시에서 기록 단위의 문서화를 이용해 정해진 정확도를 갖는 데이터만 찾을 수 있도록 되어 있다.*

정확도, 정밀도, 그리고 오류를 문서화 하는 것은 지리 데이터의 용도 적합성을 따져야 하는 경우에는 필수적이다. 문서화를 한다면 다음 사항들이 필수적으로 포함 되어야 할 것이다:

- 데이터 세트의 제목
- 데이터 출처
- 데이터 처리 내역
- 정확도
- 논리적인 일관성
- 데이터의 예상 유효 기간
- 데이터 각 문항의 정의
- 수집 방법
- 완성도
- 조건과 제한
- 관리자 정보와 연락 방법

데이터 관리자들이 이 개념들을 잘 모를 수 있으니 이들을 위해서 정의 해주는 것이 좋다. 이 중에서 개별적인 기록 보다는 데이터베이스에 있는 데이터 모음을 가리키는 일이 많다.

지리적인 정확성

지리적인 정확성은 주어진 좌표가 실제 지점과 차이가 나는 정도를 가리킵니다 (Minnesota Planning 1999). 가능하다면 좌표를 산정하는데 이용한 Geodetic Datum 을 표기 해야 한다.

데이터베이스에 각 기록의 지리적인 정확성을 표시하는 문항을 포함 시키는 것이 좋다. 이에 방법이 몇 개 존재하는데 어떤 데이터베이스에서는 암호를 사용하지만 이를 수치로 나타내는 것이 좋다 (Chapman and Busby 1994, Conn 1996, 2000, Wieczorek *et al.* 2004). 이는

특정한 용도를 위해서 데이터를 추출하는 사용자들에게 중요할 수 있다 – 예를 들어서 오차가 2000m 미만인 데이터만 원할 수 있다. 또한 지리 좌표 지정 방법 문항을 추가 하는 것도 좋다. 예를 들어서;

- 위상차 GPS 이용
- 2002 년 이전에 휴대용 GPS 기기 이용
- 1: 100,000 축적의 지도와 삼각 측정 법 이용
- 현장에서 지도를 이용한 위치 추정
- 헬기나 다른 수단으로 지도를 이용한 위치 추정
- 지점 – 반경 기법을 이용하는 지리 좌표 지정 프로그램 이용
- 안내서 이용(안내서 이름, 일자, 그리고 버전 포함).

형질의 정확성

형질의 정확성은 데이터에 묘사 되어 있는 형질이 실제와 얼마나 가까운지를 나타낸다. 이상적인 경우라면 형질의 목록을 만들고 각각의 정확성에 대한 정보가 있어야 한다. 예를 들어서,

> 기록은 경험이 많은 관찰자들이 제공한다. 추가적인 정확성은 인증된 표본과의 형질 비교를 통해서 이루어진다. 식물 기록의 40%, 양서류는 51%, 포유류는 12%, 파충류는 18%, 그리고 조류의 1%가 인증된 표본과 비교된다. (SA Dept. Env. & Planning 2002).

계통

데이터의 계통은 데이터의 출처와 현재 상태가 되기 전까지 데이터를 처리한 방법을 가리킨다. 수집 방법이 포함 될 수 있고 데이터를 검사한 방법에 대한 정보도 포함 될 수 있다. 처리 과정에 포함 될 수 있는 정보의 예시는 다음과 같다:

- 데이터 기록 방법
- 중간 처리 과정이나 방법
- 최종 생산물을 만드는데 쓴 방법
- 데이터 검증 과정

예를 들어서;

> *데이터는 20m 격자 안에서 수집 되었다. 종 개수와 다른 생태 데이터도 수집 되었다.*
> *데이터는 Twinspan 을 이용하여 비슷한 종끼리 묶는 방법으로 분류가 되었다.*

논리적인 일관성

논리적인 일관성은 데이터 내부의 논리적인 관계를 보여준다. 박물관이나 식물 표본관에서 수집하는 데이터에 다음 항목들이 상관이 없을 수 있지만 관찰 데이터의 일부와 조사 데이터에는 해당이 될 수 있다. 디지털화 된 지리 데이터의 경우에는 논리 일관성 검사를 자동으로 할 수 있다. 논리 일관성 검사에 포함 될 수 있는 항목은 다음과 같다:

- 모든 지점, 선, 그리고 폴리곤이 표시 되어 있고 중복 되어 있는 표지가 있는가?
- 선들은 노드에서 만나거나 실수로 교차가 되는가?
- 폴리곤의 경계선이 닫혀 있는가?
- 모든 지점, 선, 그리고 폴리곤이 연관되어 있는가?

논리적인 일관성은 데이터 항목간에 논리적인 관계가 있는 경우에도 해당된다. 이런 경우에는 논리적인 관계를 검사 할 때 쓰인 검사에 대한 설명을 추가해야 한다. 데이터 항목간 논리적 일관성의 예는 다음과 같다 – 한 항목에서 프로젝트 수행 기간이 '가'와 '나'

사이인데 항목 기록 기간이 그 범위 밖에 있으면 논리적인 일관성이 무너진다. 이를 검사 하는 것을 문서화 하는 것도 중요한데 점-폴리곤 검사 등이 있으며 GIS 분야에서 쓰인다. *Principles and Methods of Data Cleaning* 에 추가적인 논의를 참고 하시기 바랍니다.

완성도

완성도는 데이터의 시공간 범위를 가리킨다. 완성도의 문서화는 품질을 파악하는데 있어서 필수적인 부분이다. 예를 들면;

> *30° S 북쪽에 대해서는 완성 되었으며 30°과 40° S 사이의 지역에 대해서는 산개 되어 있는 기록만 존재.*
>
> *데이터에 주로 New South Wales 에서 1995 년 이전에 기회가 닿는 대로 수집된 기록만 포함된다.*

사용자의 관점에서 본다면 완성도는 필요한 데이터와 관련이 있다(English 1999). 다시 말하면 사용자는 수행 하려는 분석에 필요한 문항을 데이터베이스가 가지고 있는지 알아야 하며 그 문항들의 완성도를 알 필요가 있다. 예를 들어서 사용자가 시간에 따라서 변하는 형질에 대한 분석을 하려고 하는데 데이터베이스가 어느 해까지만 데이터를 갖고 있다면 데이터는 분석에 부적합 할 수 있다.

접근성

데이터가 사용자에게 가치가 있으려면 접근성이 좋아야 한다. 모든 데이터가 온라인 상으로 접근 가능한 것이 아니며 접근을 하기 위해서 관리자와 연락을 해서 허가를 얻거나 CD 와 같은 매체로 받아야 한다. 접근 조건의 문서화는 사용자에게 중요하기 때문에 데이터 품질 요소 중 하나이다. 접근성에 대한 문서화에 다음과 같은 사항들이 포함 될 수 있다:

- 데이터 관리자의 연락처
- 접근 조건
- 온라인 상으로 접근 가능한 경우 접근 방법
- 데이터 형식
- 문제점
- 저작권 정보
- 비용
- 사용 제한

시간적 정확성

시간적인 정확성은 시간과 관련된 정보 요소들이 정확한 정도를 일컫는다. 예를 들어서 "월 단위까지만 데이터가 정확함"이라고 할 수 있다. 데이터베이스가 '일' 문항에 null 값을 인정하지 않거나 값이 없을 경우에 디폴트로 1 을 기재하는 경우에 중요하다. 이런 경우에 정확성에 대한 잘못된 인식을 가질 수 있다. 기록에 수집 년만 알려져 있는데 데이터베이스가 자동적으로 1 월 1 일이라고 기재하는 경우에 더욱 그렇다. 꽃의 개화 시기나 철새의 이동을 연구하는 사용자라면 자동 기재에 대한 정보를 알아야 그런 정보를 저품질의 데이터로 분류하여 분석에서 제외 할 수 있어야 한다.

인증 과정의 문서화

데이터에 존재하는 오류에 대해서 알고 싶을 때 문서화를 참고하는 것이 중요하다. 검사를 하고 오류를 수정한 다음에 이를 문서화 하지 않으면 검사를 하는 의미가 없다. 데이터 수집인 이외의 사람이 검사를 하는 경우에 더욱 그렇다. 오류인 것처럼 보이면서 실제로 오류가 아닌 경우가 있어서 이를 수정하는 경우에 새로운 오류가 만들어지는 수가 있다. 또한 검사를 반복하는 것은 자원 낭비일 뿐이다. 예를 들면, 사용자가 데이터 품질 검사를 해서 의심이 가는 기록을 찾을 수 있다. 이들 기록을 자세히 검사하면 멀쩡한 것으로 판정 날 수 있고 outlier 으로 판정 날 수도 있다. 이 정보가 기록되지 않으면 나중에 다른 사람이 같은 검사를 해서 같은 기록을 의심하게 되어서 분석에서 해당 기록들을 제외하거나 소중한 시간을 써서 정보를 자세하게 검사 할 수 있다. 이 것은 기본적인 위험 관리의 일부이며 데이터 관리자와 사용자들은 이를 수시로 이행할 의무를 가지고 있다. 철저한 문서화의 필요성에 대해서 강조하는 것은 아무리 해도 부족하지 않다. 문서화는 사용자들이 데이터의 정체성, 품질, 그리고 잠재적인 용도에 대해서 알려준다. 또한 데이터 관리자들이 데이터 관리 하는 것을 도와준다.

문서화와 데이터베이스 설계

오류의 철저한 문서화를 보장하는 방법 중의 하나는 데이터베이스 설계와 구축 초기 과정에 이를 포함하는 것이다. 설계와 구축 과정에 이를 고려하면 품질과 정확도 문항을 나중에 추가 할 수 있다. 위치나 지리 좌표 정확성, 정보 출처, 기록 입력 담당자와 방법과 같은 문항이 이에 포함된다. 이 정보는 나중에 데이터가 특정 목적에 알맞은 것인가에 대한 물음에 답할 수 있다.

> "사용자들은 정확도나 오류에 대한 문서화가 없는 데이터를 이용할 때 그 데이터에 기반한 분석에 대해서 조심을 해야 한다." (Stribling *et al.* 2003).

Acknowledgements

Many colleagues and organisations around the world have contributed to this paper in one way or another. Some directly, some by being involved in discussions with the author over a period of more than 30 years and some indirectly through published papers or just by making their information available to the world.

In particular, I would like to particularly make mention of the staff, both past and present, of CRIA (Centro de Referência em Informação Ambiental) in Campinas, Brazil and ERIN (Environmental Resources Information Network) in Canberra, Australia who have contributed ideas, tools, theories and a sounding board that have helped the author in formulating his ideas. Their discussion of error and accuracy in environmental information over the years and the pioneering work done by them, by CONABIO in Mexico, the University of Kansas, CSIRO in Australia, the University of Colorado, the Peabody Museum in Connecticut, and the University of California in Berkeley, as well as others too numerous to mention, has helped bring us to the stage we are today in species data quality management. I thank them for their ideas and constructive criticism. In addition, discussions with Town Peterson and others at the University of Kansas, Barry Chernoff at the Wesleyan University in Connecticut, Read Beaman at Yale University, John Wieczorek and Robert Hijmans at the University of California, Berkeley, Peter Shalk and others at ETI, in Amsterdam, Stan Blum at the Californian Academy and the staff of GBIF in Copenhagen have presented me with ideas and challengers that have led to some of the ideas expressed in this paper. Any errors, omissions or controversies are, however, the responsibility of the author.

I would like to also thank those who have supplied criticisms, comments and suggestions during the editing of this document, and in particular the following members of the GBIF Subcommittee for Digitisation of Natural History Collection Data: 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.

Larry Speers of GBIF was instrumental in the commissioning of the report, and in shepherding it through all its stages.

In conclusion I would like to thank the FAPESP/Biota project in Brazil with providing me with the opportunity and support to expand my ideas on data quality management during my stay in Brazil in 2003-2004 and the GBIF organisation for supporting and encouraging the production of this report.

References

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.

Index