Open Data Platform in Korea

Global Science experimental Data hub Center

September 6, 2018
Seo-Young Noh
Contents

1. Open Science: *Connected Science*
2. Data-driven R&D Era
3. Data Infrastructure in Korea
4. Linking Data Repositories: *Practical Implementation*
5. Summary: *Responses to Requests*
Open Science: *Connected Science*
Open Science...

- OECD Principles and Guidelines for Access to Research Data from Public Funding (2006-07)

- Initial discussion of Open Science at CSTP in 2011

- Many Open Science related activities on-going (PSI, open gov data, open educational resources, MOOCS...)

- OECD produced the first Open Science report, mainly focusing on Open Access, Open Collaboration and Open Data (2015)

- Several expert groups in GSF have been formed to build advisory policy for Open Science: Research Infrastructure, Data Infrastructure for Open Science

Open science is more than open access to publications or data; it includes many aspects and stages of research processes. [...] 

Open science is a broader concept that includes
- the interoperability of scientific infrastructure
- open and shared research methodologies (such as open applications and informatics code)
- and machine-friendly tools allowing, for example, text and data mining.

Source: POLICIES TO PROMOTE OPEN SCIENCE: EVIDENCE FROM OECD COUNTRIES, Giulia Ajmone Marsan
Key features of Open Science

Main goal of Open Science:
- Provides cost-effective access to digital research data from public funding
- Enhances utilizations of research data to scientific communities as well as societies including corporate sectors

Benefits
- Easy Research → Efficiency, Removing Redundancy → Solving Contemporary Problems
- Tackling Big Problems → Enabling Big Science → Solving Problems of Humankind
- New Value Creation → Enabling Convergence → Solving Unknown Problems

Ways and Means
- Open Data, Open Access and Open Collaboration through Information and Communication Technology
- “Open Access” and “Open Collaboration” look straightforward, but “Open Data” is not so simple, requiring deep understanding the features of data

Science cannot exist in a bubble...

should work with public communities...

The Future of Science
Features of data in Open Science

Open data is the key part of Open Science:

- **Transparency** in experimental methodology, observation, and collection of data
- **Public availability and reusability** of scientific data
- **Public accessibility** and transparency of scientific communication

Data should be valuable:

- **Reusable data** (ex: observation data, knowledge database)
- **Data requiring long time** for data accumulation (ex: pathology tracking data, climate change tracking data)
- **Data requiring big budget** for data acquisition (ex: large equipment-based experiments)
- **Data requiring huge computing power** for data generation (ex: simulation data)
Interoperability – Connected Science

Why open data is matter? ... many terminologies, properties

*Transparency, Public Availability, Public Access, Reusable, Redistribution, Universal Participation*

Interoperability

Interoperability makes it possible for diverse research groups to interoperate scientific data, intermixing different datasets, *leading to open a new way for unveiled values*

**Connected Scientific Data**

- Connected Science
- Value Creation
- Social Responsibility

**Connected Data**

- Connected Devices
- Value Creation
- Social Responsibility

**Business**
Linus Tovalds opened free operating system (Open S/W), later combined with GNU project, influencing great impact.

Sharing is good, and with digital technology, sharing is easy

- Richard Stallman

Software to be distributed in a manner such that its users receive the freedoms to use, study, distribute and modify that software

GNU Project since 1984

GNU Public License
Data-driven R&D Era
Research Paradigm Shift

Data & Infrastructure are Key in Scientific Discovery

Describing natural phenomena based on **Observation**

**Modeling and Theory**

**Computing Simulation**

**Data Analysis** of tremendous data produced from large experimental facilities

---

**Research Paradigm Shift to Data Intensive Scientific Discovery**

1st Generation: Observation

Galileo's telescope

2nd Generation: Theory

Higgs Theory

3rd Generation: Simulation

Black Hole Simulation

4th Generation: Data

CERN’s CMS and ATLAS experiments → Higgs discovery

**More chance to do research with advanced equipment, higher chance to get Nobel prize**

87% of Nobel prizes have been given to researchers who produced outstanding scientific discoveries using advanced experimental equipment since 1914.

Source: The Fourth Paradigm
CERN noticed a signal like a new particle in CMS & ATLAS experiment in December 2015.

The 750 GeV diphoton excess in particle physics was an anomaly in data collected at the Large Hadron Collider (LHC) in 2015, which could have been an indication of a new particle.

However, the anomaly was absent in data collected in 2016, suggesting that the diphoton excess was a statistical fluctuation.

In the interval between the December 2015 and August 2016 results, the anomaly generated considerable interest in the scientific community, including about 500 theoretical studies.

We are in data-driven science era!!!
Our trust is in data
Data Infrastructure...that is what we need

Science relies on data, requiring infrastructure for data. Data is getting more important and growing fast.

Data Infrastructure is the one of key factors for successful science and tackling big problems of humankind.

KISTI has been in preparation for big data research era. Our mission is gradually expanding to national role for data intensive research.
Data Infrastructure:

*KISTI, Korea*
KISTI...providing powerful ICT infra. service

Unique organization in Korea, our mission is to support various R&D activities through ICT infra. service

New Supercomputing Building

25.7PFlops and 20PB ranked 11th in Top 500

Fast and secure network, providing domestic researchers with a constraint free collaborative research environment through KREONET (locally) and GLORIAD (globally)
Global Science Data Hub

Large-scale Scientific Data:
20Km CD stack with data produced per year in CERN

Collaboration with global laboratories

Data from large and high-valued research equipment

(Global)
Asia representative Data Hub

(Domestic)
Scientific data management and analysis platform service
Data Repository Infrastructure

Growing every year by ~1,200 cores and ~2PB

Best equipment procured every year

Centralized data repository model, but distributed model in near future

Interlocking with existing systems done by 100% KISTI experts
Data Repository Services (1)

1. ALICE (A Large Ion Collider Experiment)
   - To generate similar conditions that have existed a fraction of the second after the Big Bang
   - 1,655 scientists from 159 institutes, 41 countries

2. CMS (Compact Muon Solenoid)
   - To investigate a wide range of physics, including the search for the Higgs boson, new physics
   - 3,000 scientists from 172 institutes, 40 countries

3. Belle/BelleII (Japan KEK)
   - To investigate CP-violation effects and new physics
   - 428 scientists from 67 institutes, 20 countries

4. LIGO (Laser Interferometer Gravitational Wave Observatory)
   - To detect cosmic gravitational waves and to develop gravitational-wave observations
   - 1,000 scientists from 60 institutes, 16 countries

5. RENO (Reactor Experiment for Neutrino Oscillation)
   - To measure a limit on the neutrino mixing matrix parameter (Yeonggwang Nuclear Power Plant)

6. Genome Research
   - Genome data analysis for the next generation of personalized treatment: 50 Korean researchers
Data Repository Services (2)

- **Structural Biology**
  - Data management and analysis service for Transmission Electron Microscope-based research

- **Volcanic Disaster Prevention**
  - Data repository service for volcano simulation
  - Helping to build a policy for volcanic disaster response scenarios

- **CMS Tier-2 Service**
  - Unified WLCG service in Korea
  - A new Tier-2 center targeting 2018 service

- **KAGRA (Kamioka Gravitational Wave Detector, Japan, 2017)**
  - A project of the gravitational wave study group
  - Data transmission, repository and data analysis environment in conjunction with LIGO data analysis environment

- **PAL (Pohang Accelerator Laboratory, Pilot Project, 2018)**
  - Data repository and analysis service for 4th generation accelerator
  - Trial service for data convergence produced from heterogeneous large equipment for structural biology (Government Supported Funding)

- **Brain Research (Under discussion)**
  - Platform service for nationwide management, repository, transmission, and analysis of brain data
  - Extending knowledge of HEP data analysis platform service to brain data

- **RISP (Rare Isotope Science Project, RAON, under discussion)**
  - Exploration of the origins of chemical elements, structural study of new isotopes and applied medical research.
  - Discussion with Government for data center role of RAON
Korean Model for Data Intensive Research

Focusing on a centralized model for data repository at the beginning by fully utilizing ICT specialized institute like KISTI/GSDC

- Large scale research group
- Large scale research facility
- Dedicated data center

VS.

- Small scale research group
- Small or medium scale facility
- Not easy to have a dedicated data center (in size and experts)
Unified Data Analysis Environment (Centralized Model)

Advantages
1. **Pluggable Science** → Supports in unified way for various groups and equipment
2. **Data Infra. Sharing** → Reuse and full utilization of infra. saving tax-payer’s money
3. **Simple R&D Process** → Fast results from data acquisition to data analysis
Linking Data Repositories: *Practical Implementation*
A Task Force has been setup in Ministry of Science and ICT, focusing on ...

1. **Fostering** data-driven R&D communities
2. **National level** regulation for data management
3. **Scientific program development** for open data

Main activities ...

1. **Analyzing best practices** including European Open Science Cloud in EU, Big Data Hub Program in U.S
2. **Defining priority and categorization** of data intensive research fields and **new program development**
3. **Developing R&D Open Data Platform**, helping seamless data sharing, data accessing, data analysis, data linking across disciplines

**Accelerating R&D productivities through ICT-based R&D e-Transformation**
Big Picture: what we expect ...

**Expectation**
- Social Benefits
  - Public participation
- Improving R&D Productivity
  - Easy R&D
  - Fast R&D
- New Value Creation
- R&D Convergence
  - Connected R&D
- R&D Reliability
  - Reproducibility
  - Error probing

**Driving Wheels**
- AI
- Big Data
- ICT
  (Technology)
- R&D e-Transformation
- ICT utilization

**Base**
- S/W Platform
  - Unrestricted data access
- Infrastructure
  - Data acquisition
  - Digitalization
- Fostering Community
  - Main players
- Data Management
  - Data tracking
- Systemization
  - Regulations
  - Programs

**Culture:**
- Data Open & Sharing
  (Policy)
- Collaboration
- Overcoming barriers

**R&D Data Open & Sharing**
- Solving unknown problems
Stepwise Implementation: from part to all

- Covering entire data management, linking data repositories, providing easy and seamless access across various R&D
- Enabling **Open Science Service** by stepwise implementation

### 4 Pilot Projects (launched in 2018)
1. Genome Data
2. Material Data
3. Data Convergence from Large Facilities (KISTI-GSDC)
4. AI/Big Data
Connected Scientific Data – External View

Public User

Search, Access, Use

Metadata

A.I.

Data Convergence from Large Facilities

Data Flow
Request Flow

Cyber Space for Deep Learning

Unified Data Management Platform (Data Hub)

Search and Access to experiment and research data

Tracing scientific data usages

Assigning unique ID to scientific data

R&D Project Management

Tracing connections between outcomes and scientific data

Research Community

Domain Center

Domain Center

Domain Center

Domain Center

HVEM

RENO

KSTAR

Astronomical Telescope

RAON (Accelerator)

Geology - Earthquakes

Weather - Climate

Infrastructure

Public User

Research Community

Domain Center

Domain Center

Domain Center

Domain Center

HVEM

RENO

KSTAR

Astronomical Telescope

RAON (Accelerator)

Geology - Earthquakes

Weather - Climate

Infrastructure
Connected Scientific Data – Internal View

GRID-based Authentication Authorization

Access-Analysis-Sharing-Utilization

Repository for Clinical & Genomic Big Data

Big I/O

Fast I/O

Super Computer

Visualization

Data Analysis Cluster

Analysis

Security

Education

Adaption

Improvement

Medical Science Research

Third Party Analysis Service

Cyber Education Service

Interoperation between CODA and KISTI-GSDC

Interoperation between analysis result and knowledge service

Interoperation between technology development and medical related business

CODA: Clinical & Omics Data Archive

Deposit/Distribution

Publication

Project Management

Public Service

Global Service

Authentication/Authorization

Access-Analysis-Sharing-Utilization

CODA Service

Archiving

Search

Transform

Knowledge Service

DOI Repository

Data Curation & Backup

Data for Service

Data for Education

Analysis

Sharing

Utilization

Global Community

Hospitals

Institutes

Domestic Group

Researher

Seo-Young Noh
Pilot Project – Data Convergence (KISTI-GSDC)

- **Centralized data repository** is used for data collection and access
- Central data repository will be connected with Open Data Platform, which is also being developed by KISTI.
- Expect each site has **own repositories interoperable with the Open Data Platform** in future.

**Large-scale Facility**

- Experiment
- Data Convergence
- Integrative Structural Biology
- 4 Data Convergence Programs

(KISTI, KBSI, Structural Biology Community)

**Data Challenge School** (Data Convergence)

(KISTI, KBSI, Structural Biology Community)

*Transmission Electron Microscope*
Summary: Response to Requests
Summary

- Highlighted openness of access, collaborations and data
- Believed to give benefits to scientific community: Connected Scientific Data → Connected Science

For Data-driven R&D:
- Research paradigm is being shifted to data-driven scientific discovery
- Data and infrastructure are the key in scientific discovery

For Open Data Platform:
- Importance of Open Data is recognized in government level and the TF has launched four pilot projects this year (2018)
- Data repositories, setup by pilot projects, will be connected to the platform which makes them interoperable as a final goal.

Implementation of Open Data Platform is not an easy task...
The pilot projects are not big scale...

But, such a trial is a Big Step moving toward to making Open Science a reality in Korea
Responses to Requests

Identify your initiative as a stakeholder ...

- Policymakers (Government), researchers, ICT specialists are stakeholders.
- Policymakers wants to make data, generated from public funding, traceable through Open Data Platform.
- Researchers wants to use well managed ICT infrastructure for data analysis and data sharing, in order to accelerate R&D process.
- Role of ICT specialized institute like KISTI, expected from government, is well aligned with its mission – promoting Science using ICT technologies.

Analyze how your initiative addresses a multiplicity of different cultural contexts ...

- Respect differences in R&D domains and encourage scientific communities to make own standard data management plan agreed in community members.
- Structural biology community, for example, is developing a guideline including data naming convention, sharing, accessing and management policy.

Assess the proactive role of your initiative with respect to interoperability ...

- Rather than making data repositories interoperable directly(1-to-1), indirect interoperability is being considered through the Open Data Platform.
- Such an approach helps to reduce the burden of standization for all R&D. Interesting data can be accessed through OpenAPI provided by the platform. Data convergence and all R&D activities with data are left to scientist.
Thank you.