생태관측 데이터 특성 분석 및 저장소 설계 기술 보고서

한국과학기술정보연구원

www.kisti.re.kr

저자소개

변은규

한국과학기술정보연구원/선임연구원
Email : ekbyun@kisti.re.kr

김직수

한국과학기술정보연구원/선임연구원
Email : jiksoo.kim@kisti.re.kr

유진승

한국과학기술정보연구원/선임연구원
Email : jsu@kisti.re.kr

박근철

한국과학기술정보연구원/선임연구원
Email : gcpark@kisti.re.kr

황순욱

한국과학기술정보연구원/책임연구원
Email : hwang@kisti.re.kr

목 차

I. 생태관측 데이터 서비스 및 저장소 개요 ····· 3
1. 생태관측 데이터 서비스 ····· 3
2. 생태관측 서비스 ····· 3
3. 생태관측 데이터 저장소 ····· 3
II. 생태관측 데이터 서비스 동향 분석 ····· 6
1. ECN(Environmental Change Network, 영국) ····· 6
2. TERN(Terrestrial Ecosystem Research Network, 호주) ····· 8
3. NEON(National Ecological Observatory Network, 미국) ····· 13
4. 장기생태연구(한국) ····· 16
5. 기타 참고 사이트 ····· 19
III. 데이터 저장소 기술 동향 분석 ····· 21
1. 파일시스템 ····· 21
2. 관계형 데이터베이스 관리 시스템(RDBMS) ····· 22
3. 비정형 데이터베이스 시스템 ····· 24
4. 네트워크 연결 스토리지 ····· 26
5. 데이터 보관 및 백업 장치 ····· 27
IV. 생태관측 데이터 특성 분석 ····· 31
1. 생태관측 데이터의 종류 및 특성 ····· 31
2. 생태관측 데이터의 분류 기준 ····· 35
3. 요약 ····· 36
V. 생태관측 데이터 저장소 설계 ····· 38
1. 서비스 정의 ····· 38
2. 사용자 정의 ····· 39
3. 주요 제공 기능 ····· 39
4. 전체 시스템 구조 ····· 40
VI. 향후계획 및 기대효과 ····· 42
1. 향후계획 ····· 42
2. 기대효과 ····· 42
II. 참고 문헌 ····· 44

I. 생태관측 데이터 서비스 및 저장소 개요

I. 생태관측 데이터 서비스 및 저장소 개요

1. 생태관측 데이터 서비스

생태관측 데이터 서비스는 국가적 자산에 해당되는 생태분야의 관측데이터를 장기적으로 보관이 가능해야 하며, 발견되는 생태종의 증가에 따라 유연하게 대처하기 위한 플랫폼이 요구된다. 또한, 다양한 생태관련 분석 툴과의 연동을 위한 접근성과 글로벌 생태 데이터와의 비교 및 분석 연구를 위해 다수의 국가에서 공식적으로 인정받아 사용되는 표준 메타데이터 스키마 및 전송 프로토콜을 수용해야 한다.

생태관측 자료는 관측분야, 관측요소가 다양하고 관측 기관별로 다양한 종류의 원시 데이터를 수집한다. 생태 관측 자료에 대한 메타데이터 스키마를 자체적으로 정의하여 오라클과 같은 관계형 데이터베이스관리시스템(RDBMS)에 메타데이터시스템을 구축한다. 앞서 언급한 다수의 국가에서 공동 활용하기 위한 전송 프로토콜을 위해 생태분야에서도 EML(Ecological Metadata Language)이라는 메타데이터 표준 언어가 정의되어 있어 생태학자들 간에 서로의 데이터를 공유할 수 있는 기반이 마련되어 있다. 특히, 미국의 NEON(National Ecological Observatory Network)은 기본적으로 EML을 기반으로 하여서 NEON의 요구사항을 반영한 EML을 확장한 메타데이터 시스템을 구축함으로써 NSF에서 구축하는 또 다른 생태계정보 인프라인 LTER(Long Term Ecological Research Network)과의 생태관측데이터 공동 연계 활용을 모색하고 있다. 이에 우리나라도 EML 기반의 메타데이터 시스템을 구축하여 NEON, LTER 등과 연계한 글로벌 생태관측 데이터 공동 활용을 위한 전송 프로토콜을 고려한 데이터 관리가 필요하다.

2. 생태관측 서비스

생태관측 서비스는 생태관측 데이터를 통합관리하고 공유하기 위한 웹기반의 생태/환경 분야 e-science 연구 인프라 서비스라 말할 수 있다. 표준화된 메타데이터를 기반으로 연구자들을 포함한 사용자에게 생태계 정보 접근을 위한 웹기반의 통합 인터페이스를 제공하고, HTML5, 메쉬업, GIS 등, 최신의 웹 기술을 접목하여 사용자 친화적이고 유연한 사용자 연구 환경을 웹으로 제공한다.

3. 생태관측 데이터 저장소

수집된 원시 생태관측 데이터 및 가공된 데이터, 분석된 데이터 등 다양하고 용량이 큰

데이터를 저장하기 위한 공간이 반드시 필요하다. 생태관측 데이터 서비스를 제대로 제공하기 위해서는 데이터의 특성을 파악하고 그에 맞는 저장소를 설계하는 작업이 필요하다. 특히 원시데이터자체의 보관 뿐 아니라, 데이터 마이닝을 통한 유효한 정보의 추출, 통계 정보 생성, 해외 정보와의 교류, 메타데이터 카탈로그 시스템과의 연동, 웹 포탈에서의 접근의 용이성 등의 다양한 요소를 고려해야 한다. 생태관측 데이터는 수집 방법, 종류, 형식, 크기, 저장 방법, 보관 기간, 재가공 여부/방법 등이 다양하다. 다양한 데이터 저장소 기술을 파악하고 생태관측 데이터 각각에 맞는 저장 방법을 찾아 이들을 통합한 저장소 시스템의 큰 그림을 그리는 것이 중요하다.

II. 생태관측 데이터 서비스 동향 분석

II. 생태관측 데이터 서비스 동향 분석

본 챕터에서는 장기적인 생태조사 및 연구를 지원하는 국내외 사이트에 대한 동향을 분석하고 해당 시스템에서의 데이터 저장소 구조를 설명하고자 한다. 세계 여러 나라에서는 일찍부터 환경변화에 따른 생태계의 움직임에 관심을 보여 왔다. 그러한 데이터를 바탕으로 환경지표종을 개발하고, 환경변화에 대응하는 정책에 활용하고 있다.

1. ECN(Environmental Change Network, 영국)

영국 ECN(Environmental Change Network)는 생태관측 데이터 서비스 중에서 가장 오랫동안 데이터를 모아온 서비스로, 영국의 장기적인 환경 관측 및 연구를 위해 1992년부터 정기적으로 동식물군과 그들의 물리화학적 생태 환경에 대한 관측을 해오고 있다. 이렇게 관측된 데이터들은 기후변화, 공기오염 등을 포함한 환경적인 변화가 영국의 생태계에 미치는 영향을 분석하는데 활용되고 있다. 관측된 데이터들은 ECN Data Centre(http://data.ecn.ac.uk/)로 보내져서 검증 절차를 거친 후 데이터베이스에 저장되어 연구, 교육 및 비상업적인 목적을 위해 활용된다.

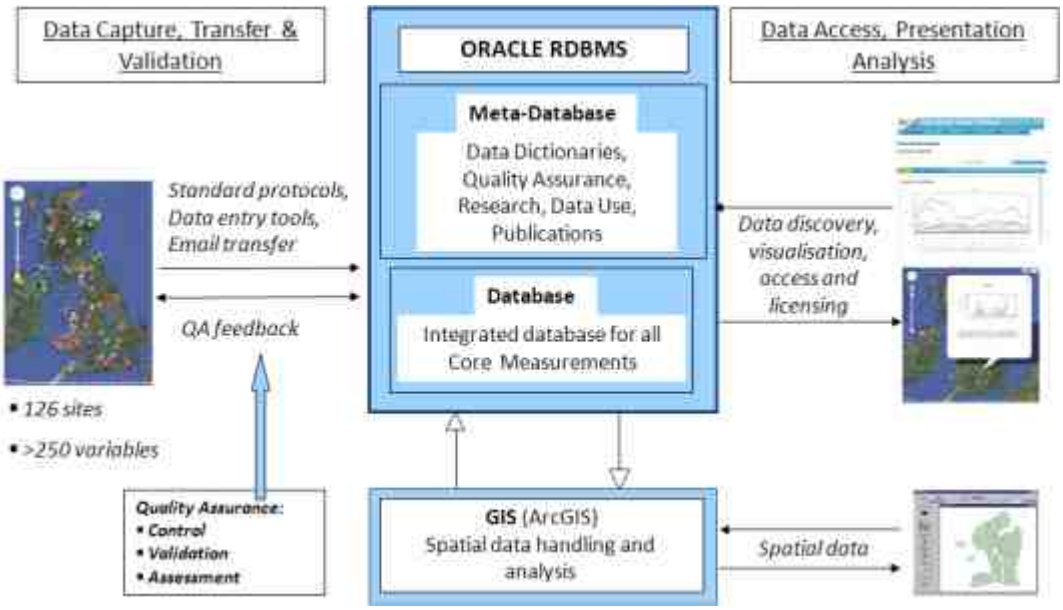


그림 3 ECN Data Management Architecture

ECN은 위의 구조도에서 보는 것과 같이 수집된 원시 데이터(Raw Data)와 이에 관련된 메타 데이터(Meta-data)를 가장 널리 쓰이는 상용 관계형 데이터베이스 시스템인

Oracle DB에 저장하고 있다. 또한, 공간 데이터 처리를 위해 Arc GIS를 활용하고 있다. 모든 데이터는 중앙 데이터베이스에서 관리되며 고속의 네트워크를 통해서 접근이 가능하다. ECN의 데이터베이스는 논리적으로 데이터 테이블과 메타 데이터 테이블로 분리되어 있으며, 데이터는 매달, 매분기, 년별로 요약된 버전(summary database)을 제공하고 있음. 이러한 요약된 데이터에 대해서는 무료로 열람이 가능하며, 원 데이터에 대해서는 특정 절차를 거쳐야 검색 및 열람이 가능하게 되어 있다. ECN에서는 데이터의 요약된 통계본의 가시화를 위해 ECN 데이터의 변수간의 상관관계를 표시할 수 있는 ECN Data Explorer와 시간의 흐름에 따른 데이터 변동 추이를 나타낼 수 있는 Time Series Viewer를 제공하고 있다.

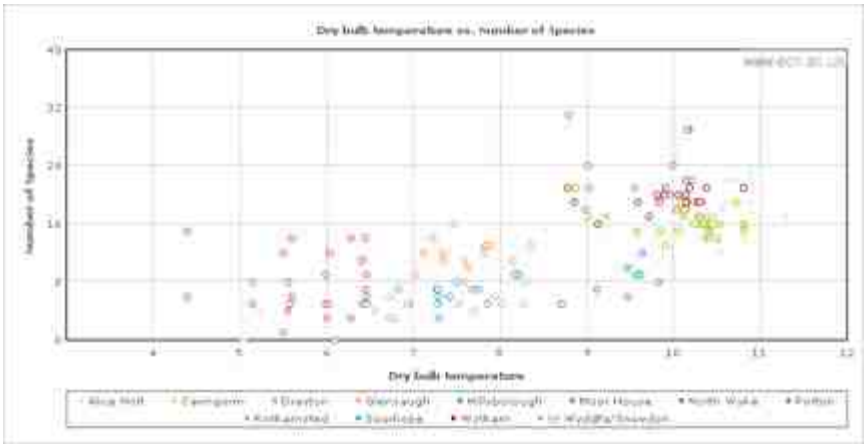


그림 4 ECN Data Explorer

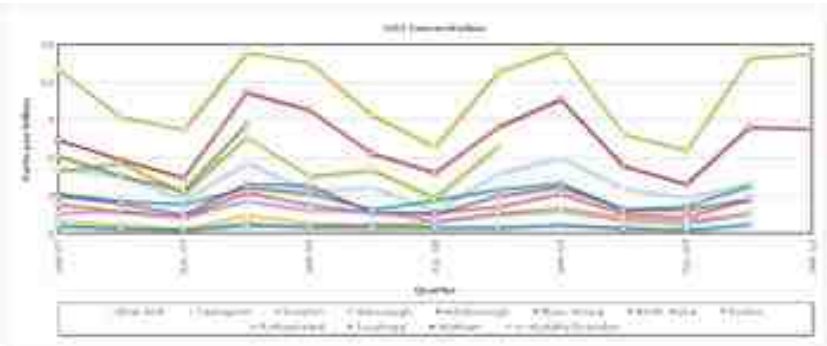


그림 5 Time Series Viewer

2. TERN(Terrestrial Ecosystem Research Network, 호주)

호주의 TERN은 협력 생태계 과학 커뮤니티를 지원하기 위한 인프라스트럭처이자 네트워크이다. 이는 장기 관측과 표준 측정 및 프로토콜을 따르며, 데이터 품질을 증진하고, 데이터 접근을 용이하도록 인터페이스를 제공한다. 다양한 사이트 네트워크로 구성된 데이터를 기반으로 연구 플랫폼을 구축하여 호주 전체 생태계 연구를 지원한다. 이는 생태 커뮤니티가 다양한 특징을 가지는 사이트에 접근하여 데이터를 조작할 수 있도록 한다. TERN은 2010년부터 호주 정부의 예산으로 진행되는 프로젝트로 AEKOS, ACEF, Soil and Land scape Grid, AusCover, OxFlux와 같은 주요 생태 사이트로 구성된다.

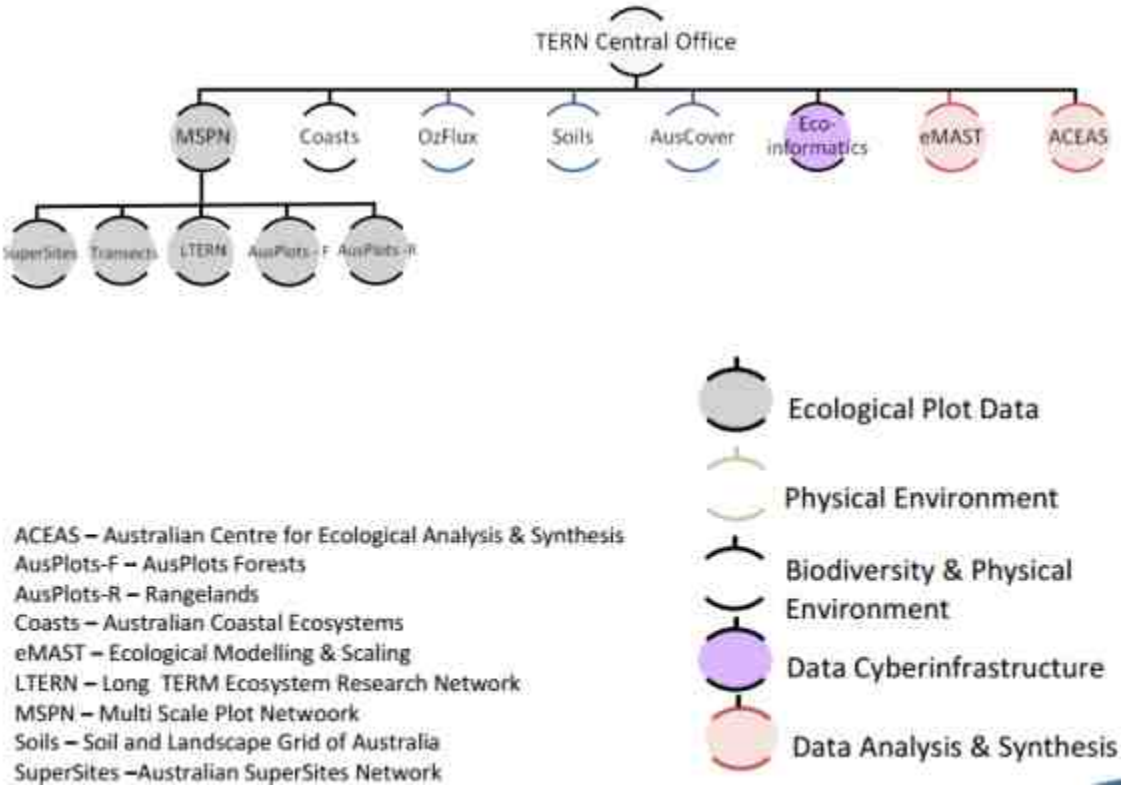


그림 7 .TERN 참여 Site들의 조직 구성

이중에는 AEKOS Facility의 Eco-Informatic와 같이 관측장비가 없는 사이버 인프라 Facility도 있고, eMAST, ACEAS와 같이 데이터 분석 및 통합을 하는 Facility도 있

다. TERN의 관측 Facility의 성격에 따라 사이트가 나뉘어져 있다. 관측 데이터는 생태, 환경, 기후, 인간 및 에너지에 걸쳐 다양한 데이터로 구성이 되며, 세부 데이터 셋은 아래 표와 같다.

생태	Plants/Animals, Vegeteation, Terrestrial ecosystem, Ecological dynamics
환경	Freshwater, Land surface and Soils, Oceans & Coasts, Water & gas
기후	Climate
인간	Agriculture, Human-nature
에너지	Energy

각 관측 Facility에서 수집된 데이터는 각각의 QA를 통해, 메타데이터와 함께 Data Portal에 모여지고, 사용자에게 검색 서비스를 통해 제공된다.

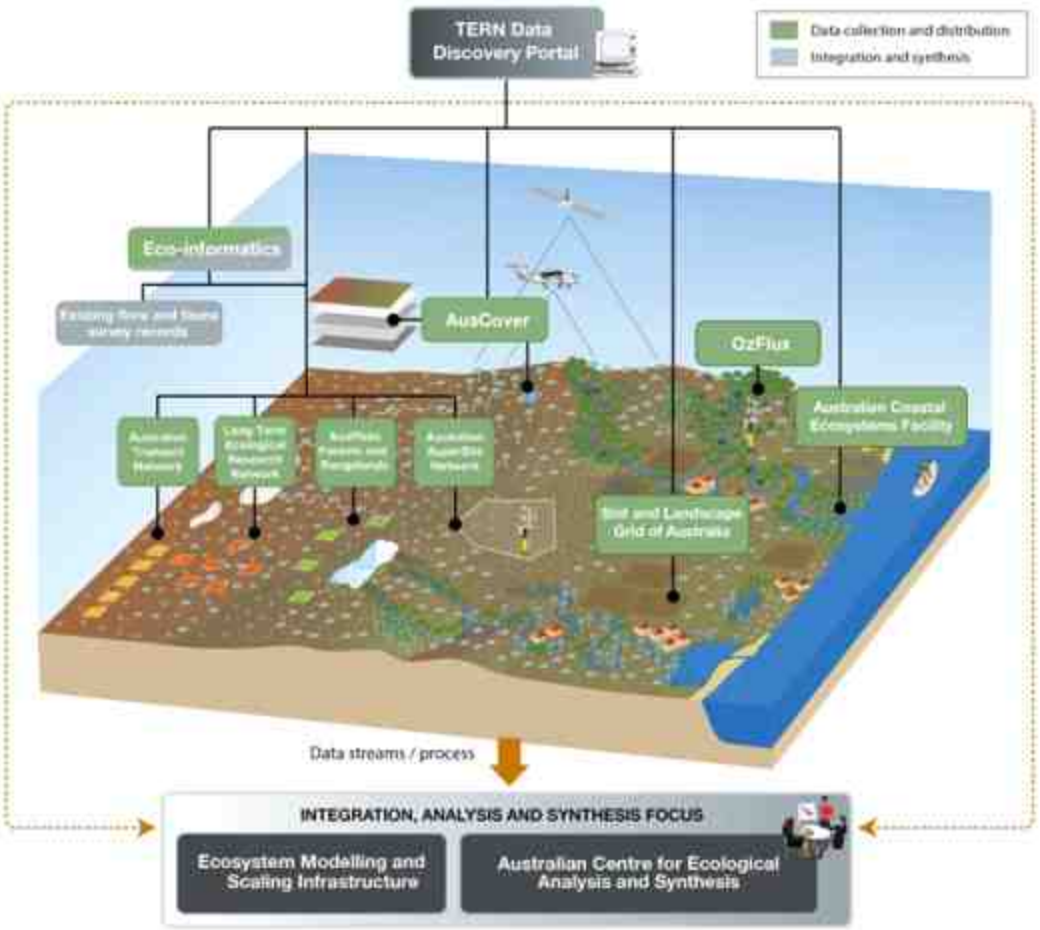


그림 8 TERN의 데이터 수집 체계

그림 9 TERN Data Discovery Portal 웹 페이지

TERN에서 정의되는 데이터셋은 저널 논문과 같이 학문적으로 인용 가능하도록 출판되어 데이터를 공유하거나 재사용하는데 활용된다. 이를 위해 DOI(Digital Object Identifier), 국제적 ISO 표준을 도입하고 있다. 호주는 국제 DataCite(데이터 인용) 컨

소시움의 회원으로, 호주 ANDS(The Australian National Data Service)에서 DOI 핸들링 서비스를 운영하고 있다. TERN의 DOI-minting 서비스는 이를 활용하여 생성되어지는 데이터셋에 DOI를 부여하여 관리한다.

TERN에서 제시하는 메타데이터의 항목은 다소 제한적인 면이 있는데, 이는 ANDS에서 공동활용하기 위해 관리되는 메타데이터 항목을 일부 따르고 있으며, 이에 대한 확장된 메타데이터를 TERN에서 운용한다. 메타데이터 항목의 예로, Data set Citation, Data set Owner(s), Associated Party, Abstract, Keywords, Geographic Coverage, Taxonomic Coverage, Contact(s), Methods Info, Access, Project Info, Additional Metadata 등으로 메타데이터가 구성된다.

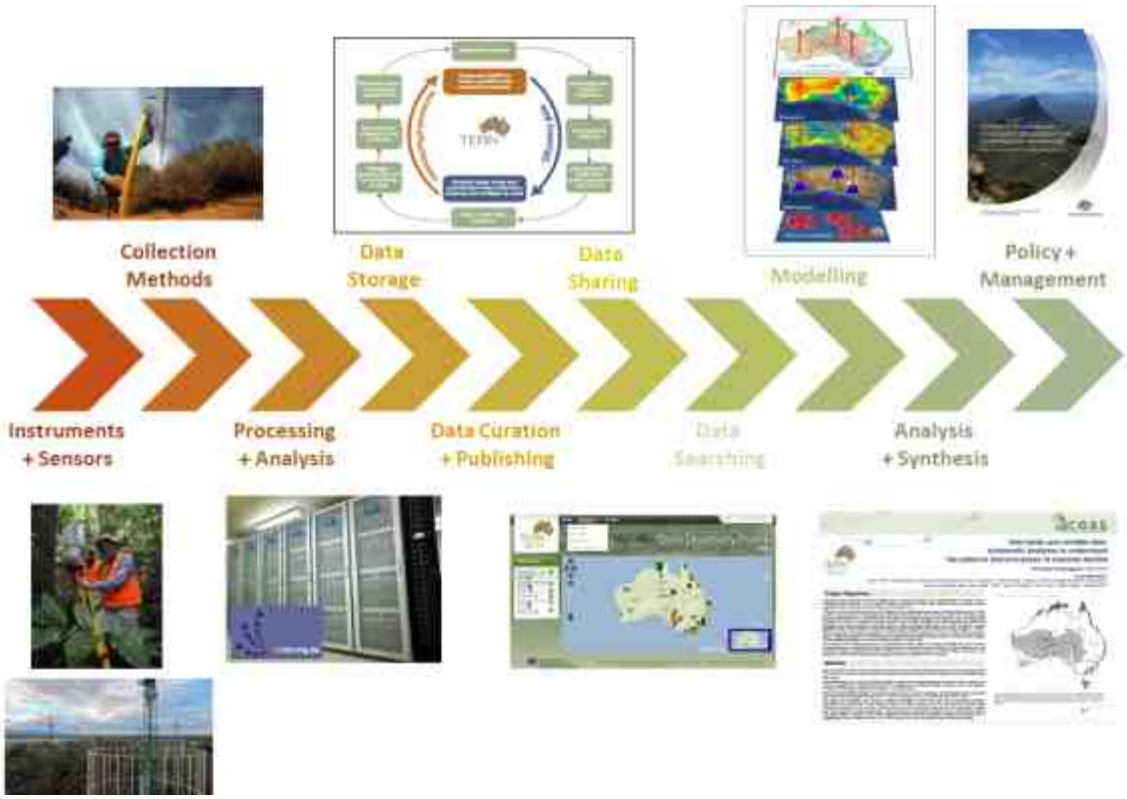


그림 10 TERN의 데이터 활용 Life-cycle

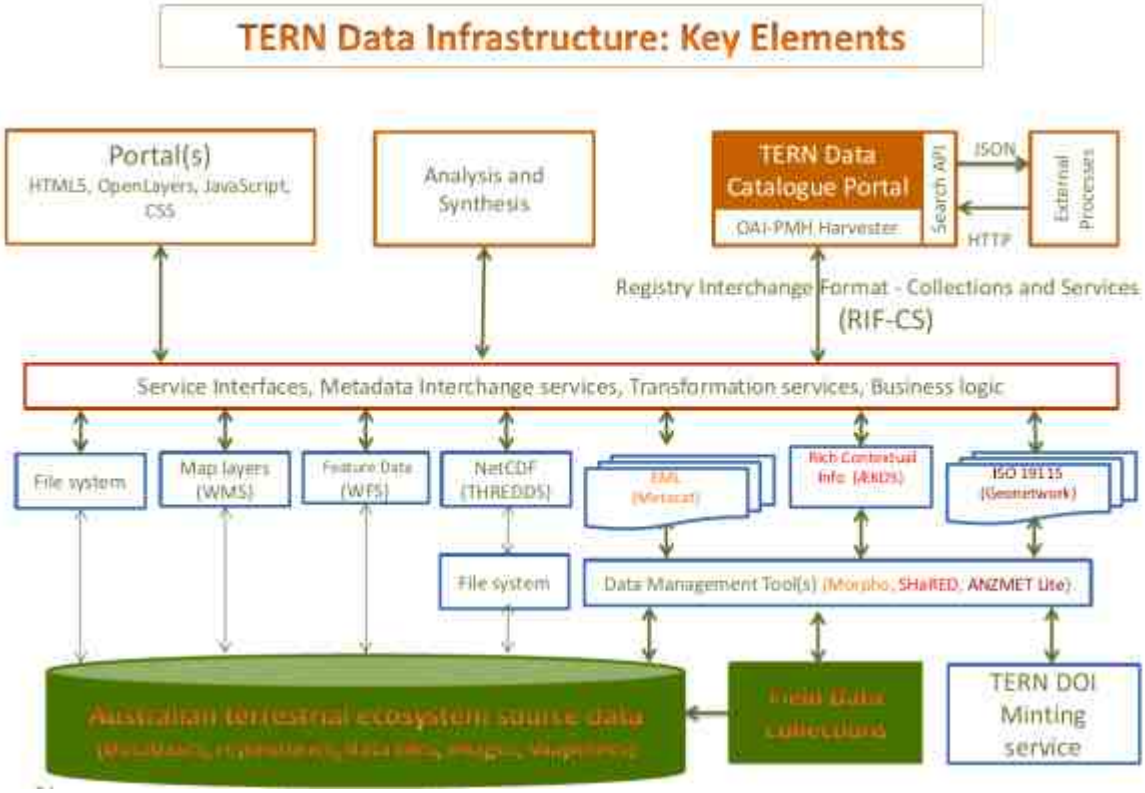


그림 11 TERN 데이터 저장소 구조

3. NEON(National Ecological Observatory Network, 미국)

미국 NEON(National Ecological Observatory Network)은 미국 대륙을 대상으로 기후변화, 토지이용변화, 외래종의 침입 등이 생태계에 미치는 영향을 파악하고 예측하기 위한 생태계 정보 네트워크 인프라(Information & Physical Infrastructure)를 제공하는 것이 목적으로 한다.

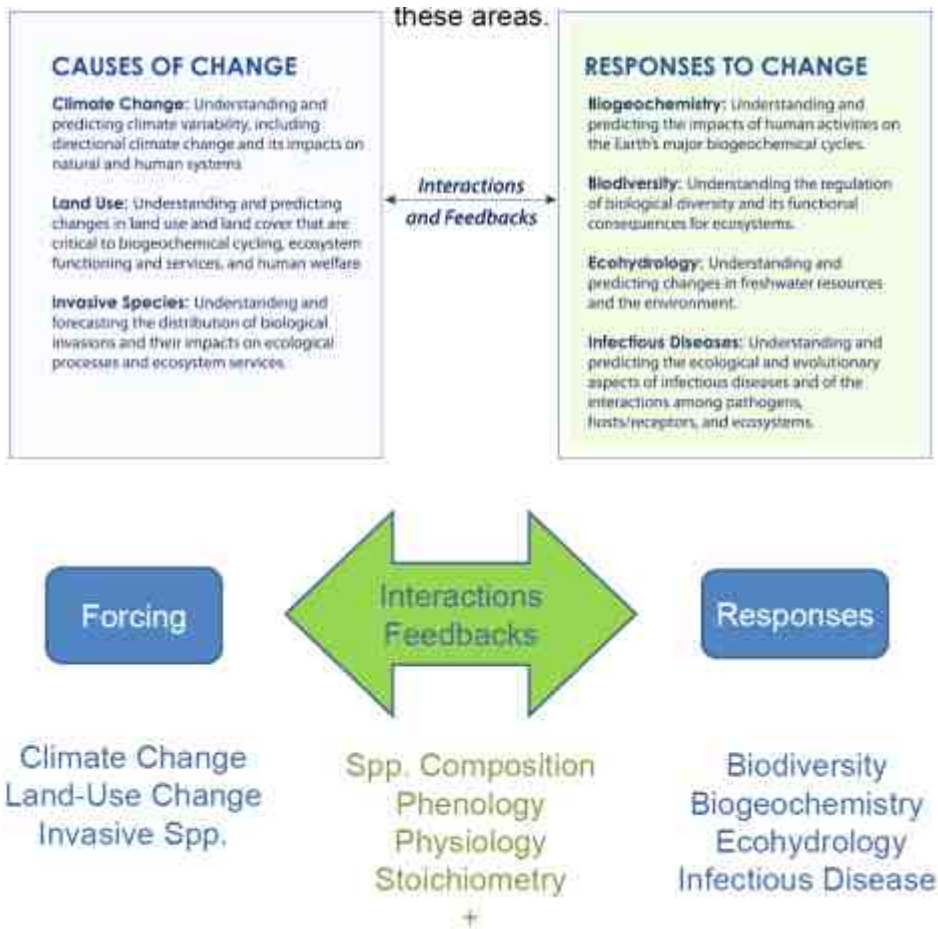


그림 13 NEON의 목적 및 역할

생태학적 데이터와 정보의 공개를 통해 과학자, 교육자, 및 정책 결정자들에게 생태계에 대한 인류의 영향을 파악하고 예측함으로써 생태학적으로 중요한 이슈들을 해결할 수 있는 토대를 마련하고자 한다. 또한 NEON Cyber Infrastructure는 이러한 NEON의 목적을 달성하기 위한 생태 데이터 수집(Production), 가공(Analysis) 및 가시화(Visualization)를 위한 플랫폼을 제공하는 것이 목표이다. 전체적인 구조는 사람에 의

해 수집되고 관측된 데이터를 처리하기 위한 Observation and Sample Management(OSM), 수집된 데이터를 가공하기 위한 모듈간의 연계 및 통신을 위한 Message Handling System(CMES), 데이터에 대한 QA(Quality Assurance) 및 운영을 위한 Operations Support System(OSS), 다양한 알고리즘을 활용하여 데이터를 가공하는 Data Processing Management System(DPMS), 가공된 데이터 및 이에 대한 메타데이터를 관리하는 Common Data Services(CDS), 과학자/교육자들에게 데이터를 공유하고 검색할 수 있도록 제공하는 Science & Education Portal로 구성되어 있다. 이 중 70%이상을 COTS/Open Source S/W로 구현하려는 계획을 가지고 있다.

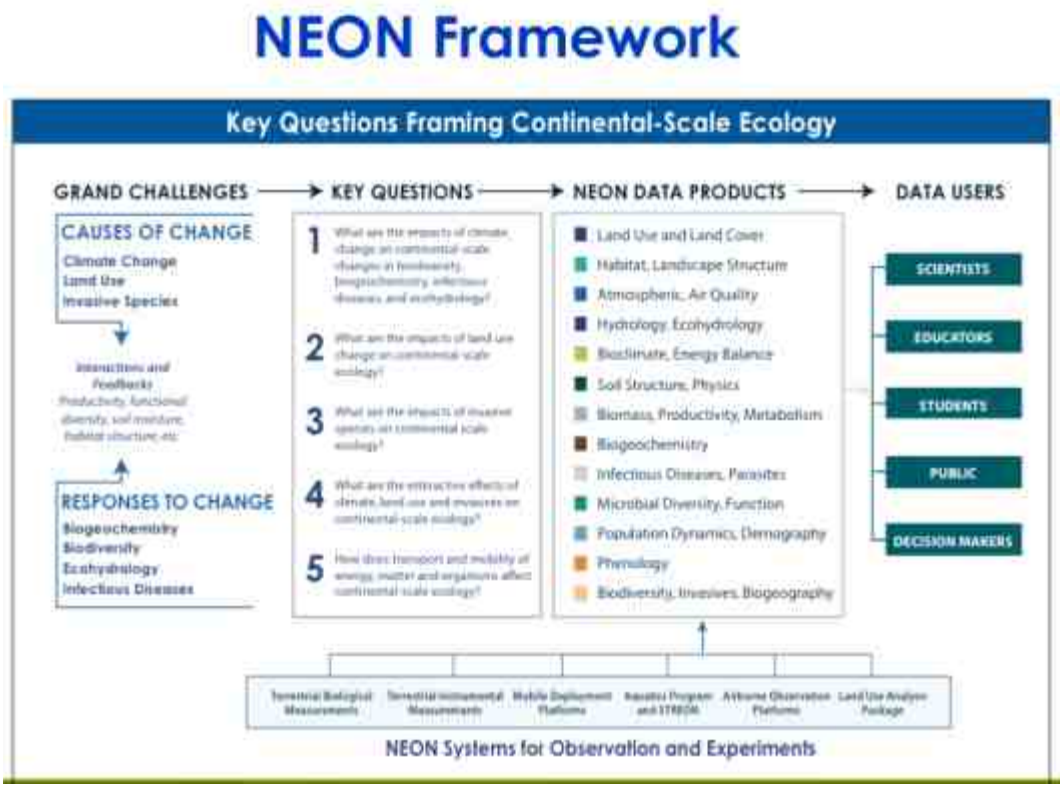


그림 14 NEON Framework

이 중 CDS가 데이터 저장소의 역할을 하고 있다. 아래 그림의 NEON Cyber Infrastructure 구조도를 보면 원시데이터 저장공간 뿐 아니라, 샘플데이터, 보정된 데이터 모두를 저장하기 위한 NEON Data Center(Boulder)가 존대하며 데이터의 장기적인 보관을 위한 아카이빙 장치도 설계에 포함되어 있다. 또한 메타데이터를 위한 저장소를 따로 두어 CDS에서 함께 관리하는 구조를 가지고 있다.

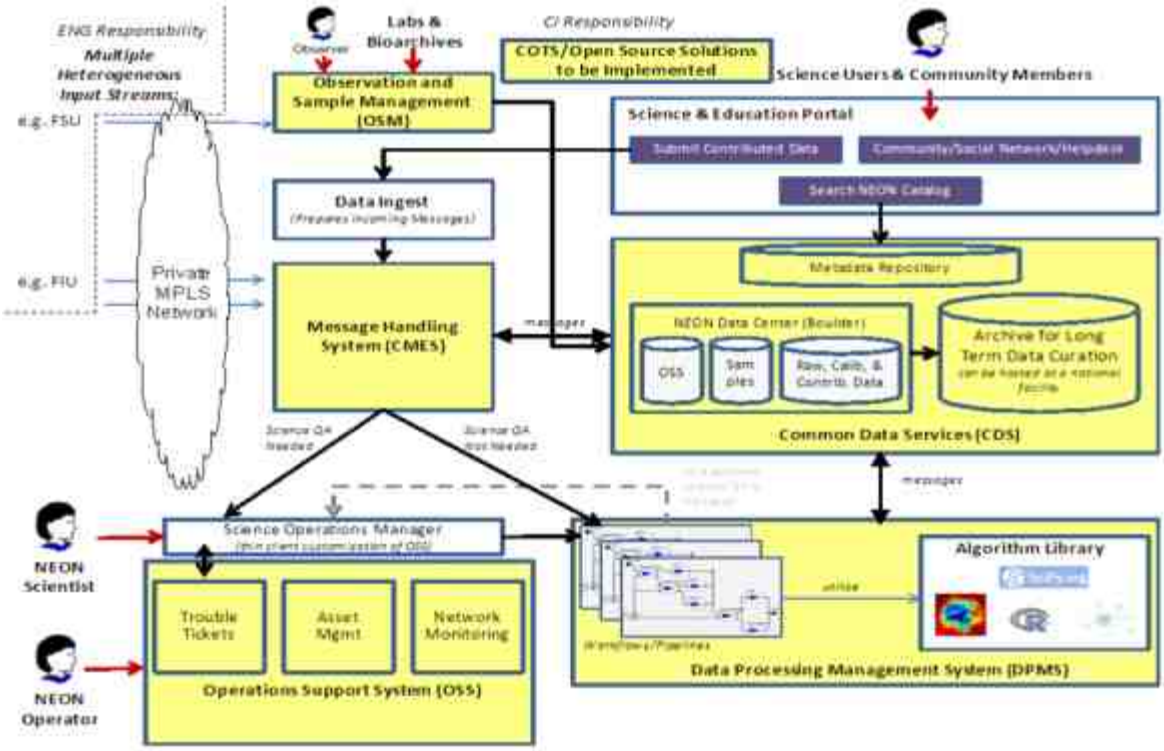


그림 15 NEON Cyber Infrastructure

NEON 프로젝트는 2004년에서 2011년에 걸치는 개념 정립 및 설계를 거쳐 2012년부터 2017년까지 미국 전역에 주요 관측소들을 설립할 예정이다.. 이후 2017년부터 2046년까지 생태 데이터를 수집 및 가공할 예정인데, 이러한 일정에 맞추어서 Cyber Infrastructure도 같이 준비해나가는 것으로 보이나, 아직까지 구체적인 설계와 사용되어질 기술 등이 정해지지 않은 것으로 보인다. NEON의 데이터 수집 시스템은 방대한 넓이의 미국 대륙을 대상으로 생태 관측 정보를 수집해야한다. 종전의 관측원이 직접 관측하는 방법과 또는 관측장비를 활용하는 방법을 모두 수용하는 FSU(Fundamental Sentinel Unit) 뿐만 아니라, 타워 형태의 대형 관측시설을 도메인마다 설치하여 장기적인 관측을 시도하고 있다. 고정된 관측시설을 보완하기 위해 MRP(Mobile Relocatable Platform)인 RTS(Relocatable Tower System)도 활용하여 추가적 관측 지역을 커버하기도 한다.

4. 장기생태연구(한국)

장기생태연구는 2004년부터 연구기반 및 시설 설치를 시작으로 조사지점 기초조사를 시작했다. 2007년부터는 사업의 정착 및 지소를 확립하고 생태분류군에 대한 심화연구를 수행하였고, 2010년부터 사업 종료, 2013년까지 생태계 장기변화 예측과 한반도 생태계 보전 대책을 수립함으로써 장기생태연구에 기여하였다.

기후 변화는 한반도 환경변화의 주요 원인 중 하나로, 자연생태계 교관 및 생물다양성의 변화가 우려되어진다. 상세하고, 과학적인 미래 기후변화 시나리오를 바탕으로 미래 기후를 전망하는 것이 생태계변화 관리체계를 확립하는 큰 요소가 될 수 있다.

기후변화에 따른 환경변화를 모니터링하고 그 결과를 추적하여 한반도에서 현재와 미래에 일어날 수 있는 환경문제에 지속적인 대응 방안을 마련하는 것이 장기생태연구의 가장 큰 목적이라 할 수 있다. 이를 위해 국가장기생태 연구에서는 생태계 변화에 대한 모니터링 자료를 지속적으로 수집할 수 있는 시스템 구축을 했다. 데이터베이스 구축은 크게 육상생태계, 담수생태계, 연안생태계, 동물생테계로 4개 대분야 80여개 이상의 지소별로 이뤄졌고, 지소별 데이터베이스 구축은 같은 분야에 대해 같은 항목으로 수집을 하여, 지역별 동일한 장소와 내용에 대해 자료의 비교분석을 가능하게 하였다.

장기생태연구는 GIS-DB를 통해 연구성과의 홍보, 자료축적을 통한 DB구축을 하였고, 연구자들 뿐만 아니라 대국민도 시각적으로 쉽게 이해할 수 있도록 개발되었다.

ILTER와 차별점은 ILTER는 전문가 집단을 위주로 운영되어, 전문가들이 개발한 Morpho, Kepler, EML을 등을 이용하기에 프로그램 다운로드, 설치 및 사용방법에 대한 전문성이 요구된다. 반면에 장기생태연구 사이트는 DB를 기반으로 PHP를 이용한 개발로 이뤄졌으며, 생태정보와 함께 공간정보의 활용으로 웹상에서 쉽게 테이블 리스트, 그래프, 통계, 메타데이터를 지도와 병행하여 결과를 볼 수 있도록 되어 있다.

국가장기생태연구 메인 페이지

조사지역 메타데이터

조사지역 연도별 통계

장기생태연구 사이트는 국내에서 처음으로 생태정보시스템 구축의 장기적 모델을 제시했다는 점에서 그 의의가 크다. 하지만, 10년동안 제한적 DB 구축의 문제점으로 장기생태연구가 국내에 국한 되는 문제점이 있었다. 장기생태연구는 ILTER라는 국제 기구를 중심으로 각국의 데이터를 서로 교환하고 비교 분석할 수 있는 체제를 갖추고 있지만, 장기생태연구 사이트는 국제표준 메타데이터인 EML의 한국어 버전이 존재치 않아 국제적 공유의 단절이 있다. 또한 개별 연구거점들에서 생산된 정보를 관리하는 책임자의 부재와 국가차원에서 종합 관리하는 관리자가 없는 것도 그들의 보고서에서 문제점으로 언급되었다.

5. 기타 참고 사이트

"http://www.ecosystem.nier.go.kr" 국가생태계정보네트워크

"http://www.alienplant.nier.go.kr" 한국의외래식물종합검색시스템

"http://www.naris.go.kr" 국가자연사연구종합정보시스템

"http://www.nature.go.kr" 국가생물종지식정보시스템

"http://www.kbif.re.kr/chm2" 한국생물다양성정보기구

"http://www.biodiv.ccbb.re.kr" 생물자원정보

"http://www.bioall.org" 국가생물자원정보포털사이트

"http://www.bric.postech.ac.kr" 생물학연구정보센터

III. 데이터 저장소 기술 동향 분석

III. 데이터 저장소 기술 동향 분석

1. 파일시스템

파일이란 이름을 가진 저장된 데이터의 집합을 말한다. IT 시스템에서 데이터를 보관하고 사용하는 가장 보편적인 형태를 말한다. 파일시스템은 계층구조를 통해 파일을 관라하는 기능을 제공한다. 주요 기능으로는 파일의 이름을 부여하여 새로운 파일을 생성하는 기능, 파일에 접근할 수 있는 권한을 설정하고 변경하는 기능, 파일을 읽거나 수정하는 기능, 파일의 생성일시/수정일시 등의 속성을 관리 하는 기능, 데이터의 안정성을 위해 동시 접근을 관리하는 기능 등이 포함된다.

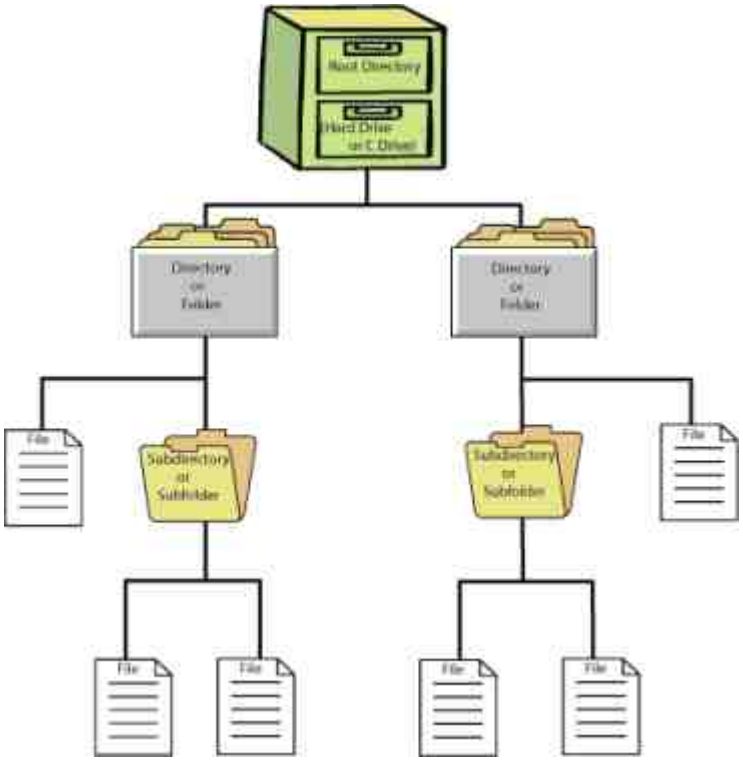


그림 19 파일시스템의 계층구조

파일에는 ASCII나 유니코드로 기술되는 인간의 언어의 내용으로 읽어질 수 있는 텍스트 파일과 이진수로 기록되는 바이너리 파일로 나뉜다. 일반적으로 동일한 데이터를 저장하는 경우 바이너리 파일이 용량이 더 작다. 생태정보 데이터의 경우 두 종류의 파일 모두가 필요하며 파일 시스템에서 이 두 종류의 파일을 모두 어려움 없이 지원한다.

생태정보 서비스에서는 대부분의 원시데이터를 파일의 형태로 파일시스템에 저장해야

한다. 또한 이후에 설명할 데이터베이스에 저장될 데이터를 제외한 모든 분석 데이터들이 파일의 형태로 저장되어야 하므로 파일시스템은 데이터 저장소 기술 중 가장 중요한 부분을 차지한다고 할 수 있다.

파일 시스템은 운영체제(OS, Operating System)과도 밀접한 관계를 가지고 있다. 호환성을 위해 Linux, Unix 계열을 운영체제를 사용하는 시스템의 범용 파일시스템은 POSIX(portable operating system interface)표준을 따른다. 가장 대표적인 파일 시스템에는 EXT4가 있다. 마이크로소프트 Windows 계열의 운영체제는 NTFS를 사용하는데 이 파일 시스템은 공개 소프트웨어가 아니므로 호환성이 떨어진다. 그 외에도 이동식 디스크 등 호환성이 중요한 경우 다양한 운영체제 및 시스템에서 접근이 가능한 FAT, exFAT 등을 사용한다. 그 외에도 네트워크 스토리지에 접근하기 위한 프로토콜 및 파일 시스템으로 NFS(network file system), CIFS(common internet file system)등도 많은 운영체제에서 쓰이고 있다.

2. 관계형 데이터베이스 관리 시스템(RDBMS)

데이터베이스는 여러 사람에 의해 공유되어 사용될 목적으로 통합하여 관리되는 데이터의 집합을 말한다. 정보를 체계적으로 축적하고 제공하는 것을 목적으로 한다. 논리적으로 연관된 하나 이상의 자료의 모음으로 그 내용을 고도로 구조화함으로써 검색과 갱신의 효율화를 꾀한 것이다. 즉, 몇 개의 자료 파일을 조직적으로 통합하여 자료 항목의 중복을 없애고 자료를 구조화하여 기억시켜 놓은 자료의 집합체라고 할 수 있다.

데이터베이스에서 가장 대표적인 개념은 관계 데이터 모델(relational data model)이라 할 수 있다. 관계 데이터모델 은 데이터 모델 중에서 가장 개념이 간단한 모델이다. 이 모델은 상대 수학적인 이론을 기반으로 한다. 현실 세계는 객체관계 그림(diagram)으로 표현되며, 개체와 그 관계는 각기 사각과 선으로 SQL 개체 관계형 데이터베이스를 지원하기 위해 SQL(Structured Query Language)가 창안되었으며, 이 언어는 수학적 관계 대수와 관계 논리(relational calculus)에 기반을 두고 있다. 데이터베이스 구조와 제약 조건을 정의하기 때문에 데이터 모델은 데이터를 조작하기 위한 연산집합을 가진다. 다시 말해, 관계 데이터모델 연산집합(a set of operations)은 관계대수로 표현되고, 그 연산은 사용자에게 여러 질의를 가능하게 한다.

```
select s.name Name, s.weight Weight,
      d.dname Name, p.name || 'Professor' ProfessorName
      from student s, department d, professor p
      where s.deptno=d.deptno
      and s.profno=p.profno(+)
      and weight < (select avg(weight)
      from student
      where deptno=(select deptno
      from student
      where name='John'));

select b.dept_name, count(*)
      from temp a, tdept b
      where b.dept_code = a.dept_code
      and a.emp_id in (select emp_id from tcom)
      group by b.dept_name;

select profno,name,sal*12+nvl(comm,0)
      from professor
      where sal*12+nvl(comm,0) >any (select avg(sal*12+nvl(comm,0)) from
professor);
```

[SQL 문을 이용한 데이터 질의 예제들]

관계형 데이터베이스 관리 시스템은 관계 데이터 모델을 이용해 데이터들의 관계를 테이블로 정리하여 관리하는 것으로 포맷이 정형화 되어 있는 데이터를 효율적/고속으로 저장/질의/분석하는 데에 강점이 있다. 그 종류로는 MySQL, PostgreSql, Oracle, Sybase, Informix, DB2 등이 있다. 그 중 MySQL, PostgreSQL 등은 공개 소프트웨어로 대부분의 리눅스 배포판에 포함되어 개인 사용자를 비롯하여 소규모 시스템에서 널리 쓰이고 있다. Oracle DB는 가장 유명한 상용 데이터베이스 관리 시스템으로 비용을 지불하더라도 대용량, 고성능이 필요한 시스템에서 널리 쓰이고 있다. 영국의 ECN에서도 메타데이터와 분석 데이터를 저장하는데 Oracle DB를 활용하고 있다.

데이터베이스 시스템은 아래 그림에서 보이듯이 매우 복잡한 구조를 가지고 있다. 위에서 설명한 파일 시스템이 운영체제의 일부로서 데이터를 파일이라는 형대로 계층적으로 관리하는 기본적인 기능을 제공하는 것에 비해, 데이터 트랜젝션이 안전하게 수행되기 위한 성질인 ACID(원자성, 일관성, 고립성, 지속성)를 보장하기 위한 알고리즘 등이 구현되어야 한다. 또한, 대용량 데이터의 처리, 다수의 사용자의 동시 처리 등을 빠르게

처리하기 위한 기술 등 최신 IT기술들이 집약되어 있다고 할 수 있다.

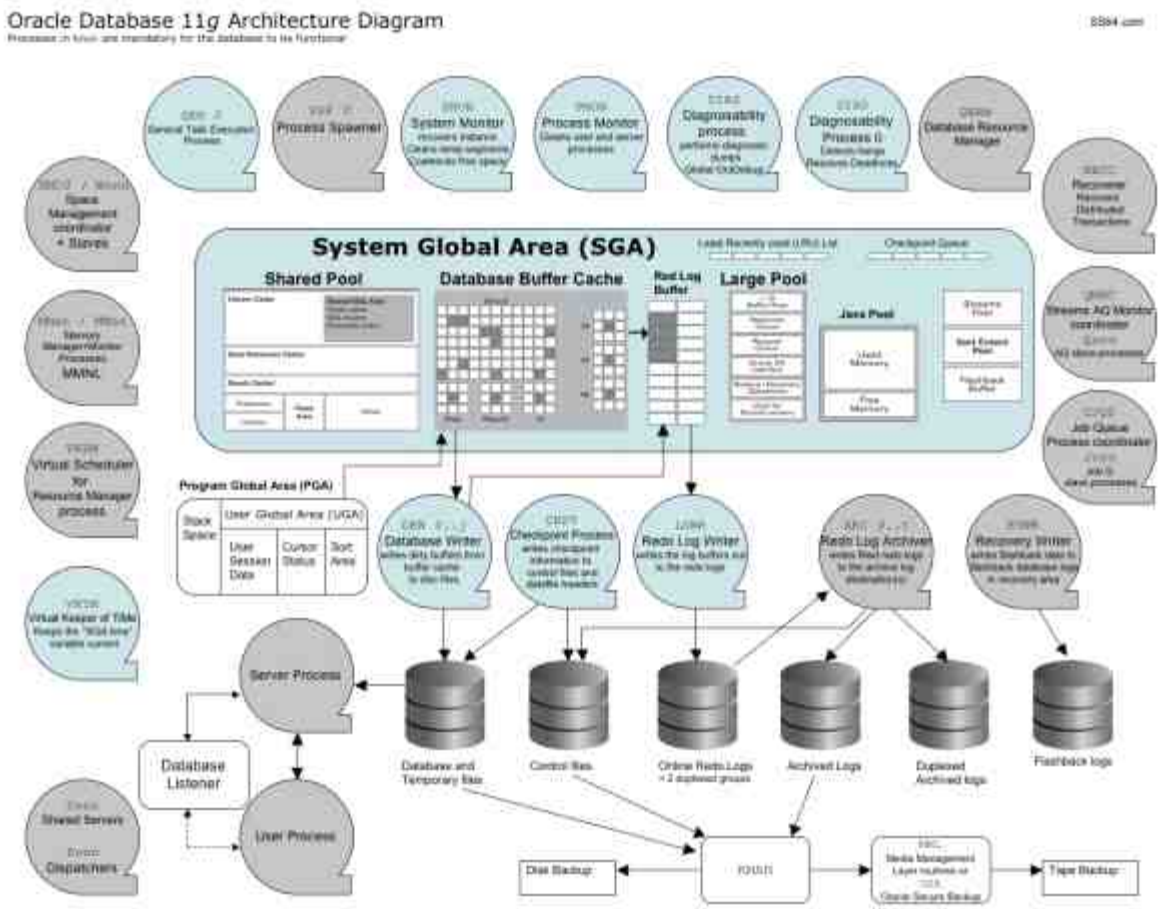


그림 20 Oracle Database 11g Architecture

3. 비정형 데이터베이스 시스템

비정형 데이터란 앞서 말한 정형데이터와 달리 일정한 규격이나 포맷을 가지지 않아서 관계형 데이터 모델로, 즉 형과 열로 표현된 테이블로 표현하기 쉽지 않은 데이터를 말한다. 예를 들면 그림, 영상, 문서 데이터 등이 이에 속한다. 실제로 수치화 되지 않은 일상생활에서 생성되는 대부분의 데이터가 비정형 데이터라고 할 수 있다. 생태정보 데이터 중에서는 텍스트, 멀티미디어 원시데이터가 이에 속한다. 특히 최근에 빅데이터 처리 기술이 화두로 떠올랐는데 SNS데이터를 비롯 인간의 일상과 관련된 많은 데이터들의 분석이 주목을 받았고 이에 따라 비정형 데이터를 저장/검색/분석하는 기술들에 대한 연구가 많이 진행됨에 따라 많은 비정형 데이터베이스 시스템들이 제안되고 개발되었다.

비정형 데이터베이스 시스템은 기본적으로 느슨한 Key-Value 쌍으로 데이터를 저장/검색한다. RDBMS보다 느슨한 제약조건을 가지고 있고 NoSQL(not only SQL) DB라 불리기도 한다. 즉, 관계형 데이터 모델로 표현 가능한 데이터 뿐 아니라 다른 대부분의 검색 가능한 데이터를 저장할 수 있다.

RDBMS vs NoSQL(Not Only SQL)

구분	RDBMS	NoSQL
장점	• 정규화된 데이터 모델 • 데이터 무결성/정합성 보장 • 작은 크기의 트랜잭션	• Web, SNS 환경의 다양한 정보를 검색및 저장
단점	• 확장성의 한계 • 클라우드 분산환경 부적합	• 데이터 무결성/정합성 보장 못함
특징	• Join을 통해 정보 연계 가능 • 트랜잭션의 안정성 확보를 위한 ACID(Atomicity, Consistency, Isolation, Durability) • 정형적인 스키마 구조	• 데이터의 수정/삭제가 거의 없고 대부분 Insert 위주 • 강한 Consistency 불필요 • 노드 추가/삭제, 데이터 분산에 유연

그림 21 RDBMS와 NoSQL의 비교

NoSQL 데이터베이스를 분류하는 접근 방식은 분류와 하위 분류와 함께 다양하다. 다양한 접근 방식으로 인해 비관계형 데이터베이스를 포괄적으로 파악하는 데에는 어려움이 있다. 기본적인 분류는 데이터 모델에 기반을 둔다. 컬럼(column) 기반의 HBase 와 Accumulo, 도큐먼트 기반의 MongoDB 와 Couchbase, 킷값 기반의 Dynamo와 Redis, 그래프 기반의 Neo4J 와 Virtuoso 등 다양하다.

구분	설명	솔루션
Key-Value 기반	Key 기반으로 Put, Delete 등의 기능 제공	HBase, Amazon SimpleDB, Azure Table Storage, Chordless, Redis, Tokyo Cabinet/Tyrant, HamsterDB,Scalien, Berkeley DB, MemcacheD, Membase(MemcacheD 기반), Mnesia, LightCloud(Tokyo Tyrant 기반), Pincaster, GenieDB 등
테이블 기반	테이블 개념 제공, 로우가 아닌 컬럼 기반으로 데이터 저장	HBase, Cassandra, Cloudata, Hypertable 등
문서 기반	JSON, XML 형태로 구조적 문서 저장	CouchDB, MongoDB, Riak, Terrastore, OrientDB, RavenDB 등
그래프 기반	그래프 형식으로 데이터를 저장 및 표현	Neo4J, Sones, InfoGrid, HyperGraphDB, Allegro Graph, Bigdata, DEX, OpenLink Virtuoso 등

그림 22 비정형 데이터베이스의 분류

비정형 데이터베이스는 빅데이터(Big data) 분석과 큰 연관을 가지고 있다. 데이터 크기가 크기 때문에 하나의 고성능 시스템으로 구축하기 보다는 저가의 노드 여러대(수십~수천) 대를 묶어서 데이터 저장과 데이터 분석을 동시에 처리하는 시스템을 구성한다. 데이터를 분석할 때 데이터를 나누어서 병렬로 처리할 수 있는 경우가 많으므로 계산자원(Computing resource, CPU + Memory)와 저장장치(HDD)가 같은 곳에 위치하면 Locality가 매우 높아져서 성능이 크게 향상된다. 미국 구글(Google)에서 제안된 GFS(Google File System)과 MapReduce 프로그래밍 모델이 이러한 패러다임을 선도하였다. 구글에서는 GFS라는 파일시스템 위에서 MapReduce를 이용한 빅데이터 분석을 하는데 원시 데이터의 저장과 분석 결과의 저장을 효율적으로 하기 위한 Key-Value 데이터베이스로 BigTable을 개발하였다. 구글의 아이디어는 논문으로 발표 외어 큰 반향을 일으켰으나 공개 소프트웨어가 아니었기에 Apache 에서는 이러한 아이디어를 공개소프트웨어로 구현하고자 하는 시도를 하여 HDFS, Hadoop, HBase를 만들어 내었다. 공개 소프트웨어의 특성상 매우 널리 퍼지고 현재까지도 발전을 거듭하고 있다.

생태정보 서비스에서는 대개 원시데이터 크기가 다양하고, 크기가 큰 경우가 많고, 데이터 마이닝 등 후처리가 필요하다. 이에 비정형 데이터베이스가 저장소에 반드시 포함되어야 하고 계산 자원과 연동을 고려한 설계가 필수적이다. 따라서 시스템 자체를 Hadoop Cluster로 구축하는 것도 고려할 필요가 있다.

4. 네트워크 연결 스토리지

생태정보 서비스의 저장소는 단순히 데이터를 보관하기 위한 공간은 아니다. 데이터를 원하는 사용자에게 제공하여야 하고 데이터의 분석을 위한 기반이 되는 곳이다. 데이터 공유를 위해서는 저장소를 위한 스토리지 하드웨어를 구축하고 이를 네트워크에 연결해야 한다. 데이터를 네트워크를 통해 전송해야 하므로 고속의 네트워크 장비가 저장소 시스템에 필수적으로 포함된다.

네트워크 연결 스토리지 기술은 이미 오래전부터 널리 쓰이고 있는 기술로 대용량 고속 데이터 전송을 위해 발전해 왔고 표준 프로토콜도 확립되어 있다. 범용의 네트워크(Internet protocol)에 연결된 저장소를 NAS(network attached storage)라고 하고, 고성능의 시스템에서 고속의 데이터 처리를 위해 데이터 공유 전용의 네트워크에 연결한 경우 SAN(storage area network)라고 한다. NAS의 경우에는 다양한 운영체제를 가

진 클라이언트(Client)를 지원하기 위해서 NFS, CIFS등의 호환성 높은 파일시스템 인터페이스를 사용하는 것이 일반적이다. 따라서 파일 단위로 원거리 저장소 서버의 데이터를 접근할 수 있다. SAN의 경우에는 파일단위 접근 이외에도 블록단위의 접근을 가능하게 하는 iSCSI 프로토콜을 사용하는 경우가 많다. 이 경우 분산 데이터베이스 등 특수 목적의 고성능 스토리지를 구축하는데 유리하다.

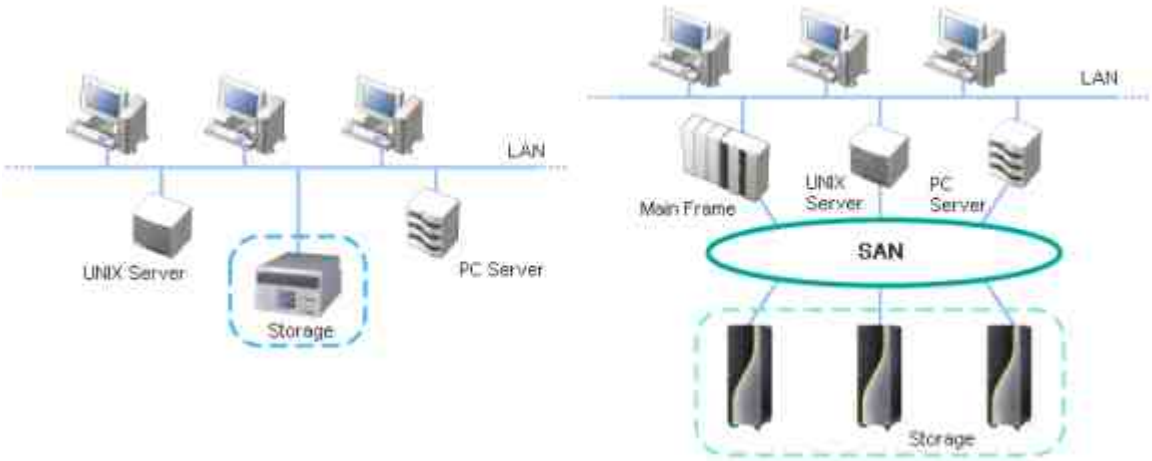


그림 23 NAS(좌)와 SAN Storage(우)의 네트워크 상 위치

생태정보 데이터 시스템의 경우 데이터의 다양성에 따른 스토리지 시스템의 다양성, 다양한 연구자 및 해외 시스템과의 연동, 메타데이터 시스템, 웹 포탈과의 연동 등을 지원하기 위해 NAS장치의 인터페이스를 반드시 제공해야 할 것으로 판단된다.

5. 데이터 보관 및 백업 장치

대용량의 데이터를 당장은 사용하지 않지만 장기간 보존해야하는 경우에는 특수한 목적의 저장장치를 이용한다. 저장 매체가 시간이 지나도 변질되지 않는 특성을 가져야 하며 재활용이 불가능하여도 되고, 습도 온도 등에 민감하지 않아서 보관에 비용이 많이 들지 않아도 되며 가격도 저렴하면 좋다. 이러한 특징에 맞춰 데이터 보관에 가장 많이 사용되는 것은 테이프 드라이브이다. 데이터가 테이프에 기록할 때는 연속된 데이터를 한번만 쓰기가 가능하고 한번 기록한 내용의 일부분만을 수정하는 것은 불가능하다. 데이터를 읽어오는 경우에도 전체 내용을 읽어오는 것과 일부분만 읽는 경우의 경우가 차이가 없으므로 Random Read 에는 매우 비효율적이다. HP, IBM, SUN(현재 Oracle)

등 고성능 대용량 시스템을 개발하는 업체에서 주로 생산한다.

그림 24 랙 용 2U 테이프 스토리지 장치 및 카트리지

데이터 백업에 테이프 드라이브 등 아카이빙 장치를 사용하는 것 외에도 서버 고장 등으로 인한 유실을 막기 위해 데이터를 이중화 하는 경우도 있다. 가장 일반적으로 쓰이는 것이 RAID(redundant array of inexpensive disk)를 사용하는 것이다. 즉 동일한 데이터를 2군데에 복사를 하던지 parity disk를 두어 하나 혹은 일부의 디스크가 고장나도 데이터를 복구 할 수 있게 하는 방식이다. RAID는 데이터의 안정성 이외에도 데이터를 여러대의 디스크에 나누어(striping) 위치시킴으로써 동시에 많은 양의 데이터를 읽어들일 때 성능을 향상시킬 수 있다는 장점도 가지고 있다. 특히 3절에서 설명한 Hadoop Cluster 의 HDFS의 경우 데이터와 계산자원 Locality 향상과 데이터 안정성을 동시에 추구하기 위해 동일한 데이터를 3군데에 중복하여 저장한다.

데이터 안정성을 높이기 위한 백업은 데이터의 중복 저장을 야기 시키므로 더 많은 하드웨어를 필요로 하고 동일한 내용을 데이터를 저장하는데 더 많은 시간을 필요로 한다. 즉 안정성과 성능/비용간에 trade-off가 발생하게 된다. 일반적으로 2배수를 필요로 하는 RAID-1대신 1/4 수준의 추가 디스크를 필요로 하는 RAID-5 와 테이프 백업을 혼합하여 사용하는 방식이 많이 사용된다.

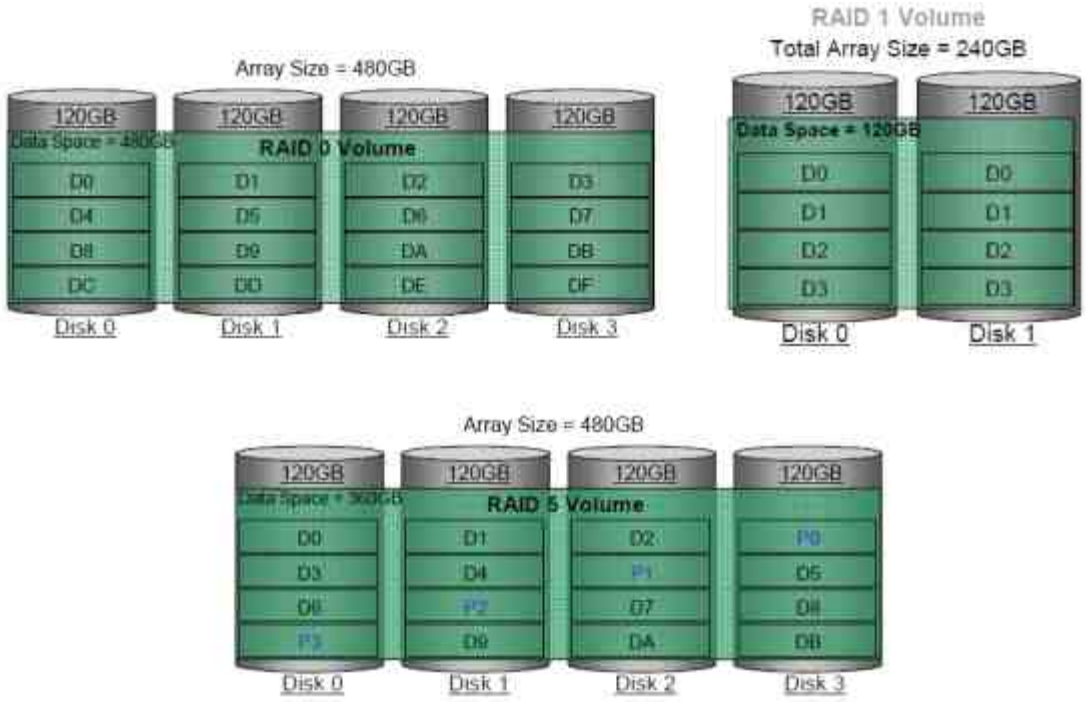


그림 25 RAID 0, 1, 5 의 구조

IV. 생태관측 데이터 특성 분석

IV. 생태관측 데이터 특성 분석

1. 생태관측 데이터의 종류 및 특성

○ 지리적으로 분산된 동종의 센서에서 주기적으로 측정된 수치(Discrete Numeric) 자료

현재 가장 널리 수집/공유/분석되는 데이터 유형 중 하나로 강수량, 온도, 풍속, 풍향 등 환경 정보 등이 이 유형에 속한다. 수치 데이터이므로 정형화된 형태로 저장하고 분석하는 게 용이하다. RDBMS에 저장하기 가장 적합한 형태의 데이터이다. 저장소에 저장되고 다른 연구에서 공유하기 위한 데이터의 형(type) 등이 표준화 되어 있는 경우가 많고 통계 분석 등의 자동화 기술도 이미 성숙되어 있다. 데이터의 크기는 측정지역의 넓이와 그에 따른 센서의 개수, 측정주기, 측정 기간 등에 따라 크게 변할 수 있고 그 양이 매우 커질 수도 있다. 그러나 일반적으로 데이터의 총량이 매우 크지는 않다. 빠른 시각화 및 정보 제공을 위해 널리 쓰이는 통계 자료를 계산자원의 유휴 시간에 미리 처리해두는 기능이 필요할 수 있다. 영국의 ECN, 호주의 TERN 에서도 이러한 데이터를 원시데이터 뿐 아니라 분석한 데이터까지 미리 저장해 두었다가 제공하는 서비스를 가지고 있다.

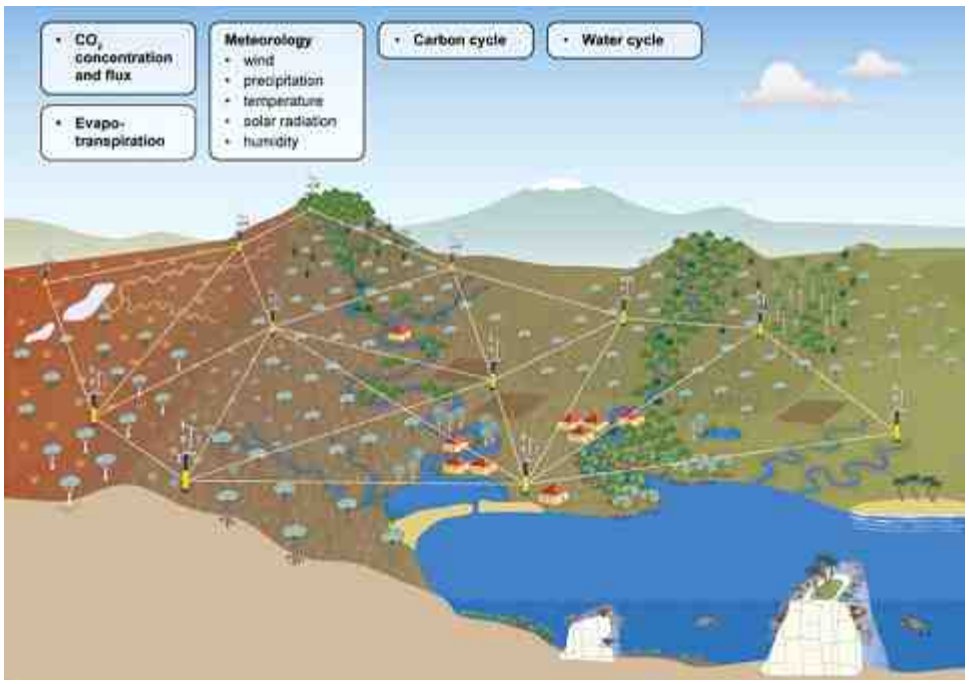


그림 26 호주 TERN의 OxFluz 센서 네트워크

그림 27 OzFlux 센서

○ 고정 설치된 센서에서 지속적으로 수집되는 신호(analog/digital stream) 데이터

위에서 기술한 주기적으로 측정되는 데이터와 유사하나 지속적으로 수집되므로 원시데이터의 크기는 더 크다. 온도나 풍향 등의 데이터와 같이 이 데이터를 샘플링 혹은 평균값을 구하여 수치 자료화 한 경우가 있다. 데이터의 크기가 크므로 센서에서 저장소까지 데이터를 전송하는 방법이 필요한데, 고속의 네트워크 장비가 있어서 실시간으로 전달하는 방법이 있고 이동식 디스크에 저장이 되고 그것을 사람이 주기적으로 수거해 가는 경우도 있다. 일반적으로 원시데이터를 직접 사용하는 경우 외에 추가로 분석 및 통계 과정을 거쳐 나온 결과를 따로 저장하는 방식을 사용할 필요가 있다.

그림 28 TERN Echo Sensor

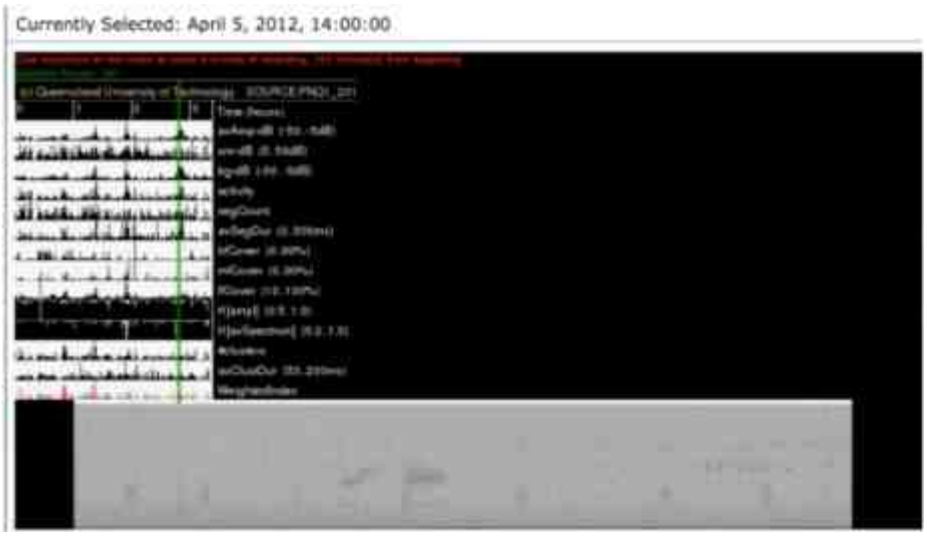


그림 29 Echo Sensor에서 수집된 데이터의 분석

TERN Echo sensor에서 수집되는 조류의 음성 데이터가 이러한 특징을 잘 반영한다. 매 센서 마다 매달 64GByte의 데이터가 수집된다. 이러한 센서가 삼각 측량을 위해 최소 3개 이상 한 사이트에 존재해야 하며 사이트의 크기가 크면 센서의 크기도 비례하여 많아지고 그에 따른 데이터의 양이 커진다. 측정된 데이터는 센서에 부착된 이동식 디스크에 저장되며 매달 수동으로 수거를 하게 된다. 수거된 데이터는 삼각 측량을 기법을 활용하여 분석되고 그를 통해 조류의 행동 특성을 파악하게 된다.

○ 이동하는 측정 장비를 이용한 멀티미디어 데이터

유인 비행기, 무인 기체(드론), 차량, 무인 로봇 등을 이용해 이동하면서, 카메라(가시광, 적외선 등), 마이크 등을 이용하여 수집한 사진, 음성, 동영상 등의 데이터가 이 분류에 속한다. 비용 등의 문제로 측정 빈도가 높진 않지만 1회의 측정에서 수집되는 원시 데이터의 크기가 매우 클 수 있다는 특징이 있다. 측정과 동시에 무선 네트워크를 통해 저장소에 전송하거나 측정이 끝나고 직접 저장소에 복사하는 방식을 사용한다. 원시 데이터 자체는 형식화 된 자료가 아니므로 이를 분석하기 위해서는 패턴인식, 경계추적 등 IT분야의 기술들을 활용해야 하고, 이에 따라 고성능의 계산 자원이 필요.

그림 30 다양한 방법을 활용한 멀티미디어 데이터 수집

○ 관측자가 직접 입력한 텍스트 자료

형식화된 데이터나 수치 정보가 아닌 자연언어로 기록되어 있는 자료가 이에 해당한다. 이 경우 상황 혹은 데이터에 대한 관측자 자신의 의견이나 설명을 기술한 경우, 혹은 연구자가 데이터를 보고 추가 메모 등을 적은 경우 등이 있을 수 있다. 텍스트 기반이고 작성 빈도수가 많지는 않아 필요한 저장 공간은 크지 않으나, 이를 자동으로 분석하기 위한 기술은 성숙되어 있지 않다. 최소한 텍스트 검색 기능 등은 제공되어야 할 것으로 판단된다.

2. 생태관측 데이터의 분류 기준

○ 종류 및 형(Type)

자연언어 텍스트, 정형화 된 수치 자료, 멀티미디어 데이터, Binary Stream 등으로 구분하여 그에 맞는 저장소를 활용해야 한다. 예를 들어 자연언어 텍스트와 멀티미디어 데이터는 NoSQL등 비정형 데이터베이스나 파일시스템을 활용하고 수치자료 등은 Oracle, MySQL 등 RDBMS 시스템을 활용한다.

○ 크기

위에서 구분한 생태정보 데이터의 특성에 따르면 바이트(Byte)단위의 작은 데이터 수백 개에 불과한 아주 작은 크기의 데이터부터 수백 기가바이트에 이르는 영상 데이터까지 크기가 매우 다양한다. 데이터는 단순한 보관 뿐 아니라 분석/가공, 시각화, 공유 등 빈번한 읽기 요구가 있으므로 각각 데이터 크기에 맞는 다양한 기술을 활용한 저장소 설계가 필요한다.

○ 전송방식

센서에 부착된 통신장비나 측정자의 모바일 장치 등 현장에서 직접 네트워크 인터페이스를 이용하여 자동으로 입력하는 경우와 대용량 데이터를 저장소 위치에서 수동으로 업로드 하는 경우를 모두 지원할 수 있어야 한다.

○ 데이터의 저장 방식

원시데이터를 파일로 저장하는 경우와 데이터베이스에 저장하는 경우로 구분할 수 있다. 일반적으로 멀티미디어 데이터나 텍스트 데이터의 원시데이터의 경우에는 파일의 형태로 저장하고, 이를 가공 및 분석한 데이터 혹은 수치로 수집된 데이터의 경우에는 데이터베이스를 활용하는 것이 적합하다.

○ 데이터의 보관 여부

원시데이터 데이터 자체를 보존 및 공유해야 하는 경우, 데이터 정제 및 분석 후 원시데이터는 삭제하는 경우, 데이터에 대한 접근이 빈번하여 고성능의 네트워크 성능을 요

구하는 경우, 데이터를 국제연구자 및 민간기관과 공유해야 하는 경우 등을 구분하여 각각에 맞는 저장 기법을 사용해야 한다. 장기 보관을 위한 테이프 장치 등과 빠른 데이터 접근을 위한 장치들의 특성은 확연이 구분된다.

○ 데이터 분석 방식

수치데이터의 경우에는 단순한 통계 등으로 처리가 가능하지만 비정형데이터를 분석하는 경우 고성능의 계산자원이 필요할 수 있다. 이 경우 계산 자원과 저장소 간의 연동을 위한 구조를 고려하여 설계해야한다.

○ 데이터 접근 권한

데이터의 접근권한은 저장소 구조에 큰 영향을 주지는 않으나 권한을 구분하여 접근을 차단하는 기능이 저장소에 추가되어야 한다.

3. 요약

종류	수집방법	크기	개수	특징
텍스트 데이터	측정자 직접 작성	작음 (10~ KB)	적음	비정형 데이터
수치 데이터	센서 자동 측정	매우 작음 (1~ Byte)	아주 많음	Type 이 정의된 데이터, 통계분석 필요
스트림 데이터	센서 자동 측정	큼 (100~ MB)	보통	정형 데이터, 샘플링 및 패턴 분석 필요
이미지	드론, 측정자 등	큼 (1~ MB)	많음	비정형 데이터, 패턴분석 등에 계산 자원 필요
영상	드론 등 이동수단 이용 측정	매우 큼 (1~ GB)	적음	비정형 데이터, 사물인식 등 분석에 고성능 계산자원 필요

표 3 생태측정 데이터의 특징 요약

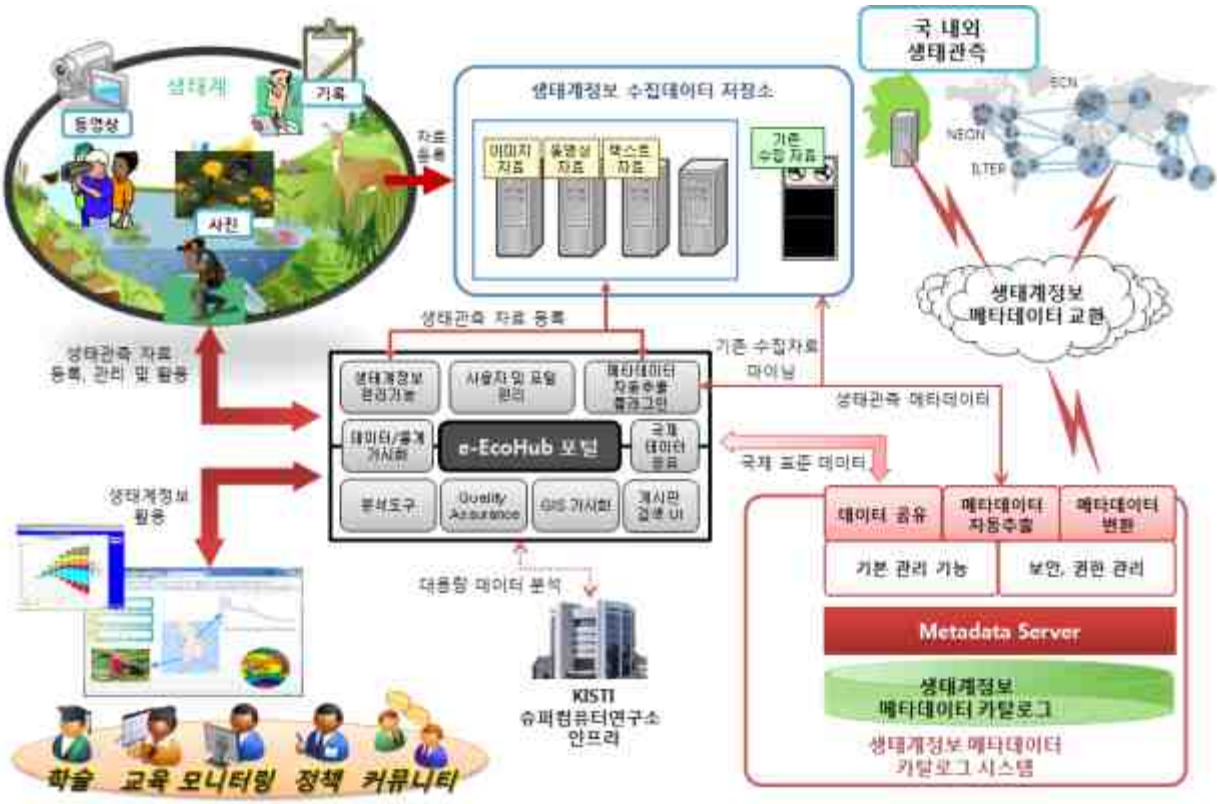
V. 생태관측 데이터 저장소 설계

V. 생태관측 데이터 저장소 설계

1. 서비스 정의

생태관측 서비스는 장기적인 생태 및 환경의 조사를 기반으로 관측된 데이터를 저장하고, 품질관리에 적합하게 데이터를 가공 및 메타데이터를 생성하는 절차를 통해 연구자가 분석할 수 있도록 사이버 인프라를 제공한다. 또한 장기적인 관측 데이터 관리와 향후 데이터 관리에 영향을 줄 수 있는 측정센서의 발달에 따른 데이터의 확장성을 고려한 시스템이 보장이 되어야한다. 연구자의 국제협력을 지원하고, 생태/환경 정책의 반영을 위해서는 국내 데이터 뿐만 아니라, 국제 데이터의 상호 교환을 통해 생태종의 유입, 국제적 선도 연구 참조 등을 위해 생태관측 메타데이터 교환을 지원할 수 있어야 한다.

데이터 저장소는 생태관측 서비스의 중요한 하부 인프라로써 자료의 저장/보관 기능 뿐 아니라 빠른 검색 기능 제공 데이터 통계/분석 등의 역할, 타 데이터 서비스와의 연동 등의 기능을 제공해야 한다.



생태관측 서비스 개념도

2. 사용자 정의

생태관측 서비스의 데이터 저장소는 데이터의 업로드 등의 특수한 경우를 제외하면 사용자들과 직접적으로 연결되지 않고 웹 포탈을 통하게 된다. 따라서 표준화되고 자동화 가능한 인터페이스의 제공이 필요하다

○ 학술연구자

생태관측 데이터를 활용하여 환경변화, 생태계 변화 요인분석 등의 연구를 수행하는 자로써 데이터를 받아 사용하는 역할 뿐 아니라 스스로 관측한 데이터를 제공하는 역할도 동시에 수행할 수 있다.

○ 교육자/학생

생물, 생태, 환경을 교육을 담당하는 교육자와 관련 학생들이 생태관측 서비스를 활용하여 교육을 받는다.

○ 정책 결정자

생태 관측 데이터를 활용하여 국가적인 환경 관련 정책을 결정한다.

3. 주요 제공 기능

○ 원시데이터 저장을 위한 NAS연동 대용량 파일 시스템
○ 수치 데이터 등 정형 데이터의 공유/분석을 위한 RDBMS 시스템
○ 비정형 데이터의 저장, 공유 및 검색을 위한 NoSQL 데이터베이스 시스템
○ 비정형 데이터의 분석 및 후처리를 위한 아파치 하둡(Hadoop) 등 빅데이터 처리 SW와의 연동 기능
○ 원시데이터를 주기적으로 분석하여 통계/정제 데이터를 생성하는 기능
○ 원거리 네트워크 장비에서 전송한 데이터를 수집하여 자동으로 저장하는 기능
○ 대용량 데이터의 저장을 위한 로컬 및 고대역폭의 네트워크를 통한 업로드 기능
○ 메타데이터 카탈로그 시스템 및 웹 포탈과의 연동을 위한 인터페이스 제공
○ 데이터를 백업 혹은 보관하기 위한 아카이빙 장치

4. 전체 시스템 구조

3절에서 나열한 기능을 포함하여 전체 시스템 아키텍처를 그리면 아래와 같다. 아래 주황색으로 표시된 4개의 요소가 필수적으로 구축되어야 하고 그 요소들 사이에 연동을 위한 네트워크 구성의 설계가 필요하다.

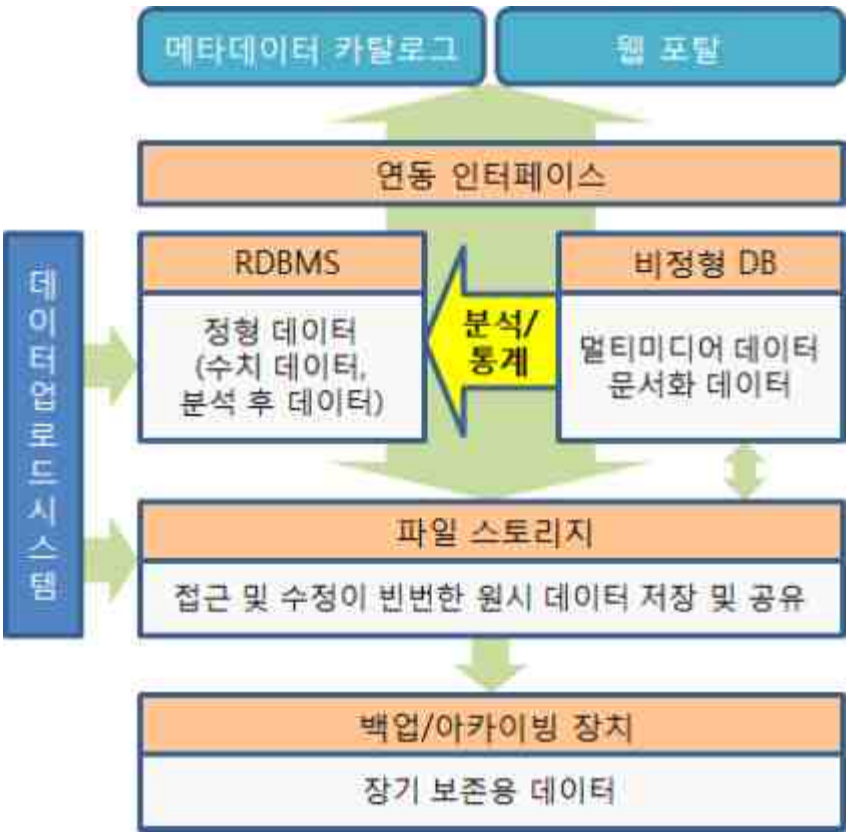


그림 32 생태관측 데이터 저장소의 필요 기능 및 아키텍처

VI. 향후계획 및 기대효과

VI. 향후계획 및 기대효과

1. 향후계획

본 연구를 통해 생태정보 서비스에서 수집하고 제공해야 하는 데이터들의 특징과 그 데이터를 효율적으로 관리하기 위한 저장소 기술을 확인하고 필요한 기술을 도출하였다. 이후에는 이 정보들을 바탕으로 저장소 시스템의 구체적인 설계를 하고 필요한 용량 및 성능을 도출 하여야 한다. 그 후, 시스템 도입 및 구축을 진행하고 웹 포탈, 메타데이터 서버, 분석용 계산 자원과의 연동 기술 개발을 수행할 예정이다.

2. 기대효과

생태분야 글로벌 사이언스게이트웨이(e-EcoHub) 구축을 통한 국내외 생태관측 데이터 공동 활용 네트워크를 구축에 활용할 수 있고 생태관측, 관측자료 관리, 모니터링, 모델링을 통한 환경 예측과 예보를 위한 정책 결정자들의 생태/환경 관련 의사결정 지원에 활용할 수 있다. IT 기반의 생태/환경 분야간의 협력을 통해 한반도 및 아시아 지역의 생태계 변화 관측 네트워크를 구축하여 지구 생태계 변화 분석 및 예측 모델에 활용할 수 있다. 또한 축적된 생태정보를 통해 환경의 상관적 관계를 연구하여 생태계를 위협하는 다양한 환경 위험을 사전에 예방하고 멸종 위기 생태종 보호에 활용함은 물론 생태 복원 및 관리할 수 있는 시스템들과 연계 활용이 가능하다. 그 밖에도 전 국토의 생태계 변화 뿐 아니라 환경오염 등을 실시간 모니터링할 수 있는 환경 네트워크에 활용하고 축적된 생태관측 정보는 신규 과학자 및 교육자 육성에도 활용이 가능하다.

VII. 참고문헌

II. 참고 문헌

[1] ECN : http://www.ecn.ac.uk

[2] TERN : http://www.tern.org.au

[3] NEON : http://www.neoninc.org/

[4] ANDS : http://ands.org.au

[5] EML : https://knb.ecoinformatics.org/#external//emlparser/docs/eml-2.1.1

[6] 안성수, et al. "생물다양성데이터 검색포탈 구축." 한국콘텐츠학회 종합학술대회 논문집 3.1 (2005): 124-128.

[7] 안부영, et al. "생물다양성 데이터교환을 위한 메타데이터 스키마 설계." 한국정보과학회 학술발표논문집 32.2 (2005): 91-93.

[8] 이슬기, et al. "GIS 를 이용한 산림성 조류의 서식지 예측 모형 및 지도구축." 한국지리정보학회지 13.1 (2010): 62-73.

[9] 안성수, et al. "KBIF 데이터 노드 구축." 한국정보과학회 학술발표논문집 31.2Ⅱ (2004): 280-282.

[10] Oracle DB : http://www.oracle.com

[11] Ghemawat, et al. "The Google file system". Proceedings of the nineteenth ACM Symposium on Operating Systems Principles (2013)

[12] Jeffrey Dean, et al. “MapReduce: Simplified Data Processing on Large Clusters”, OSDI'04: Sixth Symposium on Operating System Design and Implementation (2014)

[13] Fay Chang, et al. “Bigtable: A Distributed Storage System for Structured Data”, OSDI'06: Eighth Symposium on Operating System Design and Implementation (2016)

[14] Helen Cleugh, et al. “OZFLUX:Near-real-time measurement of CO2, water and energy fluxes: Determining the best available estimates of continental carbon and water fluxes”.

Acknowledgement:

This work is supported by both Korea Ministry of Environment as "Public Technology Program based on Environmental Policy (Grant No.: 2014000210004)" and the KISTI (Grant No. : K-16-L01-C03).

도서명: 생태관측 데이터 특성 분석 및 저장소 설계 기술 보고서

발행일: 2016년 7월 31일

발행처: 한국과학기술정보연구원