

Update report on KOBIC: Korea BioData Station (K-BDS)

Seungwoo Hwang

Korea Bioinformation Center

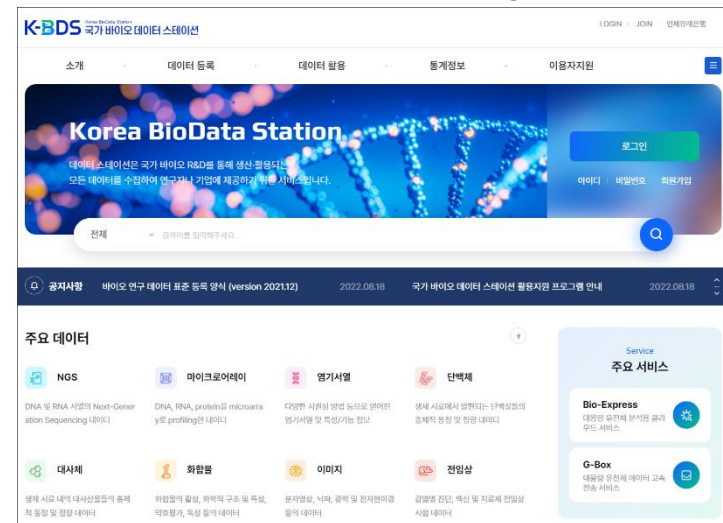


Korea Research Institute of Bioscience and Biotechnology



Introduction to Korea BioData Station (K-BDS) Project

- Goal
 - To collect **all types** of biological data (i.e., not just omics) that were produced from **all government-funded research projects**
 - To act as a data repository for both **national funders** and **international publishers**
- Current status
 - **Pilot version** opened in 2021
 - **Official version** will open next week



Why KOBIC collects data from Korea?

- **Scientific reason:** To enable data-driven research
- **Policy reason:** KOBIC's duty given from the government
 - Act on the Acquisition, Management, and Utilization of Biological Research Resources
 -  生命研究資源の確保管理及び活用に関する法律
 -  生物研究资源获取、管理和利用法
 - National Research and Development Innovation Act
 -  国家研究開発革新法
 -  国家研发创新法
 - KOBIC must collect biodata, and Korean researchers are required to submit their biodata to national repositories.

Why KOBIC developed K-BDS, a new data repository





■ Policy reason

- Korean funding agencies are requiring researchers to submit their biodata to **national** repositories.
- **Sensitive human data** should be handled in such a way that complies with relevant **Korean laws and policies**.

■ Technical and other reason

- Need a repository that can accept submission of **all types of data**, not just omics data.
- A variety of **nationwide large-scale projects** are underway, whose data must be deposited.
- **Network bandwidth**
 - Korea ↔ NCBI: 123 Mbps upload / 107 Mbps download
 - Korea ↔ K-BDS: 949 Mbps upload / 951 Mbps download

Data Management Policy (DMP) in Korea

- Major driving force for data submission to K-BDS is Data Management Policy in Korea.
- Laid out in the **Standard Guideline on Collection, Management, and Utilization of Biological Research Data**
 -  (バイオ 研究 データ 確保, 管理, 活用 標準 ガイドライン)
 -  (生命研究资源的采购、管理和使用标准指南)
 - Need to submit data in each year as well as upon completion of grant
 - Can ask for extension of the data release for up to one year and a half to funding agencies
 - Currently applied to only a limited type of projects
 - Large-scale data generation projects
 - Bio-medicine technology development project
 -  (バイオ医療技術開発事業)
 -  (生物医药技术开发项目)
 - Will be applied to a wider range of projects under law amending

Development of standard data submission forms

- Goal
 - To collect **all types** of biological data with **comprehensive context information** (i.e., metadata)
- Methodology
 - If de-facto international standards exist, benchmark them as much as possible
 - **Community-driven:** developed by several committees of ~150 researchers

- Based on **INSDC BioProject and BioSample forms**, with **some modifications**
 - Removing some fields
 - Adding some fields
 - Modifying some fields

Data submission forms: Major data







- Submission forms of **major data types** were also developed based on other existing forms

Data types	Based on
NGS read	NCBI SRA
Microarray	NCBI GEO
Nucleotide	NCBI GenBank
Proteomics	EBI PRIDE
Metabolomics	EBI MetaboLights
Chemical	NCBI PubChem BioAssay
Biolmaging	DICOM (Digital Imaging and Communications in Medicine) BIDS (Brain Imaging Data Structure) IDR (Image Data Resource) EBI BioImage EBI EMPIAR (Electron Microscopy Public Image Archive) OME (Open Microscopy Environment)

- Some modifications were also done

Data submission forms: other unstructured data

- There are many biological “data” other than these ‘omics, chemicals, or bioimages.
- Many of them can be collectively referred to as;
 - Generalist data
 - Small-science data
 - Orphan data
- Benchmarked six generalist repositories

Repository	Operated by
BioStudies  BioStudies.	EBI
Dryad  DRYAD	Dryad (US)
Figshare  figshare	Digital Science (UK)
Zenodo 	CERN (EU)
Mendeley Data  MENDELEY DATA	Mendeley (UK)
OSF(Open Science Framework)  OSF	Center for Open Science (US)

Data submission forms: other unstructured data (continued)

- The fields in the “**other**” data submission form

- 1) Keywords
- 2) Additional notes
- 3) Files (in many cases, Excel files)
- 4) Description of files
- 5) Links to external databases
- 6) Description of links

Data submission forms: Interim summary

Category	Submission forms
Common	BioProject / BioSample
Genomics	NGS read / Microarray / Nucleotide
Proteomics	Proteomics
Metabolomics	Metabolomics
Chemical	Chemical bioassay / Chemical structure / Chemical drug efficacy test / Chemical profiling
Biolmaging	MRI / PET / CT/ Ultrasonic / Optical microscopy / Electron microscopy / Cryo-EM / X-ray
Other	Generalist data

Majority **w.r.t.**
submitted file
size

Majority **w.r.t.**
submission
number

Data submission forms: Minor data

- Took one step further, and developed additional forms for **miscellaneous data**

Category	Submission forms
Medical device	In vitro diagnostics / Biomaterial medical device / Biosignal measurement device / ...
Drug development	Toxicity test / Drug repositioning / Biomarker / ...
Agriculture	Chemical pesticide / Biological pesticide / ...
Food	Food sample / Food content / Food process / Food functionality / ...
Environment	Insect distribution / ...

- **Five virtual centers** will be responsible for the five main data categories, including
 - QC of submitted data
 - Q&A service for submitters
 - Further development of submission forms
 - Further development of corresponding web page in K-BDS

Category	Data type	Primary Institute
Genomics	NGS read / Microarray / Nucleotide	KOBIC
Proteomics	Proteomics	Korea University
Metabolomics	Metabolomics	Seoul National University
Chemical	Chemical bioassay / Chemical structure / Chemical drug efficacy test / Chemical profiling	Korea Research Institute of Chemical Technology
Biolmaging	MRI / PET / CT/ Ultrasonic / Optical microscopy / Electron microscopy / Cryo-EM / X-ray	Ewha Womans University

Plans for international collaboration of each center

Category	Primary Institute	International organizations planned to collaborate
Genomics	KOBIC	ABC // DDBJ // NGDC // NCBI // EBI
Proteomics	Korea University	HUPO // ICPC
Metabolomics	Seoul National University	GNPS Massive // The Metabolomics Society // MoNA // mQACC
Chemical	Korea Research Institute of Chemical Technology	EBI ChEMBL // ACS // NCATS // SLAS // ISSX
Biolmaging	Ewha Womans University	EBI (BioImage, EMPIAR, EMDB) // RCSB PDB // LONI // Allen Institute // CCF // OpenNeuro // WMIS // NIC-HMS // NIH // Stanford

Human Data

- KOBIC established **Human Data Bank** from the Ministry of Health and Welfare   (保健福祉部)
- Made the **Provision for the Operation and Management of Human Data Bank**  (人体由来物データ銀行運営管理規定)
 (人类衍生品数据库运行管理规定)
- Established the **Data Access Committee**

■ K-BDS

- **Submitted to:** *Genomics & Informatics*
- **Title:** Introduction of the Korea BioData Station (K-BDS) for sharing biological data

■ KoNA (Korean Nucleotide Archive, a subsystem of K-BDS)

- **Submitted to:** *Genomics, Proteomics & Bionformatics*
- **Title:** Korean Nucleotide Archive as a new data repository for nucleotide sequence data

Summary

- K-BDS (Korea BioData Station) opens in December 2022
- K-BDS accepts submission of all types of biological data
- Major driving force for K-BDS submission is Data Management Policy in Korea
- K-BDS is equipped with a variety of data submission forms
- K-BDS is ready for dealing with controlled-access human data
- KOBIC is eager to bring our collaboration to the next level