이슈리포트 2014-008

과학연구에서 SW활용 사례

Case Study on Software Utilization in Scientific Research

김석원
박강민
강송희

2014.12.29

SPRi 소프트웨어정책연구소
SOFTWARE POLICY & RESEARCH INSTITUTE

○ 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 방송통신정책연구사업(R&D)의 일환으로 수행하였습니다. [14-진흥-029, 과학기술의 SW활용 현황분석 및 협력방안 연구]

○ 본 보고서의 내용은 연구진의 개인 견해이며, 본 보고서와 관련한 의문사항 또는 수정·보완할 필요가 있는 경우에는 아래 연락처로 연락해 주시기 바랍니다.

- 소프트웨어정책연구소 연구실 김석원 책임연구원(skimaza@spri.kr)

요 약 문

○ 최근 과학기술 연구는 계측장비의 발달과 시뮬레이션 데이터의 급증으로 다루어야 할 데이터가 급속도로 늘어났으며 이 데이터를 의미있는 정보로 분석하고 효율적으로 관리·공유할 수 있는 방안이 필요함

○ 이를 위해 과학계와 컴퓨터학계간의 다학제적 연구가 필수적이 되었으며 과학계와 컴퓨터학계 간 공동 연구를 통해 소프트웨어 활용의 다양한 성공 사례를 볼 수 있음

- 신경과학, 생명정보학, 천문학, 입자물리학, 환경과학 등 다양한 과학 분야에서 클라우드, 분산처리, 머신러닝 등의 소프트웨어 기술이 활용되었음

- 이 중 클라우드 기술은 과학연구에서 활발히 사용되고 있으며, 데이터 공유와 분석의 측면에서 과학연구의 효율성과 효과성을 높이고 있음

○ 선진국에서는 이러한 연구 개발의 기반 환경을 조성 중임

- 영국의 e-Science 프로그램, 유럽의 e-Infrastructure 연구, 미국의 사이버인프라 구축을 통해 과학연구에 소프트웨어 활용을 지원

- 또한, 과학자가 아닌 비전문가나 아마추어가 데이터를 수집하고 분석하는 시민과학의 형태가 소프트웨어의 발전으로 생겨났음

Summary

○ Recently data used in scientific research has increased rapidly with the development of measurement equipments and simulation environments. Thus, analyzing data into meaningful information and effective management of the enormous data are emerging issues.

○ To address the issues, **it will require an interdisciplinary research between computer science communities and academia.** There are various successes of software utilized by the scientific communities in the collaboration with computer science communities.

- Software technologies including cloud, distributed processing, and machine learning have been widely used in various scientific fields such as neuroscience, bioinformatics, astronomy, particle physics, and environmental science.

- Especially, the cloud technology is actively involved in the scientific research and enhancing the efficiency and effectiveness of scientific research in terms of data analysis and sharing.

○ In developed countries infrastructures of interdisciplinary research and development are being established.

- U.K. e-Science program, the European e-Infrastructure research project, and cyber-infrastructure in the U.S. support utilizing software for scientific research.

- Furthermore, non-professionals or amateur scientists are collecting and analyzing data by using software in the form of citizen science.

목 차

1. 과학연구의 패러다임 변화

○ 오늘날 과학기술 연구는 계측장비의 발달과 시뮬레이션 데이터의 급증으로 다루어야 할 데이터가 급속도로 증가하였음[1)]

- CERN의 대형강입자충돌기(LHC: Large Hadron Collider)에서는 매년 수십 페타 바이트의 데이터가 생성되고 이를 처리하고 분석하기 위해 약 15만대의 컴퓨터가 필요함

- 전체 게놈 시퀀싱(WGS: Whole Genome Sequencing)을 통해 수 테라바이트의 데이터가 추출됨

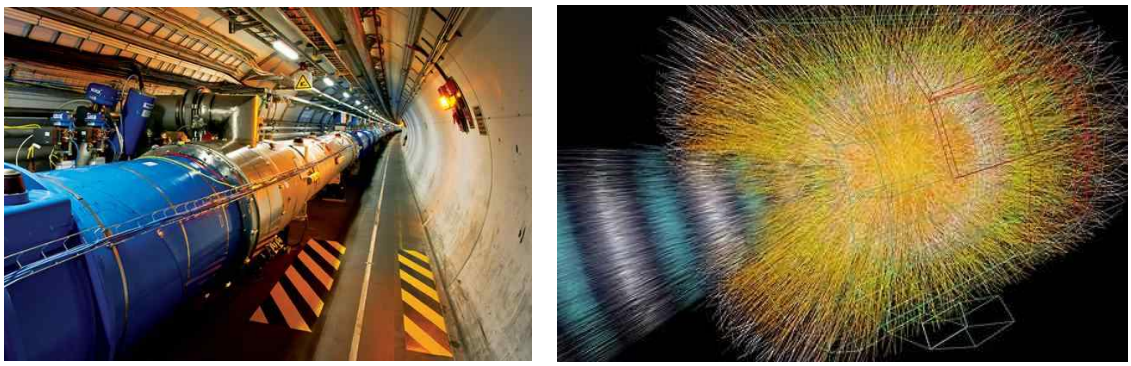
○ **(데이터의 해석)** 막대한 양의 데이터를 의미 있는 정보로 해석하는 일이 과학 연구의 중요한 비중을 차지함

- 데이터를 분류하고 정리하여 사람의 인지 능력으로 다룰 수 있는 크기로 줄이는 과정이 필요함

○ **(데이터의 관리)** 데이터의 양이 방대해지면서 데이터의 관리가 중요해 짐

- 대부분 실험 데이터가 디지털화 되거나, 컴퓨터 시뮬레이션을 통해 생성되기 때문에 파일 형태로 다양한 곳에 존재하게 됨

- 데이터를 효과적이며 효율적으로 저장하고 탐색 가능해야 함



[그림 1] (왼쪽부터) LHC, 납 충돌 시뮬레이션
(출처: www.cern.ch, www.telegraph.co.uk)

1) Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science,* 323(5919), 1297-1298.

○ **(데이터의 공유)** 방대한 데이터를 연구자간에 공유함

- 가상현실 천문대(VO: Virtual Observatory)를 통해 데이터에 자유롭게 접근하여 연구하고 새로운 발견을 VO에 등록함
- NCBI(National Center for Biotechnology Information)와 GeneBank가 생물학에서 이러한 역할을 함

○ 데이터의 해석, 관리, 공유의 측면에서 과학연구를 위한 소프트웨어가 중요한 역할을 함

- 과학계와 컴퓨터학계의 다학제간(interdisciplinary research) 협력이 필요함

○ 패러다임 변화에 따른 학제간 연구 환경 개선을 위해 선진국에서는 체계적 연구개발 기반 환경을 마련할 수 있도록 정부 차원에서 지원하고 있음

- 미국은 정부 주도로 '사이버 인프라'를 전략 과학 분야로 설정하여 고성능 컴퓨팅, 데이터 분석 및 가시화, 분산 커뮤니티 가상조직, 교육 및 인력 개발을 추진하고 있음
- 유럽연합은 FP6(6th Framework Programme for Research and Technological Development), FP7 프로그램을 통해 과학연구에 소프트웨어 활용을 지원하고 있음

○ 본 연구에서는 이러한 과학 연구에서의 소프트웨어 활용 사례를 조사하여 활용 분야, 활용된 요소 기술, 활용 방법 등을 파악하고자 함

2. 과학연구에서 SW활용 사례

(1) 신경과학 분야 사례 - CARMEN

□ 신경과학의 소프트웨어 활용 필요성

○ 신경과학은 신경시스템에 관한 과학으로 전통적으로 생물학에 속하였으나 현대에 와서는 화학, 생리학, 약학, 전산학 등을 포괄하는 광범위한 학문 분야가 됨

○ 방대한 양의 신경 데이터가 존재하며 이를 분석할 수 있는 강력한 컴퓨팅 환경이 필요함

○ 연구자들이 독자적으로 신경데이터를 수집하고 연구하는 것도 필요하지만, 광범위한 데이터간의 유기적인 연계성을 바탕으로 데이터를 해석해야 할 필요성도 있음

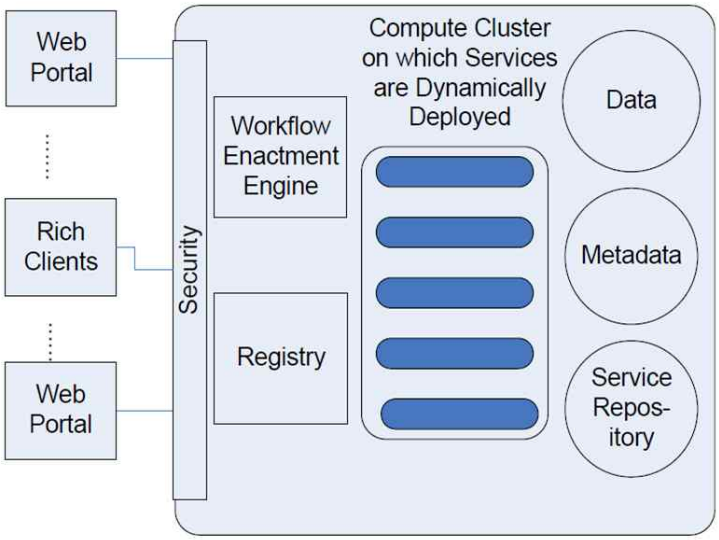
□ CARMEN 개요

○ CARMEN(Code Analysis, Repository, & Modeling for E-neuroscience)은 클라우드 기반 연구 포털 서비스로, 데이터의 저장과 분석을 위한 도구와 환경을 제공함

○ 뉴캐슬, 캠브리지 등 11개 대학에서 신경생리학자, 신경정보학자, 전산학자의 공동연구를 통해 2006년 구축됨

○ 연구자는 웹 포털을 통해 CARMEN에 접속할 수 있으며 CARMEN에 저장되어 있는 데이터를 워크플로우 엔진을 이용하여 분석할 수 있음(그림2)

- 데이터를 다운로드 받을 필요 없이 클라우드 내에서 강력한 컴퓨팅 파워를 이용하여 분석할 수 있음



[그림 2] CARMEN 구조 (출처: Watson et al., 2010)

○ 자신의 연구를 다른 연구자에게 공개하는 시점을 결정할 수 있어 연구 내용을 보호할 수 있음

□ 소프트웨어 활용성과

○ CARMEN을 통해 데이터를 검색하고 이용하며 연구 결과를 공유할 수 있음

○ 데이터를 웹상에서 분석함으로써 강력한 컴퓨팅 파워를 이용하여 테라바이트 단위의 데이터도 빠르게 분석할 수 있음

○ CARMEN을 활용한 62개의 관련 논문이 게재되었으며, 연구 결과를 공유하는 2개의 표준이 정립됨

○ 사용된 요소 기술은 웹 및 클라우드 기술

(2) 생명정보학 사례 - BIRN

□ 생명정보학의 소프트웨어 활용 필요성

○ 생명정보학은 생명공학 기술과 정보기술의 융합학문으로써 생물학 연구에 컴퓨터와 분석 소프트웨어를 활용하는 응용과학의 학문임

○ 주요 연구 분야는 서열정렬, 유전자 검색, 유전자 조합, 단백질 구조 정렬, 단백질 구조 예측, 유전자 발현의 예측 등이 있음

○ 이러한 생명정보학 연구를 위한 방대한 유기적 데이터가 분산되어 있음

○ 전체 데이터를 분석하고 해석하기 위해 여러 기종의 데이터 소스를 통합하는 동시에 처리와 분석을 용도에 맞춰서 조정할 수 있어야 하는 요구사항이 있음

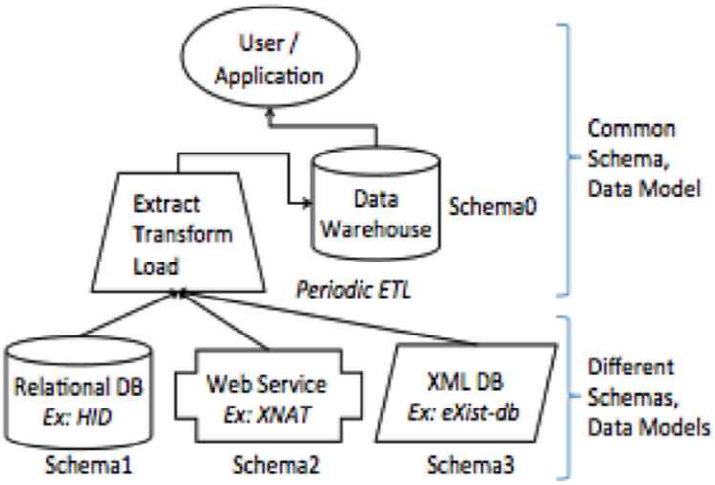
□ BIRN 개요

○ BIRN(Biomedical Informatics Research Network)은 대규모 데이터 및 컴퓨팅 자원의 처리를 다루는 분산 인프라로 Function BIRN, Morphormetry BIRN, Mouse BIRN, BIRN-CC 의 연구 분야에서 총 37개의 대학과 병원 등의 연구기관이 참여하였음

<표 2> BIRN 프로젝트의 각 연구 분야

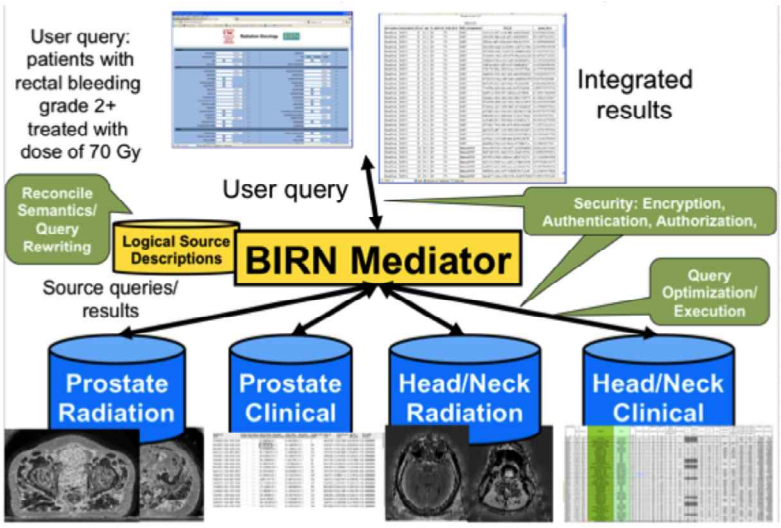
연구 분야	설명
Function BIRN	정신분열증의 원인과 치료에 대한 이해를 심화하기 위한 다중 기능성 자기 공명 도구 개발
Brain Morphometry BIRN	기억 장애나 우울증과 같은 증상의 구조적 차이와 관련된 뇌 구조의 차이를 분석하기 위해 교정 도구 및 해부학적 분석 도구를 개발
Mouse BIRN	다중 경화증, 정신분열증, 파킨슨병, 주의력 결핍 과잉 행동 장애, 뚜렛 증후군, 뇌 암과 같은 질병의 쥐 모형 연구에 초점을 둠. 이러한 신경 장애를 이해하는 데 필요한 분자 정보와 해부학적 영상을 포함하는 다양한 차원의 데이터를 집계함
BIRN-CC	공통 기술 인프라를 지원하고, 데이터 및 분석 도구를 공유하며, 직관적인 웹 포털을 제공하는 등 BIRN 협력 그룹의 공통 기술 이슈를 해결

○ BIRN은 그림3과 같이 연구에 필요한 데이터를 데이터 소스로부터 식별 및 추출(Extract)하고 공통의 스키마로 데이터를 변환(Transform)하여 하나 이상의 데이터웨어하우스에 적재(Load)



[그림 3] 데이터 수집 및 처리 흐름도
(출처: OLSON, Judith S., et al.)

○ 그림 4와 같이 BIRN 중개자가 각 이기종 데이터 소스 내 데이터에 대한 사용자 질의를 번역하여 해당 데이터 소스에 중개함



[그림 4] BIRN 중개자 활동의 예시
(출처: http://www.birncommunity.org/resources/data/)

□ 소프트웨어 활용성과

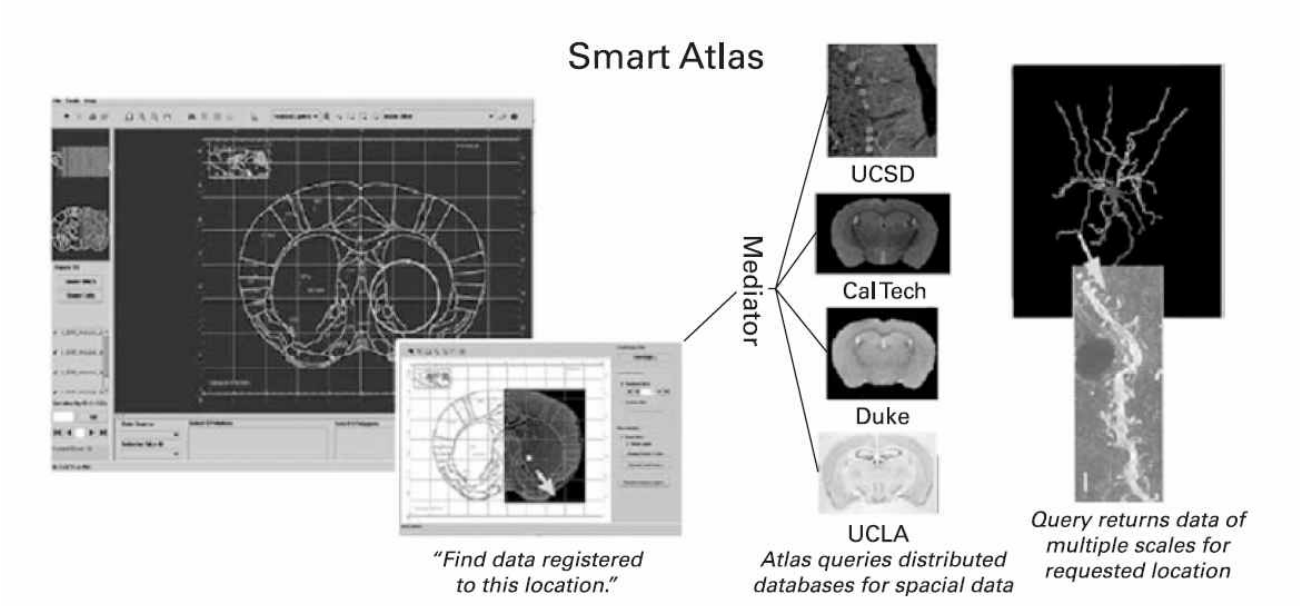
○ BIRN을 통해 다양한 연구기관 간 데이터의 효율적 공유를 위한 메타데이터를 생성하였으며, 이를 통해 이기종의 데이터 소스를 통합하고 데이터의 이관 및 공유가 용이해짐

○ 데이터의 처리와 분석을 용도에 따른 맞춤형으로 할 수 있음

○ BIRN 프로젝트는 2007년에는 1,500만개 이상 16테라바이트 규모의 파일을 수집하고 400여개 이상의 계정이 생성되어 데이터가 1,800만 번 공유되었음

○ 그림 5는 Mouse BIRN의 예시로 쥐의 뇌에 파킨슨병이나 알츠하이머병과 같은 질병과 유사한 병증을 나타낼 수 있도록 하여 새로운 치료법을 개발하는 데 활용되고 있음을 보여줌

- Mouse-BIRN의 스마트 아틀라스는 해당 뇌 부분의 해부학적 데이터를 분자학 및 구조학적 스키마로 대응시킬 수 있도록 함



[그림 5] BIRN을 활용한 쥐 스마트 아틀라스 이미징 예시
(출처: Olson, et al., 2013)

○ 사용된 요소 기술은 데이터 그리드, 데이터웨어하우스, ETL(Extract, Transform, Load), 데이터 통합, 이미지 프로세싱 등임

(3) 천문학 분야 사례 - AstroGrid

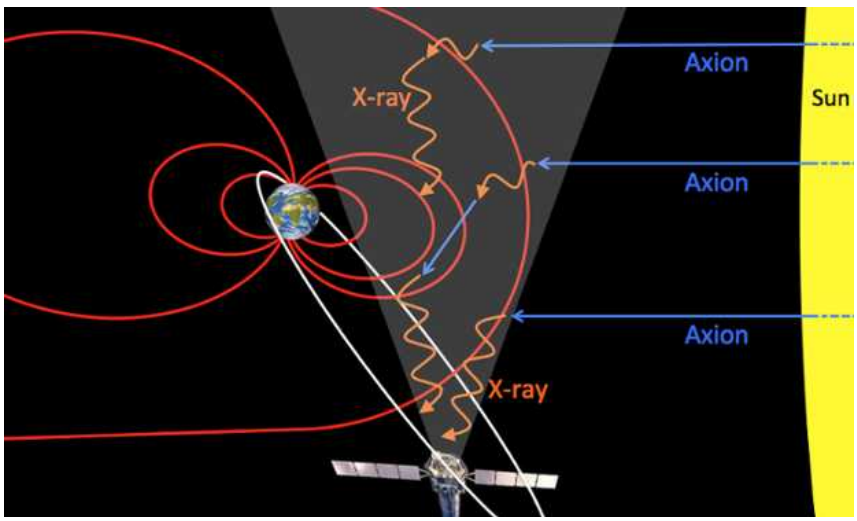
□ 천문학의 소프트웨어 활용 필요성

○ 천문학은 우주의 구조, 천체의 현상, 다른 천체 등과의 관계를 연구하는 실증적 학문

○ 1990년대에 들어 Chandra, XMM-Newton, Spitzer 등과 같은 고해상도 관측 기기들이 등장하여 데이터가 방대해 졌으며 분석이 복잡해졌음

- 예를 들어 암흑물질의 단서는 유럽 레스터대학 연구팀이 15년간 수집한 데이터에서 이상신호를 포착하여 발견함

- 그림 6은 태양으로부터 전달되는 액시온(Axion)으로 알려진 암흑물질 입자가 지구의 자기장 내에서 X선으로 변환되어 XMM-Newton에서 관찰되는 모습을 나타냄[2)]



[그림 6] 암흑물질 입자 액시온의 지구 자기장내 변화 (출처:가디언)

2) 가디언(2014.10.16), "Dark matter may have been detected - streaming from the sun's core" http://www.theguardian.com/science/2014/oct/16/dark-matter-detected-sun-axions,

□ AstroGrid 개요

○ AstroGrid는 천문학 연구를 위한 데이터 저장 공간과 분석 도구를 제공하며 국제적으로 통용되는 프로토콜과 표준을 정립하기 위해 협력하는 프로젝트임

○ AstroGrid는 자바 기반의 웹서비스를 제공하며 데이터 중심 설계를 통해 찾고자 하는 자료를 효율적으로 검색하고 분석할 수 있음

○ AstroGrid 서버측의 구성요소는 다음과 같음

- Registry: AstroGrid 서비스에서 실행가능 한 데이터와 어플리케이션 메타데이터의 리스트

- VO(Virtual Observatory) Space: 연구자들이 데이터와 워크플로우를 공유하는 저장 공간

- Data Set Access: 표준 VO 인터페이스를 통해 데이터를 게시 및 접근

- Community: SSO(Single Sign On)를 통한 사용자 관리

○ 클라이언트측의 구성은 다음과 같음

- VODesktop: VODesktop은 데이터를 검색하고 필요한 데이터를 추출해 낸 후 연구자의 컴퓨터에 다운로드 받거나 다른 분석도구에 넘겨 분석할 수 있음

- Astro Runtime: AstroGrid VO 미들웨어로 여러 VO 서비스간의 표준 인터페이스를 제공함

□ 소프트웨어 활용성과

○ 천문학계와 컴퓨터학계의 협력 커뮤니티가 만들어졌으며, 2009년까지 21개의 국제 표준과 149개의 관련 논문이 발표되었음

- AstroGrid는 2002년 International Virtual Observatory Alliance(IVOA)를 창립하는 데 기여함

- IVOA는 아르헨티나, 중국, 프랑스, 독일 등 20여개의 VO가 참여하여 데이터 표준을 정립함
- AstroGrid는 이후 유럽 전역의 VO를 잇는 EuroGrid의 구축에 기여함

○ 사용한 요소기술은 데이터 그리드, 웹, 이미지 프로세싱 등임

(4) 입자물리학 사례 - CERN ATLAS

□ 입자물리학의 소프트웨어 활용 필요성

○ 입자 물리학은 물질을 구성하는 기본적인 입자가 무엇인지, 입자 사이의 상호작용은 어떻게 일어나는지 등을 규명하여 자연현상의 본질을 탐구하는 학문임

○ 유럽 입자 물리학 연구소(CERN)는 스위스 제네바와 프랑스 사이의 국경지대에 위치한 세계 최대의 입자 물리학 연구소. 원래 명칭은 유럽 원자핵 공동 연구소(Conseil Européen pour la Recherche Nucléaire)였고 이에 따라 CERN이라 불림

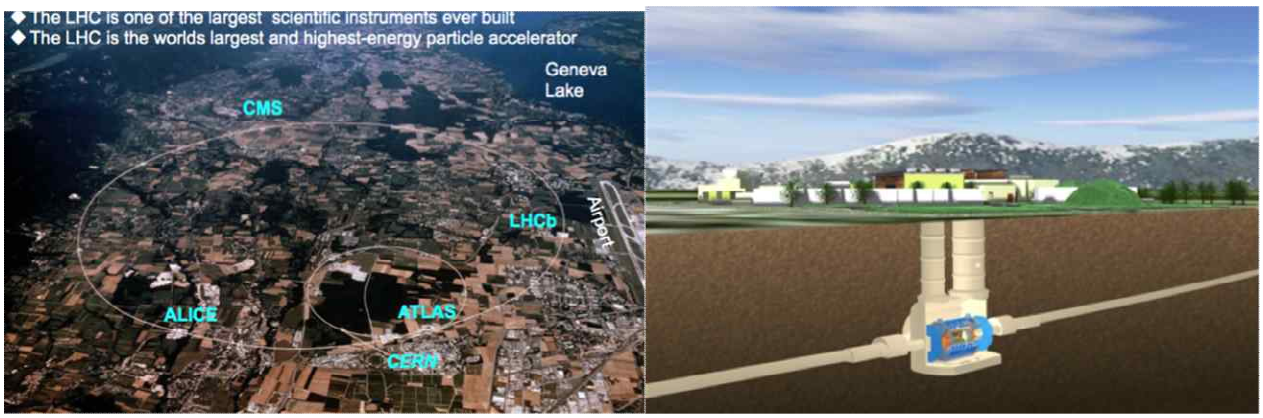
○ CERN은 설립 초기부터 입자 가속기 등을 이용해, 입자물리학 연구에 많은 기여를 하였음

- 물리학자들의 문헌 검색 및 제휴를 위하여 고안된 HTML과 월드 와이드 웹의 발상지이기도 함

○ CERN 대형강입자충돌기(LHC: Large Hadron Collider)의 ATLAS 실험에는 38개 국가의 3,000명의 과학자, 1,200명의 대학원생이 참여 중

○ ATLAS는 막대한 에너지가 필요한 입자 충돌 실험을 통해 입자물리학의 새로운 영역을 탐색하고 있음

- 2013년 힉스입자의 발견은 대표적인 성과임



[그림 7] 대형강입자충돌기와 지하에 묻힌 ATLAS 탐지기
(출처: https://developers.google.com/events/io/sessions/333315382)

○ ATLAS는 2013년 기준 140 페타바이트의 데이터와 전 세계에 걸쳐있는 수백개의 분산 컴퓨팅 센터를 다루고 있기 때문에 운영과 관리에 어려움이 있음

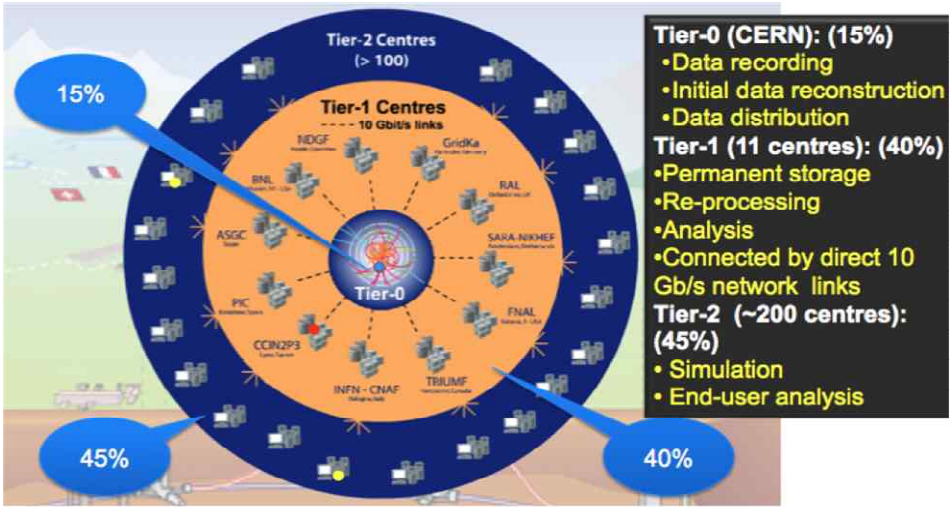
○ 또한 40 여개 이상의 국가에서 온 과학자가 협업할 수 있는 환경이 필요함

□ CERN ATLAS 개요

○ ATLAS 처리와 분석을 위해서는 작업 워크플로우와 데이터 관리의 긴밀한 통합을 바탕으로 대용량의 처리, 높은 수준의 확장성과 견고성을 확보해야 함

○ PanDA(The Production and Distributed Analysis) 워크로드 관리 시스템은 대형강입자충돌기(LHC: Large Hadron Collider) 데이터의 분산 분석을 위한 데이터 구동 작업 관리 시스템으로 ATLAS 실험 데이터의 처리와 분석을 관리하고 있음(그림 8)

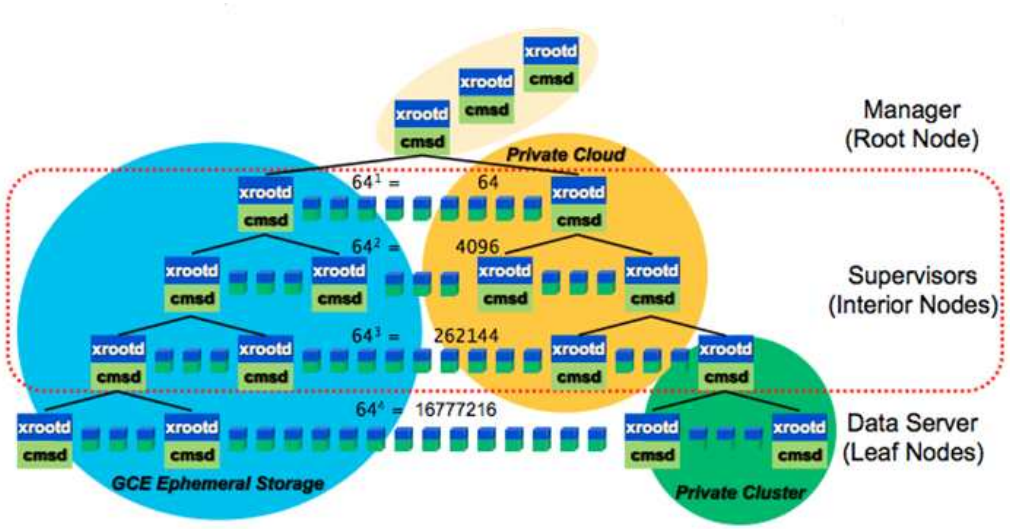
○ PanDA 워크로드 시스템은 전 세계에 분산하여 배분하고 관리 하는데서 오는 운영과 관리상의 어려움을 일부 해소하고자 퍼블릭 클라우드 도입을 검토



[그림 8] PanDA 워크로드 관리 시스템
(출처:https://developers.google.com/events/io/sessions/333315382)

○ 2012년 구글 컴퓨트 엔진의 시험운용 기간 동안 ATLAS의 일부 작업을 클라우드에 올려 8주간 시험함

- 그림 9는 ATLAS의 작업을 퍼블릭 클라우드인 구글 컴퓨트 엔진(파란색 원)과 프라이빗 클라우드(노란색, 초록색 원)가 나눠 수행하는 것을 보여 줌



[그림 9] ATLAS 작업을 나누는 방법으로 파란색이 구글 컴퓨트 엔진임
(출처:https://developers.google.com/events/io/sessions/333315382)

□ 소프트웨어 활용성과

○ 고성능 컴퓨팅에 클라우드를 사용할 수 있음을 검증함

○ 클라우드에서의 데이터 전송 및 분산 환경 실험을 통해 새로운 인프라 확보 가능성을 타진하여 인건비를 최소화하고 확장성을 최대화하는 방안을 모색할 수 있었음

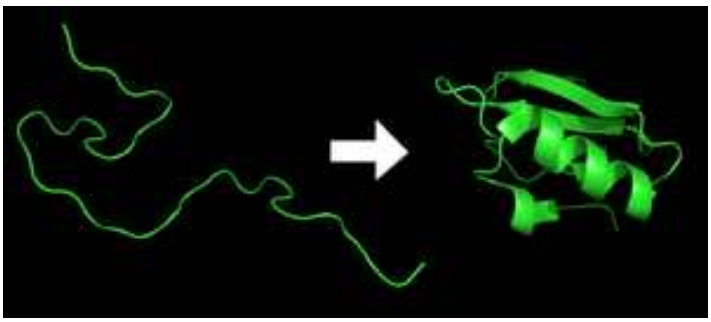
- 간편한 데이터 접근을 위해 지연 시간, 하드웨어 요구 사항 등을 테스트함
- 영구 디스크 스토리지 클러스터를 활용하여 더 나은 확장성 및 피크 성능을 확보
- 공용 네트워크를 사용하였음에도 초당 57메가비트의 높은 전송 속도를 보여줌

○ 사용된 요소 기술은 분산 파일 시스템, 고성능 분석 클러스터, 클라우드 등이 있음

(5) 분자생물학 사례 - Folding@Home

□ 분자생물학 소프트웨어 활용 필요성

○ 단백질 접힘이란 단백질 분자를 구성하는 펩티드 사슬이 공간적 배치를 통해 3차원 등 고차원 구조를 형성하는 것을 말함



[그림 10] 단백질 접힘의 전 후 모습
(출처: http://en.wikipedia.org/wiki/Folding@home)

○ 단백질 접힘의 메커니즘을 연구하면 단백질의 3차원 입체 구조를 파악하고 잘못된 접힘의 원인을 파악하여 이를 약물 설계에 응용할 수 있음

○ 단백질 접힘의 메커니즘을 연구하는 것은 수많은 컴퓨팅 자원과 고도화된 컴퓨터 알고리즘, 시행착오를 거친 학습이 필요함

□ Folding@Home 개요

○ Folding@Home은 단백질 접힘, 고도의 계산이 필요한 약물 설계나 기타 유형의 분자 역학 시뮬레이션을 위한 분산 컴퓨팅 프로젝트

- 2014년 7월 현재 179,237명이 참여하고 있으며, 전 세계에 걸쳐 다양한 과학 기관 및 연구소(노트르담 Izaguirre 랩, 버지니아 주립대학, 스톡홀름 대학, 콜로라도 주립대학, HKUST, CSULB, 템플 대학, 크로아티아 생명 과학 지중해 연구소 등)에 공유되었고, 다양한 기업(인텔, 구글, 소니, ATI, NVIDIA, 스톤하퍼 등)과 커뮤니티 봉사자들의 참여로 진행되고 있음

○ 이 프로젝트는 개인용 또는 기업용 컴퓨터의 유휴 처리 자원을 활용하여 분산 컴퓨팅 구조를 구성함

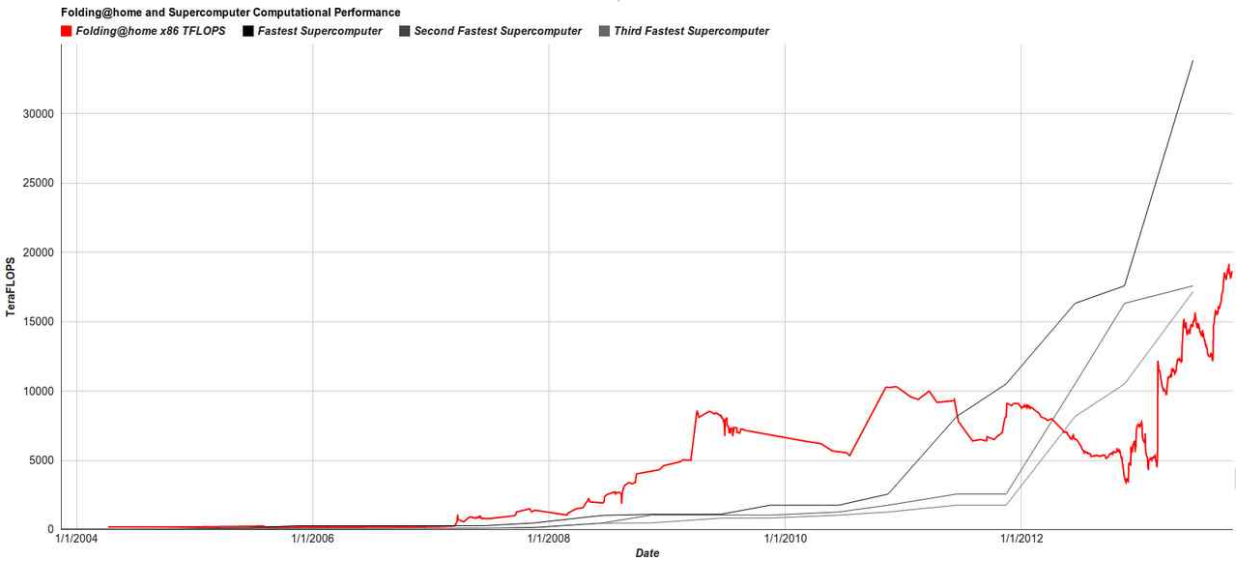
- 사용자는 부분 작업을 다운로드 받아 단백질 접힘 실험을 하고 Folding@Home 서버에 다시 업로드하며, 이 과정은 자동으로 반복됨

- Folding@Home 사용자는 개인용 컴퓨터에 클라이언트 프로그램을 설치함. 클라이언트는 백그라운드의 다른 소프트웨어 구성요소를 관리하며, 클라이언트를 통해 사용자는 접힘 실험을 중지할 수 있고, 이벤트 로그를 열람할 수 있으며, 작업 진행 상황을 체크할 수 있고, 개인별 통계를 볼 수 있음

- 이 클라이언트는 개인용 컴퓨터의 유휴 처리 자원을 활용할 수 있도록 매우 낮은 우선순위로 백그라운드에서 계속 실행되어, 정상적인 컴퓨터 사용에 영향을 주지 않음

□ 소프트웨어 활용성과

○ 2014년 7월 현재 총 179,237명이 참여하여 39,022 테라 플롭(FLOP, 1초에 38페타)의 컴퓨팅 파워를 확보하였고, 그림 11과 같이 2007년 6월부터 2011년 6월 사이에 Folding@home(적색선)은 가장 빠른 500개의 슈퍼컴퓨터(검정선)의 성능을 넘어선 바 있음. 동시간대 Folding@Home과 비슷한 프로젝트인 Seti@Home은 681 테라 플롭의 컴퓨팅 파워를 보여주고 있음[3)]



[그림 11] Folding@home 성능
(출처: http://boincstats.com/en/stats/0/project/detail/overview)

○ HKUST에서 후이 후앙의 연구 그룹은 Folding@Home을 활용하여 hIAPP(human islet amyloid polypeptide, amylin[4)])의 잘못된 접힘을 조사하기 위해 대규모 분자 역학 시뮬레이션을 수행하였으며 이를 통해 유형 II 당뇨병 환자의 95%가 hIAPP의 잘못된 접힘을 보이고 있음을 밝힘

○ 후이 후앙의 그룹은 hIAPP 의 구조적 상태를 확인하고 그들 사이

3) 보닉 통계 자료, Accessed July http://boincstats.com/en/stats/0/project/detail/overview
4) 췌장의 β세포에서 합성·분비하는 천연성 펩티드호르몬으로 인슐린과 같이 당대사를 조절한다. 수용체의 해석이나 생리적 뜻의 해명이 현재 진행되고 있으며, 이를 통해 비만인슐린 저항성을 수반하는 고혈압증이나 인슐린 비의존성 당뇨병 등의 치료제가 될 가능성이 있다.

의 전환 역학을 파악함으로써 유형 II 당뇨병 치료에 한 걸음 더 다가갈 수 있었으며, 이와 같은 결과는 2013년 미국 화학회지(JACS: Journal of the American Chemical Society)에 게재됨[5)]

○ 사용된 요소 기술은 분산 컴퓨팅, 그리드 컴퓨팅 등임

(6) 분자생물학 사례 - Foldit

□ Foldit 개요

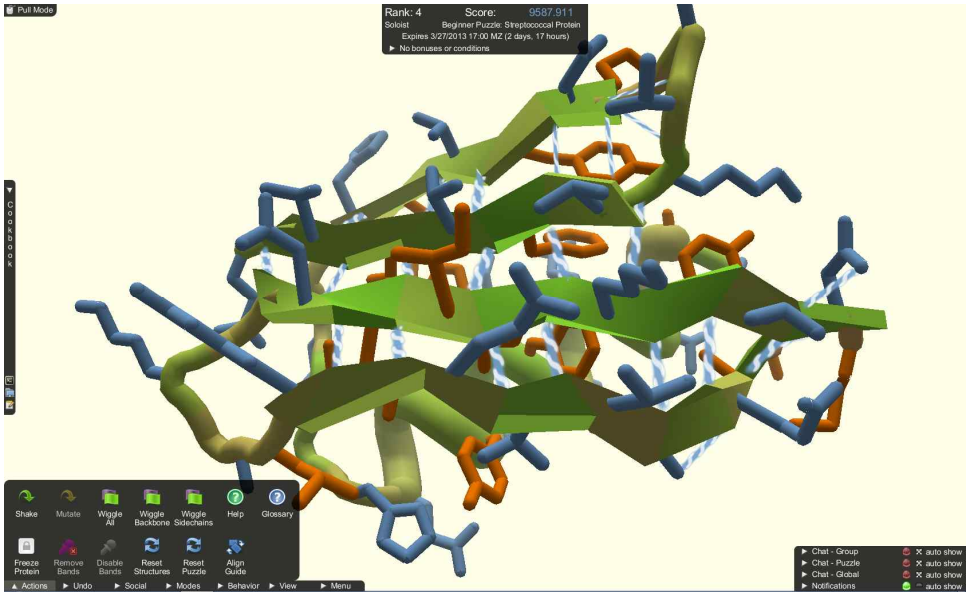
○ Foldit은 단백질 접힘을 위한 온라인 퍼즐 게임으로, 워싱턴 대학 생화학 학과와 동 대학 게임 과학 센터가 공동으로 개발함

○ Foldit은 인간의 뇌가 갖고 있는 자연적인 3차원 입체 패턴 매칭과 공간 추론 능력을 활용하여 단백질 구조 예측 문제를 해결하려는 시도를 하고 있음

○ 사용자는 Foldit에서 제공하는 다양한 도구를 활용하여 단백질 구조를 만들어 봄

○ 사용자가 만들어낸 단백질 구조는 이미 접힘 방법이 규명된 것으로 접힘의 자체 보다 사람들이 직관적으로 퍼즐을 푸는 방식을 분석하여 기존 단백질 접힘 소프트웨어가 사용하는 알고리즘을 개선하고자 함

5) Qiao, Q., Bowman, G. R., & Huang, X. (2013). Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation. Journal of the American Chemical Society, 135(43), 16092-16101.



[그림 12] Foldit의 단백질 접힘 화면
(출처: http://fold.it/portal/info/about)

□ 소프트웨어 활용성과

○ 2011년 Foldit은 에이즈를 일으키는 원숭이 바이러스의 일종인 메이슨-화이자 원숭이 바이러스(M-PMV)의 레트로 바이러스 단백질 분해 효소의 결정 구조를 해독하는 데 도움을 주었으며 과학자들이 15년 동안 해결하지 못한 이 문제를 단 10일 만에 해결함

○ 2012년 합성 화학에서 사용되는 딜스-알더 반응을 촉진하는 효소는 Foldit에서 기존의 효소보다 18배나 활동량이 많은 효소를 재설계함

○ 사용한 요소 기술은 머신러닝, 그리드 컴퓨팅, 게임화, 이미지 프로세싱 등임

(7) 분자생물학 사례 -EteRNA

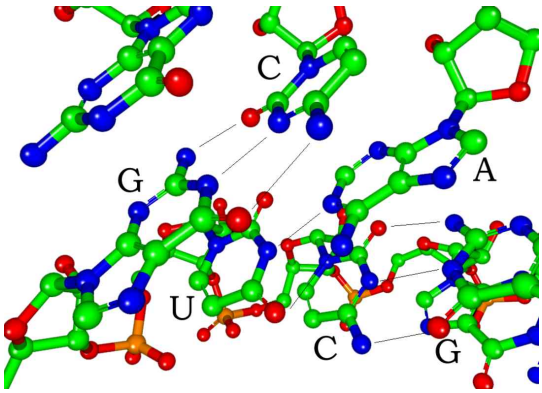
□ 분자생물학의 소프트웨어 활용 필요성

○ RNA는 단백질을 합성하는 과정에 작용하며, 아데닌(A), 구아닌(G), 우라실(U), 시토신(C)의 핵염기가 반복적으로 결합되어 있는 형태임

○ 각각의 핵염기는 구아닌(G)-시토신(C)과 아데닌(A)-우라실(U)의 형태로 결합하며, 우라실(U)-아데닌(A)-우라실(U) 등의 다양한 형태의 결합이 일어나 제한적이면서도 자유로운 입체구조를 가질 수 있음(그림 13)

○ 따라서 RNA는 핵염기 배열에 따라 자연 상에서 특정 3차원 구조로 접히는 특징을 가지며 RNA 구조에 맞는 핵염기 배열을 찾아내는 것은 질병 연구에 중요한 역할을 함

- RNA의 구조 별로 rRNA, mRNA, tRNA, siRNA 등으로 불리며 각기 다른 역할을 함



[그림 13] siRNA의 구조의 일부
(출처: http://en.wikipedia.org/wiki/RNA)

○ 하지만 특정 구조로 접히는 핵염기 배열을 찾아내는 것은 분자 간 상호작용이 복잡하여 컴퓨터 알고리즘만으로 어려우며, 알고리즘이 찾아냈다 하더라도 찾아낸 분자 배열이 실제로 원하는 구조로 접히지 않는 경우가 대부분임

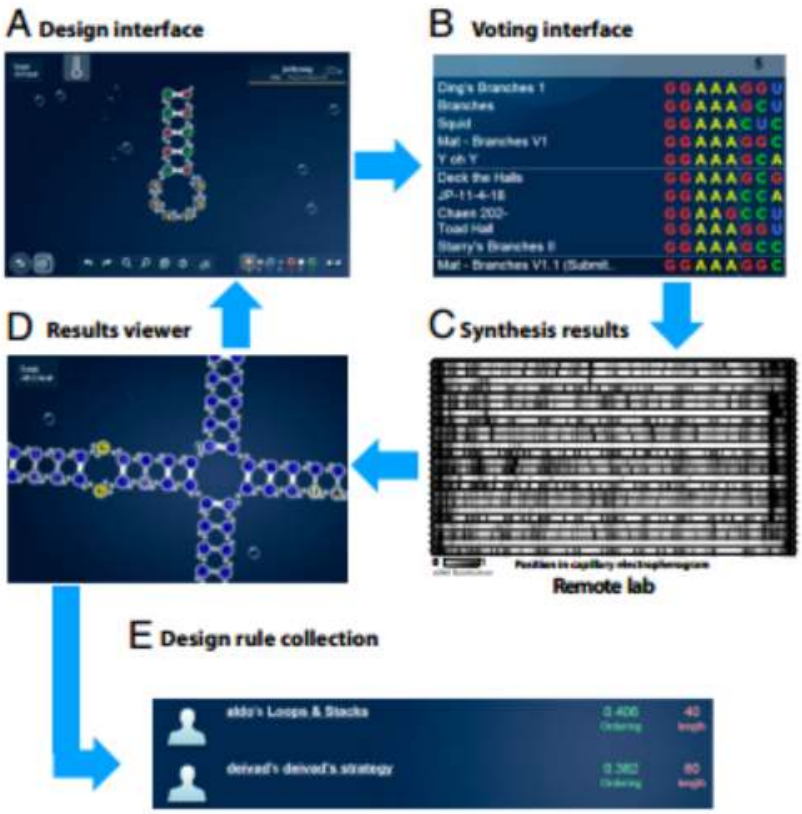
○ 기존과 다른 방법으로 RNA 구조에 맞는 핵염기 배열을 알아내고, 컴퓨터 알고리즘을 향상시킬 필요가 있음

□ EteRNA 개요

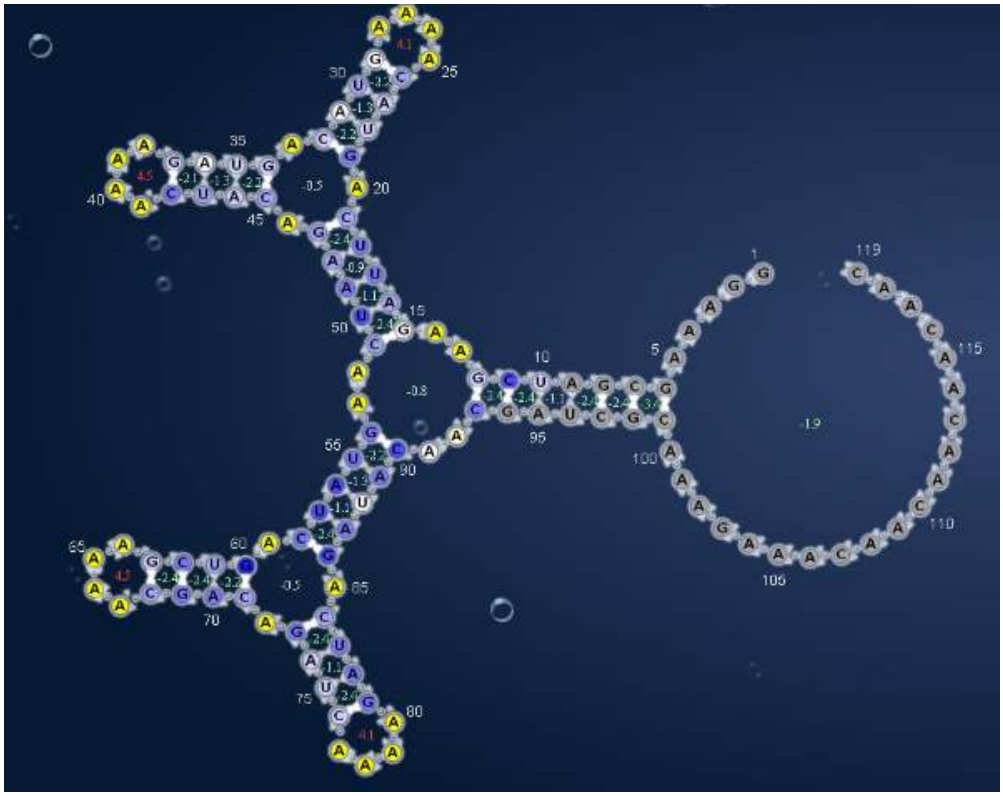
○ 온라인 퍼즐 게임을 통해 사용자가 RNA 구조에 맞는 핵염기 배열을 찾게 하는 'EteRNA' 프로젝트를 진행함

○ 사용자의 참여를 통해 RNA구조에 맞는 여러 핵염기 배열들을 찾아내고 그 중 최적의 배열에 투표하며, 이렇게 찾아진 배열은 연구실에서 실제로 합성되어 실현 가능성을 입증함(그림 14)

- 실험 결과는 다시 온라인에 게시되고 사용자는 이를 바탕으로 더욱 개선된 핵염기 배열을 찾음



[그림 14] EteRNA 워크플로우 (출처: Lee et al. 2014)



[그림 15] 사용자가 디자인한 RNA 분석 결과가 게임 내에 표시되는 화면
(출처: 사이언스온)

□ 소프트웨어 활용성과

○ 사용자가 핵염기 배열을 찾아내는 과정을 머신러닝을 통해 학습하여 이를 바탕으로 새로운 RNA 디자인 알고리즘을 만들어내었으며, 이는 기존의 알고리즘보다 더욱 좋은 성능을 냄

- 새로운 알고리즘은 기존의 알고리즘 보다 95%이상의 확률로 더 좋은 분자 배열을 찾아냄[6)]

○ EteRNA의 결과는 미국국립과학원회보(PNAS: Proceedings of the National Academy of Sciences of the United States of America) 저널에 게재되었으며, 저자에 EteRNA 사용자가 포함됨

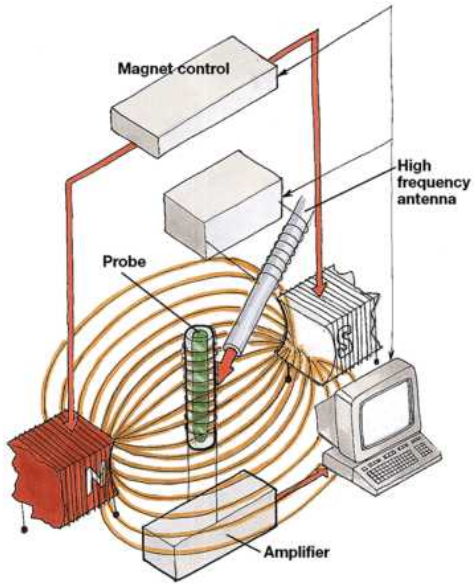
○ 사용된 요소기술은 머신러닝과 그리드 컴퓨팅임

6) 사이언스온(2014.2.5), "37,000명 온라인 게이머들이 RNA 염기배열 규칙을 풀다."

(8) 구조생물학 사례 - WeNMR

□ 구조생물학의 소프트웨어 활용 필요성

○ 핵자기 공명(NMR, Nuclear Magnetic Resonance)은 그림 16과 같이 커다란 자석 속에 놓인 단백질 등의 원자핵이 특정 주파수의 전자기파에 의해 공명하는 현상을 통해 분자의 특성을 알아내는 방법



[그림 16] NMR 분광기의 구조

(출처: http://muniche.linde.com/international/web/lg/spg/likelgspg.nsf/docbyalias/anal_nmr

○ 소각 X선 산란(SASX, Small-angle X-ray scattering)은 단백질 샘플을 X선을 통해 산란시켜 그 회절 패턴을 토대로 단백질 구조를 파악하는 방법

○ 해당 공명 신호나 회절 패턴을 토대로 분자의 구조를 계산하거나, 모델링하고 분석하는 것은 분산 클러스터 기반의 많은 컴퓨팅 파워가 필요함

□ WeNMR 개요

○ WeNMR은 구조생물학의 보완적 연구들을 가상 연구 커뮤니티로 통합하여, NMR과 SASX 데이터 분석과 구조 모델링에 필요한 계

산적 접근이 가능하도록 가상의 환경을 제공하고 있음[7]

○ WeNMR은 NMR 분석을 위한 Processing, Assignment Analysis, Structure Calculation, Molecular Dynamics Simulation, Modeling 등의 분야에서 사용되는 총 25개의 분석 도구를 하나의 시스템으로 제공하고 있음

<표 3> WeNMR에서 제공하는 NMR 분석 서비스

분류	서비스명
Processing Assignment	MDD NMR
	Auto Assign
	MARS
	UNIO
Analysis	TALOS+
	AsnisoFIT
	MaxOcc
	iCing
Structure Calculation	CS-ROSETTA
	CYANA
	UNIO
	Xplor-NIH
Molecular Dynamics Simulation	CcpNmr WMS
	AMBER
Modelling	GROMACS
	3D-DART
Various Software Tools	HADDOCK
	Format Converter
	SHIFT2X
	RCI
	Antechamber
	Preditor
	RCI
	UPLABEL
	SedNMR

(출처: www.wenmr.eu)

7) http://www.wenmr.eu

○ 또한 WeNMR은 SASX 분석을 위해 필요한 Ab Initio Shape Determination, Modeling. Instrument Access 분야에서 8가지 분석 도구를 제공함

<표 4> WeNMR에서 제공하는 SASX 분석 서비스

분류	서비스명
Ab Initio Shape Determination	DAMMIN
	DAMMIF
	GASBOR
	MONSA
Modelling	CRYSOL
	SASREF
	EOM
Instrument Access	EMBL's SAXS beamline P12 remote cccess

(출처: www.wenmr.eu)

□ 소프트웨어 활용성과

○ WeNMR은 분자구조학에서 가장 큰 연구 커뮤니티가 됨

○ 2007년부터 2011년까지 272개의 전용 CPU와 2.87테라바이트의 연구 데이터가 수집됨

- 공유 환경을 합하면 총 10,000여개의 CPU가 NMR, SASX 등의 분자분광법에 사용되고 있으며 연구데이터는 37테라바이트에 이름

○ 사용된 요소기술은 분산 처리, 클러스터, 그리드 등임

(9) 환경과학 분야 사례 - MODIS

□ 환경과학의 소프트웨어 활용 필요성

○ 환경과학은 오염 감소와 방지 등 공해문제를 중심으로 여러 환경 문제를 과학적으로 규명하려는 응용과학의 한 분야임

○ 전 세계의 환경 동태에 대한 데이터는 다양한 방법으로 수집됨

- 지구관측 위성은 매 1~2일 마다 토지, 해양, 하층부 대기 등의 데이터를 수집함

- 데이터는 위성뿐만 아니라 대지 기반의 센서로 부터도 수집됨

○ 이러한 데이터는 지구온난화, 기후 변화 감시, 대기의 연직 사운딩[8] 및 해상과 육상에서의 생물간의 상호 관계 관측 등 다양한 분야에서 유용하게 활용됨

○ NASA는 지구관측 위성에 탑재된 MODIS(Moderate Resolution Imaging Spectroradiometer)를 이용하여 수집한 데이터를 공개함

- MODIS는 일반 관측위성과 달리 36개의 광대역에 이르는 분광 파장역[9]에서 자료를 수집함

- MODIS가 수집하는 데이터는 대기 데이터, 대지 데이터 등으로 나눌 수 있으며, 또한 수집된 데이터 별로 데이터의 파동 형태가 다를 수 있음

○ MODIS에서 수집된 데이터를 효율적으로 처리하기 위해 해당 데이터를 통일된 지질학적 포맷으로 변환하여야 하며, 대기와 대지 데이터의 통합이 필요함

- 시간 소모적이며 반복적인 작업이 많이 요구됨

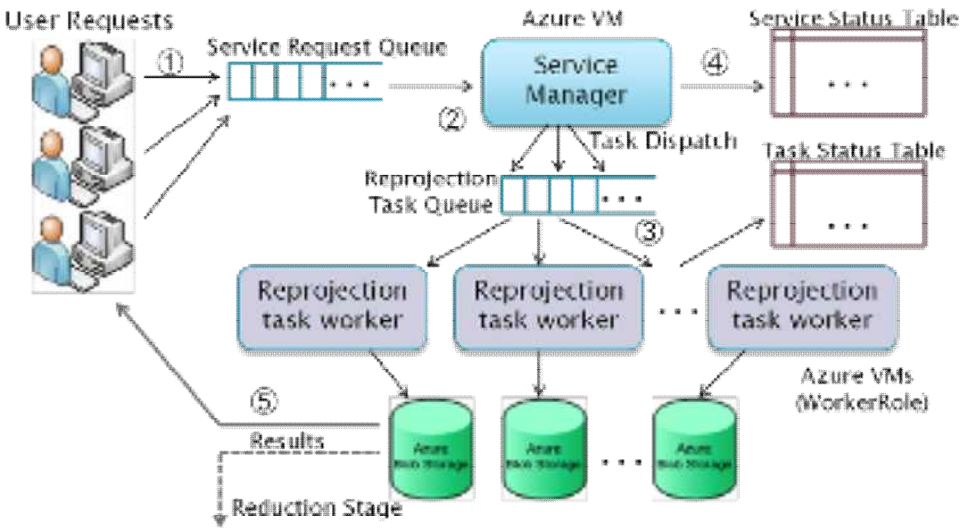
- 따라서 미대륙의 10년치 데이터를 프로세싱 하는 작업에 수만 시간의 CPU시간이 필요함

8) 직접적인 지반조사의 방법의 일종
9) 더 이상 나눠질 수 없는 단색광으로 분해된 빛의 파동이 미치는 범위

□ MODIS 데이터 통합 개요

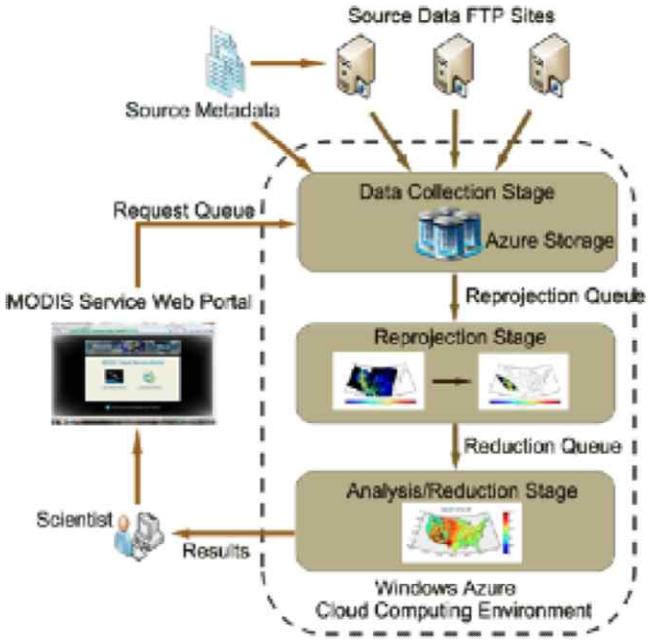
○ 데이터 파이프라인 재구성 프로젝트는 데이터 저장과 통합에 퍼블릭 클라우드 서비스인 '윈도우 애저 블랍 스토리지'를 사용하였고, 데이터 변환 및 분석에는 .NET 기반 프레임워크를 활용한 맞춤화된 애플리케이션을 사용함

○ 그림 17은 파이프라인 단계 중 변환(Reprojection) 작업을 위한 작업 스케줄링 방식을 나타낸 것으로 서비스 관리자가 변화 작업 큐에 작업을 할당하고 큐에서 순서대로 변환 작업이 실행됨



[그림 17] 재투영 작업을 위한 작업 스케줄링 방식
(출처: Li, Jie, et al., 2010)

○ 그림 18은 파이프라인 단계에 따른 시스템 구성도를 나타낸 것으로 연구자가 MODIS 웹 포털을 통해 작업을 요청하면 데이터 수집 및 통합, 변환, 분석 단계를 거쳐 의미 있는 정보가 생성됨



[그림 18] 파이프라인 단계에 따른 시스템 구성도
(출처: Li, Jie, et al., 2010)

□ 소프트웨어 활용성과

○ 클라우드 서비스를 이용해 MODIS의 데이터를 수집, 변환, 통합, 분석에 걸친 데이터의 흐름을 재구성 함

○ 데스크탑보다 저사양의 150개의 중급 애저 인스턴스를 활용하여 약 90배 빠르게 분석을 진행 할 수 있게 됨

<표 5> 데스크탑 머신과 애저 인스턴스 성능

	데스크탑	애저 인스턴스
성능	CPU: 인텔 코어2듀오 E6850@3.0GHZ 메모리:4GB 하드 디스크: 1TB SATA 네트워크: 1Gbps 이더넷 OS: 윈도우 7 RC 빌드7100(32비트)	CpU: 1.5~1.7GHZ X64 동급 프로세서 메모리:2GB 로컬 스토리지: 250GB 네트워크: 100Mbps OS:윈도우 2008 서버 x64(64비트)

(출처: Li, Jie, et al. 2010)

○ 단일 처리 및 병렬 처리/확장 수준을 파악해 보면 표5와 같음

<표 6> 1,500개의 작업을 처리하는데 걸린 시간 (단위:시간)

	MOD04_L2	MOD06_L2	MYD11_1.2.005
데스크탑	16.29	72.62	33.45
단일 인스턴스	34.21	116.19	63.56
50 인스턴스	0.76	2.25	1.12
100 인스턴스	0.40	1.20	0.61
150 인스턴스	0.30	0.85	0.44

(출처: Li, Jie, et al. 2010)

○ 사용된 요소 기술은 클라우드임

(10) 네트워크 과학 사례 - CINET

□ 네트워크 과학의 소프트웨어 활용 필요성

○ 네트워크 과학은 그래프의 추상화 이론, 알고리즘, 에이전트 기반 모델링, 시뮬레이션 등의 분야가 있으며, 생물학, 생태학, 세포 면역학, 사회과학, 보건과학, 경제, 컴퓨터 네트워크, 역학 통계, 물리학, 진화학 등에서 응용되고 있음

○ 생물학, 생태학 등과 같은 특정 과학 분야의 시스템 속성과 행동을 이해하기 위해 그래프 기반의 계산을 수행할 수 있는 소프트웨어가 필요함

○ 네트워크 연구와 특성화를 위한 다양한 알고리즘 모듈과 측정 수단이 필요하며, 대규모의 네트워크를 연구하는 특성상 고성능 컴퓨팅 파워가 필요함

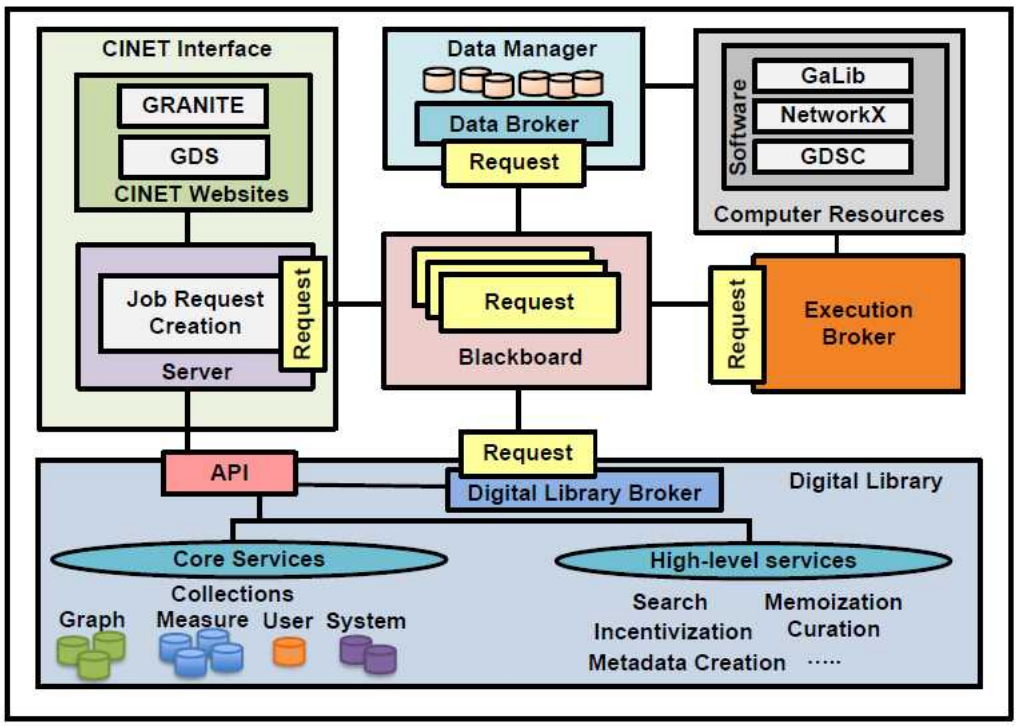
□ CINET 개요

○ CINET(A Cyber Infrastructure for Network Science)은 네트워크 과학 연구와 거대 그래프 분석을 위한 사이버인프라임

○ 백엔드 HPC(High Performance Computing) 클러스터에서 동작하는 고성능 컴퓨팅 엔진과 사용자를 연결하는 기능을 가진 미들웨어 플랫폼을 제공하고 데이터 매니저와 실행 인스턴스를 통해 사용자 작업 요청을 처리

- 계산 엔진 및 그래프 라이브러리인 Galib, Network X 구현
- 그래프 동적 시스템 계산기인 GDSC(Graph Dynamical Systems Calculator), 분산 워크플로우 관리, 시멘틱 웹 툴 구현

○ 그림 19는 CINET의 시스템으로 디지털 라이브러리, 웹 기반 인터페이스, 백엔드 시스템인 서버, 블랙보드, 데이터 브로커를 나타냄



[그림 19] 고수준 CINET 구성요소 및 상호작용
(출처: Abdelhamid et. al,, 2012)

□ 소프트웨어 활용성과

○ CINET은 개방형 과학 그리드(Open Science Grid)에 공개되어 다양한 분야의 연구자들이 무료로 CINET을 통해 연구할 수 있게 됨

○ 대상 사용자는 학생, 교사, 분야별 전문가, 연구자 등이며 HPC 시스템에 쉽게 접근할 수 없고 소프트웨어에 전문성이 부족한 사람들도 활용할 수 있도록 공용 저장소와 협업 시스템을 마련해 줌

○ 사용된 요소 기술은 클라우드(HPC as a Service)임

(11) 과학연구 관리 사례 - Taverna

□ 과학연구 관리의 소프트웨어 활용 필요성

○ 과학연구에서 사용하는 데이터와 데이터 분석에 사용되는 도구가 산재해 있어, 데이터의 유형이나 분석 도구 등을 한 번에 접근할 수 있는 과학연구 흐름 관리 시스템이 필요함

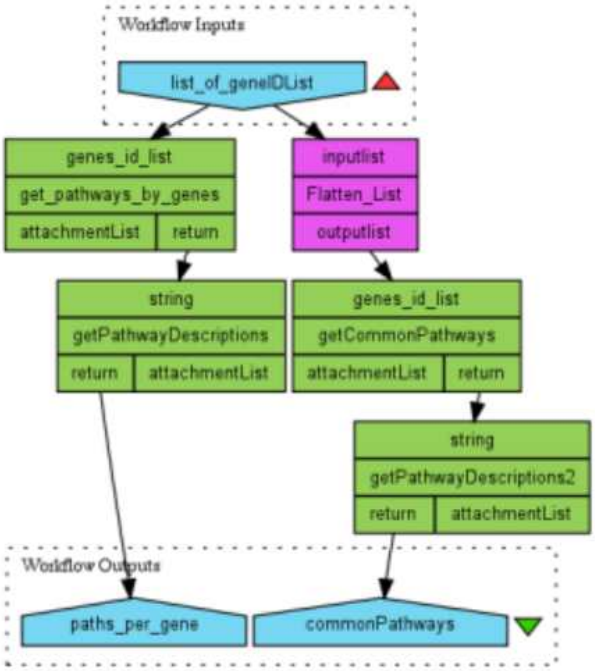
□ Taverna 개요

○ Taverna는 생물정보학(Bioinformatics), 천문학, 생물다양성 등의 연구에서 작업 흐름을 관리하고 흐름과 같이 분석을 실행 할 수 있는 프로그램임

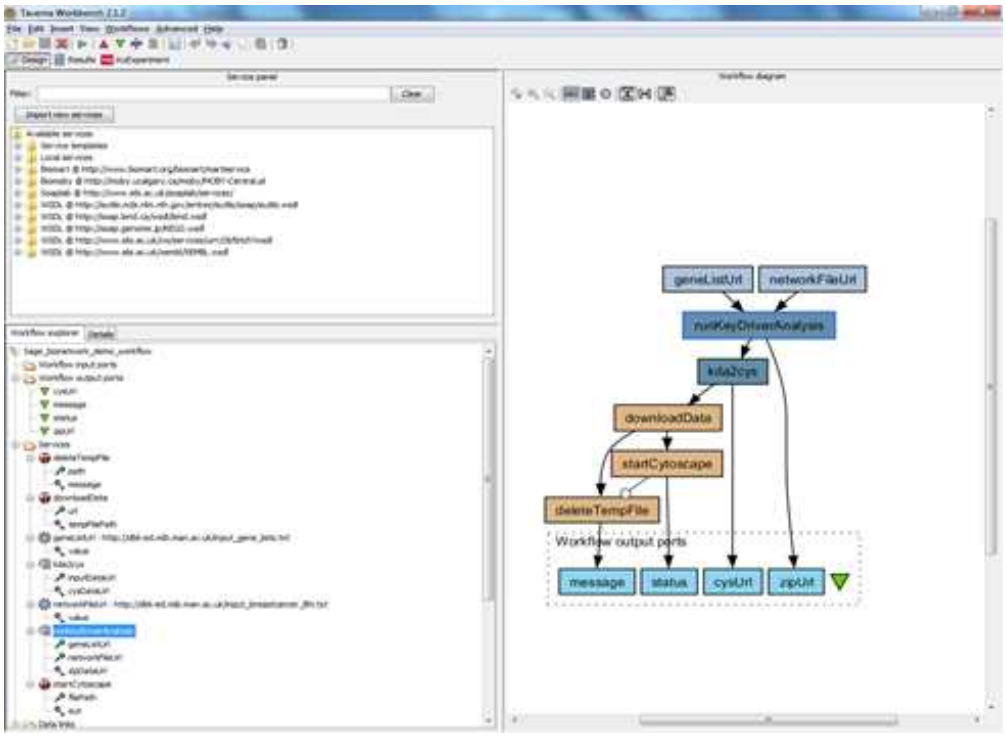
○ LGPL(Lesser General Public License)기반의 공개 소프트웨어로 자바를 기반으로 구축됨

○ Taverna는 각각의 작업에 맞는 데이터를 해당 분석 도구에 맞춰 자동으로 입력하고 작업을 통해 분석된 데이터를 다음 작업으로 전달되거나 결과로 출력됨(그림 20)

○ 연구자들은 Catalogue of Life, BioSTIF, eFamily 등 곳곳에 흩어져 있는 데이터를 한곳에서 이용할 수 있으며, 데이터의 흐름을 시각화된 도구를 이용하여 관리하고 분석할 수 있음



[그림 20] Taverna의 작업 흐름 예시(출처: Missier, Paolo, et al 2010)



[그림 21] Taverna Workbench 사용 모습
(출처:http://www.taverna.org.uk)

○ 유럽연합의 FP7과 JISC 등의 지원을 받는 myGrid 연구팀의 맨체스터 대학교의 연구팀이 구축하였음

- myGrid 연구팀은 맨체스터 대학교를 포함하여 사우스햄튼, 옥스퍼드 대학이 참여함

○ myGrid 연구팀은 Taverna를 포함하여 단백질 시퀀스 분석 도구인 utopia, 시멘틱 분석도구인 RightField 등 생물학, 천문학, 화학 등 다양한 분야의 e-Science 소프트웨어를 구축함

□ 소프트웨어 활용성과

○ Taverna를 통해 산재해있는 데이터와 분석도구에 대한 호환성을 손쉽게 확보할 수 있으며 다른 연구자가 디자인한 작업 흐름과 서비스 등을 공유할 수 있음

○ 2004년에 서비스를 시작한 이후 2014년 6월까지 9만 건이 넘는 다운로드가 이루어 졌으며, 약 380곳의 연구기관이 사용하고 있음

○ 사용된 요소 기술은 데이터 변환 및 통합 기술

(12) 사회과학 사례 - CLARIN

□ 사회과학의 소프트웨어 활용 필요성

○ 사회과학 연구자들의 연구 데이터를 공유할 수 있는 연구 인프라가 필요함

- 연구 데이터가 유실되는 경우가 많으며, 데이터의 공유에 있어 법적인 문제가 발생할 수 있음

○ 사회과학 데이터 형식이 다양하기 때문에 데이터를 등록, 분류, 검색이 어려움

□ CLARIN 개요

○ CLARIN(Common Language Resources and Technology Infrastructure)는 유럽 연구 인프라 전략 포럼(EESFRI: European Strategy Forum on Research Infrastructure)에 선정된 연구 인프라로 사회과학 연구자가 언어학과 관련된 데이터를 자유롭게 접근하고 향상된 분석 도구를 통해 연구 가능하게 함

○ CLARIN은 국제적 연구 네트워크이며, 이러한 연구 네트워크는 CLARIN-ERIC(European Research Infrastructure Consortium)에 의해 관리 및 유지 됨[10)]

- CLARIN-ERIC에는 오스트리아, 불가리, 체코, 덴마크, 에스토니아, 독일, 네덜란드, 폴란드가 참여하고 있음

○ 연구자들은 CLARIN을 통해 원하는 자료를 쉽게 찾고 이를 법적인 문제없이 쉽게 이용할 수 있도록 함

- 메타데이터 기반으로 자료를 검색할 수 있으며, 여러 국가가 가지고 있는 언어학 자료의 시멘틱 호환성을 유지하면서 여러 메타데이터를 함께 사용할 수 있는 통합 환경을 제공함

- 연구자들이 CLARIN에 자료를 업로드 하면, 이 자료들이 CLARIN을 통해 분류되어 다른 연구자들이 데이터를 쉽게 찾을 수 있도록 함

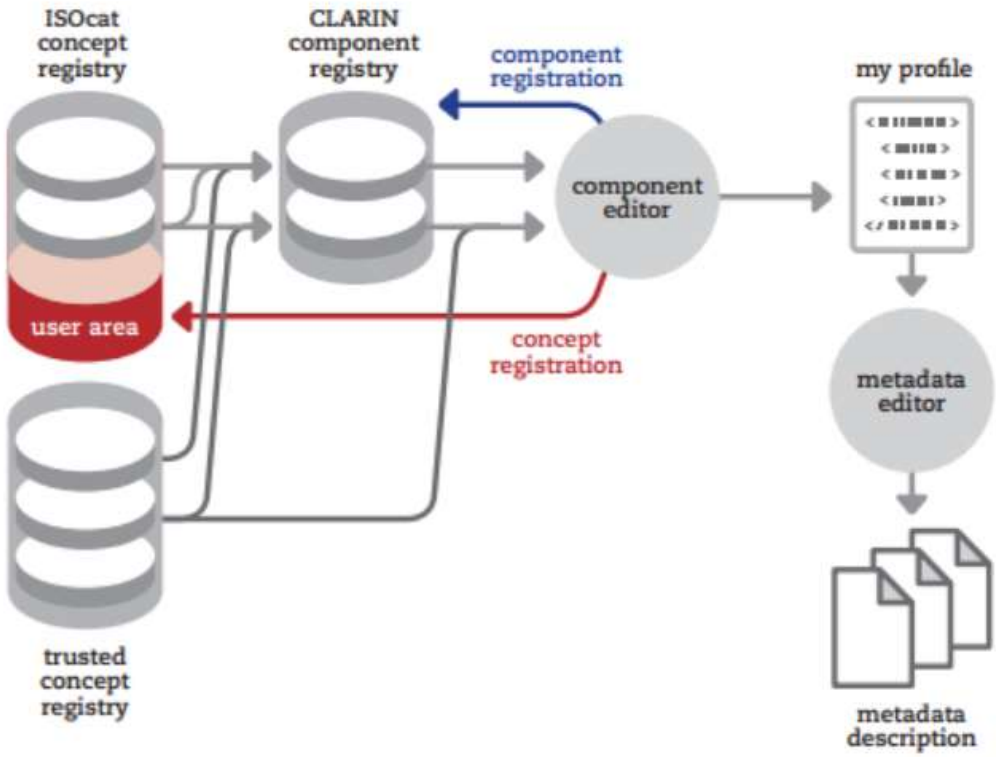
○ CLARIN은 다양한 소프트웨어 분석 도구를 제공하며 CLARIN에 저장된 데이터를 웹 기반에서 분석할 수 있어서 연구자들이 분석을 위해 데이터를 다운로드 받지 않아도 됨

- 분석 도구로 Sentence Splitters, Tokenizer, Part-of-speech Triggers, Morphological analyzer and lemmantizer, Syntax-parsers and chunker 등을 제공

10) http://www.clarin.eu

○ CLARIN 메타데이터 프레임워크는 그림 22와 같으며 표준화된 스키마(Schema)를 요구하는 것이 아니라 데이터 카테고리 레지스트리(ISOCat)를 사용하며, 이를 기반으로 사용자가 직접 메타데이터를 만들고 수정할 수 있음

- 데이터 카테고리 레지스트리를 통해 다양한 프로파일을 기준으로 검색함



[그림 22] CLARIN Component 메타데이터 프레임워크
(출처: CLARIN ShortGuide, http://www.clarin.eu)

□ 소프트웨어 활용성과

○ 언어학 자료의 호환성을 유지하면서 여러 메타데이터를 함께 사용할 수 있는 통합 환경을 구축함으로써 자료의 등록과 분류, 검색이 간편해짐

○ 웹상에서 데이터를 분석할 수 있는 도구를 제공함으로써, 방대한 양의 데이터를 개별 연구환경으로 다운로드 받지 않아도 되어 분석 시간이 단축됨

○ 개별 연구자의 데이터를 CLARIN에서 보관함으로써 데이터 유실을 방지함

○ 사용된 요소 기술은 데이터 카테고리 레지스트리(ISOCat), 데이터베이스 등임

(13) 금석학 사례 - VRE SDM

□ 금석학의 소프트웨어 활용 필요성

○ 금석학은 비석에 쓰여 있는 명문을 연구하는 학문으로 역사학, 문학, 언어학 등의 보조 과학[11]임

○ 금석학 연구는 비석이나 금속에 쓰여 있는 글자를 적외선 투사, 그림자 분석 등을 통해 해석하는 것으로 분석 방법에 따라 결과의 차이가 있기도 함

- 비석이나 금속에 쓰여 있는 문자는 가독성이 떨어지는 경우가 대부분이며, 시대에 따라 글자의 모양도 달라질 수 있음

○ 문자의 의미를 파악하기 위해 다양한 분석 방법을 사용하는 동시에 기존의 해석과 비교하는 등의 학자들 간 협력 작업이 필요함

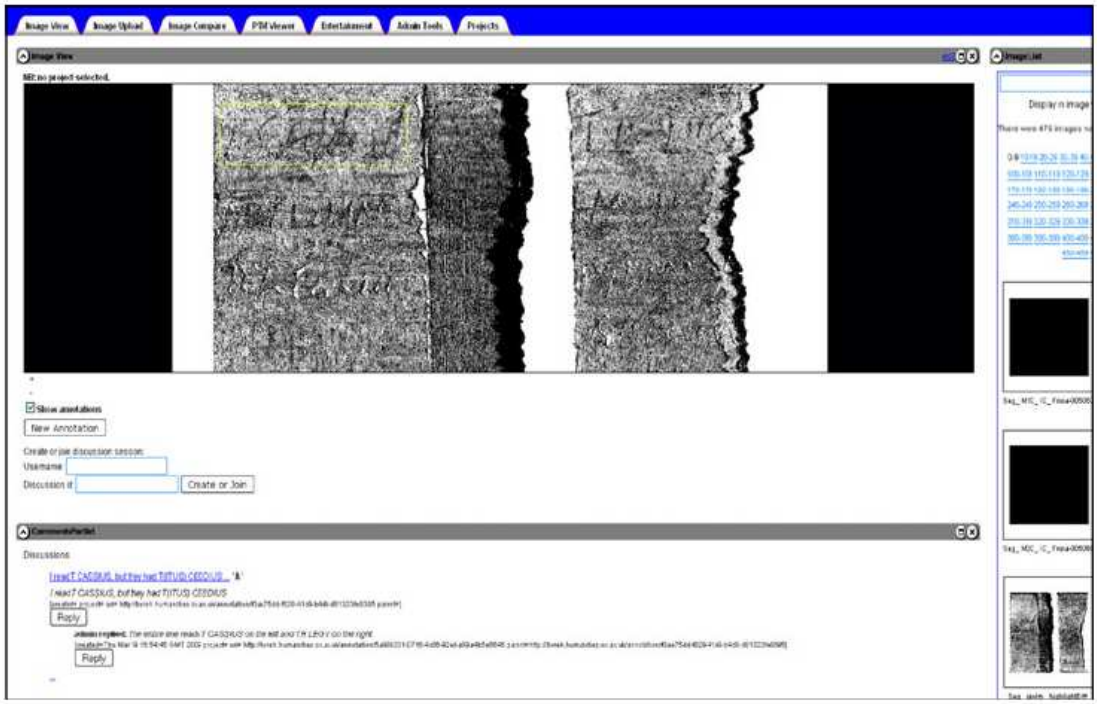
□ VRE SDM 개요

○ VRE SDM(Virtual Research Environment for Humanities for the Study of Document Manuscripts)은 금속이나 비석을 이미지화하여 저장하고, 밝기와 채도 등을 조정하여 다양한 방법으로 볼 수 있게 함

11) 보조과학은 역사적 사료를 분석하는 데 도움을 주는 과학 분야를 일컬음

○ 금석학자들 간의 협력을 강화하기 위해 이미지를 공유하고, 이미 구축되어 있는 데이터에서 분석하고자 하는 문자와 비슷한 문자열을 찾아 보여줌

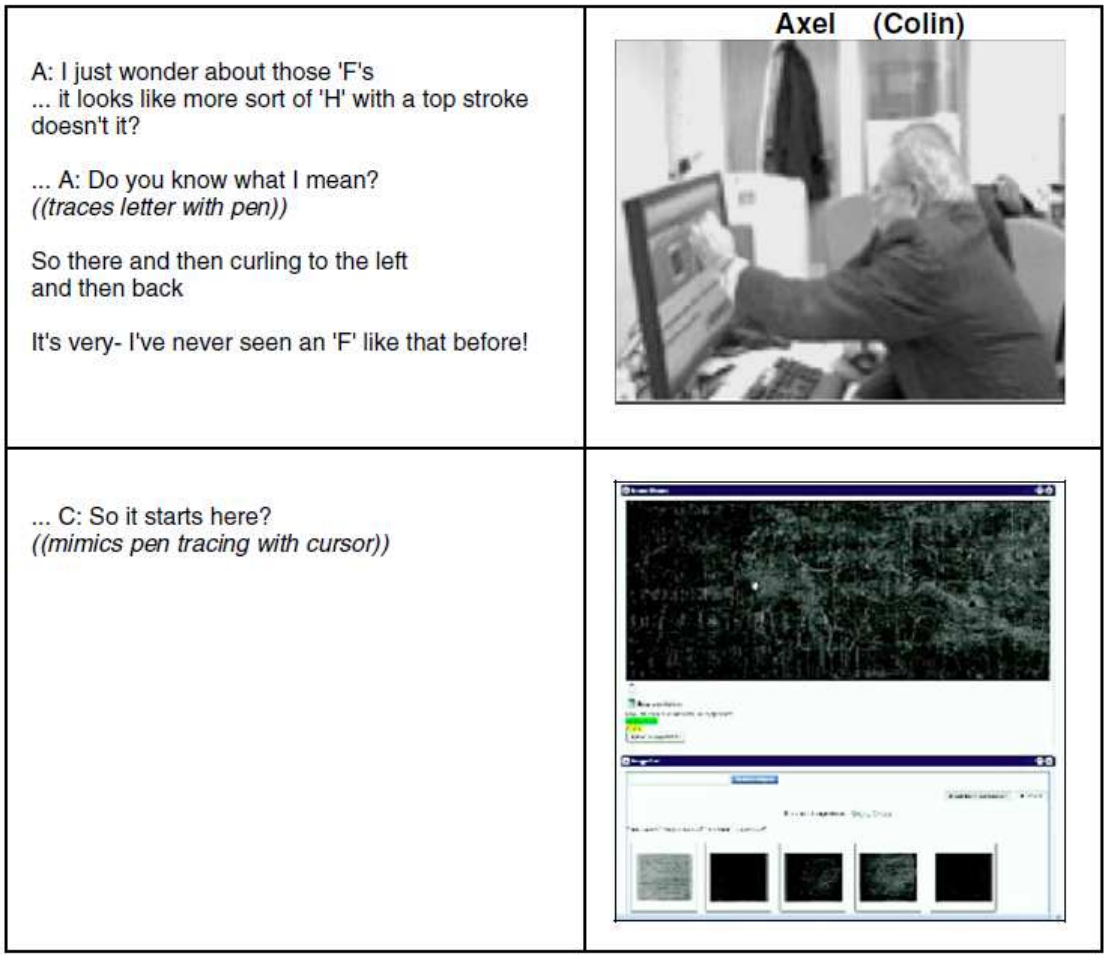
○ 그림 23은 VRE SDM의 프로토타입을 나타냄



[그림 23] VRE SDM 프로토타입
(출처:De La Flor et al., 2010)

□ 소프트웨어 활용성과

○ 기존의 연구자들은 그림 24의 상단과 같이 연구자들이 문자를 각자 해독하고 이에 대한 합의를 거쳐야 했으나, VRE SDM을 사용하면 해독하고자 하는 글자와 이미 해독된 유사한 글자를 보여줌으로써 해독 과정을 간편히 하고 해독의 정확성을 높일 수 있음



[그림 24] VRE SDM을 통해 문자를 해독하는 과정
(출처:De La Flor et al., 2010)

○ 사용된 요소 기술은 이미지 프로세싱임

3. 과학연구 SW활용 지원 현황

(1) 영국 e-Science

○ 영국에서는 1999년 John Taylor가 e-Science를 주장한 이후 과학연구에 소프트웨어를 활용하고 있음

○ 영국은 2001년부터 e-Science(2001~2011년), e-Infrastructure(2012년~) 프로그램을 통해 과학계와 컴퓨터학계의 다학제적 연구를 지원하고 있음

- (연구 지원의 목적) 과학연구에서 컴퓨터를 통해 데이터를 수집, 정리, 분석과 장기적인 데이터의 저장과 효율적인 접근을 이뤄냄으로써 과학발전을 촉진

- (e-Science 성과) 과학계와 컴퓨터학계간 커뮤니티가 형성되었으며, 소프트웨어 표준과 온톨로지(Ontology)를 설립하였음[12]. 또한 영국 전역에 26개의 e-Science 지원센터가 설립되었으며, 181개의 과학연구 소프트웨어가 제작되었음

(2) 유럽연합 e-Infrastructure

○ 유럽연합의 R&D 지원 프로그램의 일환인 e-Infrastructure 연구를 통해 과학연구에 소프트웨어 활용을 지원함

- FP6(2002~2006년) 연구 구조화 영역(Structuring the European Research Area)의 연구 인프라 프로그램을 통해 범유럽지역연구망[13]과 같은 연구자 네트워크와 그리드 인프라 구축

- FP7(2007~2013년) ICT기반의 데이터와 실험 장비들을 원격으로 제어하여 '가상실험(in-silico)' 환경 구축

12) Atkins, D. (2010). RCUK Review of e-Science: Building a UK Foundation for the Transformative Enhancement of Research and Innovation. tech. report, Eng. and Physical Sciences Research Council.

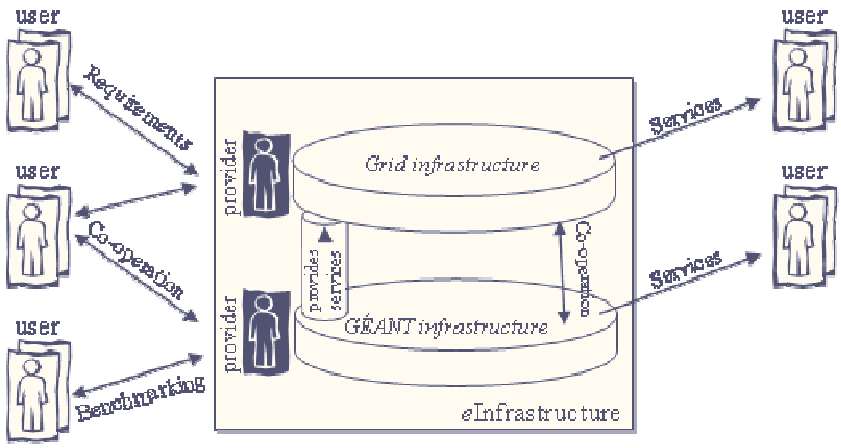
13) 범유럽지역연구망: GÉANT: European Multi-gigabit Computer Network for Research and Education

○ 2002년 유럽 연구 인프라 전략 포럼(ESFRI: European Strategy Forum on Research Infrastructure)을 설립

- 유럽 연합 국가들 간의 연구 인프라에 관련된 협약 지원 및 정책적 결정을 도와 전략적으로 연구 인프라를 구축하는데 기여

- 2006년, 2008년, 2010년 전략 로드맵을 발표하여 사회과학, 생물학, 환경, 에너지, 의료, 재료, 물리학 등의 분야에 필요한 연구 인프라 파악

○ 2009년 유럽 연구 인프라 콘소시움(ERIC: European Research Infrastructure Consortium)을 설립하여 범국가적 연구 인프라 구축에 필요한 법적인 프레임워크 수립



[그림 25] FP6에서 도입된 e-Infrastructure의 구조
(출처: http://cordis.europa.eu/ist/rn/ri-cnd/e-infrastructures.htm)

(3) 미국의 Cyber-Infrastructure

○ 대통령 직속 미국국가과학기술위원회(National Science and Technology Council) 과학기술정책분과에서 과학연구의 '사이버 인프라[14]' 관점에서 소프트웨어의 활용 방안이 논의됨

14) 1990년대 Albert Arnold Gore에 의해 '국가 정보 인프라'라는 용어가 일반화 되었고, 1998년 대통령 정책 지침에 의해 '사이버 인프라'라는 개념이 도입됨

○ 미국 연방 정부의 과학 진흥 기구인 미국국립과학재단(National Science Foundation)은 사이버 인프라를 통해 연구자들이 새로운 통찰력을 얻고 광범위하고 복잡한 문제를 해결할 수 있도록 지원하고 있음

○ 사이버 인프라는 과학연구에서 활용할 수 있는 고성능 컴퓨팅, 데이터 분석 및 가시화, 분산 커뮤니티 가상 조직, 교육 및 인력 개발 등을 포함함

- 미국국립과학재단(National Science Foundation)은 2007년 과학 연구를 위한 사이버 인프라 구축 비전[15]을 제시함
- 이후 2012년 사이버 인프라 구축을 위한 5개년 계획을 수립하고 11개 부분의 프로그램으로 수행중임

○ 미국국립과학재단(National Science Foundation)은 사이버 인프라 사업부를 통해 사이버 인프라 자원, 도구 및 서비스를 개발하고 이를 도입할 수 있도록 지원함

- 슈퍼컴퓨터, 대규모 데이터 저장소 및 디지털 과학 데이터 관리 시스템, 소프트웨어 라이브러리 및 프로그래밍 환경 등을 지원
- 사이버 인프라 시스템의 효과적인 관리, 지속적 유지 보수 생산적인 활용을 위한 교육 및 학술 교류를 지원

(4) 개방형 과학 그리드

○ 개방형 과학 그리드는 미국국립과학재단(National Science Foundation)과 미국에너지부(Department of Energy)가 공동 투자하여 과학 연구 기관에 공통 서비스와 자원을 제공하는 컨소시움으로 2004년에 설립됨[16]

15) National Science Foundation(2007), Cyberinfrastructure Vision For 21st Century Discovery
16) http://www.opensciencegrid.org/about/

- 대형강입자충돌기(LHC: Large Hadron Collider)에서 발생하는 데이터를 처리하기 위해 설립됨

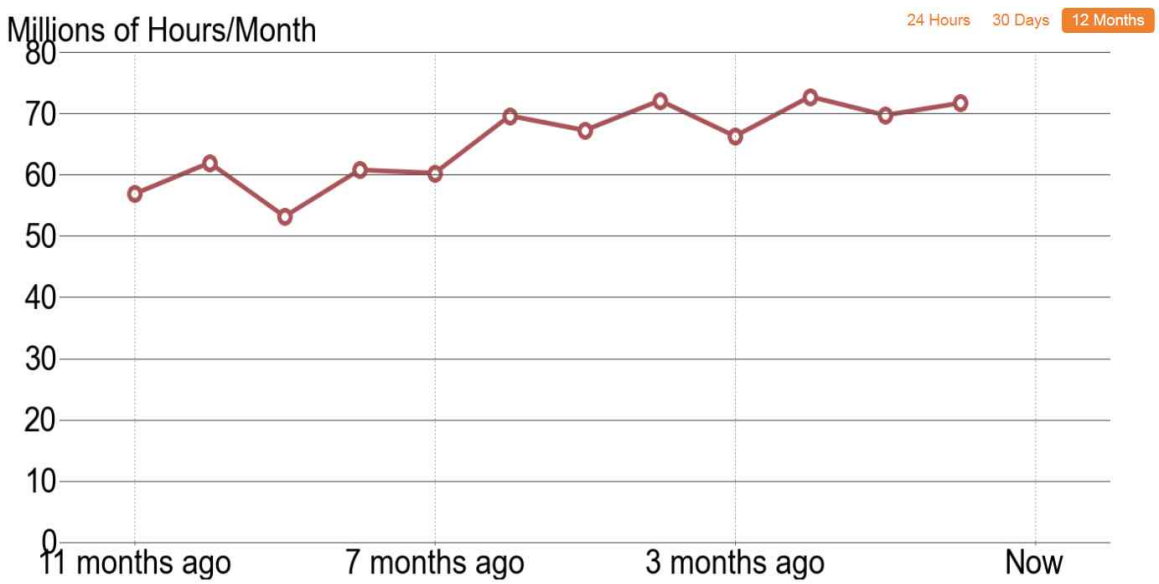
○ 미국 전역에 걸친 115개의 컴퓨터 센터와, 대학 및 국립 연구소의 연구자들 등을 연결하고 있음

- 2013년 10월부터 1년간 약 25만 테라바이트가 전송되었으며, 약 7억 8천 CPU 시간이 작업에 사용됨

○ 높은 처리량을 가지는 연산 서비스와 분산된 패브릭을 활용하며, 자원을 소유하지는 않지만 소프트웨어 및 서비스를 사용자와 자원 공유자에게 제공하여 자원의 공유를 촉진함

○ 고에너지 물리학, 나노과학, 구조생물학 등의 분야에서 사용되고 있음

- 고에너지 물리학: CMS, ATLAS
- 나노과학: nanoHUB
- 구조생물학: SBGrid



[그림 26] 개방형 과학 그리드에서 사용된 CPU 시간(출처: display.grid.iu.edu)

(5) 시민 과학

○ 시민과학(Citizen Science)은 Crowd science, Crowd-sourced science, civic science, networked science 등 다양한 이름으로 알려져 있으며, 과학자가 아닌 비전문가나 아마추어가 데이터를 수집하고 분석하는 과학 분야임

○ 시민과학은 소프트웨어의 발전으로 시민간의 협업이 가능해지고, 데이터의 공유가 일반화 되면서 활성화됨

○ 일반 대중들이 발견한 과학적 지식을 과학자가 검증하거나, 과학자가 확보해 놓은 데이터를 대중에게 공유하여 일반 시민이 과학적 발견에 도움을 주는 경우도 늘어나고 있음

- 간단한 게임(EyeWire, FoldIt, EteRNA)을 통해 새로운 과학적 지식을 발견하는 것을 돕기도 함

4. 맺음말

○ 본 보고서에서 조사한 과학 연구의 소프트웨어 활용 사례와 적용 기술에 대해서 표 6에서 요약함

○ 표 6에서 볼 수 있듯이 소프트웨어 기술 중 클라우드 기술이 과학 연구에서 활발히 사용되고 있으며, 데이터의 공유와 분석 측면에서 과학연구의 효율성과 효과성을 높이고 있음

- 클라우드 컴퓨팅, 크라우드 소싱 등 최근에 소개되는 신기술은 과학 연구에 즉시 활용되고 있음
- 이 외에도 분산 처리, 머신 러닝, 데이터베이스 분야의 최신 소프트웨어 기술이 과학연구에서 활용되고 있음
- 과학 연구에서 소프트웨어 신기술을 빠르게 적용하고 있는 것은 소프트웨어가 과학 연구의 필수 도구가 되었음을 보여줌

○ EteRNA, Foldit, Taverna 등의 사례에서 볼 수 있듯이 소프트웨어의 활용을 통해 새로운 연구 방법이 제시되기도 하며 상당한 성과를 이루고 있음

- EteRNA와 Foldit은 시민과학의 형태로 그리드 컴퓨팅 기술이 사용되었으며, Taverna에서는 데이터 흐름 관리 기술이 사용됨

○ 선진국에서는 정부 차원에서 과학계와 컴퓨터과학계의 공동 연구에 대한 비전을 제시하고 육성 및 지원 프로그램을 추진하는 등 적극적인 노력을 하고 있음

- 영국의 e-Science 프로그램, 유럽의 e-Infrastructure 연구, 미국의 사이버인프라 등을 통해 과학연구에 소프트웨어 활용을 지원하는 기반 환경 구축

○ 국내의 연구 환경도 과학계와 컴퓨터 과학계 간의 학제적 연구가 활성화될 수 있도록 정부의 지원과 산·학계의 연구 문화 개선이 필요함

<표 7> 과학연구의 소프트웨어 활용 사례 요약

사례	분야	소프트웨어 기술	시작 연도
CARMEN	신경과학	웹, 클라우드	2011
BIRN	생명정보학	데이터 그리드, 데이터웨어하우스, ETL, 데이터 통합, 이미지 프로세싱	2001
AstroGrid	천문학	데이터 그리드, 웹, 이미지 프로세싱	2001
입자물리학 클라우드	입자물리학	분산 파일 시스템, 고성능 분석 클러스터, 클라우드	2013
EteRNA	분자생물학	그리드 컴퓨팅, 머신러닝	2011
Folding@Home	구조생물학	분산 컴퓨팅, 그리드 컴퓨팅	2001
Foldit	구조생물학	머신러닝, 그리드 컴퓨팅, 게임화, 이미지 프로세싱	2010
WeNMR	구조생물학	분산 처리, 클러스터, 그리드	2010
환경과학	환경과학	클라우드	2010
CINET	네트워크과학	클라우드(HPC as a Service)	2012
Taverna	과학연구 관리	데이터 변환 및 통합 기술	2001
CLARIN	사회과학	데이터 카테고리 레지스트리(ISOCat), 데이터베이스 기술	2012
VRE SDM	금석학	이미지 프로세싱	2007

[참고자료]

국내문헌

1. 사이언스온. (2014.2.5.). 37,000명 온라인 게이머들이 RNA 염기 배열 규칙을 풀다.
2. 소프트웨어정책연구소. (2014.12.1). 과학기술의 SW활용 현황 분석 및 협력방안 연구. 경기: 미래창조과학부, 14-진흥-029.

해외문헌

1. Abdelhamid, S. E., Alo, R., Arifuzzaman, S. M., Beckman, P., Bhuiyan, M. H., Bisset, K., ... & Zhao, Z. (2012, October). Cinet: A cyberinfrastructure for network science, *2012 IEEE 8th International Conference on E-Science*. IEEE.
2. Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, 323(5919), 1297-1298.
3. Atkins, D. (2010). RCUK Review of e-Science: Building a UK Foundation for the Transformative Enhancement of Research and Innovation. Tech. Report, Eng. and Physical Sciences Research Council.
4. De La Flor, G., Jirotka, M., Luff, P., Pybus, J., & Kirkham, R. (2010). Transforming scholarly practice: Embedding technological interventions to support the collaborative analysis of ancient texts. *Computer Supported Cooperative Work (CSCW)*, 19(3-4), 309-334.
5. Qiao, Q., Bowman, G. R., & Huang, X. (2013). Dynamics of an intrinsically disordered protein reveal metastable

conformations that potentially seed aggregation. *Journal of the American Chemical Society*, 135(43), 16092-16101.

6. Guardian(2014.10.16), "Dark matter may have been detected - streaming from the sun's core"
7. Jorissen, K., Vila, F. D., & Rehr, J. J. (2012). A high performance scientific cloud computing environment for materials simulations. *Computer Physics Communications*, *183*(9), 1911-1919
8. Lee, J., Kladwang, W., Lee, M., Cantu, D., Azizyan, M., Kim, H., ... & Das, R. (2014). RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, 111(6), 2122-2127.
9. Li, J., Humphrey, M., Agarwal, D., Jackson, K., van Ingen, C., &Ryu, Y. (2010, April). escience in the cloud: A modis satellite data reprojection and reduction pipeline in the windows azure platform. *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS).* IEEE.
10. Missier, P., Soiland-Reyes, S., Owen, S., Tan, W., Nenadic, A., Dunlop, I., ... & Goble, C. (2010, January). Taverna, reloaded. *Scientific and Statistical Database Management* (pp. 471-481). Springer Berlin Heidelberg.
11. Olson, J. S., Ellisman, M., James, M., Grethe, J. S., & Puetz (2013), M. 12 The Biomedical Informatics Research Network. Scientific Collaboration on the Internet, 221.
12. Qiao, Q., Bowman, G. R., & Huang, X. (2013). Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation. *Journal of*

the American Chemical Society, 135(43), 16092-16101.

13. Watson, P., Hiden, H., & Woodman, S. (2010). e-Science Central for CARMEN: science as a service. *Concurrency and computation: Practice and Experience*, 22(17), 2369-2380.
14. Wu, W., Uram, T., & Papka, M. E. (2009, November). Web 2.0-based social informatics data grid. *Proceedings of the 5th Grid Computing Environments Workshop*. ACM.

주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.