

연구 데이터 공통표준 가이드라인

2024

한국지질자원연구원

지질자원데이터센터

머리말

본 가이드는 지질자원 분야에서 연구 데이터를 체계적으로 관리하고 활용하기 위한 지침을 제공하는 것을 목적으로 한다. 현대의 연구 데이터는 양과 복잡성이 점차 증가하고 있으며, 이를 효과적으로 관리하고 재사용 가능하게 하기 위해 데이터의 수집, 관리, 보존, 활용 전반에 걸친 표준화가 필수적이다. 이러한 특징을 가진 연구 데이터를 처리하고 공유하기 위해서는 체계적이고 일관된 기준이 필요하다.

이 가이드는 연구 데이터의 전주기적 관리 프로세스를 정의하고, FAIR 원칙(Findable, Accessible, Interoperable, Reusable)을 준수하여 연구 데이터의 검색 가능성, 접근성, 재사용성을 높이는 방법을 제시한다. 또한, 국제 단위계(SI)와 같은 표준화된 측정 단위를 사용하여 연구 결과의 신뢰성을 보장하고, 국내외 연구 기관 간 데이터 상호운용성을 높이는 것을 목표로 한다.

제1장에서는 연구 데이터의 정의와 연구 기관에서 연구 데이터를 관리하고 활용하기 위한 연구 데이터 라이프사이클을 정의한다. 이를 통해 체계적인 연구 데이터 관리의 필요성 및 표준화 과정의 기본적인 내용을 다루며, 연구 데이터의 정의 그리고 이를 효율적으로 관리하기 위한 단계별 절차와 각 단계에서 고려해야 할 표준화 요소를 설명한다.

제2장에서는 연구 데이터와 공간 정보 관리에 적용되는 주요 표준과 법적 체도를 설명한다. 여기에는 국가 표준 및 국제 표준의 적용 방법과 관련 규정들이 포함되며, 연구 데이터를 공개하는 데 있어서 의무적 준수 사항 또는 참고 사항이 될 수 있다.

제3장에서는 국제 단위계의 적용과 그 중요성을 자세히 다룬다. 연구 데이터의 일관된 표현을 위해 국제 단위계를 사용하는 것은 법적 의무 사항이므로, 본 장에서는 그 적용 방법에 대해 알아본다.

제4장에서는 연구 데이터의 원활한 관리 및 활용을 위한 FAIR 데이터 원칙과 이를 실현할 수 있는 평가 도구인 FAIRsFAIR 프로젝트에 대해 소개하고 그 이용 방법을 설명한다. 또한, 파편화된 연구 데이터를 효과적으로 패키징하는 방법을 다룬다.

제5장에서는 연구 데이터의 메타데이터에 대해 소개하고, 메타데이터 설계를 위한 기술 요소와 함께 FAIR 데이터 원칙을 충족하는 대표적인 데이터 포맷을 설명한다. 본 가이드는 지질자원 분야의 연구자들뿐만 아니라 다양한 분야에서 연구 데이터를 관리하고자 하는 모든 연구자에게 실질적인 도움을 줄 것이다. 연구 데이터의 표준화를 통해 데이터의 신뢰성을 높이고, 학문적 협업과 데이터 공유를 촉진하는 데 기여할 것이다.

작성 : 연영광 (지질자원데이터센터)

목차

1. 연구 데이터의 개념	4
1.1 정의	4
1.2 연구 데이터의 라이프 사이클	4
1.3 연구 데이터 관리와 활용	7
2. 표준개요 및 현황	9
2.1 개요	9
2.2 표준제도 현황	14
3. 국제단위계 및 적용	18
3.1 국제단위계	18
3.2 국제단위계의 적용	22
3.3 지질자원분야에서의 적용	29
4. 연구 데이터 관리 및 활용	32
4.1 FIAR data	32
4.2 FAIRsFAIR 프로젝트	35
4.3 데이터 패키징	38
5. 연구 데이터 기술 요소	46
5.1 메타데이터 표준	46
5.2 데이터 모델링	54
5.3 Fair 데이터 포맷	61
6. 참고문헌	72

1. 연구 데이터의 개요

1.1 정의

연구 데이터에 대한 활용의 필요성으로 국가차원에서 연구 데이터의 공유를 위한 큰 흐름이 진행되고 있다. 국가연구개발정보처리기준(시행 2021. 1. 1)에서는 데이터 관리계획이 명시하고 있으며, 이는 연구 데이터의 생산·보존·관리 및 공동활용 등에 관한 계획을 통해 체계적인 관리 및 활용에 대한 기본적인 규정을 포함한다. 연구 데이터는 유형적으로 정량적, 정성적 형태를 취하고 있으며, 데이터 가공 단계에 따라 원시 데이터 및 이를 해석 및 처리된 2차 가공데이터로 구분될 수 있다. 또한 기록방법에 따라 이미지, 도표, 등 다양한 기록물의 형태를 포함할 수 있다. 이에 따라 연구 데이터에 대한 목적 및 내용에 이해도를 높이기 위해 연구 데이터에 대한 기본적인 범위 및 내용들을 검토하는 것이 필요하다.

연구 데이터는 ‘연구개발과제 수행 과정에서 실시하는 각종 실험, 관찰, 조사 및 분석 등을 통하여 산출된 사실 자료로서 연구결과의 검증에 필수적인 데이터’로 국가연구개발정보처리기준(시행 2021. 1. 1)을 비롯하여 모든 국가기관의 규정에서 공통적으로 정의하여 사용되었다. 연구 데이터의 정의를 통해 그 범위를 유추하면 연구결과와 구분되어지는 관계이며, 연구결과에 대한 검증에 필요한 데이터로 유형, 가공단계 및 기록물의 형태를 별도로 구분하지 않고 있다. 따라서 세부적인 규정 및 내용들은 연구 데이터를 관리계획에서 협의에 의해 연구사업의 특성에 따라 정의될 수 있다고 볼 수 있다. 더불어 최종 연구성과의 검증 유무와 상관없이 연구사업과정에서 생산되는데이터의 재활용성을 고려하여 연구자가 자발적으로 연구 데이터로 포함하여 제공될 수 있다. 따라서 연구 데이터의 정의에 따라 연구사업의 최종 결과물의 검증을 위한 목적에 직접적으로 사용될 수 있는지 여부 등을 고려하여 연구 데이터 관리계획에서 핵심 연구 데이터 등을 구분하는 것도 좋은 전략이 될 수 있다.

1.2 연구 데이터의 라이프 사이클

연구 데이터의 체계적인 관리와 활용을 위해서는 라이프 사이클을 정의하여 각 단계별 처리과정을 정의하여 데이터에 대한 일관성 유지, 보존, 접근성 등을 확보할 수 있는 수단으로 활용할 수 있다. 연구 데이터의 라이프사이클은 프로젝트 단위, 연구

기관차원에서의 관점에 따라 달라질 수 있다. 이러한 특징에 따라 다양한 기관 및 단체에서 연구 데이터 라이프사이클이 제시되었으며 문헌상 DCC, ICPSR, IWGDD, DataONE, USGS, 등에서 정의된 내용을 볼 수 있다. 본 지침서에서는 앞서 연구 데이터의 정의에서와 같이 연구 데이터는 연구결과의 검증에 필수적인 데이터로 정의되며, 이는 분석이 종료된 이후 분석결과와 직접적 해석에 사용되는 정제된 데이터로 그 의미를 찾을 수 있다. 현행 시스템에서도 연구 데이터의 제출 단계는 연구결과가 도출된 이후 정제된 형태의 데이터에 대한 신뢰성과 재활용성을 기대 할 수 있다.

따라서 본 가이드라인에서는 기존에 제시된 연구 데이터의 라이프 사이클들의 개별 단계들을 연구 데이터 리포지터리를 운영하는 기관차원에서 이해할 수 있도록 표 1-1과 같이 재정의 하였다. 연구 데이터 라이프 사이클의 구성은 생산부터 활용 과정에서 주제에 따라 연구 데이터 생산자, 관리부터, 사용자 그리고 리포지시스템을 고려하여 그 단계별 역할 및 내용들을 반영하였다.

- 계획(Plan) : 데이터 생성 계획 및 관리 계획을 포함하여 DMP(Data Management Plan)를 작성하여 향후 데이터 관리를 위해 준비하고 활용할 수 있도록 한다. DMP에는 데이터 분류 항목과 더불어 데이터의 종류 및 용량 등을 포함할 수 있다.

- 데이터 생산(Data production) : 원시데이터를 생산하는 단계이며, 이 단계에서는 수집된 형태의 표준을 유지하기 위해 양식을 개발하고 수집 과정에서 표준을 참조하여 적용한다.

- 처리(Process) : 원시데이터 혹은 다른 소스로부터 최종 결과를 도출하기 위한 2차 데이터 생산 과정이며, 이 또한 연구 결과 검증에 요구되는 경우 연구 데이터로서의 역할을 할 수 있다. 이 과정에서는 데이터 포맷을 통일화한다.

- 제출(Submit) 데이터를 리포지터리에 제출하기 위해 데이터 패키징 규칙을 적용하여 데이터를 구성한다. 패키지에 포함되는 내용으로 데이터, 설명 및 라이선스 등이 포함될 수 있으며, 설명에서는 데이터 인용을 위한 정보, 데이터의 생성 및 작성 과정, 특수한 경우 이용 방법 등이 포함될 수 있다. 제출 과정에서는 리포지터리 상에 표준화된 메타 데이터 양식을 통해 제출된다.

- 보존(Preserve) 이 과정에서는 개별 연구자가 제출한 데이터를 검수(메타데이터의 적절한 입력, 제출 데이터의 누락 여부)에 따라 보완, 보존, 활용 처리가 진행된다.

- 장기보존(Long-Term Preserve) 데이터 선별 기준에 따라 장기 보존 여부를 결정한

다. 데이터 선별 기준은 데이터의 고유성, 가치(예를 들어 중요한 연구 결과에 대한 직접적인 데이터 관련성) 등이 될 수 있다.

- 배포(Distribute) 제출된 데이터에 대한 라이선스 정보와 함께 배포 과정이 진행되며, 리포지터리 시스템의 기능에 따라 라이선스 정보를 자동 생성하여 배포될 수 있다.

- 이용(Use) 사용자는 다운로드받은 데이터의 이용 라이선스에 따라 이용 범위 등을 확인하고, 인용 및 데이터 활용 범위를 고려하여 활용한다.

표 1-1. 연구 데이터 라이프 사이클에서 수행 주제별 역할

단계	생산자	관리부서	이용자	리포지터리
계획 (Plan)	DMP 작성	데이터 관리 계획(DMP) 수립을 지원하고, DMP 가이드라인을 제공		
데이터생산 (Produce)	표준양식 개발 및 표준 적용한 데이터 생산	수집 양식과 데이터 포맷의 표준화 지원		
처리 (process)	데이터포맷 통일화	데이터 포맷의 표준화 지원		
제출 (Submit)	메타데이터 작성 데이터 패키징를 적용	데이터 패키징 규칙 등 가이드 제공		데이터 접수, 메타데이터 양식을 통해 데이터 등록
보존 (Preserve)		데이터 검수를 통한, 보완, 보존 및 활용 처리		검수 프로세스 시스템적 구현
장기보존 (Long-Term Preserve)		데이터 선별 기준 수립, 영구보존을 위한 데이터 관리		선별된 데이터의 영구적 보존
배포 (Distribute)				데이터 라이선스 정보와 함께 검색 가능한 형태로 배포

이용 (Use)			데이터의 라이선스 정보를 확인 및 준수	
-------------	--	--	-----------------------	--

연구 데이터 라이프 사이클은 연구 데이터의 수집, 관리 및 활용을 보다 체계적으로 할 수 있는 프레임워크 역할을 하고 있으며, 연구 데이터의 수집 및 관리는 연구사업차원, 기관차원에서 라이프사이클이 다를 수 있다. 기관차원에서는 연구프로젝트 종료시점을 기준으로 정제된 데이터를 장기 보관이 필요하며, 프로젝트 진행 차원에서는 데이터 수집과정에서 데이터의 제출 및 관리가 필요하다. 따라서 연구기관에서는 연구 데이터 정책에 적합한 모델을 검토하여 수정할 필요가 있다.

본 지침서에서 소개된 DataONE는 데이터를 제출하고 관리하는 시점 측면에서 연구기관 차원에서 데이터를 관리하는데 보다 적합할 수 있으며 데이터를 관리하는 기관마다 내부적으로 적용시 부분적으로 수정 고려될 수 있다.

1.3. 연구 데이터의 관리와 활용

연구 데이터는 다양한 수집 및 생산방법, 내용적 형식 등을 가지고 있다. 또한 데이터의 특성상 국가 및 기관적 차원의 보안, 연구자의 현재시점의 지식재산권, 미래의 지식재산 발현기회, 데이터의 가치, 재활용성, 아카이빙시의 데이터 용량에 따른 비용 문제들 등 다양한 특징들이 내포되어 있기 때문에 초기 정책 수립시 연구 데이터 라이프 사이클을 고려하여 보다 정밀한 설계가 필요하다. 따라서 연구 데이터에 대한 효율적 관리를 위한 고려요소들은 다음과 같이 분야별 구성될 수 있다.

표준화 측면 : 표준의 개념은 협의적으로 기관차원에서부터 단체 및 국가 표준 등이 있으며 이는 데이터의 적용범위에 따라 그 표준화 수준을 결정할 수 있으며 표준개발은 다른 표준과 상호 참조관계를 유지해야 하기 때문에 유관 표준에 대한 검토 및 참조가 필요하다. 표준의 적용은 데이터 라이프 사이클을 고려하여 생산부터 관리 및 활용에 이르는 과정에서 적용을 통해 정보의 접근성, 이용의 효율성 및 이해도를 높이고 나아가 연계시스템간의 상호 운용성을 증대시키는 역할을 할 수 있다.

정책적 측면 : 데이터 생산부터 관리 및 활용에 이르는 전과정에서 요구되는 정책은 투명성과 합리성이 유지되어야 한다. 이러한 정책의 유형으로는 표준화, DMP 와 같은 데이터 관리, 지식재산권 보호, 데이터 활용, 데이터 아카이빙에 따른 정책이 있다.

기술적 측면 : 연구 데이터는 궁극적으로 컴퓨터에서 처리된 결과 형태이기 때문에 정보통신 분야의 기술에 의존적 형식으로 진행된다. 표준화 측면에서는 메타데이터 및 시스템을 표현하고 기술하기 위한 UML, XML ERD 등과 같은 데이터 및 데이터베이스 모델링 기술이 요구되며, 데이터 아카이빙에 따른 리포지토리 관리 및 활용은 주로 오픈 플랫폼 시스템을 도입을 고려할 수 있으며 이에 따른 오픈소스 기술들은 다음과 같다.

- CKAN (Comprehensive Knowledge Archive Network): 오픈소스 데이터 포털 플랫폼으로, 데이터 관리와 공개를 위한 종합적인 기능을 제공하며 메타데이터 관리, 데이터 저장 및 접근관리, 데이터 공유 및 협업, 데이터 큐레이션 및 아카이빙 기능들을 포함한다.

- DCAT (Data Catalog Vocabulary) : DCAT은 데이터 카탈로그를 위한 RDF 기반의 표준 메타데이터 모델로 DCAT은 CKAN의 메타데이터 스키마로도 사용되며, 두 기술은 상호 보완적인 관계를 갖는다.

- DSpace : 디지털 리포지토리 플랫폼으로, 연구 데이터와 학술 출판물 등을 관리할 수 있으며, 메타데이터 관리, 데이터 저장 및 공유, 사용자 권한 관리 등의 기능을 포함한다.

- Fedora Commons : 디지털 콘텐츠 관리를 위한 오픈소스 플랫폼으로 메타데이터 관리, 데이터 버전 관리, 데이터 보안 등의 기능을 포함한다.

- EPrints 오픈액세스 리포지토리 플랫폼으로, 연구논문과 데이터를 관리를 목적으로 한다. 주요 특징으로는 메타데이터 관리, 데이터 공유 및 협업 기능을 포함한다.

- Dataverse : 연구 데이터 관리와 공유를 위한 오픈소스 플랫폼으로, 메타데이터 관리, 데이터 파일 관리, 데이터 공개 및 공유 기능을 포함한다.

- Zenodo : CERN에서 개발한 연구 데이터 관리 및 공유 플랫폼으로, 다양한 데이터 포맷 지원, DOI 할당, 데이터 공개 및 검색 기능을 포함한다.

* 오픈플랫폼 이외 한국의 과학기술연구회 소속의 에서 주고 사용되고 있는 DataON 이 있다.

2. 표준 개요 및 현황

2.1 개요

표준이란 물건, 개념, 방법, 절차에 대한 규격, 준칙, 준거를 의미하며, 표준화란 표준 규격, 준칙, 준거를 세우기 위해 진행되는 과정으로 표준화 대상에 대하여 단순화, 통일화, 규격화 등을 수행하는 과정을 의미한다. 표준화의 궁극적인 목적은 표준화 대상에 대한 활용성을 높이는 데 있으며 세부적으로 상호 운용성과 이해수준을 향상시키고 처리비용을 저감할 수 있다.

연구 데이터의 관리 활용적 측면에서 개발이 요구되는 표준은 필요에 따라 연구 데이터 수집 및 절차, 데이터 수집 양식(필드 정의 및 포맷), 메타데이터 등을 고려할 수 있다. 이러한 표준들은 데이터 라이프 사이클과정에서 다양한 측면에서 긍정적인 효과를 얻을 수 있다. 수집과정에서의 표준은 데이터의 정확성, 일관성을 통해 신뢰성을 높일 수 있으며, 관리차원에서는 데이터 저장 검색을 용이하게 하고, 이질적 플랫폼간의 상호 운용성을 보장할 수 있다. 활용 단계에서의 표준화된 데이터는 데이터의 이해를 높이며 데이터의 통합을 용이하게 한다. 이러한 필요성에 의해 표준을 개발하는데 있어서 표준화 대상에 대한 식별도 필요하지만 표준의 적용대상, 진행정도에 따른 문서에 대한 명명법, 표준의 적용방법, 등에 대한 이해가 필요하다.

표준은 그 적용 및 참여 범위에 따라 국제, 국가, 단체 및 기관표준으로 구분될 수 있다(표 2-1). 국제표준은 전세계 대부분의 국가가 참여하여 합의를 도출한 표준으로 ITU, ISO, IEC 등에서 제정한 표준이다. 지역표준은 어느 특정지역에 소속된 국가들이 합의하여 도출한 표준으로 예로 유럽 전기통신표준협회(ETSI)가 있다. 국가표준은 국가 내의 이해당사자들 끼리 합의한 표준으로 우리나라의 경우 한국정보통신표준(KICS), 산업표준(KS)이 있다. 단체표준은 국가 내의 표준화 단체에서 제정한 표준으로 TTA표준이 이에 해당한다. 사내 표준은 법인 내에 효율적 업무를 처리하기 위해 제정하는 표준이며, 연구기관에서 자체적으로 데이터를 관리하고 활용할 목적으로 표준화를 진행하는 경우 사내표준에 해당한다.

표 2-1. 표준 참여 범위에 따른 분류

표준 종류	정의
국제표준 International Standard	전 세계 대부분의 국가가 참여하여 합의를 도출한 표준
지역표준 Regional Standard	특정 지역에 소속된 국가들이 합의하여 도출한 표준
국가표준 National Standard	국가 내의 이해당사자들이 합의한 표준
단체표준 Industry Standard	특정 산업 내의 표준 또는 협회에서 제정한 표준
사내표준 Internal Standard	법인 내에 효율적 업무를 처리하기 위해 제정한 표준

진행정도에 따른 표준문서에 대한 지칭이 달라진다. 표준은 이해당사자들과의 합의를 통해 완성되어 지며, 표준화 기구인 ISO(International Organization for Standardization)와 ITU(International Telecommunication Union)에서는 사용되는 용어는 상호 다르지만 의미는 상호 호환된다(표 2-2).

표 2-2. 표준화 진행 단계에 따른 ISO 및 ITU의 표준문서의 명명법

진행 단계	ISO	ITU	설명
초기	표준초안 (Draft proposal)	기고서 (Contribution)	표준개발을 목적으로 제출된 제안서
중간	표준안 (Draft Standard)	권고안 (Draft Recommendation)	합의된 표준안
최종	표준 (Standard)	권고 (Recommendation)	표준화 절차에 의해 심의 채택된 표준

표준의 구현 정도에 따른 분류로 그 종류는 기본표준, 기능표준, 이용자표준, 시험규격으로 나뉜다. 기본표준(Base Standard)은 ISO, ITU 등 국제 표준화 기구에서 제정한 것으로 구현의 기본이 되는 표준이다. 기능표준(Functional Standard)은 기본표준을 바탕으로 기능을 구현하기 위한 세부사항 까지 구체화된 표준이다. 이용자표준(User

Profile)은 특정분야에서 이용할 목적으로 제정된 표준으로, MAP(Manufacturing Automation Protocol), Top(Technical and Office Protocol) 등이 있다. 시험규격(Test Specification)은 표준구현 제품의 적합여부를 확인하는 규격이다.

표준의 적용방법에 따른 분류로 강제표준과 권고표준이 있다. 강제표준(Mandatory)은 의무적으로 지켜야할 표준으로 적용분야가 통신망과 같이 안정성, 신뢰성이 필요한 분야에서 법적 고시를 제정하여 적용한다. 권고표준(Recommendation)은 표준을 사용할 때의 잠재적 가치를 고려하여 권장하는 표준으로 한국정보통신표준(KICS)나 산업표준(KS)등이 이에 해당한다. 또한 표준이 공식적인 제정기구에 의해 개발된 경우 공식표준(de jure standard)이라 하며, 이외에 널리 표준처럼 사용되고 있는 규격은 사실 표준(de facto standard)이라고 한다.

표준화 단계는 아이디어 제안과 실험을 통해 표준초안을 작성하는 기초연구단계에서 시작하여, 이해당사자들이 모여 논의하고 합의를 도출하는 토의 및 합의 단계, 이를 바탕으로 제품을 실제로 구현하는 구현단계, 구현된 제품의 적합성과 상호 운용성을 검증하는 시험단계, 그리고 표준화된 제품을 실제로 사용하는 활용단계를 거친다. 마지막으로, 기술적 발전과 이용 환경의 변화에 맞추어 표준을 수정하거나 갱신하는 유지보수 단계까지 진행된다(표 2-3).

기초연구단계는 아이디어를 제안하고 이를 실험적으로 입증하는 초기 단계로, 이 단계에서 표준초안이 작성된다. 연구자는 표준화하고자 하는 기술이나 개념에 대한 아이디어를 제시하고, 실험과 검증을 통해 그 타당성을 입증하게 된다. 그 후 작성된 표준초안은 관련 표준화 기구나 단체에 제출되어 표준화 절차가 본격적으로 시작된다.

토의 및 합의 단계는 표준초안에 대해 이해당사자들이 모여 논의하고 합의점을 도출하는 과정이다. 표준화 기구나 협회는 토론의 장을 마련하며, 국내에서는 주로 정보통신기술협회(TTA)가 이러한 역할을 담당한다. 국제적으로는 ITU(International Telecommunication Union), ISO(International Organization for Standardization), IEC(International Electrotechnical Commission)와 같은 표준화 기구들이 이 과정을 주도한다. 이 단계에서 각 이해당사자들이 모여 기술적, 경제적, 법적 측면에서 논의하여 표준화 초안을 수정하고 최종 합의를 이끌어낸다.

구현단계는 협의된 표준을 기반으로 제품을 실제로 구현하는 과정이다. 표준이

제정되면, 그에 따라 즉시 제품의 개발 및 구현이 이루어진다. 이 단계에서 표준이 실질적으로 적용되며, 제품 개발자는 표준의 요구 사항을 충족하는 제품을 개발하게 된다.

시험단계는 구현된 제품이 표준에 적합한지 여부를 검증하는 과정이다. 이 과정에서는 적합성 시험(Conformance Test)을 통해 제품이 표준의 요구 사항을 충족하는지를 확인하고, 상호운용성 시험(Interoperability Test)을 통해 다양한 제품 간 서비스의 상호 운용성이 보장되는지를 평가한다. 이러한 시험은 주로 제품 개발자가 주도하며, 표준 적합성과 상호 운용성을 검증함으로써 시장에 적합한 제품을 출시할 수 있게 된다.

활용단계는 표준에 따라 개발된 제품이 실제로 사용되는 단계이다. 이 단계에서 소비자나 사용자는 표준화된 제품을 활용하게 되며, 이를 통해 제품의 일관성, 상호 운용성, 품질을 보장받을 수 있다. 표준에 맞추어 제작된 제품들은 시장에서 신뢰를 얻으며, 널리 사용되기 시작한다.

마지막으로 유지보수 단계는 기술 발전과 이용 환경의 변화에 맞추어 표준을 수정하거나 갱신하는 과정이다. 표준은 시간이 지남에 따라 새로운 기술이 등장하거나 이용 환경이 변화함에 따라 더 이상 적합하지 않을 수 있다. 이러한 경우 표준은 수정되거나 갱신되며, 더 이상 이용되지 않는 표준은 폐기된다. 유지보수 단계는 표준의 지속적인 적합성을 보장하기 위한 필수 과정으로, 기술적 진보와 표준의 이용도가 밀접하게 연계되어 있다.

표 2-3. 표준화 단계별 문서

표준화 단계	설명	표준화 문서 단계
기초연구단계 (Initial Research)	표준화 아이디어 제안 및 초안 작성	표준초안 (Draft Proposal)
토의 및 합의 단계 (Discussion and Agreement)	이해당사자 간 논의 및 표준 초안 확정	표준안 (Draft Standard)
구현 및 시험 단계 (Implementation and Testing)	표준 적용 및 적합성 여부 검증	-
심의 및 채택 단계 (Review and Adoption)	표준화 절차에 따라 표준을 심의하고 채택	표준 (Standard)

활용단계 (Utilization)	표준화된 양식이나 제품을 실제로 사용	-
유지보수 (Maintenance)	표준의 수정 및 갱신	개정 (Revision)

2.2 표준제도 현황

연구 데이터의 효율적인 관리와 활용은 단순한 기술적 문제를 넘어선 법적, 제도적 기반 위에서 이루어져야 한다. 이를 위해 국가 차원에서 마련된 표준과 국제적인 기준, 그리고 연구 데이터의 보존과 재사용을 촉진하기 위한 다양한 법적 제도가 중요한 역할을 한다. 연구 데이터 관리에 있어 이러한 법과 제도의 역할을 명확히 이해하고 적용하는 것은 데이터의 신뢰성과 상호 운용성을 보장하는 데 필수적이다.

연구 데이터 관리의 근본적인 법적 기초는 국가표준기본법과 그 시행령에서 비롯된다. 이 법은 한국 내 연구기관 및 공공기관에서 사용하는 데이터의 표준화를 위한 틀을 제공하며, 연구 데이터의 관리, 보존, 재사용을 위한 기본적인 기준을 제시한다. 국가표준기본법은 국제 표준과의 연계성을 강조하며, 특히 국제단위계(SI)와 같은 세계적으로 통용되는 표준을 사용함으로써 데이터의 일관성을 유지하는 데 도움을 준다. 이를 통해 연구 데이터는 연구기관 간의 협업과 데이터 교환에서 발생할 수 있는 문제를 최소화할 수 있다.

국가표준기본법 시행령은 보다 구체적인 표준 적용 방식을 규정하고 있으며, 이를 통해 연구기관은 연구 데이터 관리의 전 과정에서 법적 요구 사항을 준수해야 한다. 예를 들어, 연구 데이터의 수집, 저장, 보존, 그리고 활용에 이르기까지 각 단계에서 데이터의 일관성과 재사용성을 보장할 수 있도록 법적 틀을 마련한다.

연구 데이터 관리에서 국제 표준의 적용이 요구된다. 특히 연구 데이터가 국제적으로 공유되고 재사용되기 위해서는 ISO(International Organization for Standardization), IEC(International Electrotechnical Commission), ITU(International Telecommunication Union) 등에서 제정한 국제 표준을 따르는 것이 필요하다. ISO 표준은 연구 데이터 관리의 다양한 측면에서 국제적인 기준을 제공하며, 이를 준수함으로써

데이터의 상호 운용성과 재사용 가능성을 보장할 수 있다. IEC와 ITU 역시 정보통신과 전자 기술을 기반으로 한 데이터 관리 표준을 제공하며, 특히 데이터 전송 및 통신 과정에서 발생할 수 있는 문제를 해결하기 위한 국제적 기준을 제시한다.

연구 데이터가 국경을 넘어 활용되는 상황에서 국제 표준을 따르는 것은 필수적이며, 이러한 표준은 한국의 국가표준제도와 연결되어 있다. 이를 통해 국내 연구기관은 국제적인 협업 및 데이터 공유가 필요한 연구에서도 법적 요구 사항을 충족할 수 있다.

한국 내에서 연구 데이터는 TTA(Telecommunications Technology Association) 표준을 중심으로 참고할 수 있다. TTA 표준은 정보통신 기술을 기반으로 한 연구 데이터 관리의 표준을 제공하며, 데이터의 상호 운용성을 강조한다. 이를 통해 연구 데이터는 다양한 정보통신 기술 환경에서 일관되게 관리되고, 연구기관 간 데이터 교환을 원활하게 할 수 있도록 지원한다. 한편, KS(Korea Industrial Standard)는 산업 표준으로 연구 데이터 관리와는 직접적인 연관성이 높지 않지만, 일부 분야에서는 연구 데이터의 품질 보증과 일관성을 위한 참고 자료로 사용될 수 있다. 연구 데이터의 관리에는 KS보다는 ISO와 같은 국제 표준이 더 적합하며, 국제 표준을 통해 연구 데이터의 신뢰성과 재사용성을 더욱 높일 수 있다.

연구 데이터 관리계획(DMP)은 연구 데이터 관리의 필수적인 요소로 자리 잡고 있다. DMP는 연구 데이터의 생성, 보존, 활용에 대한 구체적인 계획을 수립하는 문서로, 연구자가 데이터의 전 주기에서 이를 어떻게 관리할 것인지를 명확히 정의한다. 국가연구개발정보처리기준에 따르면, 연구 데이터 관리계획은 국가 연구개발 과제에서 필수적으로 요구되며, 이를 통해 연구 데이터의 투명성, 보존성, 재사용성이 보장된다.

연구 데이터는 지식재산권의 보호 대상이 될 수 있으며, 저작권법, 지식재산권법, 그리고 국가연구개발혁신법과 밀접한 관련이 있다. 연구 데이터는 연구자의 지적 재산이므로 데이터의 소유권 및 사용권이 명확히 규정되어야 하며, 특히 시각 자료인 그림과 표는 저작권법에 의해 보호된다. 데이터의 재사용 및 공유 시 저작권 및 지식재산권 문제가 발생할 수 있으므로, 연구자는 저작물의 올바른 인용과 사용을 위해 법적 절차를 준수해야 한다. 그러나 많은 연구자들이 소유권과 권리에 대한 인식이 부족하여, 데이터의 부적절한 사용과 저작권 침해가 발생할 위험이 크다. 따라서 연구자들에게는 권리 보호와 관련된 교육 및 가이드라인이 필요하다.

연구 데이터는 공간정보 및 지질자원과 관련된 데이터를 포함할 때, 관련 법적 제도와 표준의 적용을 받는다. 공간정보 3법은 공간정보와 관련된 연구 데이터의 관리와 보존에 있어 매우 중요한 역할을 하며, 공간정보 기본법, 공간정보산업 진흥법, 공간정보의 구축 및 관리 등에 관한 법률로 구성되어 있다. 이러한 법들은 공간정보와 관련된 데이터를 효과적으로 관리하고 보존하기 위한 법적 근거를 제공하며, 이를 통해 지리적 데이터의 일관성과 활용성을 높일 수 있다. 각 부처별로 마련된 지침과 규정도 공간정보 데이터 관리에 있어 중요한 역할을 한다. 예를 들어, 국토교통부, 환경부, 국토지리정보원 등은 공간정보와 관련된 연구 데이터를 효과적으로 관리하기 위한 지침을 제공한다.

연구 데이터를 효율적으로 관리하기 위해서는 데이터베이스(DB)의 표준화가 필수적이다. 전자정부법 제50조와 공공데이터 제공 및 이용 활성화에 관한 법률 제23조는 공공기관이 관리하는 데이터베이스의 표준화를 명시하고 있으며, 이를 통해 연구 데이터는 체계적으로 관리되고 활용될 수 있다. 이러한 법적 기반을 통해 연구 데이터는 연구기관 간 통합 관리가 가능해지며, 데이터의 상호 운용성도 보장된다. 전자정부법은 공공기관이 구축하는 데이터베이스가 일정한 표준에 맞추어 관리되도록 하여, 연구 데이터의 신뢰성을 유지하고 재사용 가능성을 높인다.

데이터의 표준화는 데이터베이스 관리시스템(DBMS)에서 데이터를 효율적으로 입력, 관리, 활용하기 위한 필수적인 과정이다. 데이터베이스 설계 단계에서 표준화를 적용하면 자료의 중복성을 최소화하고 무결성을 유지하며, 검색 속도 향상과 민감 데이터의 보안을 강화할 수 있다. 이러한 설계는 구조화된 테이블과 테이블 간의 관계를 설정하여 특정 목적에 맞게 데이터베이스를 효율적으로 관리할 수 있게 한다.

연구 데이터의 DB 구축은 단순한 데이터 저장을 넘어서 데이터와 시스템 간의 융합된 활용을 고려해야 하며, 구축 이후에는 상호 운용성을 확보하는 것이 중요하다. 공공기관에서는 국가 데이터베이스 구축 시 「전자정부법」 제50조 및 시행령 제59조, 그리고 공공데이터 제공 및 이용 활성화에 관한 법률 제23조에 따라 데이터베이스의 표준화를 준수해야 한다. 또한, ‘공공기관 데이터베이스 표준화 지침’은 표준화 과정에서 고려해야 할 세부사항들을 정의하고 있어, 공공기관은 이를 바탕으로 데이터베이스

를 일관되게 관리할 수 있다.

데이터베이스 구축 과정에는 일련의 단계별 산출물이 포함되며, 이는 「전자정부법」 제8조 제1항에 근거한다. 주요 산출물은 다음과 같다:

- 엔터티 정의서: 엔터티(실체 및 객체)는 데이터베이스 테이블을 식별하기 위해 정의됨

- 속성 정의서: 각 엔터티에 포함된 속성의 이름과 설명이 포함됨

- 논리 데이터 모델 다이어그램: 엔터티 간 관계를 나타낸 정의서

- 테이블 정의서: 논리 모델에서 도출된 테이블에 대한 세부적인 정의

- 컬럼 정의서: 테이블 내 컬럼의 이름과 설명이 포함된 정의서

- 물리 데이터 모델 다이어그램: 실제 데이터베이스 테이블이 물리적으로 구현된 다이어그램

- 데이터베이스 정의서: 테이블 세트가 포함된 데이터베이스의 정의서로, DBMS 종류 및 버전이 포함됨

- 기타 산출물: 용어 개정 내용 등

표 2-4. 공공DB 표준화 점검내용(공공기관의 데이터베이스 관리지침)

구분	세부 점검 사항
데이터 표준관리체계	<ul style="list-style-type: none"> ◦ 기관의 데이터베이스 표준화를 위한 내부규정 마련 ◦ 기관의 데이터베이스 표준화 담당자 지정 ◦ 표준화 담당자의 역할과 책임 정의 ◦ 기관의 데이터베이스 표준화 절차 수립
데이터 표준수립	<ul style="list-style-type: none"> ◦ 기관의 행정코드 정의 및 코드정의서 작성 관리 ◦ 기관 표준어의 전 국가표준에 중복된 의미의 코드 및 용어가 있는지 검토 수행 ◦ 기관의 표준용어 정의 및 표준용어 정의서 작성 관리 ◦ 기관의 도메인 정의 및 도메인정의서 작성 관리
데이터 표준적용 및 현행화	<ul style="list-style-type: none"> ◦ 신규 데이터베이스의 설계 구축 시 데이터 표준(코드, 용어, 도메인 등)의 정의 ◦ 행정표준코드 관리시스템의 행정표준코드에 대한 제 개정 신청 수행 ◦ 행정표준용어에 대한 제 개정 등 신청 수행

표준화 점검 및 개선	<ul style="list-style-type: none"> ◦ 운영하고 있는 데이터베이스의 표준화 수준에 대한 주기적 점검(표준코드, 표준용어, 도메인표준 적용 수준 등) ◦ 점검 결과 표준화가 미흡한 사항에 대한 개선 계획 수립 추진(차세대 사업, 고도화 사업, 데이터 재구축 사업에 포함 등)

3. 국제단위계 및 적용

3.1 국제단위계

연구 데이터와 조사 탐사 데이터는 측정값과 이에 대응하는 단위를 사용하며, 단위의 표준은 국제단위계(SI: International System of Units)를 이용해 표기된다. 우리나라는 국가표준기본법에 따라 국제단위계의 사용을 의무화하고 있다. 국제단위계는 1960년 국제도량형총회(CGPM)에서 확립되어 과학, 기술 및 산업 분야에서 기본적으로 활용되며, 국제도량형국(BIPM)에서 관리하고 있다. 특히, 2018년 CGPM 제26차 회의에서는 플랑크 상수와 빛의 속력과 같은 상수를 고정하여 새로운 정의가 도입되었으며, 수정된 국제단위는 2019년 5월 20일에 발효되었다. CGPM의 공식 회의는 프랑스어로 진행되며, 영어권에서는 미국식 영어와 영국식 영어가 혼용되어 사용된다(예: metre와 meter). 국제단위계는 표 3-1과 같이 7개의 기본 단위로 구성되며, 기본 단위로부터 물리량을 정의하는 유도 단위가 존재한다. 2019년에는 7개의 기본 단위 중 4개의 기준이 수정되었으며, 이는 시간에 따라 변하는 물질의 양 대신 불변의 물리 상수가 적용된 것이다.

표 3-1 기본단위

기본량		기본단위	
명칭	전형적인 기호	명칭	기호
시간	초 (Second)	t	s
길이	미터 (Meter)	l, x, r 등	m
질량	킬로그램 (Kilogram)	m	kg
전류	암페어 (Ampere)	I, i	A
열역학 온도	켈빈 (Kelvin)	T	K
물질량	몰 (Mole)	n	mol
광도	칸델라 (Candela)	I _v	cd

유도단위는 기본 단위를 거듭제곱하여 곱한 형태로 정의된다. 이 곱의 수치가 1인 경우, 이를 '일관성 있는 유도단위'라고 한다. SI의 기본 단위와 일관성 있는 유도단위는 함께 일관성 있는 SI 단위 세트를 형성한다(표 3-2). 여기서 '일관성 있는'은 수치들 간의 관계가 양의 수치들 간의 관계와 동일하다는 의미이다. 일부 일관성 있는 유도

단위는 특별한 이름을 가진다. 총 표 3-3과 같이 22개의 SI 단위가 특별한 명칭을 가지고 있으며, 이는 7개의 기본 단위와 함께 SI 단위의 핵심을 이룬다. 기타 모든 SI 단위는 이 29개 중 일부의 조합이다. 7개의 기본 단위와 특별한 명칭을 가진 22개의 SI 단위가 모두 7개의 정의 상수로 구성될 수 있다. 실제로, 정의 상수의 단위는 기본 단위와 유도 단위를 포함한다.

표 3-2. 기본단위로 표시된, SI의 일관성 있는 유도단위의 예

유도량	양에 대한 전형적인 기호	기본단위로 표시된 유도단위
넓이	A	m^2
부피	V	m^3
속력, 속도	u	$m s^{-1}$
가속도	a	$m s^{-2}$
파동수, 파수	σ	m^{-1}
밀도, 질량밀도	ρ	$kg m^{-3}$
표면밀도	ρA	$kg m^{-2}$
비(比)부피	v	$m^3 kg^{-1}$
전류 밀도	j	$A m^{-2}$
자기장 세기	H	$A m^{-1}$
물질량 농도	c	$mol m^{-3}$
질량 농도	ρ	$kg m^{-3}$
광휘도	LV	$cd m^{-2}$

표 3-3. 특별한 명칭과 기호를 가진 22개의 SI 유도단위

유도량	단위의 특별한 명칭	기본단위로 표시된 단위	기타 SI 단위로 표시된 단위
평면각	라디안 (Radian)	$rad = m/m$	
입체각	스테라디안 (Steradian)	$sr = m^2/m^2$	
주파수, 진동수	헤르츠 (Hertz)	$Hz = s^{-1}$	
힘	뉴턴 (Newton)	$N = kg m s^{-2}$	N m
압력, 응력	파스칼 (Pascal)	$Pa = kg m^{-1} s^{-2}$	
에너지, 일, 열량	줄 (Joule)	$J = kg m^2 s^{-2}$	J/s
일률, 전력, 복사선속	와트 (Watt)	$W = kg m^2 s^{-3}$	W/A
전하, 전하량	쿨롱 (Coulomb)	$C = A s$	C/V

전위차(마)	볼트 (Volt)	$V = \text{kg m}^2 \text{s}^{-3} \text{A}^{-1}$	V/A
전기용량, 정전용량	패럿 (Farad)	$F = \text{kg}^{-1} \text{m}^{-2} \text{s}^4 \text{A}^2$	A/V
전기저항	옴 (Ohm)	$\Omega = \text{kg m}^2 \text{s}^{-3} \text{A}^{-2}$	V/A
전기전도율	지멘스 (Siemens)	$S = \text{kg}^{-1} \text{m}^{-2} \text{s}^3 \text{A}^2$	A/V
자기선속	웨버 (Weber)	$\text{Wb} = \text{kg m}^2 \text{s}^{-2} \text{A}^{-1}$	Wb/m ²
자기선속밀도	테슬라 (Tesla)	$T = \text{kg s}^{-2} \text{A}^{-1}$	Wb/m ²
인덕턴스	헨리 (Henry)	$H = \text{kg m}^2 \text{s}^{-2} \text{A}^{-2}$	Wb/A
섭씨온도	섭씨도 (Celsius)	$^{\circ}\text{C} = \text{K}$	
광선속	루멘 (Lumen)	$\text{lm} = \text{cd sr}$	cd sr
광조도	칸델라 (Candela)	cd	
방사능	베크렐 (Becquerel)	$\text{Bq} = \text{s}^{-1}$	
흡수선량, 커마	그레이 (Gray)	$\text{Gy} = \text{m}^2 \text{s}^{-2}$	J/kg
선량당량	시버트 (Sievert)	$\text{Sv} = \text{m}^2 \text{s}^{-2}$	J/kg
촉매활성도	카탈 (Catal)	$\text{kat} = \text{mol s}^{-1}$	

10²⁴부터 10⁻²⁴ 범위의 십진 배수와 십진 분수가 SI 단위와 함께 사용될 수 있도록 설정되어 있다. 이러한 배수 및 분수 접두어의 이름과 기호는 표 3-4에 있다. 접두어 기호는 단위 기호와 마찬가지로 주변 문장의 글자체와 관계없이 직립 형태로 사용해야 하며, 접두어 기호와 단위 기호 사이에는 빈칸이 없어야 한다. da(데카), h(헥토), k(킬로)를 제외한 모든 배수 접두어 기호는 대문자로 표기되며, 모든 분수 접두어 기호는 소문자로 표시된다. 모든 접두어 명칭은 문장 서두를 제외하고는 소문자로 작성해야 한다.

국제 단위계는 전 세계에서 통일된 기준을 제공하는 단위 체계다. 국제 단위계는 서로 일관되게 연결되어 있어서, 어떤 양의 값이 주어질 때 단위를 변환할 필요가 없다. 하지만 여전히 국제 단위계가 아닌 다른 단위도 자주 사용하고 있다. 따라서 CIPM (국제 도량형 위원회)은 일부 비국제 단위를 국제 단위계와 함께 사용하는 것을 허용했다. 이들 비국제 단위계 단위는 표 3-5에 정리되어 있다. 이런 비국제 단위계 단위를 사용할 때는 SI의 장점의 일부가 사라질 수 있다. 또한, 국제 단위계 접두어는 몇몇 비국제 단위와 함께 사용할 수 있지만, 시간과 관련된 비국제 단위와는 함께 사용되지 않는다.

표 3-4. SI 접두어

인자	명칭	기호	인자	명칭	기호
10 ¹	데카(deca)	da	10 ⁻¹	데시(dec)	d

10 ²	헥토(hecto)	h	10 ⁻²	센티(centi)	c
10 ³	킬로(kilo)	k	10 ⁻³	밀리(milli)	m
10 ⁶	메가(mega)	M	10 ⁻⁶	마이크로(micro)	μ
10 ⁹	기가(giga)	G	10 ⁻⁹	나노(nano)	n
10 ¹²	테라(tera)	T	10 ⁻¹²	피코(pico)	p
10 ¹⁵	페타(peta)	P	10 ⁻¹⁵	펨토(femto)	f
10 ¹⁸	엑사(exa)	E	10 ⁻¹⁸	아토(atto)	a
10 ²¹	제타(zetta)	Z	10 ⁻²¹	젱토(zepto)	z
10 ²⁴	요타(yotta)	Y	10 ⁻²⁴	옥토(yocto)	y

표 3-5. 국제단위계와 함께 사용되는 것이 용인된 SI 이외의 단위

양	명칭	기호	SI 단위로 나타낸 값
시간	분	min	1 min = 60 s
	시간	h	1 h = 60 min = 3600 s
	일	d	1 d = 24 h = 86 400 s
길이	천문단위	au	1 au = 149 597 870 700 m
평면각 및 위상	도	°	1° = (π/180) rad
	분	'	1' = (1/60)° = (π/10 800) rad
	초	"	1" = (1/60)' = (π/648 000) rad
면적	헥타르	ha	1 ha = 1 hm ² = 10 ⁴ m ²
부피	리터	l, L	1 l = 1 L = 1 dm ³ = 10 ³ cm ³ = 10 ⁻³ m ³
질량	톤	t	1 t = 10 ³ kg
	돌턴	Da	1 Da = 1.660 538 86 (28) × 10 ⁻²⁷ kg
에너지	전자볼트	eV	1 eV = 1.602 176 53 (14) × 10 ⁻¹⁹ J
로그비	네퍼	Np	
	벨	B	
	데시벨	dB	

3.2 국제단위계의 적용

접두어의 이용 - 접두어의 선정

· 국제단위계 측정단위 접두어는 중요하지 않은 숫자와 소수 0을 제거할 수 있도록 배수로써 표시하고자 할 경우 사용되며, 10의 거듭제곱 표시법으로 대신하여 계산 시 사용할 수 있다.

적용 예
1) 길이 (Length) - 잘못된 표기: 12 300 mm - 올바른 표기: 12.3 m
2) 길이 (Length) - 잘못된 표기: 12.3×10^3 m - 올바른 표기: 12.3 km
3) 전류 (Current) - 잘못된 표기: 0.00123 μ A - 올바른 표기: 1.23 nA

· 양을 숫자값과 단위를 사용하여 표시할 경우에는 숫자값이 0.1과 1,000사이에 위치하도록 접두어를 선정 및 사용한다. 여러 가지 형태로 표현되는 경우의 수를 최소화하기 위하여 1,000의 거듭제곱에 해당되는 접두어를 사용한다. 다만, 다음과 같은 경우는 예외로 한다.

적용 예
1) 면적 및 부피를 나타낼 경우에는 헥토(hecto), 데카(deca), 데시(dec) 및 센티(cent)를 사용한다. - 평방 헥토미터(square hectometre): $1 \text{ hm}^2 = 10000 \text{ m}^2$ - 입방 센티미터(cubic centimetre): $1 \text{ cm}^3 = 0.000001 \text{ m}^3$
2) 동일한 종류의 양에 대한 값을 수록한 표에서 또는 하나의 문맥상에서 그 값들을 비교하거나 논의할 경우에는 동일한 단위의 배수를 사용한다. - 길이를 비교할 때, 모두 미터(metre) 단위로 통일하여 사용 (10 m, 50 m, 100 m)
3) 어떤 양을 특정한 분야에서 사용할 경우에는 특정한 배수를 사용한다. - 헥토파스칼(hectopascal)은 고도계 수정치에 사용한다. 예: 1013 hPa - 밀리미터(millimetre)는 기계공학 제도에서 선형 크기에 사용한다. 예: 50 mm

접두어의 이용 - 복합단위에 대한 접두어

· 복합단위는 두 가지 이상의 단위가 조합되어 표현되는 단위로서 하나의 특별한 명

칭으로 표현되지 않는 유도단위를 말한다. 복합단위의 배수를 형성할 경우 하나의 접두어를 사용한다. 이와 같은 경우 접두어는 통상적으로 분자에 있는 단위에 붙인다. 다만 킬로그램(kg)이 분모에 오는 경우에는 예외로 한다.

적용 예
1) 길이 (Length) : mV/mm를 V/m로 변경 - 잘못된 표기: 5 mV/mm - 올바른 표기: 5 mV/mm = 5 V/m 2) 용량 (Capacitance) : kJ/g를 MJ/kg로 변경: - 잘못된 표기: 4 kJ/g - 올바른 표기: 4 kJ/g = 4 MJ/kg

접두어의 이용 - 복합접두어

· 두 개 이상의 국제단위계 측정단위 접두어를 병렬로 연결한 복합접두어는 사용하지 않는다.

적용 예
1) 길이 (Length) : 1 m/㎍를 1 nm로 변경 - 잘못된 표기: 1 m/㎍ - 올바른 표기: 1 nm (나노미터) 2) 길이 (Length) : 1 μμF를 1 pF로 변경: - 잘못된 표기: 1 μμF - 올바른 표기: 1 pF (피코패럿)

· 필요한 값이 접두어의 범위를 벗어날 경우에는 기본단위에 10의 거듭제곱을 사용하여 표시한다.

접두어의 이용 - 단위의 거듭제곱

· 접두어를 포함하고 있는 기호에 붙는 지수는 해당 단위(접두어가 붙은 단위)의 배수 또는 약수가 지수에 의해 표현된 배수만큼 거듭제곱됨을 의미한다.

적용 예
1) 길이 (Length) : $1 \text{ cm}^3 = (10^{-2} \text{ m})^3 = 10^{-6} \text{ m}^3$ 2) 주파수(frequency) : $1 \text{ ns}^{-1} = (10^{-9} \text{ s})^{-1} = 10^9 \text{ s}^{-1}$ 3) 시간당 면적의 변화(square millimeter per second) $1 \text{ mm}^2/\text{s} = (10^{-3} \text{ m})^2/\text{s} = 10^{-6} \text{ m}^2/\text{s}$

모양과 용법 - 단위기호의 표시규칙

· 단위기호의 표시규칙은 본문에 사용된 활자체와 관계없이 로마체(직립체)로 써야 한다.

· 복수의 경우에도 변하지 않으며 문장의 끝에 사용되는 경우를 제외하고 마침표를 붙이지 않는다.

· 문자로 된 단위기호는 단위명칭이 고유명사로부터 유래되어 기호의 첫 문자를 대문자로 사용하는 경우를 제외하고는 소문자로 쓴다. 접두어와 단위기호는 주위의 활자체와 관계없이 규정된 형태를 유지한다.

· 물리량을 완전한 형태로 표현할 경우 수치값과 단위기호 사이를 한 칸 띄어야 한다. 다만, 평면각의 도, 분 및 초와 십씨도의 경우에는 수치와 기호사이에 간격을 두지 않는다.

적용 예
1) 공백 추가하여 표기 : 길이 (mm) - 잘못된 표기: 35mm - 올바른 표기: 35 mm
2) 공백 없이 연속적 표기 : 온도 (°C) - 잘못된 표기: 32 °C - 올바른 표기: 32°C
2) 공백 없이 연속적 표기 : 각도 (도, 분, 초) - 잘못된 표기: 35 ° 28 '27 " - 올바른 표기: 35°28'27"

· 물리량이 형용사적인 원리로 사용될 경우 연결부호(-)를 사용할 수 있다.

적용 예
35мили 필름 : 35-mm film

- 접두어와 단위기호의 사이에는 간격을 두지 않는다.
- 단위를 표기할 경우에는 약어가 아닌 기호를 사용하여야 한다.

적용 예
1) 전류 (Current) - 잘못된 표기: The current is 10 amp. - 올바른 표기: The current is 10 A.
2) 전압 (Voltage) - 잘못된 표기: The voltage is 220 volts. - 올바른 표기: The voltage is 220 V.
3) 전력 (Power) - 잘못된 표기: The light bulb consumes 60 watts. - 올바른 표기: The light bulb consumes 60 W.
4) 길이 (Length) - 잘못된 표기: The distance is 5 meters. - 올바른 표기: The distance is 5 m.
5) 질량 (Mass) - 잘못된 표기: The weight is 70 kilograms. - 올바른 표기: The weight is 70 kg.

모양과 용법 - 단위명칭과 표기 규칙

• 단위명칭의 표기 규칙으로 철자를 모두 써야 할 경우의 단위명칭은 영어의 보통명사로 취급한다. 문장의 첫 부분에서 또는 제목과 같은 사항에 대해 대문자를 쓰는 경우를 제외하고는 단위명칭의 첫 문자는 명칭이 고유명사로부터 유래되어 그 기호를 대문자로 쓰는 경우에도 대문자로 쓰지 않는다.

적용 예
기호가 “N”일 경우에도 “Newton”이라 표기하지 아니하고 “newton”으로 쓴다. - 잘못된 표기: 50 Newton - 올바른 표기: 50 newton (기호: 50 N)

• 복수는 문법상 필요할 경우 사용한다. 예를 들어, henry의 복수로 henries를 사용하는 것처럼 복수는 규칙적으로 구성된다. 다만, 다음과 같은 불규칙 복수형태도 사용된다.

적용 예
1) 조도 (Illuminance) - lux - 단수: The light intensity in the room is 500 lux. - 복수: The light intensities in the rooms are 500 lux and 600 lux.
2) 주파수 (Frequency) - hertz - 단수: The frequency of this sound wave is 50 hertz. - 복수: The frequencies of these sound waves are 50 hertz and 100 hertz.
3) 전기 전도도 (Electrical Conductance) - siemens - 단수: The conductance of this material is 5 siemens. - 복수: The conductances of these materials are 5 siemens and 10 siemens.

· 접두어와 단위명칭의 사이에는 공백이나 연결부호(-)를 사용하지 않는다.

모양과 용법 - 곱셈 및 나눗셈에 의해 구성되는 단위

· 단위명칭에서 곱 형태의 표기

적용 예
1) 단위명칭사이에 공백을 사용하여 표기하며 연결부호(-)를 사용할 수도 있다. - 잘못된 표기: The torque is 10 newtonmetre. - 올바른 표기: The torque is 10 newton metre 또는 The torque is 10 newton-metre.
2) watt hour의 경우에는 공백을 생략한다. - 잘못된 표기: The battery capacity is 500 watt hour. - 올바른 표기: The battery capacity is 500 watthour.

· 단위명칭에서 나누기 형태 표기

적용 예
1) 빗금으로 나타내지 아니하고 단어 per를 사용한다. - 잘못된 표기: The speed of the object is 10 metre/second. - 올바른 표기: The speed of the object is 10 metre per second.

· 단위명칭에서 거듭제곱 형태 표기 형태로 단위명칭 다음에 수식어구 square(제곱) 또는 cubic(세제곱)을 사용한다.

적용 예
1) 속도 (Speed) - 잘못된 표기: The acceleration is 10 metre/second ² . - 올바른 표기: The acceleration is 10 metre per second square. 2) 부피 (Volume) - 잘못된 표기: The volume of the tank is 5 m ³ . - 올바른 표기: The volume of the tank is 5 cubic metres.

- 면적 또는 부피의 경우, 단위명칭 앞에 수식 어구를 사용할 수 있으며 면적 또는 부피에 대한 유도단위에도 적용된다.
- 복잡한 표기의 경우 혼란을 방지하기 위하여 단어 대신에 기호를 사용할 수 있다.

적용 예
1) 면적 (Area) - 문장에서: The area is 150 square millimetres. - 간단한 설명: The area is 150 mm ² . - 수식에서: $A = 150 \text{ mm}^2$ 2) 부피 (Volume) - 문장에서: The volume of the container is 2 cubic metres. - 간단한 설명: The volume is 2 m ³ . - 수식에서: $V = 2 \text{ m}^3$ 3) 면적에 대한 유도 단위 (Derived unit for area) - 문장에서: The power density is 50 watt per square metre. - 간단한 설명: The power density is 50 W/m ² . - 수식에서: Power density = 50 W/m ²

- 단위기호의 수학적 표기 : 곱 형태의 표기

적용 예
1) 토크 (Torque) 의 표현에서 newton metre를 Nm 또는 N·m로 표시 - Torque = 10 Nm 또는 Torque = 10 N·m

· 단위기호의 수학적 표기 : 나누기 형태 표기

적용 예
1) 속도의 경우 다음과 같은 방법을 사용한다. (m/s 또는 $m \cdot s^{-1}$ 또는 $\frac{m}{s}$) - 속도 (Speed) : $10 m/s$ 또는 $10 m \cdot s^{-1}$ 2) 모호성을 피하기 위하여 괄호를 사용하지 않는 경우, 동일한 수식에 두개 이상의 빗금을 사용할 수 없다. - 잘못된 표기(엔탈피, Enthalpy): $500 J/mol/k$ - 올바른 표기(엔탈피, Enthalpy): $500 J/(mol \cdot k)$ 또는 $500 J \cdot mol^{-1} \cdot k^{-1}$ 또는 $(500 J/mol)/k$

· 동일한 수식에 기호와 단위명칭을 조합하여 사용할 수 없다.

적용 예
1) 잘못된 표기 - 에너지 밀도 (Energy Density) - The energy density is 500 joules/kilogram. - The energy density is 500 joules·kg ⁻¹ . 2) 올바른 표기 - 에너지 밀도 (Energy Density) - The energy density is 500 joules per kilogram. - The energy density is 500 J/kg. - The energy density is 500 J·kg ⁻¹ .

숫자 - 숫자의 표현

· 소수점 표시는 온점(.)을 사용하며 반점(,)도 사용할 수 있다. 1 미만의 숫자를 표기할 경우 숫자 다음에 소수점을 찍는다.

· 반점(,)은 숫자를 분리하는데 사용하지 않는다. 숫자는 소수점을 기준으로 하여 좌측 및 우측으로 세 자리씩 묶어 사용하며 각 묶음간의 분리를 위하여 약간의 간격을 둔다. 묶음간의 간격은 대략 문자 'i'의 폭이어야 하며, 인쇄형태에 따라 단어 간에 사용되는 폭이 변동될지라도 간격의 폭은 일정하여야 한다.

적용 예
73 655 7 281 2.567 321 0.133 47

· 숫자의 곱을 나타내는 표시는 십자꼴(\times) 또는 가운데 점을 사용한다. 다만, 가운데

점을 곱의 표시로 사용할 경우에는 동일한 수식에 찍은 온점을 소수점으로 사용할 수 없다.

· 단위기호에 양의 상태에 관한 사항을 표시하기 위하여 문자를 추가할 수 없다. 따라서 “megawatts electrical(힘)” 을 MWe로 표시하거나, “volts ac” 를 Vac로 표시하거나 “kilojoules thermal(에너지)” 를 kJt로 표시할 수 없다. 압력에 관한 경우 절대압력과 계기압력을 구별하기 위해 사용되는 약어 “psia” 와 “psig” 와 동등한 국제단위계를 만들지 않았다. 문맥이 내용상 의심스러운 경우에는 압력에 관한 단어를 적절히 수식어로 사용하여야 한다.

적용 예
“계기압력 13kPa에서(at a gauge pressure of 13 kPa)” 또는, “절대압력 13kPa에서(at an absolute pressure of 13 kPa)”

3.3 지질자원분야에서 적용

지질자원 분야에서 사용되는 단위 체계는 국제단위계(SI)를 기반으로 하거나, 오랫동안 사용되어온 비국제 단위들이 혼재되어 있다. 특히 지질학, 화학, 생물학 등의 분야에서는 관례적으로 비국제 단위가 여전히 사용되는 경우가 많다. 그러나 국제단위계는 법적으로 권장되는 표준 단위 체계이므로, SI 단위와 비국제 단위 간의 관계를 명확히 하거나, 필요한 경우 단위 변환 정보를 제공하는 것이 바람직하다.

표 3-6은 지질자원분야에 대하여 사용하는 대표적인 측정항목(단위)에 대하여 국제단위 여부를 식별하기 위한 목적으로 목록을 정리하였다. 또한 SI로 변환 가능한 단위에 대하여도 별도로 표기 하였다.

표 3-6. 지질자원 연구 데이터에서 주요 사용되는 단위체계

항목	측정 항목	단위	SI 유무 및 변환	설명
암석	연령 측정	Ma	1 Ma = 3.1536×10^{13} 초	과거 지질 시점(백만 년)을 의미하는 단위
		ka	1 ka = 3.1536×10^{10} 초	과거 지질 시점(천 년)을 의미하는 단위
	원소 농도	ppm	1 ppm = 1×10^{-6} kg/kg	백만 분의 일을 의미하는 단위
	화학 조성	wt%	1 wt% = 0.01 kg/kg	중량 백분율을 의미
	시료 질량	kg	SI	킬로그램을 의미하는 단위
	박편 두께	μm	SI	마이크로미터를 의미하는 단위
고생물(화석)	화석 연대	Mya	1 Mya = 3.1536×10^{13} 초 전	과거로부터 백만 년을 의미하는 단위
		Ma	퇴적물 두께 m 또는 초(왕복주시)	왕복주시는 음원에서 발진된 파가 대상 물체 또는 반사계면까지 갔다가 수진기에 돌아올 때까지의 시간
	화석 밀도	g/cm ³	SI	세제곱 센티미터당 그램을 의미
광물(광미, 합성 화합물, 폐자원)	결정 크기	mm	SI	밀리미터를 의미하는 단위
	전기 전도도	S/m	SI	미터당 지멘스를 의미하는 단위
토양	단위중량	kN/m ³	SI	세제곱 미터당 킬로뉴턴을 의미
육상시추 코어	시추 깊이	m	SI	미터를 의미하는 단위
	입도 분포	mm	SI	밀리미터를 의미하는 단위
해저암석 코어	밀도	g/cm ³	SI	세제곱 센티미터당 그램을 의미
	함수율	%	비율 단위	무차원 단위
	전단강도	kPa	SI	킬로파스칼을 의미하는 단위
	음파 전달 속도	m/s	SI	초당 미터를 의미하는 단위
	퇴적층 두께	m	SI	미터를 의미하는 단위
해저퇴적물 코어	전단강도	kPa	SI	킬로파스칼을 의미하는 단위
	대자율	SI	비율 단위	외부 자기장에 의해 자화되는 정도를 의미하는 무차원 값
	투과율	Darcy	1 Darcy =	투과성을 측정하는 단위

			9.869×10^{-13} m ²	
	연령	ka	SI	천 년을 의미하는 단위
	퇴적층 두께	m	SI	미터를 의미하는 단위
지하수/지 표수	이온 농도	mg/L	SI	리터당 밀리그램을 의미하는 단위
	동위원소 비율	‰	1 ‰ = 0.001	천분율을 의미하는 무차원 단 위
	농도	µg/L	SI	리터당 마이크로그램을 의미하 는 단위
	농도	ppb	1 ppb = 1×10^{-9} kg/kg	십억 분의 일을 의미하는 단위
공통	밀도	g/cm ³	SI	세제곱 센티미터당 그램을 의 미하는 단위
	축척	예) 1:1000	비율 단위 (무 차원)	지도에서 사용되는 무차원 단 위
	카메라 해상도	MP	1 MP = 1×10^6 픽셀	메가픽셀을 의미하는 단위
	인쇄 해상도	dpi	인치당 점 수	인치당 점 수를 의미하는 단위
	스캔 해상도	µm	SI	마이크로미터를 의미하는 단위

4. 연구 데이터의 관리와 활용

4.1 FAIR Data

FAIR 데이터 원칙은 2016년에 발표된 FAIR 데이터 원칙은 네덜란드에서 열린 Lorentz Center Workshop에서 처음 제안되었으며, 주로 유럽연구이사회(European Research Council, ERC)와 유럽 오픈 사이언스 클라우드(EOSC) 같은 유럽 내 연구 기구와 이니셔티브에 의해 확산되었다. 이 원칙은 특히 연구 데이터의 상호운용성과 재사용성을 촉진하기 위한 국제적인 데이터 관리 가이드라인으로 자리 잡았으며, 유럽이 이를 주도적으로 채택하고 다양한 프로젝트를 통해 구체화하였다.

이 원칙은 디지털 자산에 대한 검색가능성, 접근성, 상호운용성 및 재사용성을 개선하기 위한 목적을 가지고 있으며, 데이터 자원의 사람의 관리 한계를 벗어나는 규모로 증가하고 있기 때문에 기계가 데이터를 처리할 수 있도록 하는 것을 목표로 한다. 이를 위해 데이터 측면에서는 디지털 자산으로서 재사용성을 높이기 위한 방법으로 사람과 기계가 이해할 수 있는 데이터로 표현하는 원칙으로 정의될 수 있다. 이러한 목적으로 FAIR는 검색성, 접근성, 상호운용성, 자사용성을 기준을 제시하고 있다. 데이터는 가독성 정도에 따라 인간가독, 기계가독, 인간-기계가독형 데이터로 구분될 수 있다(표 4-1).

표 4-1. 가독 주체에 따른 데이터의 분류

인간 가독	인간-기계 가독	기계 가독
이미지 형태의 문서 별도로 인코딩 되지 않은 이미지 지도	CSV, EXCEL 파일 GeoJSON Shapefile	바이너리 데이터 등 센서로부터 수신된 원시데이터

FAIR 데이터에 대한 기준은 각 이니셜(F. A. I. R)에 따라 원칙들이 정의되어 있다. 각 원칙에서 사용되는 용어에서 주로 메타데이터에 적용되는 내용이나 데이터와 메타데이터의 구분이 명확치 않을 때 (메타) 데이터로 정의되어 있으며, 데이터에만 해당하는 경우에는 데이터로 정의하여 설명되어 있다.

가. 찾기 가능한 (Findable) : 데이터를 (재)사용하는 첫 번째 단계는 데이터를 찾는 것이

다. 메타데이터와 데이터는 사람과 컴퓨터 모두 쉽게 찾을 수 있어야 한다. 기계가 읽을 수 있는 메타데이터는 데이터 세트와 서비스를 자동으로 검색하는 데 필수적이므로 Fair data 원칙 프로세스의 필수 구성 요소이다.

F1. (메타)데이터에는 전 세계적으로 고유하고 지속적인 식별자가 있어야 한다.

F2. 데이터는 풍부한 메타데이터로 설명되어야 한다. (아래 R1에 정의됨)

F3. 메타데이터는 묘사하려는 데이터의 식별자를 분명하고 명시적으로 포함되어야 한다.

F4. (메타) 데이터는 검색 가능한 자원에 등록되거나 인덱스 되어야 한다.

나. 접근 가능한 : (Accessible) : 사용자가 필요한 데이터를 찾으면 인증 및 권한을 포함한 해당 데이터의 액세스 방법을 알 필요가 있다.

A1. (메타)데이터는 표준화된 통신 프로토콜을 사용하여 식별자로 검색 가능해야 한다.

A1.1 프로토콜은 개방적이고 무료이며 보편적으로 구현할 수 있어야 한다.

A1.2 프로토콜은 필요한 경우 인증 및 권한 부여 절차를 허용해야 한다.

A2. 메타데이터는 해당 데이터를 더 이상 사용할 수 없어도 액세스할 수 있어야 한다.

다. 상호 운용 가능한 : (Interoperable)

I1. (메타)데이터의 지식 표현을 위해 잘 형성되고, 접근 가능하며, 공유되며, 광범위하게 적용 가능한 설명 언어를 사용합니다.

I2. (메타)데이터는 FAIR 원칙을 따르는 어휘를 사용합니다.

I3. (메타)데이터에는 다른 (메타)데이터에 대한 식별 가능한 참조 정보가 포함되어야 한다.

라. 재사용 가능한 : (Reusable)

R1. 메타(데이터)에는 정확한 관련 속성이 풍부해야 합니다.

R1.1 (메타)데이터는 명확하고 액세스 가능한 데이터 사용 라이선스와 함께 게시됩니다.

R1.2 (메타)데이터는 기록에 연결됩니다.

R1.3 (메타)데이터는 규율에 따라 커뮤니티 표준을 충족합니다.

4.2 FAIRsFAIR 프로젝트

FAIRsFAIR는 유럽연합의 Horizon 2020 프로그램의 일환으로 시작된 대규모 프로젝트로, 이 프로젝트는 데이터 관리와 관련된 표준, 도구, 교육, 정책 등을 개선하여 연구 커뮤니티가 FAIR 원칙을 더 쉽게 적용할 수 있도록 돕는 것을 목표로 하고 있다. 따라서 개념적 원칙 수준에 머무는 것이 아니라 실제로 FAIR 원칙들을 실행할 수 있도록 정책, 교육, SW 도구 등을 제공하고 있다.

FAIR 원칙에 따라 FAIRsFAIR Data Object Assessment Metrics는 연구 데이터의 생애 주기 동안 FAIR 원칙(검색 가능, 접근 가능, 상호운용 가능, 재사용 가능)을 준수할 수 있도록 평가 기준을 제공하는 데 중점을 둔다. 여기서 연구 데이터 객체라는 용어는 데이터 세트로서, 데이터, 메타데이터, 정책 및 절차와 같은 문서들로 구성된다. 해당 명세서에서는 연구 데이터 객체가 얼마나 FAIR한지 평가할 수 있는 기준을 제시하고 있다. 이 평가지표는 총 4번에 걸쳐 갱신되었는데 최종 버전에서는 17가지의 평가요소들을 포함한다.

이 평가 요소들의 식별자는 4개의 구성 요소로 나누어져 있으며, 각각의 식별자는 특정 평가 기준을 나타낸다. 예를 들어, FsF-F1-01D라는 식별자는 다음과 같은 의미로 분해할 수 있다:

- FsF: FairsFair 프로젝트를 의미한다.
- F1: FAIR 원칙의 특정 항목(예: F1, 데이터를 쉽게 찾을 수 있음)을 나타낸다.
- 01: 해당 평가 요소의 로컬 ID를 의미한다.
- D: 평가할 리소스 유형(데이터 또는 메타데이터)을 나타낸다.

이 항목들은 FAIRsFAIR Data Object Assessment Metrics에서 사용되는 15가지 평가 지표를 나타내며, 연구 데이터의 FAIR 원칙 준수 여부를 평가하는 데 사용된다. 각 항목은 식별자와 해당 평가 기준으로 구성되어 있다. 다음은 각 항목에 대한 설명이다:

FsF-F1-01D: 데이터는 전역적으로 고유한 식별자를 할당받아야 한다.

FsF-F1-02D: 데이터는 영구적 식별자를 할당받아야 한다.

FsF-F2-01M: 메타데이터는 데이터를 찾을 수 있도록 설명적 핵심 요소(창작자, 제목, 데이터 식별자, 발행자, 발행일, 요약, 키워드)를 포함해야 한다.

FsF-F3-01M: 메타데이터는 설명하는 데이터의 식별자를 포함해야 한다.

FsF-F4-01M: 메타데이터는 기계에 의해 검색 가능한 방식으로 제공되어야 한다.

FsF-A1-01M: 메타데이터는 데이터의 접근 수준과 조건을 포함해야 한다.

FsF-A1-02M: 메타데이터는 표준화된 통신 프로토콜을 통해 접근 가능해야 한다.

FsF-A1-03D: 데이터는 표준화된 통신 프로토콜을 통해 접근 가능해야 한다.

FsF-A2-01M: 데이터가 더 이상 이용할 수 없는 경우에도 메타데이터는 계속 제공되어야 한다.

FsF-I1-01M: 메타데이터는 공식 지식 표현 언어를 사용하여 표현되어야 한다.

FsF-I2-01M: 메타데이터는 시맨틱 리소스를 사용해야 한다.

FsF-I3-01M: 메타데이터는 데이터와 관련된 엔터티 간의 연결 링크를 포함해야 한다.

FsF-R1-01MD: 메타데이터는 데이터의 내용을 명시해야 한다.

FsF-R1.1-01M: 메타데이터는 데이터 재사용에 대한 라이선스 정보를 포함해야 한다.

FsF-R1.2-01M: 메타데이터는 데이터의 출처 정보(Provenance Information)를 포함해야 한다.

FsF-R1.3-01M: 메타데이터는 해당 연구 커뮤니티에서 권장되는 표준을 따라야 한다.

FsF-R1.3-02D: 데이터는 해당 연구 커뮤니티에서 권장되는 파일 형식으로 제공되어야 한다.

각 평가지표에 대한 명세는 기술, 평가 및 설명에 대한 세부 항목으로 표 4-2와 같은 테이블 형태로 구성되어 있다.

FAIR 데이터 원칙은 기본적으로 기계가 데이터 이해할 수 있는 형태로의 공유된 데이터에 대하여 접근 및 이해하여 사용자에게 정보를 제공하여 활용성을 높이도록 하는 것이다. 따라서 이러한 평가를 돕기 위해 FAIRsFAIR 프로젝트에서는 온라인 자동 평가도구인 F-UJI (<https://www.f-uji.net/>) 와 Post 서비스 제공을 위해 Python 소스를 공개하고 있다. (<https://github.com/pangaea-data-publisher/fuji>)

- 온라인 자동 산정 도구로서 F-UJI 도구는 온라인 리소스의 URL을 입력하면 평가요소에 대한 내용음 자동 산정하여 점수형태로 그 결과를 도출하게 된다.

표 4-2. FAIRSFAR 평가지표의 구성요소

기술 (DESCRIPTION)	
지표 식별자 (Metric Identifier)	각 평가지표를 구분하기 위한 코드이다. 예를 들어 FsF-F1-01D는 특정 평가 항목을 나타낸다. 이 코드는 해당 지표가 어떤 FAIR 원칙에 속하는지, 그리고 평가 대상이 데이터인지 메타데이터인지를 구분한다.
설명 (Description)	해당 평가지표의 세부적인 설명을 제공한다. 이 항목은 무엇을 평가하는지, 어떤 기준에 따라 데이터를 분석할지를 명확히 정의하며, 사용자가 평가 지표의 목적과 평가 방법을 이해할 수 있도록 돕는다.
배경 (Background)	이 평가지표가 만들어진 이유와 필요성을 설명하는 항목이다. 특정 평가 기준이 왜 중요한지, 그리고 이 지표가 데이터 관리 및 FAIR 원칙 준수에 어떤 기여를 하는지를 설명한다.
FAIR 원칙 (FAIR Principle)	해당 평가지표가 FAIR 원칙의 어느 부분을 평가하는지 나타낸다. 예를 들어, “F1“은 데이터를 쉽게 찾을 수 있도록 검색 가능성(Findability)을 평가하는 지표임을 의미한다.
코어트러스트 실 제휴 (CoreTrustSeal Alignment)	이 항목은 해당 평가지표가 CoreTrustSeal 인증 기준과 얼마나 일치하는지를 보여준다. CoreTrustSeal은 연구 데이터 저장소의 신뢰성을 평가하는 국제 인증이므로, 이와의 일치 여부는 데이터 저장소가 FAIR 원칙을 충족하는지 평가하는 데 중요한 요소이다.
평가 (ASSESSMENT)	
요구사항(들) Requirement(s)	이 항목에서는 해당 지표를 충족하기 위해 데이터가 갖춰야 할 조건과 요구사항을 명시한다. 예를 들어, 데이터에 고유한 식별자가 있어야 하거나, 메타데이터에 접근 조건이 명확히 기재되어야 하는 등의 구체적인 요구사항이 포함된다.
준수 수준 (Compliance Level)	데이터나 메타데이터가 해당 평가지표를 얼마나 충족하고 있는지를 나타내는 수준이다. 각 수준별 준수 정도에 따른 점수를 0.5부터 2까지 배정하여 준수도를 점검한다.
방법 (Method)	평가지표를 적용하여 데이터를 평가하는 구체적인 방법을 설명한다. 데이터나 메타데이터가 어떻게 평가되어야 하며, 어떤 절차와 도구를 사용해 평가할 수 있는지에 대한 지침을 제공한다. 예를 들어, 자동화된 도구를 통해 메타데이터의 가독성을 평가하는 방법 등을 포함할 수 있다.
설명 (COMMENTS)	
관련 자료 (Related Resources)	해당 평가지표에 대한 참고 자료나 관련 문서를 제공하여, 사용자가 더 깊이 이해하거나 추가적인 정보를 얻을 수 있도록 돕는다. 표준문서, 기술 가이드라인, 연구 자료 등이 여기에 포함될 수 있다.

- 오픈소스 형태로 공개된 F-UJI는 Python 기반과 Docker 기반 설치를 지원하며, 사용자는 데이터를 분석할 수 있도록 설정 파일을 수정해 자신만의 평가 환경을 구성할 수 있습니다. 또한, GitHub API와 통합하여 데이터의 소프트웨어 리포지토리까지 평가할 수 있다.

4.3 데이터 패키징

FAIRsFAIR의 공식적인 평가 기준에는 데이터 패키징 규칙이 포함되어 있지는 않지만, 메타데이터 일관성, 버전 관리, 파일 형식 등 데이터의 재사용성을 높이는 패키징 방식을 따르는 것이 간접적으로 중요한 역할을 할 수 있다. 연구 데이터는 생산 주기와 수집 대상, 목적에 따라 파편화될 수 있다. 이러한 파편화된 데이터는 동일한 연구자에게서도 발생할 수 있으며, 데이터가 시공간적으로 통합되어 방대한 양이 될 경우에는 제출과 활용을 위해 분할화가 필요하다. 데이터의 일관성을 유지하려면 리포지터리 제출 단계에서 패키징 규칙을 준수해야 하며, 이에 따라 파일 저장 방식, 네이밍 규칙, 메타데이터 형식에 대한 규칙이 필요하다. 본 절에서는 연구 데이터를 제출하거나 관리할 때 요구되는 데이터 패키징 방법에 대해 구체적인 기준을 제시해 보고자 한다. 표 4-3은 데이터 구성형태에 따른 패키징 방법들이 포함되어 있다.

표 4-3. 데이터 구성 형태 따른 및 방법론

데이터 구성 형태	패키징 방법
하나의 주제로 다양한 데이터가 획득되는 경우	하나의 대상 혹은 주제에서 다양한 데이터가 도출될 경우 하나의 데이터셋으로 패키징 수행(예; 암석시료로부터 다른 분석 시험이 수행된 경우 시료와 분석결과를 함께 패키징 하는 것을 고려할 수 있음)
동일한 주제의 데이터가 시기적 혹은 지역적으로 획득되는 경우	패키징 단위는 데이터 생산자에 의해 결정할 수 있으며, 시기적, 혹은 지역적(공간적)으로 패키징할 수 있으며, 일관성이 유지되도록 생산 부서에서 규칙을 정할 수 있음
데이터셋이 큰 경우	유통 가능한 형태로 데이터 세트를 자르는 경우 통합을 위한 방법론이 명시되어야 함

연구 데이터를 데이터세트로 패키징하기 위해서 생산된 연구 데이터와 함께 요구되는
부수적인 데이터는 다음과 같다.

- 생산된 연구 데이터 : 연구 사업을 통해 획득된 데이터
- 설명용 데이터 : 데이터를 설명하기 위한 데이터(readme.txt)
- 라이선스 데이터 : 데이터에 대한 라이선스(license.txt)

하나의 주제로 다양한 데이터가 획득되는 경우

패키징 원칙: 동일한 주제 내에서 다양한 데이터가 획득될 경우, 각 데이터 소스와 그
분석 결과를 하나의 패키지로 묶어야 합니다.

예시: 암석 시료에서 다양한 분석이 수행된 경우, 시료 정보와 각각의 분석 결과를 함
께 패키징.

```
rock_sample_analysis/  
├── sample_info.csv  
├── chemical_analysis.csv  
├── physical_properties.csv  
└── documentation/  
    ├── readme.md  
    └── metadata.json
```

동일한 주제의 데이터가 시기적 혹은 지역적으로 획득되는 경우

패키징 원칙: 시기적 또는 지역적으로 데이터를 패키징할 때 일관성을 유지하기 위해
생산 부서에서 규칙을 설정해야 합니다.

예시: 특정 지역에서 시간별로 수집된 환경 데이터.

```
environmental_data/  
├── region1/  
│   ├── 2024-01/  
│   │   ├── air_quality.csv  
│   │   ├── water_quality.csv  
│   │   └── soil_samples.csv  
│   └── 2024-02/  
│       ├── air_quality.csv  
│       └── water_quality.csv
```

```

|       |─── soil_samples.csv
|─── region2/
|       |─── 2024-01/
|       |       |─── air_quality.csv
|       |       |─── water_quality.csv
|       |       |─── soil_samples.csv
|       |─── 2024-02/
|       |       |─── air_quality.csv
|       |       |─── water_quality.csv
|       |       |─── soil_samples.csv

```

데이터세트가 큰 경우

패키징 원칙: 데이터 세트가 너무 큰 경우, 유통 가능한 크기로 데이터를 자르고, 이후 통합을 위한 명확한 방법론을 제공해야 합니다.

예시: 대규모 기상 데이터 세트

```

climate_data/
|─── part1/
|       |─── temperature_2024-01.csv
|       |─── precipitation_2024-01.csv
|       |─── metadata_part1.json
|─── part2/
|       |─── temperature_2024-02.csv
|       |─── precipitation_2024-02.csv
|       |─── metadata_part2.json
|─── integration_guide.md

```

연구 데이터의 명명 방식에 일정한 규칙을 적용하면 이름 중복, 정체성 혼란, 향후 이름 변경 및 정렬 작업 등의 문제를 방지할 수 있다. 연구 데이터의 이름 지정 방식 시 다음과 같은 규칙을 적용할 것이 필요하다.

- 파일과 폴더의 이름은 의미 있고 가능한 한 명료해야 함
- 데이터/샘플의 이름은 해당 데이터/샘플의 내용을 반영하고, 고유하며 설명적 (descriptive)이어야 함
- 파일 유형(file type), 파일 버전 정보, 소프트웨어 정보 또는 데이터 수집 위치와 같은 모든 파일과 관련된 반복되는 메타데이터 등 기타 관련 정보를 readme.txt 파일에 작성

- 날짜 입력 시 표준 형태 사용
- 일관된 버전 넘버링 체계(version numbering scheme)를 통해 새 버전의 존재와 해당 데이터의 변경 여부를 추적할 수 있으며, 이전에 사용한 버전과 현재 작업 중인 버전을 명확하게 정할 수 있다.

버전 넘버링 체계와 관련하여 호주의 연구 데이터를 서비스하고 있는 ANDS(Australian National Data Service)에서는 표 4-4와 같이 적용하고 있다.

표 4-4. 넘버링 시스템 예시 1)

구분	내용
버전 넘버링 시스템 1	<ul style="list-style-type: none"> <input type="checkbox"/> 데이터 버저닝을 위해 Major.Minor (예: V2.1)의 두 부분으로 된 넘버링 규칙을 적용함 <input type="checkbox"/> Major 부분은 범위, 상황 또는 사용 목적의 변화를 일으킬 수 있는 데이터셋의 내용 또는 형태의 변경을 나타내며 다음과 같이 개정될 경우 Major 부분의 번호를 갱신함 <ul style="list-style-type: none"> ▪ 컬렉션에 상당한 양의 새로운 데이터 항목이 추가되거나 삭제됨 ▪ 시간 또는 공간 기준선(spatial baseline)이 변경되어 데이터 값이 변경됨 ▪ 추가적인 데이터 속성(data attributes)이 도입됨 ▪ 데이터 생성 모델(data generation model)의 변경 ▪ 데이터 항목 포맷이 변경됨 <input type="checkbox"/> Minor 부분은 기존 데이터 항목보다 품질이 향상된 것을 나타내며, 초기 컬렉션의 사용 목적 또는 범위에 영향을 미치지 않고 다음과 같이 개정될 경우 Minor 부분의 번호를 갱신함 <ul style="list-style-type: none"> ▪ 데이터 속성의 이름 변경 ▪ 기존 데이터의 에러 수정 ▪ 일부 매개 변수 조정하여 데이터 생성 모델 재실행
버전 넘버링	<input type="checkbox"/> 단순한 하나의 숫자를 적용하여 데이터의 개정 및 버전을 나타냄

1) ANDS. (n.d.). Data versioning. Retrieved November 14, 2019 from <https://www.ands.org.au/working-with-data/data-management/data-versioning>

시스템 2	(예: V1, V2)
버전 넘버링 시스템 3	<input type="checkbox"/> 천문학 또는 해양 과학과 같은 일부 분야에서 데이터 제공자는 일반적으로 원시 데이터를 다양한 수준(예: 레벨 0, 레벨1, ...)의 버전으로 처리하며, 과학적 해석 및/또는 데이터 품질 제어 수준에 따라 레벨이 증가하여 적용 <input type="checkbox"/> 예를 들어, NASA는 다음과 같은 범주를 사용하여 각 버전의 데이터가 원시 데이터와 가까운 정도를 표현함 <ul style="list-style-type: none"> ▪ 레벨 0 원시 위성 데이터 ▪ 레벨 1 좌표 보정된 원시 데이터 ▪ 레벨 2 지구 물리 매개변수 변환된 원시데이터 ▪ 레벨 3 재샘플링, 공간 평균화, 시간 경과에 따라 보간/평균화 ▪ 레벨 4 데이터는 모델 출력 또는 하위 수준 데이터의 분석 결과

데이터의 설명

리포지터리에 제출된 데이터의 재사용성을 높이기 위해서는 데이터에 대한 데이터 즉 메타데이터가 필요하다. 메타데이터는 데이터 관리-검색을 위한 형태와 실제 데이터 활용을 위해 설명된 형태로 구분할 수 있음. 표 3-1-5.는 지구과학 분야에서 적용성을 검토할 수 있는 메타데이터 표준 현황이다.

연구 데이터를 효과적으로 사용하기 위한 제출된 데이터와 함께 설명 메타데이터가 포함되어야 한다. 데이터 설명을 위한 메타데이터는 분야별 데이터의 특성이 다르기 때문에 이에 대한 표준은 없으나 기술해야 할 항목들은 다음과 같이 정리될 수 있다.

readme.txt
<input type="checkbox"/> 데이터 파일과 함께 데이터 파일을 해석하는데 필요하거나 설명할 수 있는 도큐멘테이션 파일을 제공해야 함 <input type="checkbox"/> 도큐멘테이션 파일의 예로는 코드북, 데이터 수집 도구, 요약 통계, 프로젝트 요약 및 데이터 관련 출판물 목록이 포함될 수 있음 <input type="checkbox"/> 그 외에도 다음과 같은 내용을 포함할 수 있음 ²⁾ <ul style="list-style-type: none"> ▪ 프로젝트 배경 및 목적 ▪ 방법론에 관한 정보

- 사용한 소스
- 관련된 연구
- 샘플링 절차
- 데이터세트의 내용과 구조
- 콘텐츠의 설명과 파일명 목록
- 데이터를 작업하거나 읽을 때 필요한 도구 또는 소프트웨어
- 데이터의 알고 있는 오류 또는 약점에 관한 설명
- 프로젝트와 관련되거나 결과로 출판된 출판물에 대한 레퍼런스
- 레코드, 데이터 변환 또는 형식 변경(format change)에 관한 문서

리포지터리에 제출된 데이터를 법적 테두리 안에서 활용하기 위해서는 유·무료 정책 수립과 이에 따른 라이선스를 정의해야 하며 라이선스에는 데이터의 권한 및 사용 조건을 명시해야한다. 라이선스에서 명시될 데이터 소유권은 데이터 관리 및 사용뿐만 아니라 데이터에 대한 제어 및 권한을 의미한다.

소유권은 주요 연구자, 후원 기관, 펀딩 기관 및 참여하는 모든 대상과 관련 될 수 있으며, 연구 데이터를 기탁하기 전에 해당 연구 데이터의 소유권 문제(소유권, 저작권 및 지적재산권 등)를 검토하여야 한다. 데이터 라이선스는 데이터 생산기관의 정책에 따라 가변적으로 설정가능하며 일반적으로 참고할 수 있는 공개용 라이선스는 크리에이티브 커먼즈(Creative Commons)³⁾, 공공저작물에서 사용할 수 있는 공공누리, 이외에 일반적인 특수 제약조건을 포함하여 공개 라이선스를 설정할 수 있다. 영국 지질조사소인 BGS의 경우 별도의 공개 라이선스 정책을 수립하여 사용하고 있다.

license.txt

- 라이선스 유형 명시하여 데이터 사용 시 적용범위를 명시할 수 있음
 - 크리에이티브 커먼즈 라이선스는 공개 콘텐츠에 널리 사용되며, 버전 4.0은 데이터 라이선싱을 명시적으로 고려하였음
 - BGS 출판물은 영국 연구회인 NERC(Natural Environment Research Council)의 정책에 따라 유 무료정책을 적용하고 있으며 기관에서 생산된 데이터의 성격과 사용 상황에 따라 달리하고 있음

2) ESSA(n.d.). ESSA guidelines for depositors. Retrieved

3) <https://creativecommons.org/>

□ 인용문구 명시 :데이터 인용은 일반적으로 다음과 같은 7가지 요소로 구성된다.⁴⁾

- 저자
- 공개일
- 데이터 제목
- 버전 또는 에디션 번호
- 아카이브 또는 배포자
- 영구 식별자
- 액세스 날짜 및 시간
 - 데이터를 작업하거나 읽을 때 필요한 도구 또는 소프트웨어
 - 데이터의 알고 있는 오류 또는 약점에 관한 설명
 - 프로젝트와 관련되거나 결과로 출판된 출판물에 대한 레퍼런스
 - 레코드, 데이터 변환 또는 형식 변경(format change)에 관한 문서

권장 파일 형식

연구 데이터의 지속적인 액세스와 잠재적인 재사용을 고려하여 데이터 파일의 형식은 가능한 기계 가독형이면서, 가능한 공개된 포맷을 사용해야 데이터에 대한 활용성을 제고할 수 있다. 지구과학분야의 범용적 데이터 파일포맷과 관련하여 NGDC(NGDC Acceptable digital formats)⁵⁾에서 권장하고 있는 파일 형식은 다음 표 4-5와 같다.

표 4-5. 권장 연구 데이터 파일 포맷

데이터 유형	권장 형식(Preferred format)	파일 확장자
지질공학-환경 데이터	Association of Geotechnical and Geo-environmental Specialists (preferably version 3.1 or 4.0)	.ags
지구물리 데이터	Log ASCII Standard	.las
	Seismic data	.sgy
	Sidescan sonar data	.xtf
일반 과학 데이터	Microsoft Excel files	.xls, .xlsx
	Comma-separated value files	.csv
	Data files (with readme file on software)	.dat

4) USGS. (n.d.). Data Citation Standard. Retrieved September 5, 2019 from <https://www.usgs.gov/products/data-and-tools/data-management/data-citation>

5) NGDC. (n.d.). Acceptable digital formats. Retrieved August 19, 2019 from <https://www.bgs.ac.uk/services/NGDC/preferredDigitalFormats.html>

데이터 유형	권장 형식(Preferred format)	파일 확장자
	Tab delimited data file	
	Portable document format (PDF, PDF/A)	.pdf
	Extensible mark-up language	.xml, .json
텍스트	Microsoft Word document	.doc,.docx
	Text file (plain/ASCII)	.txt
	Rich text format	.rtf
프레젠테이션	Microsoft Powerpoint presentation	.ppt
GIS/공간정보	ESRI shapefile/MapInfo/QGIS files/GeoTIFF	
데이터베이스	Microsoft Access database	.aacdb
	Oracle export	
	MySQL export format	
	SQLite format	
이미지	Tagged image format	.tif
	Joint Photographic Exports Group	.jpg
	Portable Network Graphics	.png
	Drawing Interchange Format (AutoCAD)	.dxf
비디오	Apple Quick Time Movie	.mov
	Audio Video Interleaved	.avi
	Digital Moving Picture Exchange Bitmap	.dpx
	Moving Picture Experts Group	.mp4

5. 연구 데이터 기술요소

5.1 메타데이터 표준

Dublin Core Metadata

더블린 코어(Dublin Core)는 디지털 자원의 메타데이터를 표준화하기 위한 메타데이터 요소 집합으로, 정보 검색과 데이터 관리를 용이하게 하기 위해 사용된다. 더블린 코어는 다양한 디지털 자원, 예를 들어 텍스트, 이미지, 오디오, 비디오 등을 기술하고 설명하는 데 널리 활용되며, 단순하면서도 확장 가능한 구조로 다양한 분야에 적용된다. 5-1은 더블린코어 메타데이터 요소이다.

표 5-1. 더블린코어 메타데이터 요소

Element	Description
Title	자원에 부여된 이름
Creator	자원의 내용을 만드는 데 일차적인 책임이 있는 주체
Type	자원의 성격이나 장르
Contributor	자원의 내용에 기여한 주체
Publisher	자원을 이용 가능하게 만든 주체
Date	자원이 현재 형태로 제공되기 시작한 날짜
Language	자원의 지적 내용이 표현된 언어
Format	자원의 물리적 혹은 디지털 형식
Description	자원의 내용에 대한 설명
Subject	자원의 내용과 관련된 주제
Relation	관련 자원에 대한 참조
Identifier	자원에 대한 명확한 참조
Rights	자원에 대한 권리 정보
Source	현재 자원이 파생된 자원
Coverage	자원의 내용적 범위 또는 기간

DCMI Abstract Model(DCAM)은 더블린 코어 메타데이터 이니셔티브에서 정의한 메타데이터 구조의 추상 모델로, 메타데이터 요소들 간의 의미론적 관계와 사용 방법을 설명하는 프레임워크다(그림 5-1). 이 모델은 다양한 환경에서 더블린 코어 메타데이터가 일관되게 사용될 수 있도록 하며, 상호 운용성을 유지하면서 정확하게 표현될 수 있도록 돕는다.

DCAM의 주요 개념으로는 자원을 설명하는 메타데이터 요소들의 집합인 설명

(Description)이 있다. 여러 자원을 설명하는 설명들의 집합은 설명 세트(Description Set)라고 한다. 속성(Property)은 자원의 특정 속성을 나타내며, 이 속성에 대한 구체적인 값을 값(Value)이라고 한다. 예를 들어, 제목(Title)은 자원의 속성이며, “Dublin Core“는 그 속성에 해당하는 값이다.

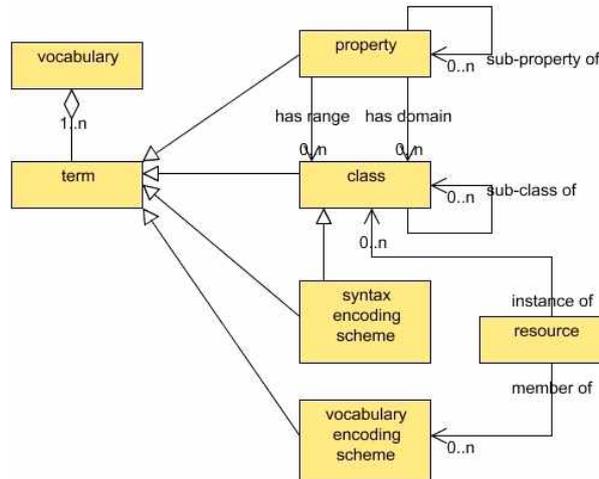


그림 5-1. DCMI: DCMI Abstract Model (출처 : Powell, 2007)

DataCite Metadata Schema

DataCite Metadata Schema는 연구 데이터의 식별과 인용을 지원하기 위한 메타 데이터 표준으로, 디지털 자원의 발견, 접근, 활용, 그리고 지속 가능한 보존을 촉진하기 위해 개발되었다. 특히 연구자, 출판사, 기관 등이 연구 데이터를 쉽게 찾고, 식별하고, 재사용할 수 있도록 돕는 데 중점을 두고 있다. DOI(Digital Object Identifier)를 통해 데이터에 고유한 식별자를 부여하여 데이터의 접근성과 추적 가능성을 강화하는 것이 주요 목표다.

DataCite Metadata Schema는 연구 데이터의 체계적 관리와 재사용을 촉진하기 위해 설계되었다. 특히 데이터에 대한 DOI를 할당하여 연구 데이터의 인용을 용이하게 하고, 이를 통해 연구 성과를 정량화할 수 있도록 지원한다. 이 스키마는 다양한 데이터 유형과 학문 분야에서 사용되며, 연구 데이터의 투명성, 접근성, 그리고 학술적 인용을 보장하는 중요한 역할을 한다.

DataCite Metadata Schema는 연구 데이터에 대한 풍부한 메타데이터를 제공하기 위해 여러 요소를 포함하고 있다. 필수 요소와 선택적 요소로 나뉘며, 데이터에 대한 충분한 정보를 제공할 수 있도록 구성된다.

표 5-2. DataCite 메타데이터 요소

Element	Description
Identifier	연구 데이터의 고유 식별자 (주로 DOI)
Creator	데이터를 생성한 연구자나 단체
Title	데이터의 이름 또는 제목
Publisher	데이터를 공개한 기관이나 출판사
PublicationYear	데이터가 처음 공개된 연도
ResourceType	데이터의 성격을 정의 (예: Dataset, Text)
Subject	데이터가 다루는 주제 또는 연구 분야
Contributor	데이터 제작에 기여한 개인 또는 단체
Description	데이터의 내용을 구체적으로 설명
Rights	데이터에 대한 사용 권리 및 라이선스 정보
Version	데이터의 특정 버전을 명시
Language	데이터가 작성된 언어
FundingReference	데이터를 지원한 연구비 제공 기관

DataCite의 스키마를 이용한 샘플 XML 변환 코드이며 이다. DataCite 메타데이터는 연구 데이터를 고유하게 식별하고, 체계적으로 관리하고, 다양한 시스템 간에 데이터를 원활히 교환하기 위한 목적으로 사용된다. 이 형식은 데이터의 상호 운용성을 보장하고, 연구 데이터를 쉽게 검색하고 인용할 수 있도록 한다.

```
<resource xmlns="http://datacite.org/schema/kernel-4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://datacite.org/schema/kernel-4
http://schema.datacite.org/meta/kernel-4.3/metadata.xsd">
  <identifier identifierType="DOI">10.1234/example-doi</identifier>
  <creators>
    <creator>
      <creatorName>Hong Jil-Dong</creatorName>
    </creator>
  </creators>
  <titles>
    <title>Sample Research Data</title>
  </titles>
  <publisher>Data Repository</publisher>
  <publicationYear>2024</publicationYear>
  <resourceType resourceTypeGeneral="Dataset">Text</resourceType>
  <subjects>
```

```

<subject>Physics</subject>
</subjects>
<contributors>
  <contributor contributorType="Researcher">
    <contributorName>Kim Man-Su</contributorName>
  </contributor>
</contributors>
<descriptions>
  <description descriptionType="Abstract">This dataset contains sample research data for physics
  experiments.</description>
</descriptions>
<rightsList>
  <rights rightsURI="https://creativecommons.org/licenses/by/4.0/">Creative Commons Attribution 4.0
  International</rights>
</rightsList>
</resource>

```

ISO 19115 - Geographic Information Metadata

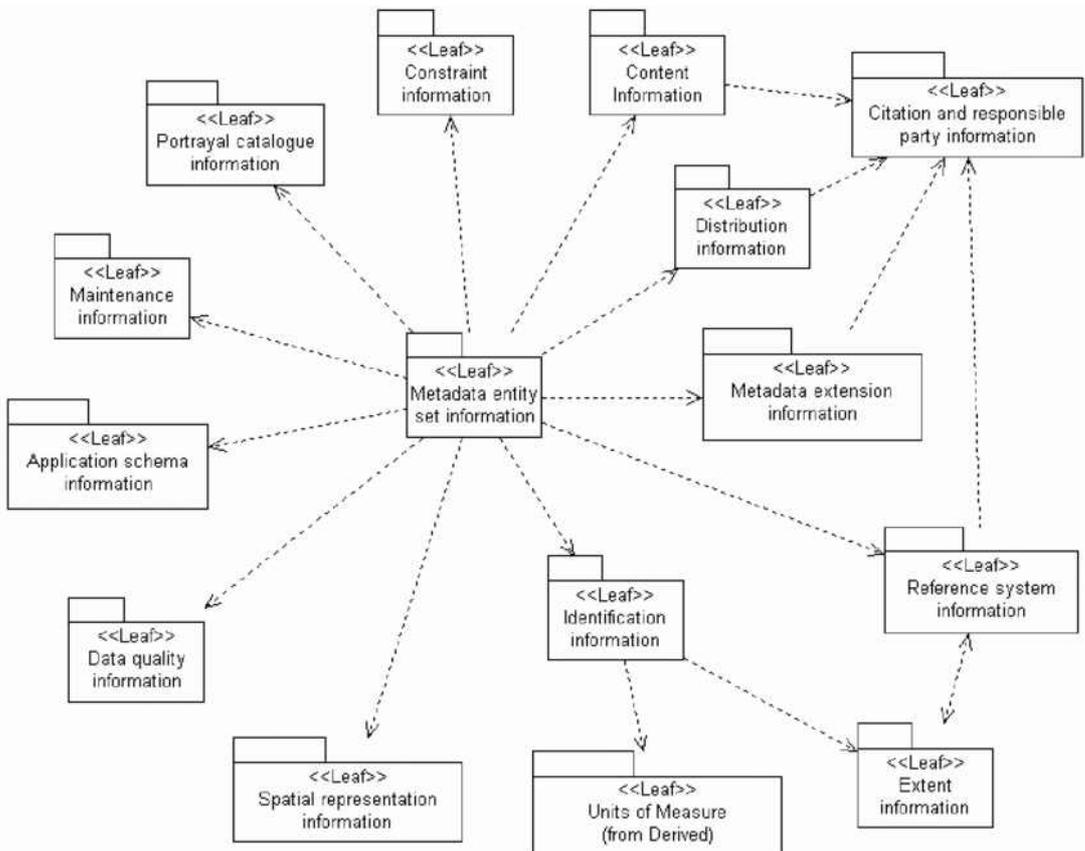


그림 5-1. ISO 19115 Package Diagram (출처 : Kresse, 2004)

ISO 19115는 지리 정보(공간 데이터)의 메타데이터 표준으로, 지리 정보 데이터

를 효율적으로 관리, 교환, 검색 및 재사용할 수 있도록 하기 위해 제정된 국제 표준이다. 이 표준은 공간 데이터의 의미, 출처, 품질, 구조, 책임 소재 등을 설명할 수 있는 메타데이터 요소를 정의한다.

ISO 19115는 2003년에 처음으로 국제표준화기구(ISO)에서 발표되었다. 이 표준은 전 세계적으로 공간 데이터를 생성, 관리, 배포하는 다양한 기관과 조직들이 데이터를 표준화된 방식으로 설명할 수 있게 하기 위해 개발되었다. GIS(Geographic Information System, 지리정보 시스템)의 보급과 함께, 전 세계적으로 공간 데이터의 수요가 폭발적으로 증가하면서 이러한 데이터를 체계적으로 관리하고 공유할 필요성이 커졌다. ISO 19115는 이러한 배경 속에서 등장한 표준이다. 표준이 처음 발표된 이후, 공간 데이터의 특성과 기술 발전에 맞추어 여러 차례 개정이 이루어졌다. 가장 큰 개정은 ISO 19115:2014로, 이전 버전의 표준을 확장하여 보다 복잡하고 다양한 공간 데이터를 다룰 수 있게 만들었다.

ISO 19115의 구성은 그림 5-2와 같이 UML 형식으로 패키지 다이어그램 형식으로 구성되어 있으며 각 패키지 내에는 해당 주제를 중심으로 식별하거나 설명할 수 있는 다양한 속성들이 구성된다. 대표적인 패키지에 대한 설명은 표 5-3과 같이 구성된다.

표 5-3. ISO 19115 메타데이터 패키지 요소

패키지 구성요소	설명
식별 정보 (Identification Information)	공간 데이터 세트에 대한 기본적인 식별 정보로, 데이터의 제목, 설명, 목적 등을 포함
제작 정보 (Data Quality Information)	데이터의 품질을 평가하는 데 필요한 정보로, 정확성, 완전성, 일관성 등이 포함
공간 참조 정보 (Spatial Reference Information)	공간 데이터를 참조하는 좌표계와 투영법 등의 정보를 포함
배포 정보 (Distribution Information)	데이터 세트의 배포 방법과 사용 권한, 접근 방법 등을 설명
책임 정보 (Responsible Party Information)	데이터 생성자, 관리자, 배포자 등의 책임 기관 또는 인물에 대한 정보
메타데이터 자체 정보 (Metadata Information)	메타데이터 자체에 대한 정보로, 생성 날짜, 작성자, 표준 버전 등이 포함
공간 및 시간적 범위 정보 (Spatial and Temporal Extent Information)	데이터가 적용되는 지리적 및 시간적 범위에 대한 정보

Schema.org

Schema.org는 웹에서 구조화된 데이터를 표현하기 위한 메타데이터 표준으로, 검색 엔진과 애플리케이션이 정보를 더 잘 이해하고 처리할 수 있도록 설계되었다. 연구 데이터가 점점 더 많은 학문적 및 실무적 중요성을 가지면서, 이러한 데이터를 웹에서 쉽게 찾고 접근할 수 있도록 지원하는 Schema.org의 역할은 더욱 부각되고 있다. 연구자들이 데이터를 공유할 때 Schema.org 메타데이터를 사용하면 검색 엔진에서 연구 데이터를 더 쉽게 검색할 수 있으며, 학술 포털과 데이터베이스에서도 빠르게 접근할 수 있다. Schema.org는 연구 데이터를 구조화된 방식으로 표현할 수 있는 능력을 제공한다. 이를 통해 데이터의 제목, 창작자, 발행일, DOI와 같은 주요 정보들이 명확하게 드러나며, 연구 데이터의 검색성과 활용성이 크게 향상된다. 또한, Schema.org는 다른 메타데이터 스키마와도 함께 사용되며, 예를 들어 DataCite나 Dublin Core와 결합함으로써 연구 데이터 관리와 인용의 효과성을 높일 수 있다.

이 표준은 단순히 구조화된 데이터 표현에 그치지 않고, 다양한 시스템에서 데이터를 교환하고 사용할 수 있도록 상호 운용성까지 제공한다. 학술 출판사, 연구 데이터 저장소, 학술 검색 엔진 등이 Schema.org 메타데이터를 사용하여 데이터 간 통합을 원활하게 이루고, 이를 통해 데이터의 접근성과 활용 가능성이 증대된다. Schema.org는 연구 데이터를 설명하기 위한 Dataset이라는 유형을 제공하여 연구 데이터를 구조적으로 표현할 수 있게 한다. 이를 통해 연구 데이터에 대한 중요한 정보를 명확히 제시할 수 있으며, 검색 엔진이 이를 쉽게 인식하게 함으로써 데이터 검색 및 접근성을 높인다.

표 5-4. Schema.org 메타데이터 구성 항목

항목	설명
제목 (name)	데이터셋의 제목
설명 (description)	데이터셋의 설명
창작자 (creator)	데이터의 창작자(연구자)
공개일 (datePublished)	데이터가 공개된 날짜
식별자 (identifier)	데이터셋의 고유 식별자(DOI 등)
발행자 (publisher)	데이터 발행자(기관)
라이선스 (license)	데이터 사용에 대한 라이선스 정보

다음은 석회암 데이터에 대한 Schema.org 스키마 예시다. 이 예시는 Dataset 유형을 사용해 연구 데이터의 메타데이터를 JSON-LD 형식으로 표현하였다. 이 메타데이터는 레코드 항목을 웹 페이지에 동적으로 로드하도록 하여 검색 엔진이 데이터를 이해하여 인덱싱을 용이하게 한다.

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Dataset",
  "name": "석회암 데이터",
```

```

“description”: “홍길동이 촬영한 석회암에 대한 이미지와 관련 데이터. 이 데이터는 한국의 특정 지역에서 채취된 석회암의 구성과 특성을 설명한다.“,
“creator”: {
  “@type”: “Person“,
  “name”: “홍길동“,
  “affiliation”: {
    “@type”: “Organization“,
    “name”: “한국지질자원연구원“
  }
},
“datePublished”: “2024-01-15“,
“identifier”: “https://doi.org/10.1234/limestone-data“,
“license”: “https://creativecommons.org/licenses/by/4.0/“,
“keywords”: [“석회암“, “지질학“, “암석 데이터“, “홍길동“],
“publisher”: {
  “@type”: “Organization“,
  “name”: “한국지질자원연구원“
},
“spatialCoverage”: {
  “@type”: “Place“,
  “name”: “한국“,
  “geo”: {
    “@type”: “GeoCoordinates“,
    “latitude”: “37.5665“,
    “longitude”: “126.9780“
  }
},
“temporalCoverage”: “2023-06-01/2023-12-31“,
“url”: “https://example.org/dataset/limestone“
</script>

```

DCAT (Data Catalog Vocabulary)

데이터 카탈로그에 대한 메타데이터를 구조화하고 표현하기 위한 표준 형식으로, 데이터 세트의 검색, 접근, 재사용을 용이하게 하기 위해 설계된 RDF 기반의 메타데이터 스키마다. DCAT은 주로 공공 데이터와 연구 데이터 포털에서 사용되며, 다양한 데이터 카탈로그 간의 상호 운용성을 높이고, 데이터 공유와 통합을 촉진하는 역할을 한다.

DCAT은 데이터 세트와 관련된 정보(예: 제목, 발행일, 작성자, 라이선스 등)를 구조적으로 표현하여 데이터가 저장된 시스템 간의 데이터 교환을 지원한다. 이를 통해 다양한 데이터 소스에서 데이터 세트를 보다 쉽게 발견하고 활용할 수 있다. 또한 DCAT은 데이터 카탈로그 내의 데이터 세트뿐만 아니라 데이터 세트 간의 관계, 배포

형식, 접근 방법 등도 정의하여 사용자가 데이터의 출처와 메타데이터를 명확히 이해할 수 있도록 한다.

DCAT은 특히 유럽연합과 같은 국제적 프로젝트에서 공공 데이터를 공유하는 데 많이 사용되며, 데이터 포털 간에 메타데이터의 일관성을 유지하고 상호 운용성을 보장하기 위한 표준이다. 다음은 석회암 시료 데이터에 대한 암석 시료 데이터 세트에 대한 DCAT (Data Catalog Vocabulary) 예시이다. 이 예시는 JSON-LD 형식으로 작성되었다.

```
{
  "@context": {
    "@vocab": "http://www.w3.org/ns/dcat#",
    "dcterms": "http://purl.org/dc/terms/"
  },
  "@type": "Catalog",
  "dcterms:title": "한국지질자원연구원 암석 시료 데이터 카탈로그",
  "dcterms:description": "한국지질자원연구원에서 제공하는 다양한 암석 시료 데이터의 카탈로그.",
  "dcterms:publisher": {
    "@type": "Organization",
    "dcterms:name": "한국지질자원연구원"
  },
  "dcterms:issued": "2024-01-01",
  "dcterms:modified": "2024-03-01",
  "dcterms:language": "ko",
  "dcterms:license": "https://creativecommons.org/licenses/by/4.0/",
  "dataset": [
    {
      "@type": "Dataset",
      "dcterms:title": "석회암 시료 데이터",
      "dcterms:description": "태백 석회암 시료에 대한 데이터",
      "dcterms:issued": "2023-06-01",
      "dcterms:modified": "2024-02-01",
      "dcterms:identifier": "https://example.org/dataset/limestone-sample",
      "dcterms:publisher": {
        "@type": "Organization",
        "dcterms:name": "한국지질자원연구원"
      },
      "dcterms:license": "https://creativecommons.org/licenses/by/4.0/",
      "dcat:distribution": {
        "@type": "Distribution",
        "dcat:accessURL": "https://example.org/dataset/limestone-sample/download",
        "dcat:mediaType": "application/json"
      }
    }
  ]
}
```

```
}  
]  
}
```

5.2 데이터 모델링

데이터의 표준화와 상호운영성을 확보하기 위해서는 체계적인 구조적 설계가 필수적이며, 이를 위해 데이터 모델링 도구는 데이터를 시각적으로 명확하게 표현하고 구현할 수 있는 기반을 제공한다. 데이터베이스 설계를 위한 ER 모델(Entity-Relationship Model)은 데이터의 주요 개체(Entity)와 관계(Relationship)를 시각적으로 모델링하며, 주로 관계형 데이터베이스에서 사용된다. UML(Unified Modeling Language)은 객체지향적 관점에서 시스템과 데이터를 설계하는 범용 모델링 언어로, 소프트웨어 시스템의 아키텍처와 데이터 흐름을 시각적으로 표현한다. 한편, RDF(Resource Description Framework)와 OWL(Web Ontology Language)은 메타데이터 관리와 링크드 데이터 설계를 위한 도구로, RDF는 주어-서술어-목적어(Triple) 구조로 데이터 간의 관계를 표현하고, OWL은 RDF를 확장하여 더 복잡한 의미적 관계와 논리적 제약을 정의하는 온톨로지 언어다.

ER 모델(Entity-Relationship Model)

연구 데이터 관리와 상호운영성을 확보하는 데 중요한 데이터베이스 설계 도구로, 데이터를 구조화하여 개체(Entity)와 그 관계(Relationship)를 시각적으로 모델링하는 데 사용된다. 연구 데이터는 대규모의 복잡한 데이터 셋을 포함하며, 이러한 데이터를 효율적으로 관리하고 분석하기 위해 ER 모델을 활용하면 데이터의 구조를 명확하게 정의하고, 다양한 시스템 간의 상호운영성을 향상시킬 수 있다. 특히 관계형 데이터베이스 시스템에서 ER 모델은 데이터를 테이블로 변환하여 저장하고 관리하는 기초를 제공하므로, 연구 데이터 관리에서 데이터 무결성, 중복 방지, 효율적인 질의 처리 등이 가능해진다.

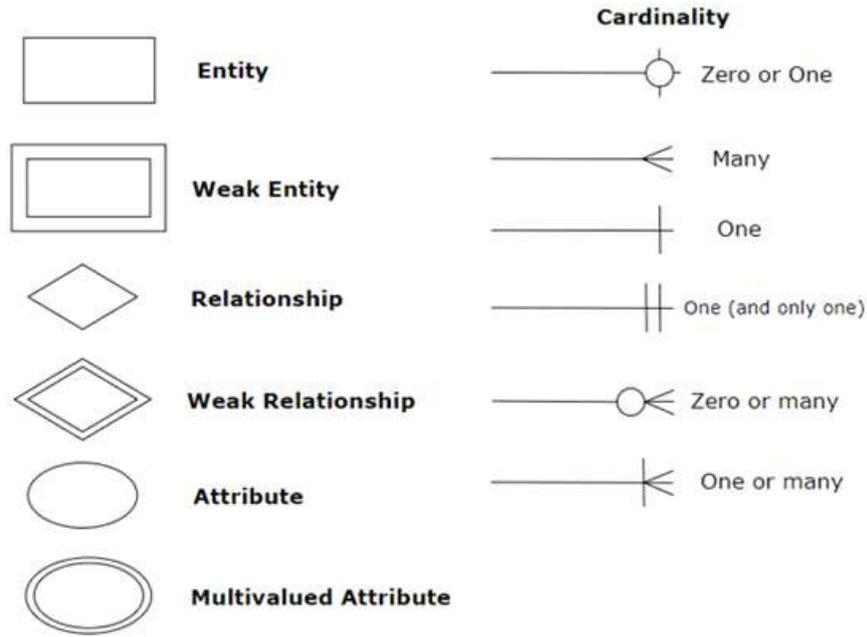


그림 5-2. ER 다이어그램의 개체 및 관계 표기법

ER 모델의 구성 요소는 크게 엔터티(Entity), 속성(Attribute), 관계(Relationship)로 나뉜다. 각 구성 요소는 데이터베이스 구조 설계의 핵심이며, 연구 데이터의 다양한 요소를 체계적으로 표현하는 데 사용된다. 엔터티는 데이터베이스에서 관리하고자 하는 주요 객체나 개체를 의미한다. 연구 데이터의 맥락에서 엔터티는 연구 대상이나 실체, 예를 들어 실험, 연구자, 장비, 샘플 등 다양한 실제 데이터를 반영하는 개체로 간주될 수 있다. 엔터티는 연구에서 저장하거나 관리해야 하는 중요한 정보를 담고 있다. 엔터티는 명사적 개체로서, 데이터베이스에서 구체적인 데이터를 저장할 테이블에 해당하며, 각 엔터티는 고유의 속성과 식별자를 가지며 독립적으로 존재하는 개체를 의미한다. 예를 들어, 연구 프로젝트에서 '연구자', '실험', '데이터셋' 등이 엔터티로 정의될 수 있다. 속성은 엔터티가 가지는 특성이나 정보를 의미한다. 연구 데이터에서 속성은 각 엔터티에 대한 구체적인 데이터를 정의하는 데 사용된다. 속성은 엔터티에 속하는 정보를 설명하며, 데이터베이스 내에서는 필드로 저장된다. 속성은 엔터티의 세부 정보를 제공하는 요소로, 엔터티가 가지고 있는 특성이나 상태를 나타낸다. 속성은 데이터의 값을 가지며, 이를 통해 엔터티를 구체적으로 설명할 수 있다. 예를 들어 '연구자' 엔터티의 경우, 속성으로는 '이름', '소속', '연구 분야' 등이 포함될 수 있다. '실험' 엔터티의 경우, '날짜', '결과' 등이 속성이 될 수 있다. 관계는 엔터티 간의 상호작용이나 연관성을 나타낸다.

연구 데이터에서 관계는 서로 다른 엔터티 간의 데이터 흐름이나 의존성을 나타내며, 이를 통해 연구 데이터 간의 상호 연관성을 명확하게 표현할 수 있다. 관계는 일대일, 일대다, 다대다와 같은 기수성을 가진다. 관계는 두 개 이상의 엔터티 간의 연결을 정의하며, 이 연결은 엔터티 간에 어떤 일이 발생하거나, 어떻게 데이터를 공유하는지를 설명한다. 관계는 데이터를 통합적으로 관리하는 데 매우 중요한 역할을 한다. 예를 들어, '연구자'와 '실험' 간의 관계는 '연구자가 실험을 수행한다'로 정의될 수 있으며, '실험'과 '데이터셋' 간의 관계는 '실험 결과로 데이터셋을 생성한다'로 설명될 수 있다.

UML(Unified Modeling Language)

시스템 및 데이터를 다객체지향적 관점에서 설계하기 위한 범용 모델링 언어로, 소프트웨어 시스템의 구조와 동작을 시각적으로 표현하는 데 사용된다. 연구 데이터 관리와 상호운영성 측면에서도 UML은 매우 유용하다. 연구 데이터는 다양한 요소와 복잡한 상호작용을 포함하고 있으며, 이를 효율적으로 관리하기 위해 UML을 활용하면 데이터를 체계적으로 구조화하고, 시스템의 아키텍처와 데이터 흐름을 명확히 설계할 수 있다. UML은 객체 지향 시스템 설계에 최적화되어 있어, 복잡한 연구 데이터 처리 및 시스템 설계에 적합하며, 다양한 도메인에서 데이터를 표준화하고 상호운영성을 높이는 데 기여한다.

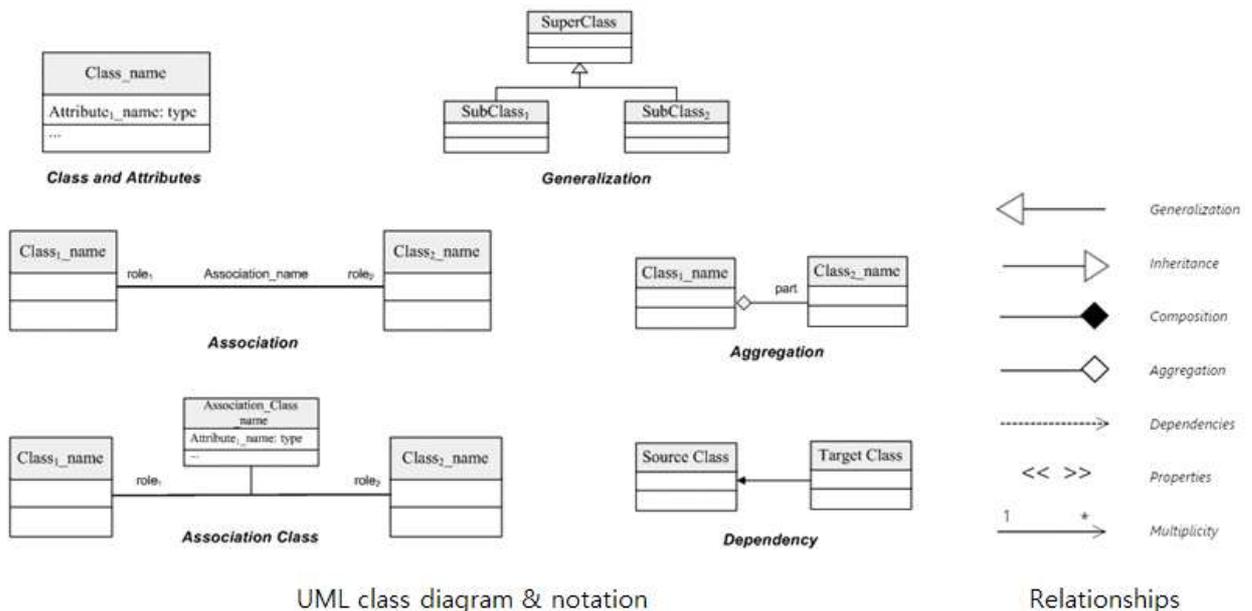


그림 5-3. 클래스 다이어그램과 관계 표기법

UML은 다양한 다이어그램을 통해 시스템의 구조와 동작을 시각적으로 표현하며, 주요 구성 요소로는 클래스(Class), 속성(Attribute), 관계(Relationship), 다이어그램(Diagram) 등이 있다. 이들 구성 요소는 시스템의 객체, 속성, 그리고 객체들 간의 상호작용을 명확하게 나타내며, 연구 데이터 시스템에서도 이러한 요소들을 활용하여 데이터를 설계하고 관리할 수 있다.

클래스는 객체의 틀 또는 설계도를 나타내며, 시스템에서 관리해야 하는 데이터를 정의하는 역할을 한다. 연구 데이터 관리에서 클래스는 주요 연구 개체나 데이터를 나타낼 수 있다. 각 클래스는 속성(Attribute)과 메서드(Method)를 가지며, 객체의 상태와 동작을 정의한다. 클래스는 데이터베이스에서 테이블과 유사한 역할을 하며, 연구 데이터의 특정 유형을 표현한다. 예를 들어, '연구자(Researcher)' 클래스는 연구자의 이름, 소속, 연구 분야 등의 속성을 포함할 수 있으며, 연구자에 관련된 행동(예: 실험 수행)을 메서드로 정의할 수 있다. 속성은 클래스에 속하는 세부 정보를 설명하는 요소로, 각 클래스의 특성을 정의하는 데 사용된다. 연구 데이터에서 속성은 연구자의 이름, 실험의 날짜 등 구체적인 데이터를 나타낸다. 속성은 클래스가 관리하는 데이터의 구체적인 특성이다. 연구 데이터 시스템에서 각 클래스는 여러 속성을 가질 수 있으며, 이를 통해 데이터의 세부 사항을 정의할 수 있다. '연구자' 클래스의 속성은 '이름(Name)', '소속(Affiliation)', '연구 분야(Field of Study)' 등이 될 수 있다. '실험' 클래스는 '날짜(Date)', '결과(Result)' 등의 속성을 가질 수 있다. 관계는 클래스 간의 연관성을 나타내며, 두 개 이상의 클래스가 어떻게 상호작용하는지 설명한다. 연구 데이터 시스템에서 관계는 연구자와 실험, 실험과 데이터셋 간의 연결을 정의하는 데 사용된다. UML에서는 주로 상속(Inheritance), 연관(Association), 집합(Aggregation), 포함(Composition) 등의 관계를 표현한다. 관계는 클래스 간의 연결을 나타내며, 시스템에서 객체들 간의 상호작용을 정의하는 중요한 요소다. 이를 통해 시스템 내에서 데이터 간의 의존성이나 상호작용을 명확히 설명할 수 있다. '연구자'와 '실험' 클래스 간에는 '연구자가 실험을 수행한다'는 연관 관계가 있을 수 있으며, '실험'과 '데이터셋' 클래스 간에는 '실험 결과로 데이터셋을 생성한다'는 포함 관계를 정의할 수 있다.

UML에서는 대표적으로 클래스 다이어그램(Class Diagram), 시퀀스 다이어그램(Sequence Diagram), 활동 다이어그램(Activity Diagram) 등이 있으며, 연구 데이터 시스

템을 시각적으로 표현하는 데 활용된다. 클래스 다이어그램은 클래스와 클래스 간의 관계를 시각적으로 표현한다. 연구 데이터 관리에서는 연구자의 구조, 실험 과정, 데이터셋 간의 관계 등을 명확하게 나타낼 수 있다. 시퀀스 다이어그램은 객체 간의 상호작용을 시간의 흐름에 따라 표현한다. 연구 데이터 시스템에서 작업 흐름이나 데이터 처리 과정을 시각적으로 모델링하는 데 사용될 수 있다. 활동 다이어그램은 시스템 내에서 수행되는 활동이나 작업의 흐름을 표현한다. 연구 데이터를 처리하는 과정에서 발생하는 다양한 활동을 시각화할 수 있다.

UML은 다양한 소프트웨어 시스템과 데이터베이스 설계, 소프트웨어 아키텍처에 사용되며, 이를 통해 시스템의 구조와 동작을 시각적으로 표현하는 데 매우 유용하다. 소프트웨어 개발 라이프사이클에서 요구 분석, 설계, 구현에 이르기까지 UML은 전반적인 과정에 사용된다.

RDF(Resource Description Framework)

주로 웹 상에서 데이터를 구조화하고 메타데이터를 표현하는 표준 포맷이다. RDF는 연구 데이터 관리와 상호운용성 측면에서 중요한 역할을 하며, 데이터를 트리플(주어-서술어-목적어) 형태로 표현하여 데이터 간의 관계를 명확하게 정의한다. 연구 데이터는 여러 소스에서 발생하며, 이를 체계적으로 관리하고 다른 시스템과 상호작용하기 위해 RDF를 활용하면 데이터 간의 의미적 관계를 표현하고 상호연결성을 강화할 수 있다. RDF는 특히 시맨틱 웹과 같은 분야에서 데이터를 의미적으로 연결하고 공유하는 데 사용되며, 연구 데이터 관리 시스템에서도 데이터의 상호운용성과 통합을 촉진한다.

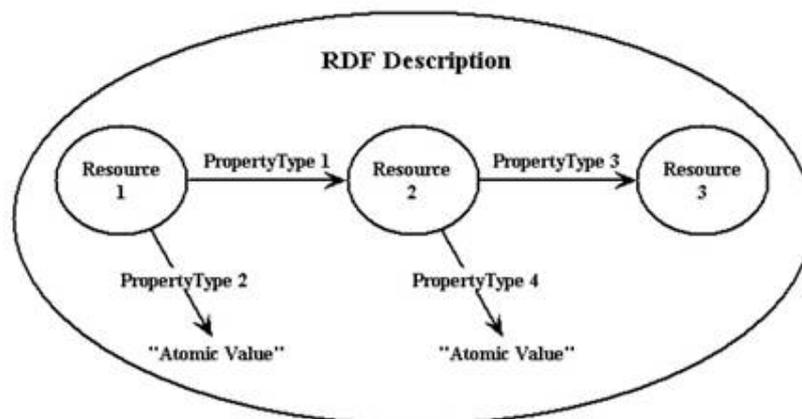


그림 5-4. RDF(Resource Description Framework)의 개요 (출처 : Danjuma, 2013)

RDF의 구성 요소는 크게 주어(Subject), 서술어(Predicate), 목적어(Object)로 나뉘며, 이를 통해 데이터 간의 관계를 트리플 구조로 표현한다. 이 트리플 구조는 데이터의 상호 연결성을 표현하는 기본 단위로 사용된다. 주어(Subject)는 관계의 주체를 나타내며, RDF에서 설명하는 개체나 리소스를 의미한다. 연구 데이터 관리에서 주어는 데이터의 주된 대상이나 엔티티를 나타낸다. 예를 들어, 연구 논문이나 데이터셋 자체가 주어로 사용될 수 있다. 주어는 URI(Uniform Resource Identifier)를 통해 고유하게 식별되며, 이를 통해 웹 상의 다른 데이터와 연결될 수 있다. 서술어(Predicate)는 주어와 목적어 간의 관계를 설명하는 역할을 한다. 서술어는 주어에 대해 특정한 속성이나 관계를 설명하는 요소로, 주로 메타데이터를 정의하는 데 사용된다. 서술어는 주어와 목적어를 연결하는 역할을 하며, 이들을 통해 데이터 간의 관계를 명확히 설명할 수 있다. 예를 들어, 연구자가 논문을 작성했다는 관계를 설명할 때, '작성자'가 서술어 역할을 하며 주어(논문)와 목적어(연구자)를 연결한다. 목적어(Object)는 주어에 대한 특정 값을 제공하거나, 주어와 연관된 다른 리소스를 나타낸다. 목적어는 주어와 서술어를 통해 설명되는 대상이나 값을 나타내며, 리터럴 값(예: 텍스트, 숫자) 또는 다른 리소스(다른 주어)가 될 수 있다. 예를 들어, '연구자는 논문을 작성했다'라는 문장에서, '논문'이 목적어 역할을 하며, 이는 연구자가 작성한 결과물을 나타낸다.

RDF는 트리플 구조를 통해 데이터를 표현함으로써, 데이터 간의 의미적 연결을 가능하게 한다. 연구 데이터 관리에서 RDF는 다양한 데이터 소스를 통합하고, 데이터 간의 관계를 시맨틱하게 표현함으로써 상호운용성을 극대화할 수 있다. RDF는 주로 시맨틱 웹과 같은 환경에서 사용되지만, 연구 데이터 관리에서도 메타데이터를 명확히 정의하고 데이터를 체계적으로 구조화하는 데 매우 유용하다.

RDF는 연구 데이터와 메타데이터를 관리하는 데 있어 중요한 역할을 하며, 특히 데이터의 의미적 관계를 표현하는 데 강점을 가지고 있다. 이를 통해 연구 데이터는 표준화되고, 시스템 간의 상호작용이 원활해지며, 다양한 데이터 소스 간의 상호 연결성을 확보할 수 있다. RDF는 URI를 통해 전 세계적으로 데이터를 고유하게 식별하고 연결할 수 있기 때문에, 연구 데이터의 재사용성과 상호운용성을 위한 도구로 사용된다.

OWL(Web Ontology Language)

RDF 기반의 온톨로지 언어로, 데이터 간의 의미적 관계와 논리적 제약을 더 구체적으로 정의하는 데 사용된다. 연구 데이터 관리와 상호운용성 측면에서 OWL은 복잡한 데이터 간의 의미적 연결을 표현하고, 데이터의 계층적 구조를 명확히 정의할 수 있는 도구로 중요한 역할을 한다. OWL은 RDF의 확장판으로, 더 정교한 데이터 표현과 추론 기능을 제공하여 연구 데이터를 시맨틱하게 모델링하고 관리할 수 있게 해준다.

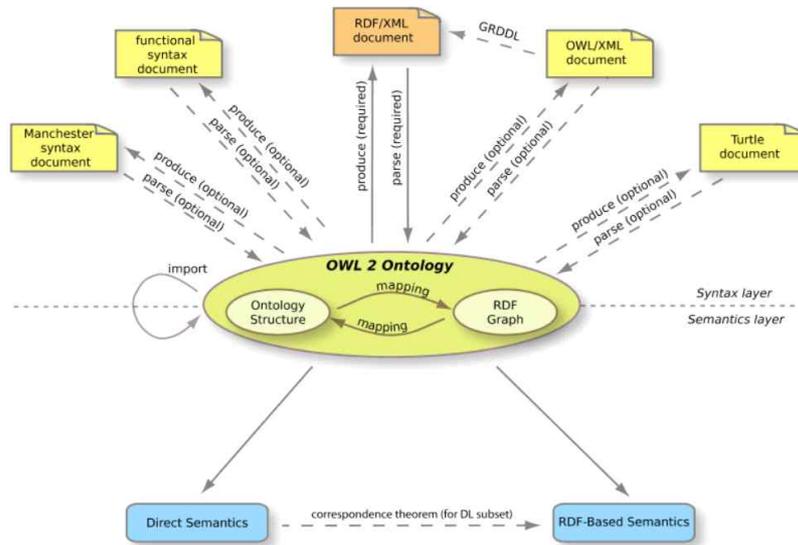


그림 5-5. OWL의 구조 (출처 :World Wide Web Consortium. 2012))

OWL의 핵심 구성 요소는 클래스(Class), 개체(Individual), 속성(Property), 관계(Relationship)로 나뉜다. 이러한 요소들은 연구 데이터 간의 논리적 구조를 정의하고, 연구 데이터베이스에서 복잡한 관계를 효과적으로 표현하는 데 사용된다. 클래스(Class)는 데이터의 집합이나 범주를 정의하는 역할을 한다. 연구 데이터 관리에서 클래스는 특정 데이터 유형이나 개념을 나타내며, 예를 들어 '연구자', '논문', '실험'과 같은 데이터 유형을 정의할 수 있다. 클래스는 데이터의 범주화를 통해 데이터 간의 유사성을 묶어 주며, 상위 클래스와 하위 클래스 간의 관계를 설정하여 데이터의 계층적 구조를 표현할 수 있다. 예를 들어, '사람' 클래스는 '연구자'나 '학생' 클래스의 상위 클래스가 될 수 있다. 개체(Individual)는 클래스에 속하는 특정 인스턴스 또는 실제 객체를 의미한다. 연구 데이터에서 개체는 실질적인 데이터를 나타내며, 클래스에 속한 구체적인 항목을 표현한다. 예를 들어 'John Doe'는 '연구자' 클래스에 속하는 개체로 정의될 수 있다. 개체는 클래스의 특성을 상속받으며, 클래스 간의 관계에 따라 상호작용할 수 있다. 속성

(Property)는 클래스나 개체 간의 관계나 특성을 정의한다. OWL에서 속성은 객체 속성(Object Property)과 데이터 속성(Data Property)으로 나뉜다. 객체 속성은 두 개체 간의 관계를 정의하며, 데이터 속성은 개체와 특정 데이터 값 간의 관계를 나타낸다. 예를 들어 '연구자' 클래스에서 '논문을 작성했다'는 객체 속성이 될 수 있고, '이름'이나 '생년월일' 같은 속성은 연구자의 개인 정보를 설명하는 데이터 속성으로 사용될 수 있다. 관계(Relationship)는 클래스나 개체 간의 상호작용이나 연관성을 나타낸다. OWL에서는 클래스 간의 상속 관계뿐만 아니라 상호배타성, 동등성 등의 복잡한 관계를 정의할 수 있다. 연구 데이터 관리에서 관계는 다양한 데이터 간의 연관성을 표현하며, 데이터를 논리적으로 연결하는 데 중요한 역할을 한다. 예를 들어, '연구자'가 '논문'을 작성한다는 관계는 연구 데이터 간의 상호작용을 표현하는 데 사용할 수 있다.

OWL의 중요한 기능 중 하나는 추론이다. OWL은 정의된 클래스, 개체, 속성 간의 관계를 기반으로 자동으로 새로운 정보를 추론할 수 있다. 예를 들어, 특정 연구자가 실험을 수행했다는 사실이 주어지면, 그 연구자가 해당 연구 데이터를 생성했다는 추가 정보를 자동으로 도출할 수 있다. 이를 통해 연구 데이터는 더 풍부한 의미적 정보를 제공하게 되고, 상호운용성 및 데이터의 재사용성이 극대화된다. OWL은 연구 데이터 관리에서 복잡한 데이터 구조와 논리적 제약을 표현하는 데 매우 유용하며, 특히 데이터 간의 계층적 관계나 의미적 연결을 표현할 때 큰 강점을 가지고 있다. 이를 통해 연구 데이터는 더 정교하게 관리되고, 데이터 간의 의미적 일관성을 유지할 수 있다. OWL은 RDF와 함께 사용되며, URI를 기반으로 전 세계적으로 데이터 간의 상호 연결성을 강화하고, 연구 데이터의 상호운용성과 재사용성을 촉진하는 데 사용할 수 있다.

5.3 FAIR 데이터 포맷

FAIR 원칙(Findable, Accessible, Interoperable, Reusable)은 사람과 기계가 모두 이해할 수 있는 데이터 포맷을 지향하며, 데이터를 효율적이고 체계적으로 구조화할 수 있는 다양한 데이터 형식을 사용한다. 이러한 원칙은 데이터를 보다 쉽게 검색하고 접근하며, 시스템 간 상호운용성을 보장하고 재사용 가능하게 하려는 목적을 가진다. 이를 통해 데이터는 특정 도메인뿐만 아니라 다양한 환경에서도 일관성 있게 관리되고 해석될 수 있다.

데이터를 표현하는 대표적인 구조화된 포맷으로는 CSV, XML, JSON, RDF, GeosciML 등이 있다. 각각의 포맷은 데이터의 성격에 따라 적합한 역할을 하며, CSV는 단순한 테이블 형식의 데이터 교환에 적합하고, XML과 JSON은 계층적이거나 다양한 데이터 구조를 표현하는 데 사용된다. RDF는 데이터 간의 의미적 관계를 표현하는 데 유리하며, GeosciML은 지질학적 데이터를 위한 특화된 형식으로 복잡한 지질 정보를 관리하는 데 적합하다.

CSV(Comma-Separated Values)

데이터를 간단하고 효율적으로 저장하는 텍스트 기반의 파일 형식으로, 각 데이터 항목을 쉼표로 구분한다. 주로 데이터를 테이블 형태로 저장하는 데 사용되며, 행은 데이터의 레코드, 열은 속성을 나타낸다. CSV는 데이터베이스, 스프레드시트, 프로그래밍 언어 등에서 널리 사용되며, 대량의 데이터를 간단하게 저장하고 교환하는 데 매우 유용하다.

CSV의 주요 구성 요소는 행, 열, 구분자로 나뉜다. 이러한 요소들은 데이터를 간결하게 표현하고 저장하는 데 적합한 구조로 구성된다. 행은 CSV 파일에서 데이터의 한 레코드를 나타낸다. 각 행은 고유한 데이터 항목을 저장하며, 연구 데이터 관리에서 한 명의 연구자 정보나 하나의 실험 결과를 하나의 행에 담을 수 있다. 예를 들어, 실험 결과나 연구자의 정보를 기록할 때 각 행은 하나의 독립적인 데이터 항목을 저장한다. 열은 행에서 각 데이터 항목의 속성을 나타낸다. CSV 파일의 각 열은 특정 데이터를 설명하며, 동일한 속성 집합을 가진 데이터들이 행으로 나열된다. 예를 들어, 연구자의 이름, 소속, 연구 분야 등이 각각의 열로 나타날 수 있다. 열은 데이터의 속성을 정의하고, 데이터베이스나 스프레드시트 프로그램에서 쉽게 관리된다. 구분자는 CSV 파일에서 데이터 항목 간의 구분을 담당한다. 기본적으로 쉼표가 사용되지만, 필요에 따라 탭(TSV 형식)이나 세미콜론을 구분자로 사용할 수도 있다. 데이터 항목에 쉼표가 포함될 경우 해당 데이터는 따옴표로 묶여 처리된다. 이는 CSV 파일의 유연성을 높이고, 다양한 상황에서 데이터를 효과적으로 저장할 수 있도록 한다.

CSV는 데이터 관리에서 간결하고 가벼운 형식이기 때문에 대규모 데이터를 다룰 때 매우 효율적이다. 특히 데이터베이스에서 데이터를 추출하거나 다른 시스템으로 전송할 때 CSV는 데이터 교환 포맷으로 사용될 수 있다. 이 파일 형식은 사실상의 표준으로,

다양한 애플리케이션에서 지원되며, 상호운용성이 높다. 스프레드시트 프로그램(예: Excel), 데이터베이스 관리 시스템, 통계 분석 소프트웨어에서 CSV는 중요한 파일 형식으로 사용된다.

CSV의 장점 중 하나는 파일 크기가 작고, 복잡한 구조가 없다는 점이다. 이는 대량의 데이터를 빠르게 처리하고 저장하는 데 유리하다. 또한 CSV 파일은 텍스트 기반이므로 사람이 읽고 수정할 수 있으며, 데이터의 간단한 수정과 관리가 가능하다. 그러나 CSV 파일은 기본적으로 2차원 구조만을 지원하며, 복잡한 계층적 데이터나 관계형 데이터를 표현하는 데는 한계가 있다. 또한 CSV 파일 자체에는 메타데이터가 포함되지 않기 때문에, 데이터의 의미나 구조를 설명하려면 별도의 문서화가 필요하다.

CSV는 데이터의 구조를 단순하게 표현하고, 시스템 간 데이터 교환을 효율적으로 수행할 수 있는 포맷이다. 특히 연구 데이터 관리에서 대용량 데이터를 저장하거나 다른 시스템으로 전송하는 데 자주 사용된다. 상호운용성은 뛰어나지만, 복잡한 데이터 구조나 메타데이터 처리가 필요한 경우에는 다른 포맷을 보완적으로 사용해야 한다.

XML(eXtensible Markup Language)

데이터를 구조화하여 저장하고 교환하는 데 사용되는 마크업 언어로, 다양한 응용 프로그램 간의 상호운용성을 높이는 데 중요한 역할을 한다. XML은 태그를 사용하여 데이터를 계층적으로 구조화하며, 데이터의 의미를 명확하게 표현할 수 있다. 연구 데이터 관리에서 XML은 복잡한 데이터 구조를 효율적으로 표현하고, 데이터를 다른 시스템과 교환할 때 널리 사용된다.

XML의 주요 구성 요소는 요소(Element), 속성(Attribute), 텍스트 내용(Content), 루트 요소(Root Element)로 나뉜다. 이러한 구성 요소들은 XML 파일의 데이터 구조를 명확하게 정의하고, 계층적 관계를 표현하는 데 사용된다.

요소(Element)는 XML 문서의 기본 단위로, 데이터를 태그로 감싸서 표현한다. 각 요소는 시작 태그와 종료 태그로 이루어져 있으며, 그 안에 다른 요소나 텍스트 데이터를 포함할 수 있다. 연구 데이터 관리에서 요소는 다양한 데이터 항목을 표현하는 데 사용될 수 있다. 예를 들어, 연구자, 실험, 데이터셋 등을 각각의 요소로 정의하여 데이터를 저장할 수 있다.

속성(Attribute)은 요소에 대한 추가 정보를 제공하는 데 사용된다. 속성은 요소의 시작 태그 내에 정의되며, 해당 요소의 특성을 설명하는 역할을 한다. 연구 데이터 관리에서 속성은 연구자 이름, 실험 날짜, 데이터 형식과 같은 정보를 저장하는 데 유용하다. 속성은 요소와 밀접하게 관련된 부가 정보를 제공함으로써 데이터를 보다 상세하게 설명할 수 있다.

텍스트 내용(Content)은 요소 안에 포함된 실제 데이터를 의미한다. 요소가 데이터를 감싸고 있으며, 그 안의 텍스트 내용이 실제 데이터 값으로 사용된다. 예를 들어, 연구자의 이름, 실험 결과 등의 실제 데이터를 텍스트 내용으로 포함시킬 수 있다. 텍스트 내용은 단순한 문자열일 수도 있고, 복잡한 계층 구조 내에서 여러 요소와 함께 저장될 수 있다.

루트 요소(Root Element)는 XML 문서의 최상위 요소로, 모든 다른 요소를 포함하는 역할을 한다. XML 문서는 반드시 하나의 루트 요소를 가져야 하며, 이 루트 요소 안에 다른 모든 데이터가 계층적으로 배치된다. 연구 데이터 관리에서 루트 요소는 전체 데이터를 포괄하는 개념으로, 예를 들어 '연구 데이터'라는 루트 요소 아래에 연구자, 실험, 결과 등의 요소가 포함될 수 있다.

XML은 데이터의 계층적 구조를 명확하게 정의할 수 있기 때문에 복잡한 데이터 모델을 표현하는 데 매우 적합하다. 연구 데이터 관리에서는 실험 데이터, 연구 결과, 메타데이터 등을 구조화하여 저장하고, 다양한 시스템 간에 데이터를 교환하는 데 사용된다. XML은 다양한 소프트웨어와 호환되며, 특히 웹 환경에서 데이터를 주고받는 표준 형식으로 자리잡고 있다.

XML의 장점은 데이터의 의미를 태그로 명확히 정의할 수 있다는 점이다. 이를 통해 데이터를 쉽게 해석하고, 다양한 시스템에서 일관된 방식으로 처리할 수 있다. 또한 XML은 확장성이 뛰어나며, 데이터를 추가하거나 수정할 때 기존 구조를 유지하면서 새로운 데이터를 쉽게 추가할 수 있다. 그러나 XML은 데이터의 표현이 비교적冗長하여, 큰 규모의 데이터를 처리할 때 파일 크기가 커지고, 성능이 저하될 수 있다는 단점이 있다.

결론적으로, XML은 연구 데이터 관리에서 복잡한 데이터 구조와 메타데이터를 정의하는 데 매우 유용한 포맷이다. 계층적 데이터 모델을 표현할 수 있고, 다양한 시스템 간에 데이터를 상호운용할 수 있게 해준다. XML은 연구 데이터의 구조와 의미를 명확하게

정의할 수 있는 도구로, 데이터의 일관성과 재사용성을 높이는 데 기여한다.

JSON(JavaScript Object Notation)

데이터를 구조화하여 저장하고 전송하는 데 사용되는 경량의 데이터 교환 형식이다. JSON은 사람과 기계가 모두 쉽게 읽고 쓸 수 있는 형식으로, 주로 웹 애플리케이션에서 서버와 클라이언트 간 데이터를 주고받을 때 널리 사용된다. 연구 데이터 관리에서도 효율적인 데이터 교환을 위해 자주 활용되며, 객체와 배열을 기반으로 데이터를 계층적으로 표현할 수 있어 복잡한 데이터 구조를 단순하고 명확하게 표현하는 데 적합하다.

JSON의 주요 구성 요소는 객체(Object), 배열(Array), 키(Key), 값(Value)로 나뉜다. 이러한 구성 요소는 데이터를 명확하게 구조화하고, 다양한 형태의 데이터를 효율적으로 저장하는 데 사용된다. 객체(Object)는 JSON 데이터의 기본 단위로, 중괄호로 감싸인 키-값 쌍의 집합으로 이루어진다. 각 객체는 특정 데이터를 표현하며, 다양한 속성(키)과 그에 해당하는 값을 포함할 수 있다. 연구 데이터 관리에서는 객체를 사용하여 연구자, 실험, 데이터셋 등의 데이터를 표현할 수 있다. 예를 들어, 하나의 연구자가 객체로 표현될 수 있으며, 그 안에 이름, 소속, 연구 분야 등의 정보가 키-값 쌍으로 저장된다. 배열(Array)은 객체 또는 값들의 목록을 나타내며, 대괄호로 감싸진다. 배열은 여러 개의 데이터를 순서대로 저장할 수 있는 구조로, 연구 데이터에서 실험 결과나 여러 연구자 정보를 저장하는 데 유용하다. 예를 들어, 여러 명의 연구자 정보를 배열로 표현하거나, 특정 실험에서 얻은 여러 결과 값을 배열로 저장할 수 있다. 키(Key)는 객체 내에서 데이터의 속성을 정의하는 요소로, 고유한 식별자 역할을 한다. 각 키는 값과 쌍을 이루며, JSON 구조에서 데이터를 명확하게 식별하고 접근할 수 있게 한다. 연구 데이터 관리에서는 키를 사용해 데이터 항목을 정의하고, 해당 데이터에 접근할 수 있도록 한다. 예를 들어, 연구자의 이름, 소속, 연구 분야 등이 각각 키로 정의될 수 있다. 값(Value)은 키에 대응하는 실제 데이터를 나타낸다. 값은 문자열, 숫자, 불리언, 객체, 배열 등 다양한 형태로 나타낼 수 있으며, 연구 데이터의 구체적인 정보를 제공한다. 예를 들어, 연구자의 이름이나 실험 결과 값이 JSON에서 값으로 저장된다. 값은 데이터의 구체적인 내용을 표현하며, 키를 통해 접근된다.

JSON은 계층적 데이터를 명확하게 표현할 수 있는 유연한 구조를 가지고 있으며,

다양한 웹 애플리케이션과 데이터베이스에서 쉽게 처리할 수 있다. 연구 데이터 관리에서는 실험 데이터, 연구 결과, 메타데이터 등을 계층적으로 표현하고 저장할 수 있는 효율적인 포맷으로 사용된다. JSON은 특히 웹 기반 시스템에서 데이터를 주고받는 데 매우 적합하며, 클라이언트와 서버 간의 상호작용에서 널리 사용된다.

JSON의 장점은 간결한 구조와 유연성에 있다. JSON은 사람이 읽고 쓰기 쉬운 형식을 가지고 있어 데이터를 쉽게 이해할 수 있으며, 데이터의 구조를 명확하게 정의할 수 있다. 또한, 대부분의 프로그래밍 언어에서 JSON을 쉽게 파싱하고 생성할 수 있어 상호운용성이 뛰어나다. 그러나 JSON은 메타데이터를 직접 포함하지 않기 때문에 데이터를 설명하거나 추가 정보를 제공하기 위해 별도의 메타데이터가 필요할 수 있다.

GeoJSON

지리 공간 데이터를 표현하기 위한 JSON 기반의 포맷으로, 위치 정보와 관련된 데이터를 효율적으로 저장하고 교환하는 데 사용된다. 주로 웹 애플리케이션이나 GIS(지리정보시스템)에서 공간 데이터를 처리하고 시각화할 때 널리 활용된다. 연구 데이터 관리에서도 지리적 위치 정보가 포함된 데이터를 저장하고 처리하는 데 적합한 형식으로, 간결하고 쉽게 읽을 수 있는 구조를 가진다.

GeoJSON의 주요 구성 요소는 객체(Object), 속성(Properties), 지오메트리(Geometry)로 나뉜다. 이러한 요소들은 지리적 데이터를 효과적으로 표현하고, 좌표 기반 데이터를 체계적으로 저장하는 데 사용된다.

객체(Object)는 GeoJSON의 기본 단위로, 특정 지리적 데이터를 표현한다. 각 객체는 지리적 형상과 그와 관련된 속성 데이터를 포함할 수 있으며, 지리적 위치를 나타내는 점(Point), 선(LineString), 다각형(Polygon) 등의 형상 정보를 저장한다. 예를 들어, 연구 데이터 관리에서 특정 실험이 수행된 위치나 데이터 수집 지점을 GeoJSON 객체로 표현할 수 있다.

지오메트리(Geometry)는 객체의 위치 정보를 저장하는 요소로, 좌표를 기반으로 지리적 형상을 나타낸다. 점, 선, 다각형과 같은 기본적인 지리 공간 개념을 정의하며, 다중 지오메트리(GeometryCollection)를 사용하여 복수의 형상을 하나의 객체로 표현할 수 있다. 연구 데이터에서 지오메트리는 실험 현장이나 특정 지역의 경계 정보를 좌표로 저장하는 데 활용될 수 있다.

속성(Properties)은 객체에 대한 추가 정보를 제공하는 요소로, 지리적 형상에 부가적인 설명이나 특성을 부여한다. 연구 데이터 관리에서는 속성을 통해 위치와 관련된 데이터를 저장할 수 있다. 예를 들어, 실험 위치의 이름, 연구자의 이름, 실험 날짜 등을 속성으로 정의할 수 있다. 속성은 객체와 밀접하게 연관된 부가 정보를 제공함으로써 지리 공간 데이터를 보다 상세하게 설명할 수 있다.

GeoJSON은 좌표 기반의 지리 공간 데이터를 효율적으로 표현할 수 있는 포맷으로, 공간적 관계를 포함한 복잡한 지리 데이터를 직관적으로 표현할 수 있다. 연구 데이터 관리에서 GeoJSON은 특정 위치에서 수행된 실험 데이터, 지리적 분포를 나타내는 데이터셋 등을 관리하는 데 적합하다. 또한, GeoJSON은 다양한 웹 기반 지리 정보 서비스(GIS) 및 지도 서비스에서 지원되며, 위치 기반 데이터를 시각화하는 데 널리 사용된다.

GeoJSON의 장점은 간결한 데이터 표현과 유연성에 있다. JSON의 장점을 그대로 계승하면서도 지리적 데이터를 쉽게 구조화할 수 있어, 다양한 시스템에서 쉽게 처리되고 시각화될 수 있다. 특히 웹 맵핑 애플리케이션에서 좌표 기반 데이터를 시각적으로 표현하는 데 매우 적합하다. 그러나 GeoJSON은 대규모 데이터를 처리할 때 파일 크기가 커질 수 있으며, 좌표 시스템이 기본적으로 WGS84(위도, 경도)로 제한된다는 한계가 있다. GeoJSON은 지리적 위치 정보를 포함한 연구 데이터 관리에서 효율적이고 직관적인 데이터 포맷으로, 지리적 형상과 속성을 함께 표현할 수 있어 공간 데이터를 다루는 데 적합하다. GeoJSON을 사용하면 지리 공간 데이터를 시각화하고 공유하는 데 있어 상호운용성과 재사용성을 극대화할 수 있다.

RDF (Resource Description Framework)

웹 상에서 데이터를 구조화하고 메타데이터를 표현하기 위한 표준 포맷이다. RDF는 데이터를 트리플 형태로 표현하며, 데이터 간의 관계를 명확하게 정의하는 데 사용된다. RDF는 특히 시맨틱 웹과 링크드 데이터에서 데이터를 의미적으로 연결하고 상호운용성을 높이는 데 중요한 역할을 한다. 연구 데이터 관리에서도 RDF는 데이터의 의미적 관계를 표현하고, 데이터를 표준화하여 상호운용성을 강화하는 데 자주 사용된다.

RDF의 주요 구성 요소는 주어(Subject), 서술어(Predicate), 목적어(Object)로 나뉜다. 이러한 구성 요소는 데이터를 트리플 구조로 표현하여 정보 간의 관계를 명확하게 설명하는 데 사용된다. 주어(Subject)는 RDF 문장에서 데이터의 주체를 나타내며, 설명하고자

하는 대상이나 리소스를 의미한다. 연구 데이터 관리에서 주어는 연구자, 데이터셋, 실험과 같은 주요 개체를 나타낼 수 있다. RDF에서 주어는 주로 URI(Uniform Resource Identifier)를 사용해 고유하게 식별되며, 이를 통해 웹 상의 다른 데이터와 연결될 수 있다. 서술어(Predicate)는 주어와 목적어 간의 관계를 설명하는 역할을 한다. 서술어는 주어에 대해 특정한 속성이나 관계를 설명하는 요소로, 연구 데이터의 특성을 나타내는 메타데이터를 정의할 때 사용된다. 예를 들어, 서술어는 연구자와 실험 간의 관계나 연구 데이터의 속성을 설명하는 데 사용될 수 있다. 목적어(Object)는 주어에 대한 특정 값을 제공하거나, 주어와 연관된 다른 리소스를 나타낸다. 목적어는 주어와 서술어를 통해 설명되는 대상이나 값을 의미하며, 연구 데이터에서 연구 결과나 실험 조건 등을 목적어로 사용할 수 있다. 목적어는 리터럴 값(문자열, 숫자 등) 또는 또 다른 리소스(다른 주어)가 될 수 있다.

RDF는 이러한 트리플 구조를 통해 데이터를 표현함으로써, 데이터 간의 관계와 의미적 연결을 가능하게 한다. 연구 데이터 관리에서 RDF는 다양한 데이터 소스와 시스템 간에 상호운용성을 보장하고, 데이터 간의 연관성을 명확히 정의하는 데 중요한 역할을 한다. RDF는 또한 지식 그래프와 같은 구조에서 데이터를 표현하고, 데이터를 통합하는 데 사용될 수 있다.

RDF의 장점은 데이터를 의미적으로 표현하고 연결할 수 있다는 점에 있다. RDF는 단순한 데이터 저장을 넘어서, 데이터 간의 복잡한 관계를 표현할 수 있기 때문에 연구 데이터 관리에서 매우 유용하다. 또한 RDF는 시맨틱 웹 기술을 지원하므로, 데이터를 웹 상에서 상호 연결하고 검색할 수 있는 강력한 기능을 제공한다. 그러나 RDF는 비교적 복잡한 구조를 가지며, 대규모 데이터를 처리할 때 성능 저하가 발생할 수 있다는 한계도 있다. RDF는 연구 데이터 관리에서 데이터의 의미적 관계와 구조를 표현하는 강력한 도구로, 특히 시맨틱 웹 및 링크드 데이터 환경에서 데이터를 상호운용 가능하게 만들고, 표준화된 형식으로 저장하여 재사용성을 극대화할 수 있다. RDF는 연구 데이터를 명확하게 정의하고 연결할 수 있는 효율적인 포맷으로, 다양한 시스템 간의 데이터 교환을 가능하게 한다.

GeosciML(Geoscience Markup Language)

지구과학 데이터를 표준화하여 교환하고 표현하기 위한 XML 기반의 포맷이다. 주

로 지구과학 정보를 저장하고 공유하는 데 사용되며, 다양한 지질 데이터 소스 간 상호운용성을 높이는 역할을 한다. GeosciML은 지구과학 특성, 지층, 암석, 광물, 지질 경계 등과 같은 지구과학 개념을 모델링하여 데이터 교환을 지원하며, 지구과학 데이터의 일관된 표현을 보장하는 표준 포맷으로 자리잡고 있다.

GeosciML의 주요 구성 요소는 지질 객체(Geologic Object), 속성(Properties), 관계(Relationships)로 나뉜다. 이러한 요소들은 지구과학 데이터를 구조화하여 명확하게 표현하고, 데이터를 다양한 시스템에서 일관성 있게 처리할 수 있도록 설계되었다. 지질 객체(Geologic Object)는 GeosciML에서 지구과학 개념을 정의하는 기본 단위로, 지구과학 요소나 특성을 나타낸다. 연구 데이터 관리에서 지질 객체는 특정 지층, 암석 유형, 지질 구조 등을 표현할 수 있다. 예를 들어, 지질 단위, 광물 자원, 단층 등의 지구과학 요소가 지질 객체로 정의될 수 있다. 각 객체는 고유한 특성과 관계를 가지며, 해당 지질 정보를 구조화하여 저장한다. 속성(Properties)은 지질 객체의 특성이나 속성을 설명하는 요소로, 객체에 대한 구체적인 데이터를 제공한다. 연구 데이터 관리에서는 각 지질 객체에 속성들이 할당되어 구체적인 정보를 나타낸다. 예를 들어, 지질 단위의 이름, 암석의 화학적 성분, 지층의 두께 등이 속성으로 포함될 수 있다. 속성은 지질 객체를 설명하는 부가적인 정보를 제공하여 데이터의 의미를 명확하게 한다. 관계(Relationships)는 지질 객체 간의 상호 연관성을 설명하는 역할을 한다. GeosciML에서는 객체 간의 관계를 정의하여, 지구과학 데이터가 어떻게 연결되어 있는지 설명할 수 있다. 연구 데이터 관리에서 관계는 서로 다른 지구과학 특성 간의 연결을 표현하며, 예를 들어 특정 지질 단위가 다른 단위와 어떻게 접촉하고 있는지, 지층 간의 경계는 어떻게 이루어졌는지를 설명할 수 있다.

GeosciML은 지구과학 데이터를 계층적으로 표현할 수 있어, 복잡한 지구과학 개념을 명확하게 모델링할 수 있다. 연구 데이터 관리에서 GeosciML은 지구과학 데이터의 표준화를 지원하고, 다양한 시스템 간의 데이터 교환을 원활하게 한다. 특히 지리 정보 시스템(GIS)과 연계하여 지질 데이터를 시각화하고 분석하는 데 사용되며, 지질 데이터의 상호운용성과 재사용성을 극대화할 수 있다.

GeosciML의 장점은 지구과학 데이터를 명확하게 구조화하고, 이를 다른 시스템과 공유할 수 있다는 점에 있다. GeosciML은 복잡한 지질 데이터 구조를 일관되게 표현할 수 있기 때문에 지질 연구나 자원 탐사에서 매우 유용하다. 또한, XML 기반으로 설계되

어 다양한 웹 애플리케이션이나 데이터베이스 시스템에서 쉽게 활용될 수 있다. 그러나 GeosciML은 XML 기반이기 때문에 데이터의 표현이 다소 장황해질 수 있으며, 대규모 데이터를 처리할 때는 성능 저하가 발생할 수 있다는 단점이 있다.

GeosciML은 지구과학 데이터 관리와 교환에서 표준화된 포맷으로, 지질 데이터의 의미적 일관성을 보장하고 상호운용성을 높이는 데 중요한 역할을 한다. GeosciML은 연구 데이터를 명확하게 정의하고 다양한 지구과학 정보를 체계적으로 저장할 수 있는 효율적인 포맷으로, 지구과학 데이터의 통합과 분석을 가능하게 한다.

가. 지질 단위(GeologicUnit)의 GeosciML 표현 (XML 형식)

상부 백악기층을 정의하며, 구성 되는 속성으로 사암과 이암이 포함되며, 지질 시대 (700 Ma)과 두께(100미터)에 대한 GeosciML에 대한 표현은 다음과 같다.

```
<gsml:GeologicUnit xmlns:gsml="http://www.opengis.net/gsml/4.1">
  <gml:name>상부 백악기층</gml:name>
  <gsml:rank>Formation</gsml:rank>
  <gsml:composition>
    <gsml:CompositionPart>
      <gsml:lithology>사암</gsml:lithology>
      <gsml:lithology>이암</gsml:lithology>
    </gsml:CompositionPart>
  </gsml:composition>
  <gsml:age>
    <gsml:GeologicAge>
      <gsml:beginning>700</gsml:beginning>
    </gsml:GeologicAge>
  </gsml:age>
  <gsml:thickness>
    <gsml:Thickness>100</gsml:Thickness>
  </gsml:thickness>
</gsml:GeologicUnit>
```

</gsml:GeologicUnit>

나. 단층(Fault)의 GeosciML 표현 (XML 형식)

단층을 정의하며, 그 구성 속성으로 단층의 이동 방향(남서쪽), 경사각(30도), 단층의 유형(역단층)에 대한 GeosciML에 대한 표준은 다음과 같다.

```
<gsml:GeologicFeature xmlns:gsml="http://www.opengis.net/gsml/4.1">
  <gml:name>남서향 역단층</gml:name>
  <gsml:GeologicStructure>
    <gsml:type>Fault</gsml:type>
    <gsml:movementDirection>남서</gsml:movementDirection>
    <gsml:dip>
      <gsml:Angle>30</gsml:Angle>
    </gsml:dip>
    <gsml:senseOfMovement>역단층</gsml:senseOfMovement>
  </gsml:GeologicStructure>
</gsml:GeologicFeature>
```

6. 참고 문헌

Albertoni, R., Browning, D., Cox, S., Gonzalez-Beltran, A. N., Perego, A., & Winstanley, P. (2024). The W3C data catalog vocabulary, version 2: Rationale, design principles, and uptake. *Data Intelligence*, 6(2), 457-487.

Ammann, N., Nielsen, L. H., Peters, C. S., & de Smaele, T. M. (2019). Datacite metadata schema for the publication and citation of research data.

Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., ... & Stall, S. (2020). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments.

Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., & Stein, L. A. (2004). OWL web ontology language reference. *W3C recommendation*, 10(2), 1-53.

Booch, G., Rumbaugh, J., & Jacobsen, I. (1997). UML: unified modeling language. Versão.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (1997). Extensible markup language (XML). *World Wide Web Journal*, 2(4), 27-66.

Butler, H., Daly, M., Doyle, A., Gillies, S., Hagen, S., & Schaub, T. (2016). The geojson format (No. rfc7946).

Candan, K. S., Liu, H., & Suvarna, R. (2001). Resource description framework: metadata and its applications. *Acm Sigkdd Explorations Newsletter*, 3(1), 6-19.

Candan, K. S., Liu, H., & Suvarna, R. (2001). Resource description framework: metadata and its applications. *Acm Sigkdd Explorations Newsletter*, 3(1), 6-19.

Danjuma, S., Zhou, J., Mei, H., Aliyu, A., & Waziri, U. (2013). Design, Analysis and Implementation of Semantic Web Applications. *International Journal of Computer Science Issues (IJCSI)*, 10(6), 110.

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., ... & White, A. (2020). FAIRsFAIR data object assessment metrics. DOI: <https://doi.org/10.5281/zenodo.4081213>.

Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., ... & Diepenbroek, M. (2021). From conceptualization to implementation: FAIR assessment of research data objects. *Data Science*

Journal, 20(1), 1-14.

Gnaiger, E. (2020). The elementary unit—canonical reviewer's comments on: Bureau International des Poids et Mesures (2019) The International System of Units (SI).

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: evolution of structured data on the web. *Communications of the ACM*, 59(2), 44-51.

Habermann, T. (2019). Mapping ISO 19115-1 geographic metadata standards to CodeMeta. *PeerJ Computer Science*, 5, e174.

International Bureau of Weights and Measures, Taylor, B. N., & Thompson, A. (2001). The international system of units (SI). Gaithersburg, MA, USA: US Department of Commerce, Technology Administration, National Institute of Standards and Technology.

Jacobson, L., & Booch, J. R. G. (2021). The unified modeling language reference manual.

Jerome, J. A. (2003). ISO 19115: Metadata standard proposed element set. UK GEMINI.

Kresse, W. (2004). Standardization of geographic information. University of Applied Sciences, Neubrandenburg.

Mäs, S., Henzen, D., Bernard, L., Müller, M., Jirka, S., & Senner, I. (2018, June). Generic schema descriptions for comma-separated values files of environmental data. In *The 21th AGILE International Conference on Geographic Information Science*.

Pan, Jeff Z. "Resource description framework." *Handbook on ontologies*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. 71-90.

Patel-Schneider, P. F. (2014). Analyzing schema.org. In *The Semantic Web-ISWC 2014: 13th International Semantic Web Conference*, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13 (pp. 261-276). Springer International Publishing.

Powell, A., Nilsson, M., Naeve, A., Johnston, P., & Baker, T. (2007, June 4). DCMI abstract model. DCMI. <https://dublincore.org/specifications/dublin-core/abstract-model/>

RDA FAIR Data Maturity Model Working Group, B. (2020). FAIR Data Maturity Model: specification and guidelines. Research Data Alliance.

Scheffler, M., Aeschlimann, M., Albrecht, M., Bereau, T., Bungartz, H. J., Felser, C., ... & Draxl, C. (2022). FAIR data enabling new horizons for materials research. *Nature*, 604(7907), 635-642.

Sen, M., & Duffy, T. (2005). GeoSciML: development of a generic geoscience markup language. *Computers & geosciences*, 31(9), 1095-1103.

Song, I. Y., & Froehlich, K. (1994). Entity-relationship modeling. *IEEE potentials*, 13(5), 29-34.

TTA, 2005, 표준화의 개요, 표준화 백서 4-9.

Thompson, A., & Taylor, B. N. (2008). Use of the international system of units (si). NIST Special Publication, Gaithersburg.

Weibel, S. L., & Koch, T. (2000). The Dublin core metadata initiative. *D-lib magazine*, 6(12), 1082-9873.

Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). Dublin core metadata for resource discovery (No. rfc2413).

World Wide Web Consortium. (2014). Data catalog vocabulary (DCAT).

World Wide Web Consortium. (2012). OWL 2 web ontology language document overview.

Yu, J. J. (2017). 공간정보 표준화 동향. *Information and Communications Magazine*, 34(4), 47-52.

Zhou, M. (2022, June). A review of Javascript object notation in data analysis. In 2022 8th international conference on humanities and social science research (ICHSSR 2022) (pp. 1771-1774). Atlantis Press.

da Silva Santos, L. O. B., Burger, K., Kaliyaperumal, R., & Wilkinson, M. D. (2023). FAIR data point: a FAIR-oriented approach for metadata publication. *Data Intelligence*, 5(1), 163-183.

국제단위계

김은정, & 남태우. (2012). 연구데이터 수집에 영향을 미치는 요인 분석. *정보관리학회지*, 29(2), 27-44.

김주섭, 김선태, & 최상기. (2019). 연구 데이터 관리 및 서비스를 위한 핵심요소의 기능적 요건. *한국문헌정보학회지 (Journal of the Korean Society for Library and Information Science)*, 53(3), 317-344.

리포지터리

심원식. (2015). 국가 차원의 연구데이터 관리체계 구축을 위한 로드맵 제안. *한국문헌정보학회지*, 49(4), 355-378.

윤석찬. (2002). 국제단위계 (SI) 의 이해와 올바른 사용. *전기저널*, 309, 58-62.

윤종민, 통신, 방송의, 융합, 선점을, 전략적, ... & 수요. (2001). 정보통신 표준화사업 관리의의미와 중요성. *TTA Journal*, 76, 33-36.

이호성. (2018). Redefinition of the SI base units: its background and

principle. Journal of the Korean Society for Precision Engineering, 35(4), 365-373.

정낙삼. (1997). 국제단위계 (SI) 해설. 전자파기술, 8(2), 66-74.

항공정보 및 항공지도 등에 관한 업무기준[국토교통부고시 제2021-1190호, 2021. 10. 27.]