과학기술데이터 공유 및 통합

건국대학교 정갑주 · 김동광 · 이종현
국민대학교 황선태

1. 필요성

e-Science 연구 환경은 전통적 환경에서 존재하는 컴퓨팅자원, 응용소프트웨어, 과학기술데이터, 실험장비 등의 공유와 통합을 구현하고, 이를 기반으로 다양한 기관 및 과학자들 간의 협업 방식의 새로운 연구기법 개발을 지원하는 것을 목표로 한다. 대부분의 연구들은 아직 컴퓨팅자원간의 공유 및 통합에 초점이 주어져 있고, 다른 자원들에 대한 연구는 아직 전 세계적으로 미흡한 수준이다. 본 논문에서는 **과학기술데이터 공유 및 통합**에서 고려되어야한 문제들 제시하고, 현재 연구 중인 해결 방안들에 대해서 설명하고, 대표적 해외 사례들을 간략히 소개한다.

효과적 데이터의 공유와 통합은 다음과 같은 관점에서 매우 중요하다.

- **중복 작업의 최소화.** 첨단과학 연구를 수행하기 위해서는 많은 물질적, 시간적 비용이 요구된다. 예를 들어, 단백질 분자 시뮬레이션 경우, 슈퍼컴퓨터를 이용해도 수개월이 걸리는 경우도 많다. 시뮬레이션 과정 및 결과 데이터가 공유가 되면, 불필요한 중복 실험을 피할 수 있게 되고, 이를 통해 보다 신속한 연구진행이 가능하고, 고가의 연구 장비의 효율적 이용이 가능하다.
- **다학제간 공동 연구 촉진.** 다른 학문분야에서 유사한 연구대상을 연구하는 경우는 매우 흔하다. 예를 들어, 의학과 생명공학, 생명공학과 환경공학 등의 경우 많은 유사 연구가 진행되고 있다. 그러나 학문 분야 간 교류가 제한적이어서 연구결과의 공유 또는 상호보완이 매우 제한적으로 이루어지고 있다. 다양한 분야 데이터들을 효율적으로 공유·통합할 수 있는 e-Science 연구 환경이 구축되면, 다학제간의 공동 연구를 촉진할 수 있다.
- **데이터 관리는 궁극적으로 지식관리.** 과학기술 연구에서 궁극적으로 모든 지식은 데이터로 저장된다. 따라서 효과적 데이터 관리는 지식의 축적, 지식의 전수·공유, 지식의 발전 등 첨단연구 활성화에서 반드시 요구된다.
- **시스템화된 데이터 관리를 통해 많은 단순 반복적 데이터 관리 수작업의 제거 가능.** 과학연구에서 연구자들이 실제 소모하는 대부분의 시간은 첨단연구문제에 대한 고민보다 잡다한 실험 데이터들에 대한 단순 반복적 관리 수작업에 소모된다. 따라서 이러한 데이터 관리 수작업들이 시스템화 되면 연구자들이 대부분의 시간을 첨단연구문제 해결에 사용할 수가 있고, 이는 곧 첨단 연구 활성화에 기여하게 된다.
- **데이터 분석 중심의 첨단연구 기법 개발을 촉진.** 많은 실험 데이터들이 축적이 되고, 공개 및 공유가 되면, 과거에는 불가능했던 다양한 복합 분석 기법들의 개발이 가능해 진다. 예를 들어, 유사한 연구결과들에 대한 비교 분석을 통해 공통적 유형을 파악하는 연구기법 등이 가능하다. IT 분야에서는 이러한 연구기법을 일반적으로 데이터 마이닝 (data mining)이라 하고, 실제 이미 다양한 학문분야에서 이러한 기법이 적용되고 있다.
- **실험 데이터 기반의 효과적 과학기술 전문가 양성 교육 프로그램 개발이 가능.** 첨단 과학기술 실험은 비용이 매우 높고 시간이 많이 요구되고, 소수의 과학자들에게만 실험 기회가 주어지기 때문에 실제 실험을 통한 다수의 전문가 양성은 현실적으로 단기간에 불가능하다. 그러나 실험 데이터들이 축적되고 공개가 되고, 이러한 실험 데이터들을 통한 교육 프로그램 (예, 가상실험 기반 교육프로그램)이 개발되는 경우 다수의 전문가들을 짧은 시간과 적은 비용으로 양성할 수가 있다.

2. 공유 · 통합 유형 및 도전과제

과학기술데이터 의 공유·통합 작업은 다양한 목적을 가지고 다양한 방식으로 진행되고 있다. 이러한 공유와 통합의 유형을 요약해 보면 다음과 같다.

- **대형 실험 결과의 공유.** 연구 결과로서 얻어지는 데이터의 크기가 거대하거나, 한 번의 실험을 위해서 많은 시간을 들여야 하는 연구 분야에 대한 데이터 공유·통합 시스템에 구축되고 있다. 예를 들어 분자 시뮬레이션과 같은 분야는 한 번의 실험을 위해서 수개월의 시간이 요구된다. 이러한 데이터들이 공유될 경우, 실험시간 및 비용이 엄청나게 절약될 수 있다. 구체적인 예로서, BioSimGrid 프로젝트를 들 수 있다[7].
- **지역적으로 분할된 데이터의 공유와 통합.** 대기과학, 지질과학, 천문학 등의 분야의 데이터들은 연구 특성상 어쩔 수 없이 지역적 제약을 가지게 된다. 따라서 특정 지역 연구팀들은 그 지역의 데이터에 집중하게 되고, 다른 지역 데이터에 대해서는 접근이 어렵게 된다. 이러한 분야에서는 여러 지역 데이터들 공유·통합할 수 있는 연구환경이 요구된다. 대표적인 예로서, GEON 프로젝트를 들 수 있다[6].
- **관련 학문 분야간 데이터 공유 및 통합.** 동일 또는 유사 연구대상에 대해서 여러 학문분야에서 연구를 수행하는 것이 매우 많고, 이 경우 다른 학문분야의 연구결과를 참고하면, 보다 효율적 연구진행이 가능하다. 하지만 학문분야간 교류가 제한적인 경우가 많다. 따라서 여러 학문분야간 데이터 공유를 위한 연구환경이 구축되면, 기존의 연구가 보다 효율적이 될 수 있고, 더 나아가 기존 환경에서 불가능했던 연구가 가능할 수 있다. 대표적인 예로서, BIRN 프로젝트를 들 수 있다[8].
- **실험 장치에서 생성되는 데이터 공유.** 고가의 실험장비를 이용하거나, 또는 실험 환경을 구축하기 위해서 많은 시간적 물질적인 비용이 소요되는 연구분야들이 있다. 이 연구 분야에서는 대 다수의 연구자들은 이러한 장비들을 사용할 기회를 얻지 못한다. 이러한 장비들을 이용한 실험 결과 데이터들이 공유될 수 있는 경우, 훨씬 많은 연구자들이 관련 연구들을 할 수가 있고, 또한 이들 데이터를 이용해서 교육이 가능하게 된다. 또한 가상실험과 같은 기법들을 이용해서 실제 장비를 사용하지 않고도, 실제 실험에 참여한 경험을 얻을 수가 있다. 구체적인 예로서, NEES 프로젝트를 들 수 있다[9].

이러한 과학기술데이터의 공유와 통합을 효과적으로 지원하기 위해서는 아래의 기술적 문제점들이 해결되어야 한다.

- 해당 과학기술분야 특성을 효과적으로 반영하는 **데이터 모델** 지원
- 기관별로 분산되어 있고, 다양한 포맷으로 저장되어 있고, 고속 전송이 요구되는 데이터를 위한 대규모 **분산 데이터 저장소** 지원
- 여러 데이터 모델로 저장된 데이터들에 대한 **효과적 검색** 지원

3. e-Science 데이터 그리드

e-Science 연구 환경에서 구체적으로 데이터 관리를 담당하는 환경은 데이터 그리드이다. e-Science 데이터 그리드 구축 시 고려되어야 할 사항들은 다음과 같다. 데이터 관리 방법에 대한 연구는 오랜 기간에 걸쳐서 비즈니스 분야를 대상으로 개발되어 왔다. 그러나 **과학기술데이터들은 비즈니스 데이터들과는 다른 특성을 갖고 있어 새로운 데이터 관리 기법 개발이 요구된다.** 과학기술데이터 관리를 위해서 요구되는 사항은 아래와 같다.

3.1 데이터 모델 관리

데이터 그리드에서 데이터를 통합 관리하기 위해서는 이를 표준적으로 표현하고 관리할 수 있는 데이터 모델의 정의가 필요하다. 그리드에서 관리하는 데이터는 크게 두 가지로 분류 된다.

- **데이터.** 연구과정에서 얻어지는 관측 데이터 (예, 센서 데이터 또는 동영상)와 이러한 데이터를 분석하여 얻어지는 분석결과, 마지막으로 이러한 분석결과를 토대로 작성된 보고서 또는 논문 등 다양한 종류의 데이터을 의미한다. 즉, 연구 과정을 통해서 생성되는 모든 자료들을 의미한다. 따라서 과학기술데이터들은 전통적 관계형 데이터 모델만으로는 표현이 어렵다.
- **메타데이터(데이터모델).** 데이터를 설명하는 데이터 (즉, 데이터에 관한 데이터)이다. 과학기술데이터들이 실험 전 과정을 통해서 생성되기 때문에, 이들 데이터들에 대한 메타데이터는 매우 중요하다. 예를 들어, 실험장치의 관측데이터가 생성된 경우, 해당 실험시설 이름, 실험자, 시간, 실험조건, 분석결과 등 실험 결과를 설명하는 정보가 메타데이터가 될 수 있다. 메타데이터의 가장 중요한 역할은 데이터 검색 시에 가능한 검색조건으로 사용되는 것이다. 사용자들은 실제 데이터에 대한 지식 없이도 메타 데이터를 통해서 검색이 가능하다. 따라서 사용자 입장에서는 메타데이터에 포함이 안 된 정보를 통해 실험 데이터를 검색 하는 것은 매우 어렵다고 보아야 한다.

기존 시스템 환경에서는 데이터를 관리하기 위해서

관계형 데이터베이스를 이용하는 것이 일반적인 방안이다. 테이블간의 관계를 ER 다이어그램과 같은 데이터 모델 정의 기법으로 표현하고, 데이터들을 SQL언어를 이용해서 관리한다. 하지만, **과학기술데이터는 이러한 테이블 형식으로 표현하기에는 데이터 유형이 너무 다양하고, 데이터 구조가 복잡하다. 더욱이 새로운 실험기법 또는 실험장비가 지속적으로 개발되기 때문에, 데이터 모델이 지속적으로 확장되어야 한다.** 이러한 특징으로 인해서, 관계형 데이터베이스를 과학기술데이터 관리에 적용하는 데에는 기술적 어려움이 많이 발생한다.

구체적인 과학기술데이터의 예로서, 토목분야 e-Science 환경인 NEES의 데이터 모델은 다음과 같다. 이 환경은 지진에 대한 건축물의 영향을 연구를 목표로 한다. 이 환경에서는 다양한 실험 장비를 이용하고, 수행하는 실험에 따라서 실험 환경과 측정데이터가 계속 바뀌게 된다. 따라서 이런 다양한 실험들에 통일되게 적용할 수 있는 데이터 모델 개발은 매우 어렵다. 따라서 NEES에서는 최대한 공통적 특성을 반영하는 **참조 데이터 모델**을 정의하고, 실험시설 또는 실험유형별로 수정 또는 확장해서 사용하는 전략을 채택하였다. 그림 1에서 참조 데이터 모델을 제시하였다. 이 참조 데이터 모델은 Site Specifications Database, Project Description, Domain Specifinc Models, Common Elements, Data의 계층으로 구분된다. 각각의 계층에는 다음과 같은 데이터 요소들이 존재한다. 다양한 실험 사이트(Site A, B, C)와, 각 사이트에는 다양한 실험 장비들(Equipments)와 연구원들(People)이 있다. 각 연구원들은 새로운 실험(Experiments)을 수행할 것이고, 실험의 횟수에 따라서 다양한 결과 데이터들이 생성되어 지게 된다. 최종적으로 생성되는 데이터들은 파일 저장소나 데이터베이스에 저장될 수 있다.

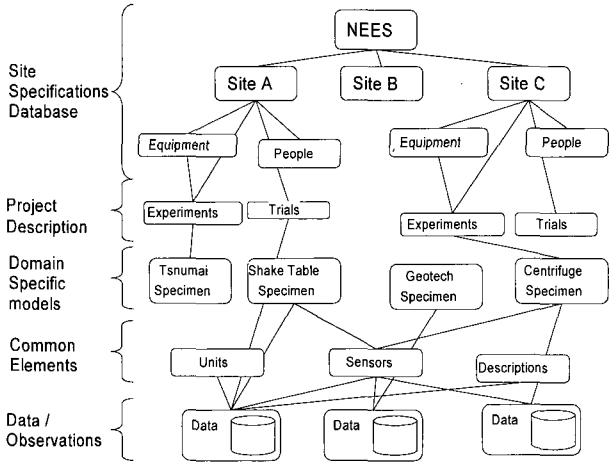


그림 1 NEES 참조 데이터 모델 개요

이러한 데이터 모델들을 표현하기위한 대표적 기법들로 ER 모델, 객체지향모델, XML 기반 모델, RDF 기반 모델 등이 존재한다. ER 모델과 객체지향모델은 오랜 기간 데이터 모델링 및 소프트웨어 개발에서 사용되어온 기법이다. XML 기반 모델링은 최근 제안된 방법으로 다양한 분야에서 활발하게 연구가 진행 중에 있다. XML 기반 모델링에서 XML Schema는 구체적으로 데이터모델의 구조, 내용, 의미 등을 정의하기 위해서 개발된 방법이다. RDF는 데이터 모델에서 구성요소간의 의미적 관계를 보다 체계적으로 표현하기위해서 제안된 기법으로 메타데이터를 보다 상세하게 표현할 수 있게 한다. 보다 체계화된 데이터 모델링을 위해서 Ontology에 대한 연구도 활발하게 진행되고 있다[1-5].

과학기술데이터를 위한 구체적인 데이터 모델 예로서 Ontology of Science[6], Sensor ML[7], NEES 참조 데이터 모델[8] 등을 들 수 있다. 현재 다양한 과학기술분야에서 이와 유사한 데이터 모델들이 개발되고 있다.

3.2 데이터 색인 및 검색 시스템

위에서 설명한 것과 같은 그리드 시스템의 다양한 형태의 데이터들을 효율적으로 통합하고 공유하기 위해서는 이질적인 데이터에 대해서 통합 색인 및 검색이 필수적으로 요구된다.

- **색인(indexing).** 데이터 등록 시 입력되는 메타데이터 정보를 검색 테이블 (인덱스 테이블)에 등록하여 검색 시 신속하게 찾을 수 있게 하는 작업.
- **검색.** 메타데이터 조건의 질의 시 인덱스 테이블에서 조건에 만족되는 실험 데이터들을 찾아주는 작업. 구체적인 질의 예로 "실험시설=3차원 조석수조 & 시간=2005년10월20일" 조건을 만족하는 실험 결과를 검색하라"를 들 수 있다.

그러나 과학기술 데이터들은 연구하는 주제 및 조건에 따라서 데이터 모델이 변화하는 특징이 있다. 많은 연구 활동을 하면서 이미 수행했던 연구와 비슷한 형태의 실험을 수행하게 되는데, 이때에 기존에 수행했던 연구의 데이터 모델을 수정함으로서 새로운 실험을 위한 데이터 모델이 만들어 진다. 데이터 모델의 변화는 크게 3가지로 정리해 볼 수 있다. **1) 새로운 실험 요소의 추가, 2) 기존 실험 요소의 제거, 3) 기존 실험 요소의 수정**이다. 이를 정리하면 그림 2와 같다. 이처럼 과학기술데이터들에서 효율적 색인 및 검색을 지원하기 위해서는 아래 기능이 추가로 요구된다.

- **새로운 데이터 모델의 지속적 지원.** 실험 시설별로

개별 데이터 모델, 그리고 실험 유형별로도 다양한 데이터 모델이, 그리고, 더 나아가 새로운 장비의 도입 시에도 새로운 데이터 모델이 요구될 수 있다. 즉, 편리하게 새로운 데이터 모델을 추가하고 기존의 모델을 확장할 수가 있어야 한다. 그림 2이 이러한 요구사항을 설명하고 있다.

- **다양한 메타데이터 모델 기반의 검색.** 다양한 메타데이터들이 존재할 경우, 데이터 검색 시에 어떤 메타데이터를 근거로 질의를 해야 할 지가 분명하지 않다. **사용자 입장에서는 메타데이터에 대한 정확한 지식 없이도 검색이 가능해야 한다.** 그림 3은 사용자의 편리성을 최대한 반영하는 검색 시스템 아키텍처를 제시하고 있다.

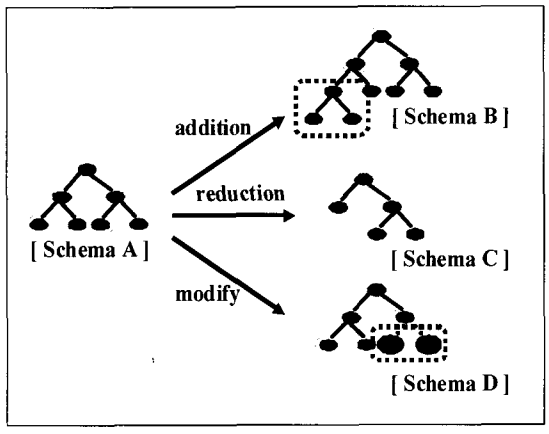


그림 2 데이터 모델 변경 지원

그림 2에서 제시된 통합 색인 및 검색 시스템에서는 각 실험 센터에서 각자의 실험 모델에 맞도록 데이터 저장소와 색인 시스템을 구축하고, 이 시스템을 이용해서 수행된 실험 데이터에 대한 데이터 모델을 정의하고 색인 정보를 관리할 수 있다. 이 색인정보를 이용해서 연구원들은 실험 데이터를 검색할 수 있다. 각각의 실험 기관에서 관리되고 있는 색인 정보와 데이터, 데이터 모델들은 중앙의 통합 데이터 저장소와 통합 색인 시스템에 의해서 통합되고 관리 될 수 있다. 이렇게 구축된 통합 색인 정보를 이용해서 다른 기관의 연구원들이 그리드 인프라를 통해서 구축되어 있는 전체 시스템에 대해서 통합 검색을 수행 할 수 있게 되는 것이다.

이렇게 구축된 통합 색인 시스템은 데이터공유 및 통합 활발하게 이루어 질 수 있게 하는 중요한 요소이다. 하지만, 현재까지 구축된 그리드 시스템에는 이와 같은 색인 및 검색에 대한 표준화된 관리 시스템이 존재하지 않으며, 활발한 데이터 공유 및 통합을 위해서는 이런 통합 색인 및 검색 시스템에 대한 연구가 필요하다.

최근 들어 Ontology 기반의 데이터 통합기법은 다양하게 연구되고 있다. 대표적인 방법으로 통합 온톨로지를 이용방법과 온톨로지의 관계에 대한 학습(heuristic)을 이용하는 방법들이 있다. 통합 온톨로지를 이용하는 방법은 각각의 도메인에 종속되어 있는 온톨로지들을 확장해서 표준화된 vocabulary를 구축을 통한 방법이다. 이에 대한 대표적인 프로젝트로는 SUMO(Suggested Upper Merged ontology)[9]와 DOLCE[10]가 있다. 학습을 통한 통합 기법은 데이터베이스와 XML Schema를 맵핑하는 것과 비슷하게 각 온톨로지에 대한 추론을 통해서 온톨로지의 관계를 분석하여 통합하는 방법을 의미한다. 이러한 방법으로는 PROMPT, ANCHORPROMT, IF-Map과 같은 기법들이 있다[11,12].

3.3 통합 데이터 저장소

다양한 과학 분야에서 생성되는 과학기술데이터들은 다양하다. 기존에는 이런 데이터들을 각 연구기관의 환경에 맞는 데이터베이스, 대형 저장소 시스템, 네트워크 접속 저장장치와 같은 파일 시스템에 저장하였다. 그러나 각 실험 기관에서 생성되고 관리되는 다양한 데이터들을 효율적으로 통합하고 공유하기 위해서 통합 관리할 수 있는 분산 데이터 저장소의 구축이 요구된다. 이를 위해서 **다양한 데이터 모델을 정의할 수 있고, 이 모델을 기반으로 데이터들을 공유할 수 있는 환경 구축이 필요하다.**

또한 데이터 그리드에서는 활발한 공유를 위해서 데이터들을 통합 검색할 수 있는 서비스도 제공되어야 한다. 사용자가 찾고자 하는 데이터에 대한 구체적인 정보 없이도 접근이 가능하도록, 실제 파일이 존재하는 물리적인 위치 정보를 이용하는 것이 아니라, 이런 파일들을 관리해서 얻어지는 정보(메타 데이터)들의 논리적인 위치 정보를 이용해서 데이터를 검색 할 수 있는 서비스를 제공하여야 한다. 데이터 검색에 필요한 정보뿐만 아니라 데이터에 대한 간략한 설명들을 기술하는 메타데이터를 통해서 사용자들에게 찾고자 하는 데이터를 검색할 수 있는 서비스를 제공하였다.

연구원들은 연구를 통해서 생성된 데이터들을 분석하기 위해서 다양한 방식의 분석 툴을 이용한다. 이와 같이 데이터 관리를 위해서 연구원들은 데이터가 저장되어 있는 저장소에서 분석을 위해서 다른 시스템으로 데이터를 복제하거나 이동하기도 한다. 연구 데이터를 이동하기 위한 데이터에 대한 철저한 접근 권한 제한 및 인증 서비스를 제공하고, 빠른 네트워크 성능과 안정적인 편리한 데이터 전송 기능 및 데이터 보안 기술 등이 요구 된다.

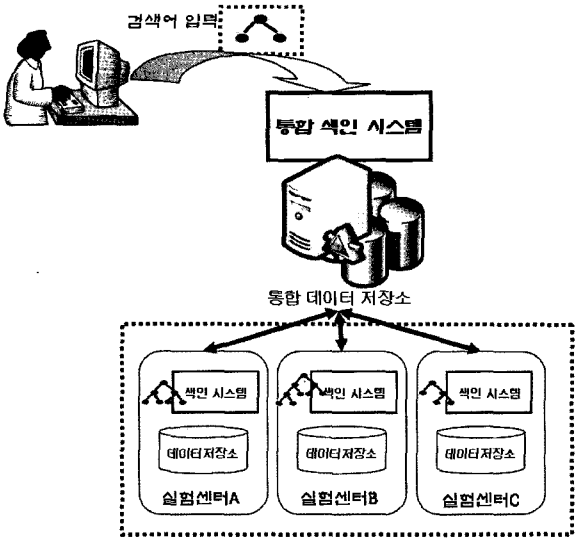


그림 3 e-Science 데이터 그리드 구조

그림 4는 데이터 그리드의 일반적인 구조를 보여준다. 각 사이트에는 이질적인 데이터 저장소(파일 저장소, 데이터베이스, 테이프 등)들이 존재한다. 이들을 통합하기 위해서 Integrated Data Access Manager을 이용해서 데이터 저장소에 저장되어 있는 데이터들에 대해서 색인을 구축하고 검색할 수 있도록 관리 시스템을 구축한다. Integrated Data Access Manager를 통해서 통합 관리되는 데이터들은 데이터 그리드의 Core 서비스에 의해서 그리드 환경에 적합하도록 관리된다. 각 Core 서비스는 사용자별 데이터에 대한 접근 권한을 관리하는 Access Control과 데이터들에 대한 통합 색인 정보 및 메타 데이터들을 관리하기 위한 Data Management, 그리드 시스템의 자원을 관리하기 위한 Resource Manager로 구성된다.

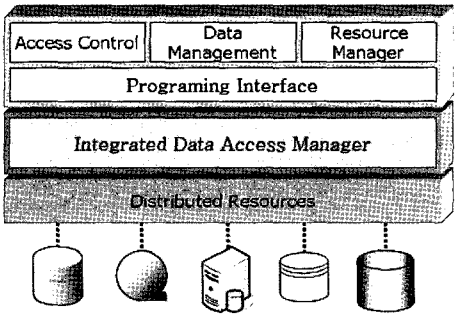


그림 4 데이터 그리드 기본 구조 예

하지만 현재 데이터 그리드 구축을 위한 시스템 도구 및 모델은 앞에서 언급한 것과 같은 다양한 데이터 모델에 대한 통합 및 검색을 위한 도구가 아직 많이 존재하지 않고, 존재하는 도구는 제한적 기능을 제공하고 있다. 따라서 앞에서 언급되었던 문제들의 대부분은 이러한 도구들로 해결이 매우 어렵다. 대표적 구축 도구는 아래와 같다.

3.3.1 GridFTP

GridFTP는 계층적 데이터 스토리지 시스템, 디스크, 스토리지 브로커(Storage Broker)와 같은 이기종 저장 장치로 구성된 데이터 저장소를 다른 컴포넌트와 효율적인 상호 작용할 수 있는 공개 인터페이스 통해서 대용량의 파일을 접근할 수 있는 기능을 제공하는 것을 목표로 디자인된 데이터 접근 및 전송을 위한 그리드 시스템의 서비스이다. GridFTP는 데이터 전송의 표준으로 사용되는 표준 FTP 프로토콜을 확장하여 보안성을 높이고, 안정적이고 빠른 데이터 전송 및 신뢰성이 높은 데이터 전송을 위한 기능을 추가한 서비스이다. 또한 다른 그리드 서비스들 간의 연결을 쉽게 하게 위해서 GridFTP 서비스는 확장성 있고 각 파일에 대한 엄격한 접근 권한 제어를 위한 강력한 인증 및 보안성을 제공하는 Grid Security Infrastructure (GSI)를 이용한다. 이와 같이 데이터 그리드는 GridFTP를 이용해서 파일 기반의 데이터들을 보안과 안정성, 신뢰성을 보장하는 전송, 접근 서비스로 사용된다.

3.3.2 OGSA-DAI(Data Access & Integration)

OGSA-DAI[13]는 그리드 환경에서 서로 다른 source의 Data들을 통합 질의 할 수 있는 기능을 제공한다. 그리드 상에는 다양한 형태의 데이터 저장 방법(Relational Database, XML Database , File, etcs)이 존재하기 때문에, 이 시스템을 통해서 사용자는 다양한 데이터 저장소에 대한 통합 질의를 수행하고, 통합된 결과를 얻을 수 있다. OGSA-DAI는 크게 2개의 서비스로 구성 되어있다. 데이터 저장소에 대해서 직접적으로 질의를 수행하고 결과를 얻어내는 DSR(Data Service Resource)과 이 DSR을 사용자들이 검색하고 사용자가 요청한 질의를 적당한 DSR이 수행 할 수 있도록 관리하는 Data Services로 구성되어 있다(그림 5).

현재 OGSA-DAI는 주목받는 웹서비스 표준인 WS-I (Web Services Inter-operability)와 WSRF(Web Services Resource Framework)에 대한 호환성을 제공하기 위해서 두 가지의 버전으로 제공되고 있다.

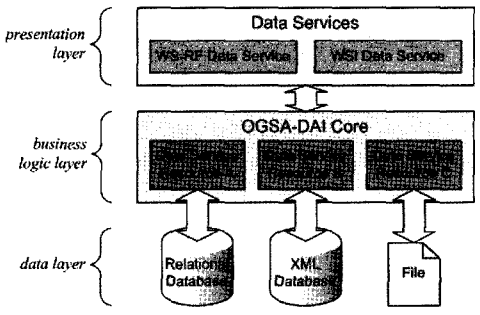


그림 5 OGSA-DAI 구조

3.3.3 SRB(Storage Resource Broker)

SRB(Storage Resource Broker)는 이질적인 데이터 자원들을 동일한 인터페이스를 통해서 접근하고 데이터 관리할 수 있는 기능을 제공하는 분산 데이터 관리를 위해서 SDSC(San Diego Supercomputer Center)에서 만든 클라이언트/서버 미들웨어 시스템이다. SRB는 메타 데이터 카타로그 관리 시스템인 MCAT을 이용함으로서 물리적인 위치와 이름에 상관없이 논리적인 이름과 위치 정보로 사용자들이 쉽게 접근하고 검색할 수 있도록 구성되어 있다. 데이터를 관리를 위해서 공유, 복제등의 기능을 제공하며, 이를 사용하는 사용자들은 다양한 데이터들을 상위 레벨의 API를 이용해서 동일한 방식으로 접근할 수 있다(그림 6).

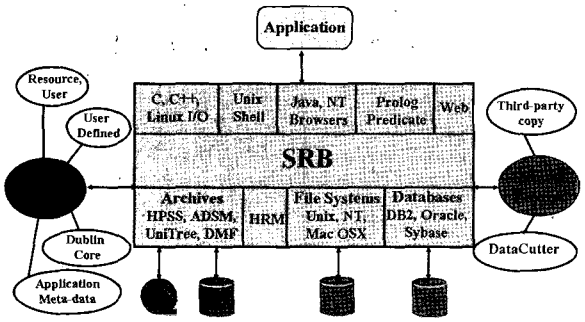


그림 6 SRB 아키텍쳐

4. 관련 연구

현재 전 세계적으로 진행 중인 과학기술데이터 공유 및 통합 연구들은 아직 위에서 언급된 문제들에 대해서 제한적 해결만을 시도하고 있는 수준이다. 대표적인 사례들은 다음과 같다.

4.1 GEONgrid

Geoscience Network(GEON)[14]은 지구 과학분야의 과학자들이 지질 정보와 지역적 특성, 이에 대한 분석/연구 결과를 상호간에 효율적으로 공유하기 위한 사이버 인프라스트럭처를 구성하기 위한 목표로 구축된 그리드 시스템이다. GEON에서의 데이터는 지구 과학의 물질인 신성암(Pluton)과 같은 암석 데이터의 특징을 3D모델을 이용해서 데이터 표현을 구축한 정보들과, 암석과 중력에 대해서 시맨틱하게 통합된 데이터를 저장한다. 이처럼 지역적으로 분산되고 다양한 시스템으로 구성되어 있는 데이터들을 효과적으로 공유하기 위해서 GEON은 분산 시스템 방식을 이용한 통합 시스템을 구축한 것으로, SOA(Service Oriented Architecture)모델을 이용하였다.

이런 웹 서비스들을 이용해서 다양한 사용자들은 지역에 상관없이 접근성, 사용성을 높일 수 있으며, 각각의 서비스를 조합하여 포탈을 구축하고 포탈을 통해서 데이터 및 서비스를 검색하여, 사용자가 원하는 방식으로 조합하여 사용할 수 있도록 되어있다. 포탈의 서비스들을 이용해서 공유되는 데이터들에 대한 분석 및 공동 연구를 수행할 수 있는 모델로 구축되어 있다. 또한 데이터 공유를 위해서 서비스 기반의 웹 포탈을 이용해서 공동 연구 과학자들에 의해서 구축된 데이터들을 효율적으로 통합 검색하기 위해서 Ontology를 이용해서 데이터들을 표현하였다. Ontology는 데이터들을 Object로 정의하고 Object들 간의 관계의 표현(e.g. Class, subClass, part-of, 등)을 통해서 지식(Knowledge)을 구축하는 방법으로, GEON에서는 지구 과학 분야에서 생성되는 다양한 형태의 데이터들을 표현하기 위한 데이터 모델을 설계하는데 이용 하였다. 이렇게 통일된 형태의 데이터 표현 모델을 구축하는 것은 표준적인 데이터 표현으로 인해 키워드 매칭을 통한 검색에 비해 효율적인 검색과 좀 더 사용자가 원하는 검색 결과를 얻을 수 있기 때문에, 활발한 데이터 공유가 이루어 질수 있다.

4.2 BioSimGrid

BioSimGrid 프로젝트[15]는 영국의 e-Science 프로젝트의 일환으로, 단백질 모션 시뮬레이션과 같은 실험은 단백질과 관련된 다양한 형질의 구조를 이해하는데 매우 중요하기 때문에 바이오 커뮤니티들이 보다 쉽게 대규모 분자 시뮬레이션의 결과에 접근할 수 있도록 하는 것을 목표로 한다.

이 프로젝트는 앞에서 설명했던 긴 작업 시간의 실험을 통해서 생성되는 거대한 크기의 결과 데이터가 생성되는 바이오 가상 시뮬레이션과 같은 성격을 가지고 있기 때문에, 이런 특성의 데이터들을 공유하기 위해서 BioSimGrid는 데이터 그리드에 초점을 맞추고 있다.

이런 거대한 데이터의 공유 문제를 해결하기 위해서 이 프로젝트는 **1) 데이터 저장소의 최소화, 2) 추상화된 데이터 계층, 3) 데이터 위치의 투명성, 4) 데이터 전송률의 최대화** 등의 세부 목표를 통해서 데이터 공유를 제공하려고 하였고, 이런 BioSimGrid의 기본적인 아키텍처가 그림 7에 나타나 있다.

BioSimGrid는 데이터를 관리하고 공유하기 위해서 데이터와 데이터베이스 자원을 연관된 장소에 분산시키고, 통일된 데이터 접근을 위해서 BioSimGrid의 그리드 미들웨어는 이러한 데이터 자원들을 투명성 있게

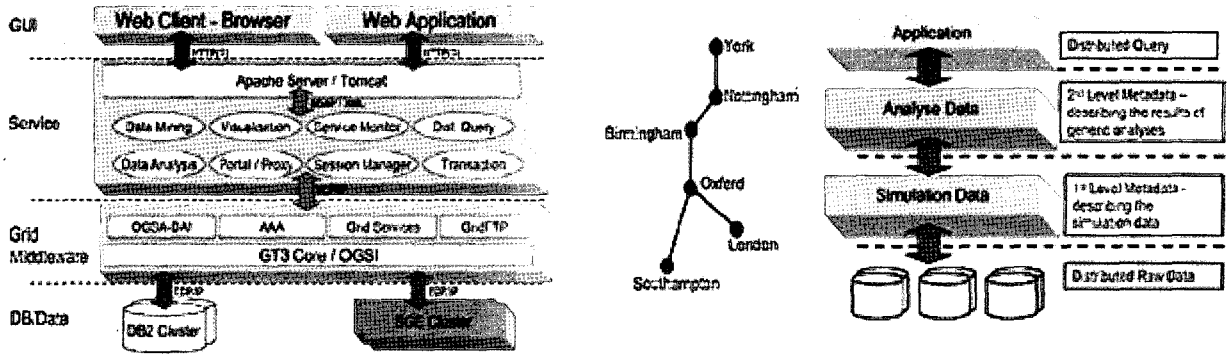


그림 7 BioSimGrid의 아키텍쳐와 데이터베이스 구성

접근할 수 있도록 서비스를 구축하였다. 이와 연관된 애플리케이션들은 규격화된 미들웨어 컴포넌트를 사용할 뿐만 아니라, 분산 서비스로써 독립적으로 개발될 수 있으며, 시스템에서 통합된다. 이러한 기능들은 시스템의 신축성(scalability)과 호환성(flexibility)을 증가 시킨다.

4.3 BIRN

BIRN(Biomedical Informatics Research Network)[16]은 2001년에 미국의 National Institutes of Health's National Center for Resource Resource에 의해서 시작되었고, 의학분야의 과학자들이 질병에 대한 치료 방법들을 고성능의 컴퓨터와 데이터 통합 기법과 다양한 컴퓨팅 기술들을 이용하여 위한 협업 환경을 구축하고, 상호간의 실험 결과 및 질병 정보들을 공유할 수 있는 인프라 구축을 목표로 하고 있다.

그러나 이 프로젝트는 동일한 연구 주제인 뇌를 연구하지만 서로 다른 연구 방법을 연구원들 간에 공통의 주제 의식을 가지고 서로간의 데이터를 공유하기 위한 문제점을 해결해야 되었다. 이 시스템에서 중요하게 다르고 있는 뇌에 대한 데이터는 크게 3가지 분야로 나눠 볼수 있다. **1) 인간의 뇌의 관련된 질병과 형태와의 관계에 대한 연구(Human Structure BIRN), 2) 정신분열증에 대한 이미징 분석(FIRST BIRN), 3) 질병에 걸린 쥐의 뇌에 대한 다양한 방식의 분석 자료(Mouse BIRN)** 등이다. 이런 세 가지 연구 분야에 대해서 의학 과학자들은 서로 다른 다양한 실험 방법을 이용하기 때문에 다양한 종류의 실험 결과가 나오게 된다. 실험 방식의 다양성으로 인해서 생성된 데이터들의 분석 방법도 다양하기 때문에, 표준적인 데이터분석 및 표현 방법의 부재로 실험 결과 데이터를 공유하기가 어렵다. 이런 다양한 방식으로 분석이 필요한 데이터들을 효과적으로 공유하기 위해서 BIRN은 개별적인 데이터 분석 및 표현 방법을 필요로 하는 데이터의 통합과 공유를 위해서 Wrapper/Mediator 기술을 이용해서 문제를 해결하였다. Wrapper를 통해서 뇌에 대한 연구분야의 과학자들이 실험을 통해서 얻어지는 실험 결과를 분석하고 표현하는 개별적인 방법들을 각각 시스템화하고, 이들을 표준화된 결과물을 산출하도록 하였다. 이렇게 번역된 결과 산출물은 중앙의 Mediator에 의해서 사용자에게 통합되어 보여지게 되어 사용자가 쉽게 다른 분야의 실험 데이터를 접근하고 분석 할 수 있도록 시스템에 구축되었다(그림 8).

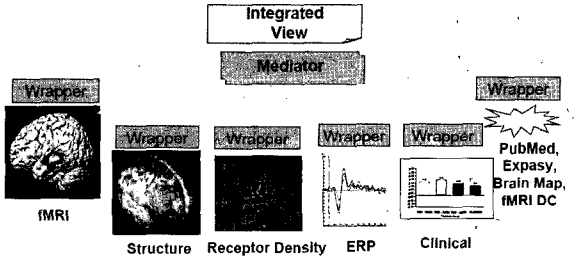


그림 8 BIRN의 Mediator기반 통합

4.4 NEESgrid

NEESgrid(The Network for Earthquake Engineering Simulation)[17]는 건축 분야의 기술자와 연구자들이 지진의 영향을 받는 건축물들에 대해서 고가의 실험하는 장비 및 실험을 통해서 얻어지는 결과를 그리드 기반 시스템을 통해서 공유 인프라 구축을 목표로 하고 있다. 건축물에 대한 지진의 영향에 대한 연구 결과 데이터뿐만 아니라 고가의 가상 시뮬레이션 장비를 원격에서 공유하여 실험을 수행할 수 있도록 하기 위해서 NEESgrid는 원격제어(Tele-operation)과 원격 모니터링(Tele-presentation)과 관련된 서비스들을 구축하였다.

이 프로젝트는 실험 장비를 원격에서 제어하고 모니터링하기 위해서 다양한 서비스들을 구축하였다. 실험 장비에 대해서 실시간 모니터링 및 센서 데이터 추출을 위해서 Data Turbine이라는 서비스를 이용하고, 원격에서 이루어지는 실험 정보인 모니터링 영상 정보와 센서에서 추축한 시그널 정보들을 보관할 수 있도

록 구축하였다. 이와 같이 장비를 이용한 실험의 특성에 맞는 서비스를 구축하였다. 또한 원격 실험 시설을 제어하기 위한 서비스도 제공하고 있다. 이를 위해서 제어하고 결과 센서 데이터를 추출하기 위해서 가장 많이 사용되는 소프트웨어(LabView)를 이용해서 실험 장소에 원격 제어 서비스를 생성하고, 원격지에서는 이 서비스를 이용해서 장비를 제어 할 수 있도록 구성하고 있다.

데이터 공유를 위해서는 참조 데이터 모델을 정의하고, 이 모델을 기반으로 NEESCentral[18]이라는 중앙 데이터 저장소를 구축하였다. 이 저장소에는 실험 결과뿐만 아니라, 협업에 필요한 자료, 분석된 데이터들을 공유할 수 있게 하였다.

5. 결 론

본 논문에서는 미래의 첨단과학 연구에서 매우 중요한 과학기술데이터 공유 및 통합에 관련된 문제들, 이러한 문제점들에 대한 접근방법들과, 구체적인 해외 사례들을 설명하였다. **과학기술데이터는 우리에게 익숙한 비즈니스 데이터와는 다른 특성을 많이 갖고 있다.** 예를 들어, 다양한 실험환경 및 실험데이터에 대한 모델링이 요구되고, 지속적 진화가 지원되어야 하고, 이질적 데이터들 간의 통합이 요구되고, 복잡한 시스템 환경에서 정보 보호가 요구된다. **따라서 단순히 기존 기법 및 도구를 약간의 수정을 통해 적용하는 것은 매우 어렵다고 보아야 하고, 효과적 과학기술데이터 공유 및 통합을 위해서는 종합적이고 체계적 접근이 요구된다.**

현재 전 세계적으로 다양한 과학 분야에서 데이터 공유 및 통합 노력이 진행 중이 있으나, 이에 비해서 국내에서는 아직 이러한 노력이 미비한 수준이다. 향후 국내에서도 해외에서 주목받을 수 있는 연구들이 많이 진행되기를 기대한다.

6. 관련연구

[1] Database Design Using Entities and Relationships, Chen, P. P., S. B. Yao (ed.), Principles of Data Base Design, Prentice-Hall, NJ, pp.174-210, 1985.

[2] J. Arlow and I. Neustadt. UML and the Unified Process: Practical Object-Oriented Analysis and Design, Addison-Wesley Pub Co., Boston, MA, 2001.

[3] XML(Extensible Markup Language), http://www.w3.org/XML

[4] XML Schema, http://www.w3.org/XML/Schema

[5] Resource Description Framework(RDF), http://www.w3.org/RDF

[6] R. Benjamins, D. Fensel, and A. G. Perez. "Knowledge Management through Ontologies," Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management, Basel, Switzerland, 1998.

[7] M. Botts. Sensor Model Language(SensorML) for In-situ and Remote Sensors, OpenGIS Interoperability Program Report, OGC 02-026, Open GIS Consortium Inc, 2002.

[8] Reference NEESgrid Data Model[TR-2004-40] (2004), Jun Peng , Kincho H. Law.

[9] Towards a standard upper ontology, I. Niles and A. Pease , In The 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).

[10] Sweetening wordnet with DOLCE, A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. AI Magazine, 24(3):13.24, 2003.

[11] The PROMPT suite: Interactive tools for ontology merging and mapping, N. F. Noy and M. A. Musen. International Journal of Human-Computer Studies

[12] IF-Map: an ontology mapping method based on information flow theory, Y. Kalfoglou and M. Schorlemmer. Journal on Data Semantics, 1(1):98.127, Oct., 2003.

[13] OGSA-DAI, http://www.ogsadai.org.uk/

[14] GEON, http://www.geongrid.org

[15] BioSimGrid, http://www.biosimgrid.org

[16] BIRN http://www.nbirn.net

[17] NEESgrid, http://it.nees.org

[18] NEESCentral, https://central.nees.org.

정 갑 주

1984. 2 서울대학교 컴퓨터공학과(학사)
1986. 2 서울대학교 컴퓨터공학과 인공지능(석사)
1996. 2 New York University, Computer Science(박사)
1995. 12~1997. 8 University of Florida, Post Doc.
1997. 8~2001 건국대학교 컴퓨터공학과 조교수
2001~현재 건국대학교 인터넷미디어공학부 부교수
관심분야 : Grid Computing & e-Science, Data Integration
E-mail : jeongk@konkuk.ac.kr

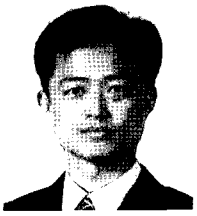
김 동 광

2005. 2 건국대학교 컴퓨터공학과(학사)
2005. 3~현재 건국대학교 컴퓨터공학과 (석사과정)
관심분야 : Grid Computing, Semantic Data Integration
E-mail : walhalla@gcslab.konkuk.ac.kr

이 종 현

2003. 2 건국대학교 컴퓨터공학과(학사)
2005. 2 건국대학교 컴퓨터공학과(석사)
2005. 3~현재 건국대학교 컴퓨터공학과 박사과정
관심분야 : Grid Computing, Linux Kernel Programming
E-mail : lejohy@gcslab.konkuk.ac.kr

황 선 태



1985 서울대학교 컴퓨터공학과(학사)
1987 서울대학교 컴퓨터공학과(석사)
1996 Manchester University(PhD)
1997~현재 국민대학교 컴퓨터학부 부교수
관심분야 : e-Science,그리드시스템,PSE, 공개소프트웨어
E-mail : sthwang@kookmin.ac.kr