

[About the Logo]

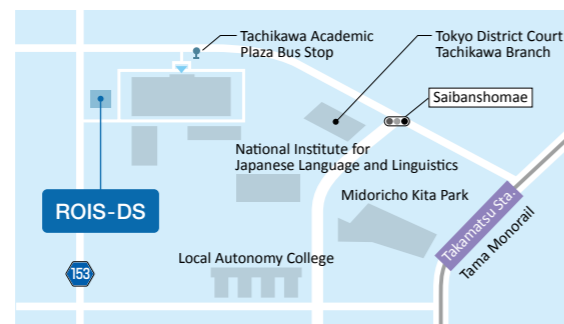


The letters "DS," the initials for "data science," are angled in a way that represents the tilt of the Earth on its axis, and the perimeter represents the Center's four research institutes.

The overall image represents the development of an organization that does not fit into set parameters but rather changes its shape as it develops, while the familiar silhouette suggests an organization contributes to society.

Joint Support-Center for Data Science Research (ROIS-DS)

10-3 Midori-cho, Tachikawa-shi, Tokyo 190-0014
URL: <https://ds.rois.ac.jp/en/>



For more details, visit our website

Inter-University Research Institute Corporation
Research Organization of Information and Systems

Joint Support-Center for Data Science Research

ROIS-DS

Data for All

Brochure 202412

Contact

Data Science Promotion Section Tel: +81-42-512-9254 email: [ds_suishin\(at\) rois.ac.jp](mailto:ds_suishin@ois.ac.jp)

2024-2025

Joint Support-Center for
Data Science Research
(ROIS-DS)

Director

Masaru Kitsuregawa



Data for All ~ Pioneering a New Future with Data ~

We are surrounded by vast amounts of data. Data science as a toolkit for analyzing and interpreting data is becoming an increasingly important means for addressing the complex challenges facing modern society and for enabling more effective and sustainable decision-making.

Since the launch of ChatGPT in 2022, the technologies underlying generative artificial intelligence (AI) have continued to advance at an accelerating pace, generating ever more diverse and bountiful data. This has been accompanied by the emergence of a range of new problems, including hallucinations, ethical issues, use of inappropriate expressions, and rights violations. ROIS will continue its mission to pioneer a new future together with you in order to address social issues based on our philosophy of "Data for All."

First, as an organization that develops novel services in anticipation of future demand, ROIS promotes innovative and forward-thinking initiatives through flexible management. In addition to continuing to provide data sharing and data analysis support to academia, ROIS aims to become an incubator organization focused on data science that spans diverse academic fields.

In November 2023, we launched the Center for Juris-Informatics. The Center is working to develop juris-informatics, a new academic field based on two disciplines, law supported by AI and the law governing AI, and is working to become a global center for juris-informatics by establishing databases related to juris-informatics and investigating the usage of these databases.

In April 2024, we launched the Center for Research and Development on Data Lake, the mission of which is to systematically collect a wide variety of high-quality, large-scale data and to establish a framework for providing data in a safe, secure, and sustainable manner. In June 2024, the Center was selected by the Cabinet Office's Cross-ministerial Strategic Innovation Promotion Program to develop a medical data-management system to promote the utilization of medical data and medical LLM/LMMs.

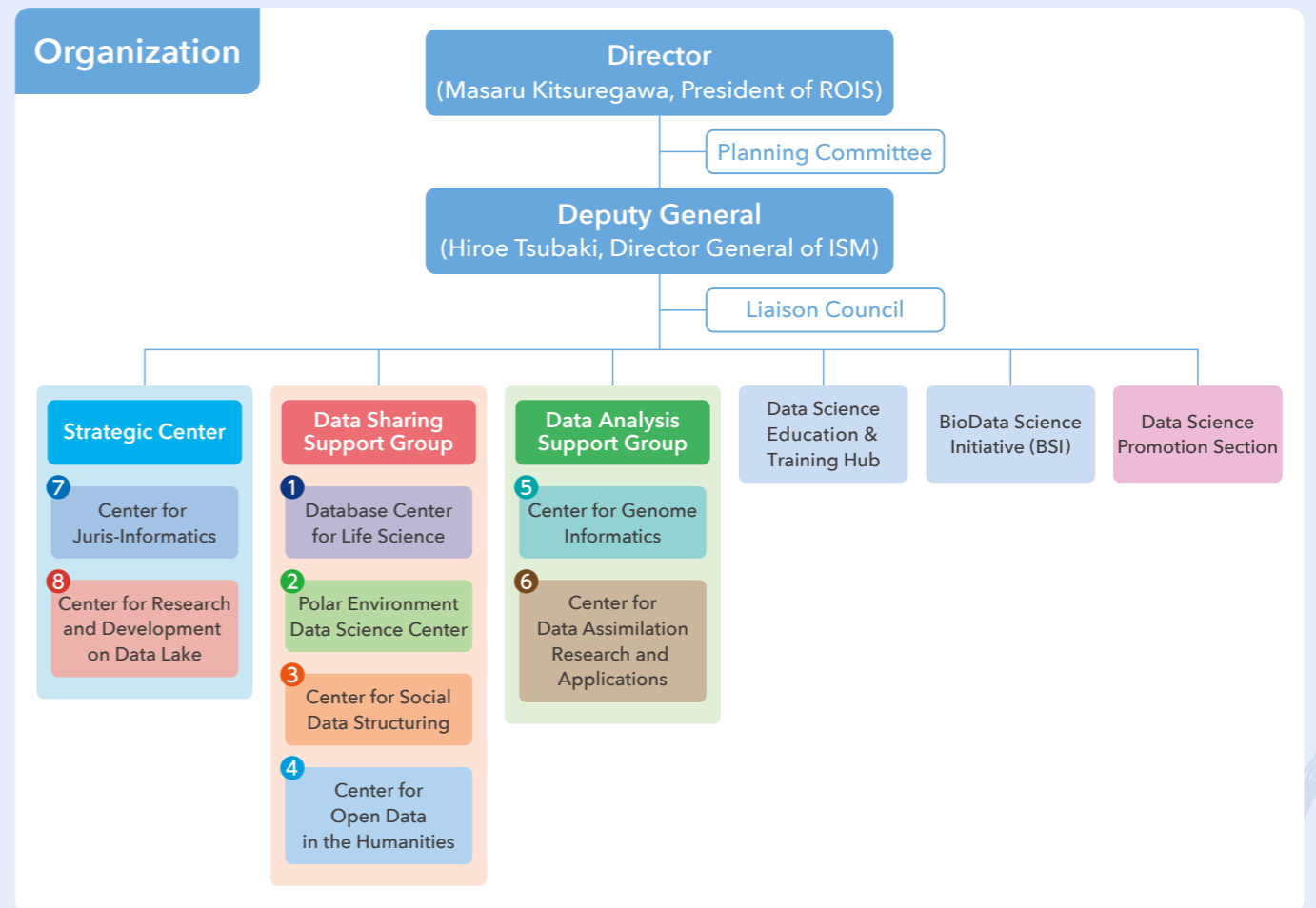
The activities of such centers enable us to directly assimilate feedback from the front lines of research and to promote the exploration of new scientific fields that directly address societal needs, further contributing to our goal of serving as a critical cornerstone for innovative research through the rapid and efficient utilization of resources.

Our goal is to maximize the potential of our institution to support both domestic and international research based on the following three pillars: creating unprecedented "new services," fostering "human resources" capable of provide such services, and establishing an "environment for collaborative research" centered on data science.

We endeavor to work even harder together with all of you to create an environment where we can learn and create together, and to successfully carry out and complete a wide range of innovative projects. We look forward to your continued support and guidance.

Contents

▶ Message from the New Director	01	▶ Data Analysis Support	
▶ Organizational Structure	02	Center for Genome Informatics (CGI)	09
▶ Outline / About the Strategic Center	03	Center for Data Assimilation Research and Applications (CARA)	10
▶ Hub Activities	04	▶ Strategic Center	
▶ Data Sharing Support		Center for Juris-Informatics (CJI)	11
Database Center for Life Science (DBCLS)	05	Center for Research and Development on Data Lake (DLRD)	12
Polar Environment Data Science Center (PEDSC)	06	▶ The Collaboration Program	13
Center for Social Data Structuring (CSDS)	07	▶ DS Promotion and Others	14
Center for Open Data in the Humanities (CODH)	08		



1 Database Center for Life Science (DBCLS) p.05

Promoting open science in the life science field and R&D for life science database integration

3 Center for Social Data Structuring (CSDS) p.07

Maintaining databases on social survey data, official statistics microdata, and social big data for university researchers as well as creating communities for data usage to promote empirical research for solutions to various social challenges, including the environment, public security, and the economy

5 Center for Genome Informatics (CGI) p.09

Supporting data analysis to obtain biologically significant information from a large amount of genome and transcriptome data, making full use of cutting-edge bioinformatics technology

7 Center for Juris-Informatics (CJI) p.11

Based on two disciplines, law supported by AI and the law governing AI, the Center for Juris-Informatics is working to develop a new academic field called "juris-informatics" and to create a global research center on juris-informatics by establishing databases related to juris-informatics and investigating the usage of these databases

2 Polar Environment Data Science Center (PEDSC) p.06

Promoting resource-sharing to provide valuable data as well as analytical support for data on changes in the polar environment and Earth system over a long-time axis, from the past to the present

4 Center for Open Data in the Humanities (CODH) p.08

Creating a new academic field for humanities based on data science (digital humanities), as well as forming and enhancing research hubs beyond organizational boundaries by promoting data-centric openness

6 Center for Data Assimilation Research and Applications (CARA) p.10

Solving problems in various scientific fields and industries by applying data assimilation technique

8 Center for Research and Development on Data Lake (DLRD) p.12

The center is working to establish a framework for systematic collection and provision of high-quality, diverse, large-scale data that is safe, secure, and sustainable, with the aim of accelerating AI research, especially, generative AI models of a wide range of fields, including medicine and big science

Support Project (Data Sharing Support)

- Life science data sharing support project
- Polar environmental science data sharing support project
- Human and social science data sharing support project
- Humanities open data sharing support project

Human Resource Development (Training of Data Scientists)

- Training Program for Advanced Research Educators in Data Science

Support Project (Data Analysis Support)

- Genomic data analysis support project
- Data fusion computing support project

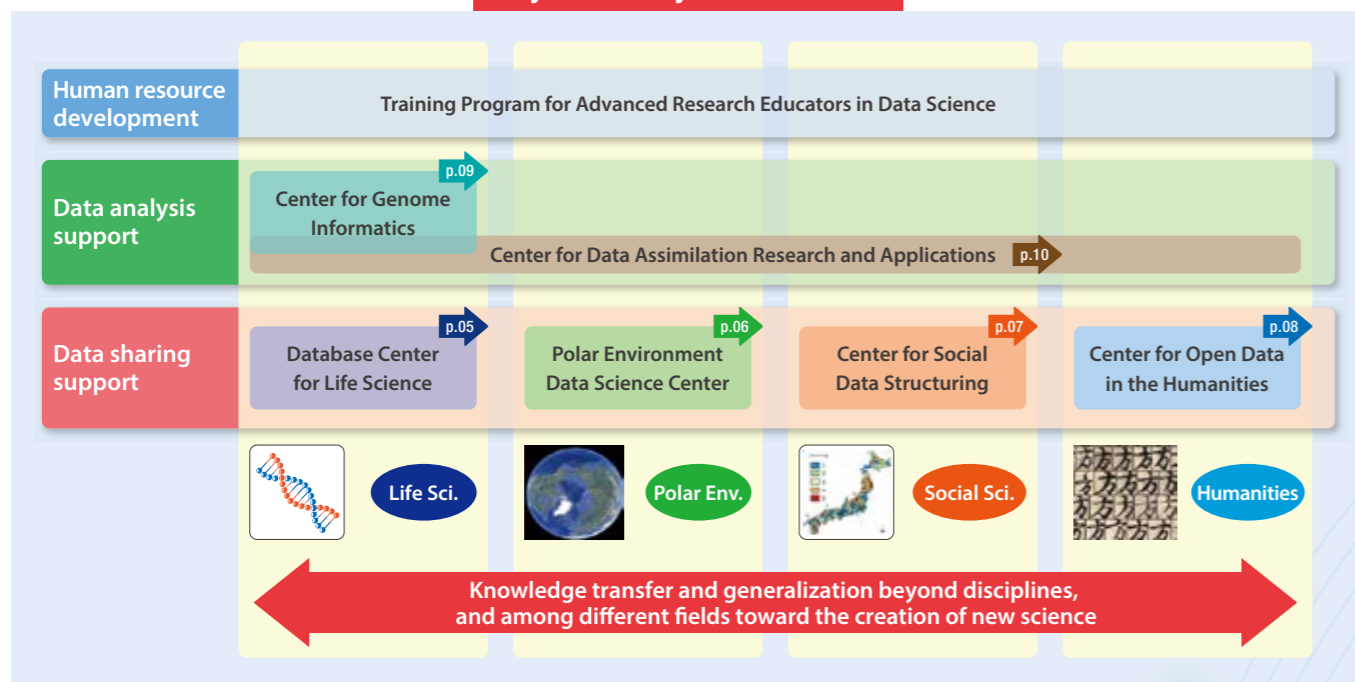
The Collaboration Program (ROIS-DS-JOINT)

- Joint research program
- Joint research meeting

See p.13 for details

All researchers of universities, etc., who wish to conduct data science research through collaboration

Various forms of support for joint use & joint research



About Strategic Center

Strategic Center is an innovative entity within ROIS-DS that was established in 2024 with the mission of creating new services and technologies that contribute broadly to society through advanced and multifaceted research activities aimed at addressing a wide variety of social issues and national strategic targets. The center aims to develop an agile and flexible high-quality research environment through the utilization of external funds. It will spearhead the creation of new services to be provided by DS and promote research that responds rapidly and with precision to the diverse demands of society.

Data Science Education and Training Hub

Training young researchers from universities and other institutions nationwide to become advanced research educators in data science

Recent years have seen a rapid increase in the number of universities establishing data science departments and faculties. One serious problem that has emerged in this context is a shortage of faculty members with expertise in statistics, which is the foundation of data science. It was to address this issue that the Consortium for Training Experts in Statistical Sciences (core institution: The Institute of Statistical Mathematics) was established in 2021 as a collaborative initiative involving the Joint Support-Center for Data Science Research (ROIS-DS) and universities across Japan. The consortium provides training to young researchers in various academic fields to nurture university faculty who are able to give lectures targeting master's degree students and to provide guidance in the use of statistics in research.

The Data Science Education and Training Hub provides young university faculty members who have been educated at the consortium with special training and necessary support in statistical science education and research mentorship, with the goal of fostering educators capable of providing high-level statistical education and research mentorship and conducting even more advanced research using statistics. In addition, faculty members of the hub serve as mentors in the Human Resource Development Program for Data Science Researchers administered by the Graduate University for Advanced Studies (SOKENDAI) to train young researchers at SOKENDAI's parent institutes to become data science researchers.



Statistics Professors Training Program Interim Report Meeting (June 14, 2024)

BioData Science Initiative (BSI)

Promoting the "Systematization of Knowledge" in the life sciences and creating a foundation for data science

The BioData Science Initiative (BSI) was launched in April 2022 as a cross-disciplinary research hub that is jointly administered by ROIS-DS, the National Institute of Genetics, and other ROIS member institutes. Biodata refers to both quantitative and qualitative information about the characteristics, functions, and states of organisms, as well as the structure and dynamics of biological populations, interactions among organisms, and interactions between organisms and their environments. BSI is tasked with systematizing biological knowledge by standardizing the wide range of available biodata, including a large amount of genomic information on various species obtained under diverse experimental conditions as well as information published in academic articles. As the core institute in Japan and Asia for registering, accumulating, organizing, standardizing, and providing biodata in an integrated manner, BSI aims to establish a research infrastructure that contributes to the advancement of biodata science and provides biodata to a wide audience through domestic and international collaborative networks. BSI also serves as a data registration hub for field-specific repositories and also promotes open science through its function as a centralized contact point for researchers. BSI, which is home to the greatest concentration of biodata scientists in Japan, promotes the integration of AI and the life sciences via the human resources assembled there and cultivates biodata scientists.

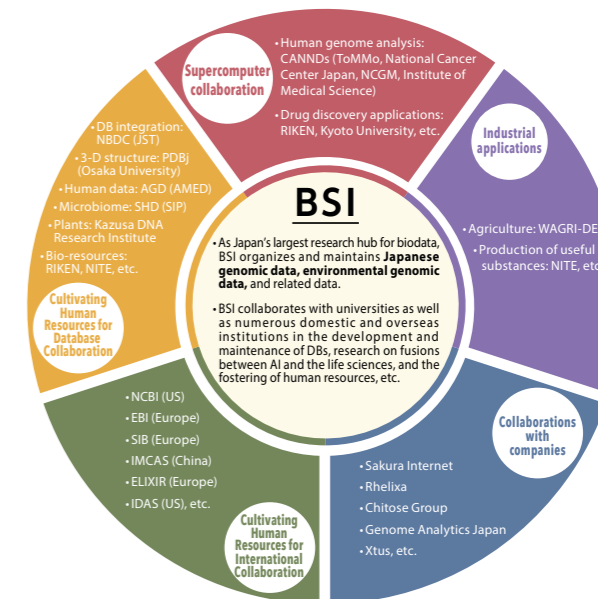


Figure: BSI and BSI's external network

Database Center for Life Science (DBCLS)

Promoting Open Science in the Life Science Field through Database Integration

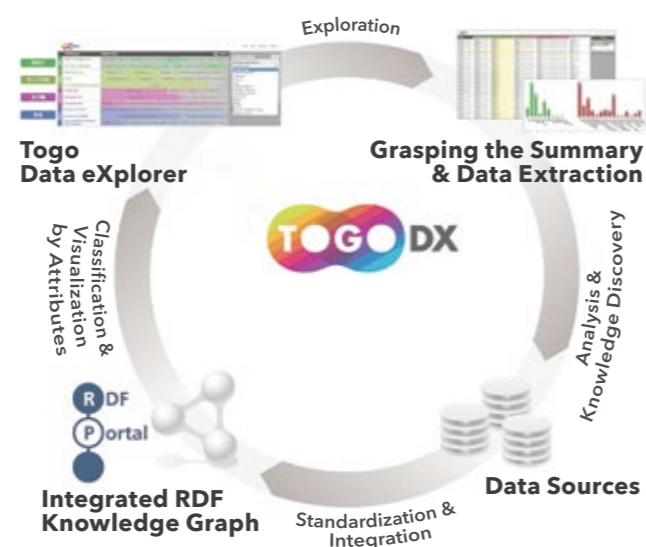
To promote open science in the life science field, DBCLS conducts research and development on database integration so that universities, research institutions, and other organizations throughout the country can utilize and apply, in a centralized manner, the diverse and rapidly growing number of databases they operate. We focus on creating knowledge graphs that standardize the terms used to describe data and their classification systems (ontologies), developing technical tools that extract effectively necessary information from these created knowledge graphs, using efficient literature information management, and so on.



Yuji Kohara, Director of DBCLS

Creating an Environment for Integrative Utilization of Databases

To construct a knowledge graph infrastructure that enables integrative utilization of life science databases, DBCLS assists with the conversion of databases into the Resource Description Framework (RDF) format, aggregates these databases, and provides access through the RDF Portal. In addition, DBCLS is developing various applications to utilize the knowledge graph infrastructure. For example, TogoDX/Human is a one-stop application for exploring information on humans that can be used to perform integrative analysis of results obtained from targeted searches. The application can also provide useful information for interpreting and discussing experimental results. DBCLS plans to expand TogoDX to other species while continuing to enhance data on humans.



→ Data analysis platform powered by TogoDX, a framework for exploring the knowledge graph which integrates various databases (applied to human data)

TogoDX/human [<https://togodx.dbcls.jp/human/>]



→ Group photo of participants at the international BioHackathon, held at Tsuchiyu Onsen, Fukushima Prefecture, August 2024

Collaborating with Domestic and Overseas Research Organizations to Realize Database Integration

For over 10 years, DBCLS has hosted various hackathons aimed at deepening the external collaborations that are critical for establishing and utilizing the knowledge graph infrastructure. This includes the international BioHackathon, wherein collaborative development work is carried out intensively in a week-long camp format. Such efforts have resulted in the establishment of rules, mechanisms, and tools for interoperability of data, which have been adopted by both domestic and overseas organizations and have advanced efforts toward international standardization. From FY2022 to FY2024, DBCLS has taken over the operation of the RDF Portal, TogoVar (the Integrated Database of Japanese Genomic Variation), and NBDC Human Database, which were previously operated by the NBDC Program Office of the Japan Science and Technology Agency. DBCLS also collaborates with institutes that analyze Japanese genomes and diseases as well as those that investigate environmental microorganisms.



Location : Univ. of Tokyo Kashiwanoha-campus Station Satellite 6F
178-4-4, Wakashiba, Kashiwa-shi, Chiba 277-0871 JAPAN
URL: <https://dbcls.rois.ac.jp/index-en.html>



Polar Environment Data Science Center (PEDSC)

To contribute to global environmental research, we promote data science and the publication and sharing of various scientific data obtained from the Arctic and Antarctic Regions

The Center aims to promote the dissemination and shared use of a wide variety of scientific data obtained through observation and research activities in both the North and South Polar Regions and to generate greater research output by providing data science support to universities and other external communities. On the international stage, in addition to serving as the central hub for data activities in the polar science and Japan's national data center for polar observation and research, PEDSC seeks to contribute to global environmental research by creating new polar science based on data.

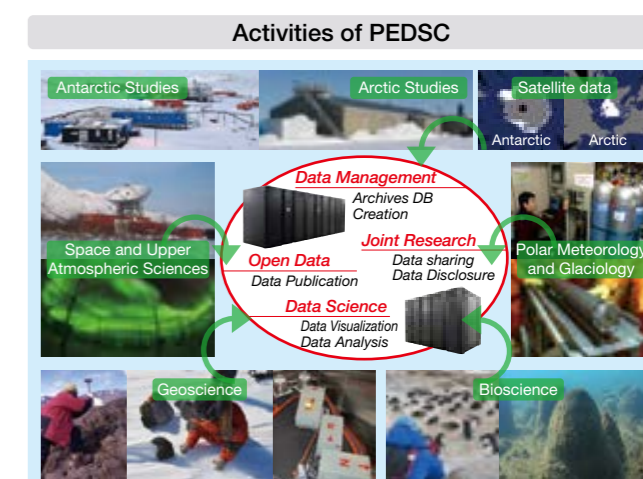


Akira Kadokura, Director of PEDSC

Data Handled by PEDSC

The data handled by PEDSC encompasses all fields of polar science. A wide range of observations and research activities focusing on various aspects ranging from the upper atmosphere to oceans, snow and ice, geosciences, and biology are being conducted in both polar regions. These efforts have resulted in the collection of diverse data, including digital data recorded on various media and data from samples that have been collected and stored.

After collection, these data undergo various forms of processing and are subjected to analyses to yield physically meaningful data, which can then be used to produce scientific results. To generate reliable scientific results, the data must be securely stored and protected from loss, degradation, and tampering. Additionally, the reliability of the data must be assured by making the data available to everyone and ensuring that the results are reproducible.

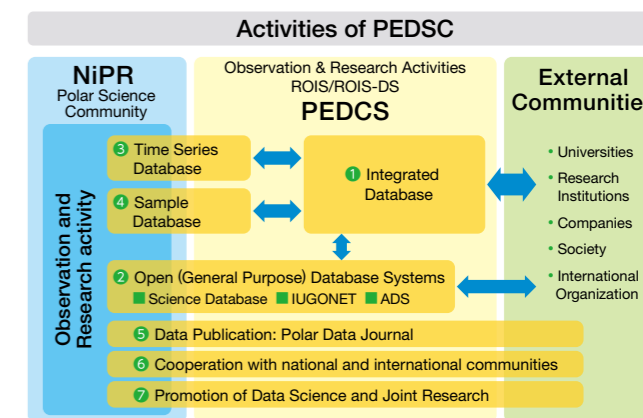


→ PEDSC handles data obtained from all fields through scientific observation and research activities in both the Arctic and the Antarctic Regions.

PEDSC Activities

Meanwhile, in the case of research on issues such as global environmental change, entirely new results can be generated by simultaneously using a wide variety of data from multiple fields. In such cases, it is necessary to handle metadata, such as the location information and attribute information of various data, in a unified manner. In other cases, data from one field may be applied to a completely different field, leading to the generation of new results or values that were not anticipated.

Therefore, it is critical that the data be open and easy to locate. PEDSC provides support for activities related to the processing, analysis, storage, sharing, publication, and collaborative use of polar science data.



→ PEDSC bridges the various data activities between polar science and external communities.

<http://pedsc.rois.ac.jp/en/activity>



Location: Data Science Building 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-0014 Japan
URL: <http://pedsc.rois.ac.jp/en/>



Center for Social Data Structuring (CSDS)

To Contribute to Solving Social Issues by Maintaining and Sharing Various Data Obtained from Society

CSDS's mission is to promote empirical research aimed at resolving societal issues and to develop foundational technologies that contribute to the management and utilization of "social data" by organizing and providing these data to various stakeholders. "Social data" refers to the spectrum of data related to various phenomena that occur in society that are obtained using various measures and means, including social survey data targeting individuals or organizations, public microdata based on statistical surveys conducted by government agencies or public entities, and social big data collected in real-time using various devices that measure social activity.



CSDS conducts the following projects and initiatives in three main groups according to the fields from which data are available.

Social Survey Related Project

The project implements data collection by forming a nationwide joint research network and maintaining and disseminating social survey data. It promotes projects on the maintenance and publication of large-scale academic research data inherited from the Institute of Statistical Mathematics, research planning and data sharing of joint surveys conducted with researchers of other organizations, and research and educational activities of the compliance-related issues associated with the administration of the social surveys (Fig. 1).

Official Microdata Project

The project is responsible for the maintenance of official statistics microdata and the development of resource-sharing systems, research and development for online data analysis systems, operation of an on-site analysis facility (Fig. 2), and so on. It promotes research projects such as microdata security related to the secure publication of microdata, the promotion of secondary use of official statistics, the project for synthesis credit risk database in the economic and financial sector, and evidence-based policymaking using government statistics. It also functions as the secretariat of Research Consortium for Official Microdata of Japan.

Social Big Data Project

The project maintains methods for managing shared data and platforms, which are needed for research activities that use social big data among researchers from different organizations. In addition, it conducts joint research using social big data. In projects aiming to develop a data-sharing infrastructure in the real world, we work with municipalities and other organizations to empirically develop a data-sharing infrastructure system that can efficiently collect, monitor, and analyze the everchanging real-world conditions in social infrastructure such as roads and traffic.



Fig. 1: Lifecycle of Data

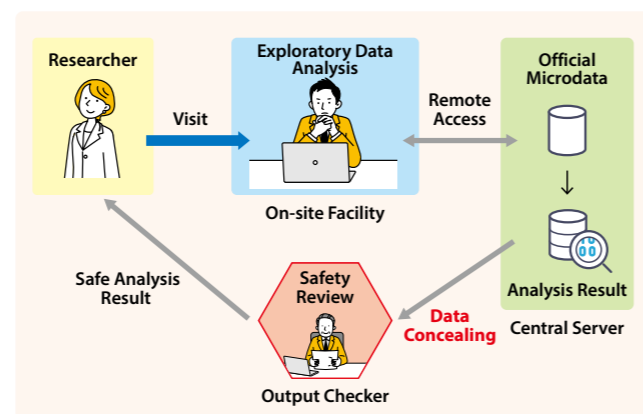


Fig. 2: On-site analysis process (See p.14 for details on the use of facilities)

Center for Open Data in the Humanities (CODH)

Promoting Open Science and Digital Innovation in the Humanities: Development of Data-Driven Humanities and Humanities Big Data

The Center for Open Data in the Humanities (CODH) aims to promote open science and digital transformation in the humanities. CODH is working to promote "data-driven humanities" that promise to revolutionize research methods in the humanities through the introduction of the latest data-driven technologies in informatics and statistics and to promote the use of "humanities big data" (i.e., big data generated in the humanities) by researchers in other disciplines through its promotion of research on the use of AI and cross-disciplinary historical research. As the result of such efforts, CODH has created various research resources that are available to the public.



miwo:
App for AI Kuzushiji Recognition

An app that uses AI to quickly convert images of characters written in cursive into modern Japanese characters. Available free of charge for iOS and Android.

IIF Curation Platform

Introducing to the world of the IIF a new concept of curation and realizing a user-driven IIF platform.

Collection of Facial Expressions (KaoKore)

Using the IIF Curation Viewer, faces that appear in artwork are cut out and gathered for use in art history research.

Edo Big Data

Historical big data on the city of Edo, (Tokyo) including commerce, tourism, and individuals, are linked with geographical data and integrated for data-driven analysis of the past world.

Bukan Complete Collection

Creating an information platform for daimyo (feudal lord) families and shogunate officials based on the comprehensive analysis of Bukan, a bestseller series for 200 years during the Edo period.

Rekiiske

Sharing knowledge and experience related to historical materials to promote material-based research in various fields.



Center for Genome Informatics (CGI)

CGI supports informatics analysis to extract biologically meaningful information from large amounts of genome and transcriptome data

The advancement of next-generation sequencing (NGS) technology has led to wide-spread use of NGS-based whole-genome analyses in various fields of life science. However, given that NGS data comprises an enormous amount of fragmentary sequence data, biological expertise as well as bioinformatics knowledge and technologies are essential for analysis of these data. CGI engages in research and development of information science technologies for high-precision analysis of large-scale genome data, supports the analysis of real-world data, and fosters the development of skilled personnel for this purpose.



CGI provides support to researchers at universities, research institutions, and so on for performing various kinds of genomic data analyses (Fig. 1).

<p>• De novo genome sequencing</p> <p>De novo genome assembly Use NGS data to construct genomes of novel organism species without reference sequence</p> <p>Genome annotation Identify and annotate gene location on genomic sequence and exon-intron structures</p>	<p>• Transcriptome analysis</p> <p>Gene structure/expression analysis Conduct gene structure identification and gene expression quantitative analysis through <i>de novo</i> assemble mapping of RNA-seq</p> <p>non-coding RNA analysis RNA secondary structure prediction, targeted retrieval of miRNA, etc.</p>
<p>• Genome resequencing</p> <p>Whole genome resequencing Compare sequence reads of whole genomes to reference genomes and detect single nucleotide variations and structural variations</p> <p>Targeted genome sequencing/epigenetics analysis Exom, RAD-seq, ChIP-seq, HiC-seq, etc.</p>	<p>• Metagenomic analysis</p> <p>Metagenomic assembly <i>De novo</i> assemble metagenomic sequence reads</p> <p>Species classification, gene prediction Metagenomic sequence clustering, gene prediction, pathway analysis, etc.</p>

Fig.1: Genome data analysis performed at our center

Although the genomic data handled by CGI consist primarily of NGS sequence data, the research objectives and experimental conditions vary greatly from study to study. The subjects of research also range widely from mammals and other vertebrates to insects, plants, fungi, and prokaryotes. An appropriate analytical method must be selected according to genome size, structure, and evolutionary background. Leveraging the extensive experience of its faculty, CGI provides analysis support that is both flexible and precisely tailored to the specific needs of each research project.

In addition to developing various analysis pipelines (e.g., the genome annotation pipeline in Figure 2 and genome resequencing pipeline) aimed at streamlining analysis support, CGI is developing new analytical methods (e.g., gene prediction techniques, RNA-seq assemblers, and metagenomic species classification methods) to create an environment that provides cutting-edge analytical methods.

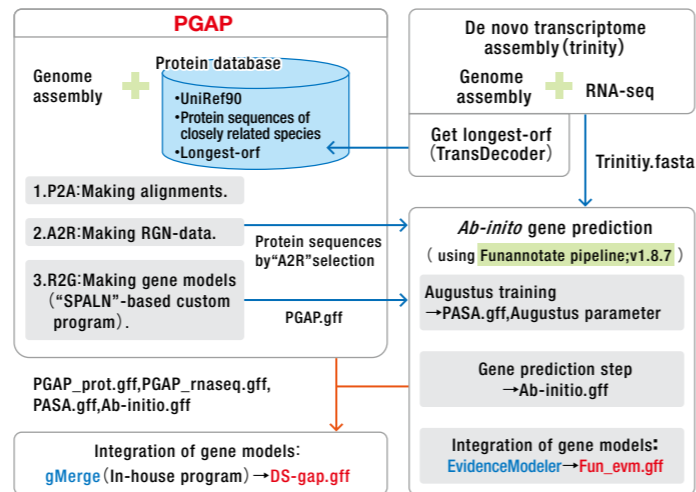


Fig.2: Genome annotation pipelines

Center for Data Assimilation Research and Applications (CARA)

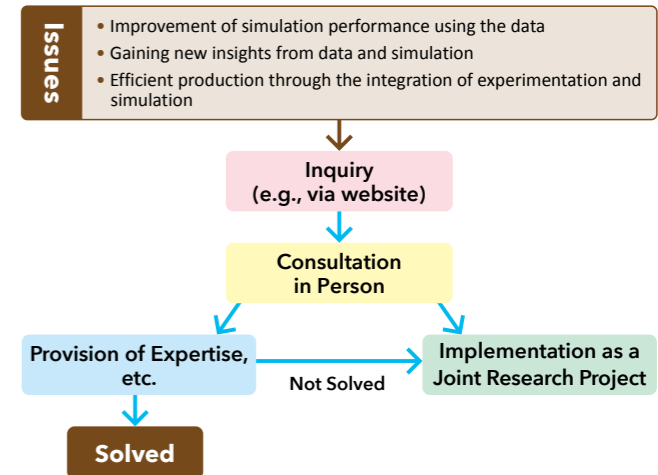
Problem-solving through the integration of simulation and observational data: Want to keep trying a simulation that doesn't work?

Data assimilation refers to the process of integrating observational data with numerical simulations. Data assimilation enables researchers to develop data assimilation systems that are capable of highly accurate predictions as well as emulators that can significantly reduce computation time. The experienced researchers at CARA provide support to personnel in a wide range of scientific and industrial fields who are facing challenges using simulations through the provision of expertise and know-how regarding techniques for integrating data and simulations.



Services Offered by CARA

CARA provides consultation services for data assimilation research as well as advice and technical guidance during meetings and troubleshooting support. CARA can assist with everything from data assimilation methods based on statistical science to consultations on the application of data assimilation in real-world settings and collaborative research opportunities. If you are unsure how to begin implementing data assimilation or have already completed data assimilation but are uncertain about the next steps, CARA is happy to help and looks forward to your inquiries.

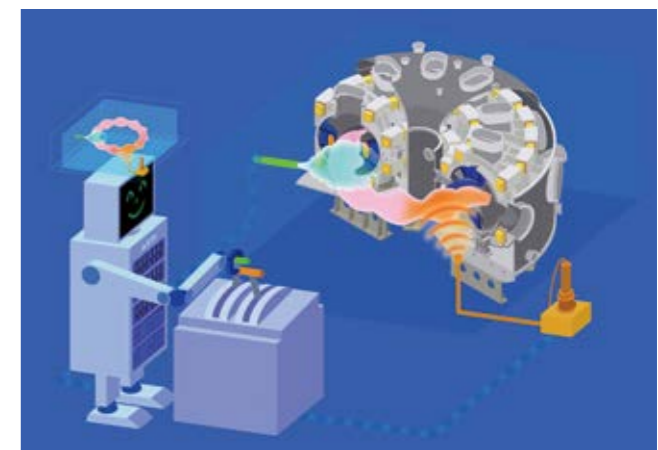


Please read the website and contact us by email for a research consultation.

Research Overview

- Data assimilation system: development and methods
- Model evaluation and parameter estimation in cellular and developmental biology, using data assimilation
- Application of data assimilation and related statistical methods in upper-atmospheric physics
- Data assimilation for the magnetosphere-ionosphere system
- Tsunami data assimilation for improved tsunami forecasting: spatiotemporal regression and applications to urban-rural issues

CARA members belong to the faculties of the Institute of Statistical Mathematics, National Institute of Genetics, and the National Institute of Polar Research and can provide consultations, both theoretical and practical, based on data assimilation methods rooted in statistical science as well as opportunities for collaborative research.



A conceptual image of digital twin control, where real plasma is controlled through a virtual plasma recreated on a computer via data assimilation.

CGI
Center for Genome Informatics

Location: National Institute of Genetics 1111 Yada, Mishima-shi, Shizuoka 411-8540 Japan
URL: <https://genome-info.nig.ac.jp/> (Japanese website)

(Japanese website)

CARA
Center for Data Assimilation Research and Applications

Location: The Institute of Statistical Mathematics 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8562 Japan
URL: <http://daweb.ism.ac.jp/cara/en/>

Center for Juris-Informatics (CJI)

The IT and AI Era of Justice: Urgent need to establish juris-informatics as a new academic discipline that combines technology and the law

Research background

▶ AI technologies such as ChatGPT that use large-scale data are beginning to be used in intellectual work. Major law firms outside Japan are increasingly utilizing GPT; however, in Japan, the Code of Civil Procedure was amended only in 2022, and efforts toward IT transformation have just begun. Action is urgently needed if Japan is to catch up with global technological standards. Meanwhile, the spread of machine learning has led to the emergence of new legal issues related, for example, to the protection of personal information and copyright infringement. To create AI that can be trusted, it is essential to foster legal professionals who understand both legal norms and AI technologies—ideally individuals who comprehend both fields. Accordingly, it is necessary to establish a new academic discipline that integrates AI and the law.



Ken Satoh, Director of CJI

Establishment of Juris-Informatics as a Discipline

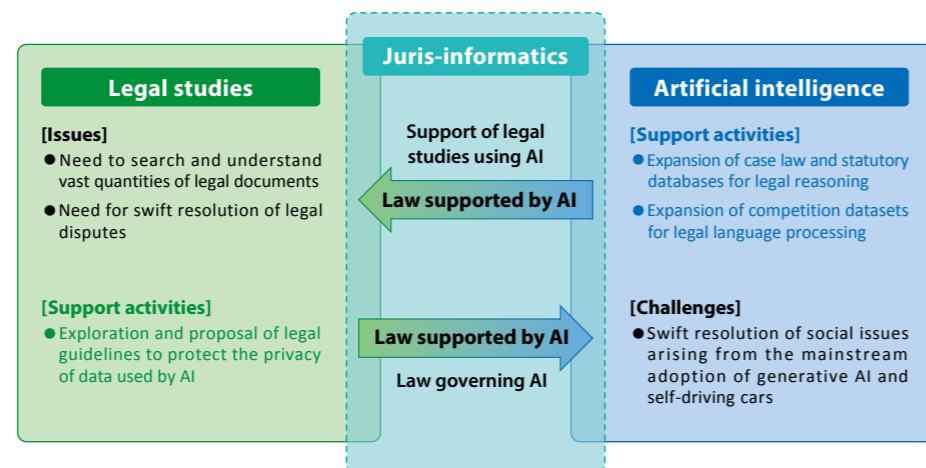
▶ To this end, CJI is implementing law supported by AI and the law governing AI. In addition to enabling the processing of large volumes of legal documents, CJI seeks to establish juris-informatics as a new research field with a mission to address legal and social issues arising from the mainstream adoption of generative AI, self-driving cars, and other related issues.

(1) Law supported by AI

In addition to investigating the manners in which legal data are collected, managed, and utilized, CJI will expand case law and statute databases that support legal reasoning. Furthermore, CJI aims to enable the search of large volumes of legal documents and support the resolution of legal issues in the field of law. To support competition in various tasks related to the collection and management of various information, including the laws and ordinances in different countries (original texts, logical formula, etc.), case summaries, and grounds for verdicts, CJI will work to develop the technologies needed for each task and will encourage participation by individuals from diverse fields in order to build a new juris-informatics research community.

(2) Law governing AI

CJI works with legal scholars to investigate methods for addressing legal issues related to data science and artificial intelligence, and proposes legal guidelines to protect the privacy of data used by AI. CJI also seeks to swiftly address and resolve social issues arising from the widespread adoption of generative AI and self-driving cars. Furthermore, CGI is developing technologies with legal and ethical norms in mind and supports the education of data scientists.



Center for Research and Development on Data Lake (DLRD)

Creation of a New Information Infrastructure for a Data-Driven Society: Establishment and operation of a unified information infrastructure to ensure access to ever-evolving AI models and increasingly diverse content

Background and Mission of DLRD

▶ AI technology has advanced rapidly over just the last few years, with services such as ChatGPT and other generative AI already having a profound impact on the functions of modern society. From the perspective of users, AI models that learn from large and diverse datasets should have characteristics such as ease of understanding, transparency, and assurance of safety and security. DLRD's mission is to develop and administer a unified information infrastructure that clearly indicates the provenance of the models and data and ensures easy access to the most recent content.



Miyuki Nakano, Director of DLRD

(1) The Data Lake Information Infrastructure

The Data Lake facilitates collection and management of content by providing a unified infrastructure that will enable users to access and utilize diverse data, both structured and unstructured, via a single platform. DLRD will also administer and provide easy access to AI models that have been trained on data in the same manner as the content. The centralized management environment for both content and models established by DLRD will provide a framework to promote data utilization and AI technology development, and lay the groundwork for our nation's future information infrastructure with the following functions:

- Continuous accumulation and provenance management of content, establishment of vector indexing technology.
- Ongoing operation of models and management of provenance information, including that of data used for model training and the learning process, as well as provision of cutting-edge technologies related to model operation.
- As an inter-university research institute, build a robust system supported by security technologies to provide a secure and safe utilization environment for advanced research by industry, academia, and government.
- Identification of societal issues associated with data utilization, including legal issues such as the protection of personal information and copyrights, and social ethical issues related to life, and providing a safe and secure environment for the use of AI technologies and content.

(2) A Medical Data-management System to Promote Utilization and research and development related to Medical LLMs

In addition to promoting medical research, there is increasing recognition of the importance of sharing medical information, leveraging generative AI, and ensuring the provision of appropriate medical technologies in both academic and societal settings. DLRD is in the process of concretizing the Data Lake infrastructure and has begun research and development on a medical data-management system (Figure 1), with funding provided through the FY2023 supplementary budget of the Strategic Innovation Promotion Program. The goal is to deploy a 10-peta-scale large-scale data utilization infrastructure and create a large-scale medical LLM/LMM model management system together with a medical database. Additionally, DLRD will systematically collect and appropriately manage the training data, which will include medical data.

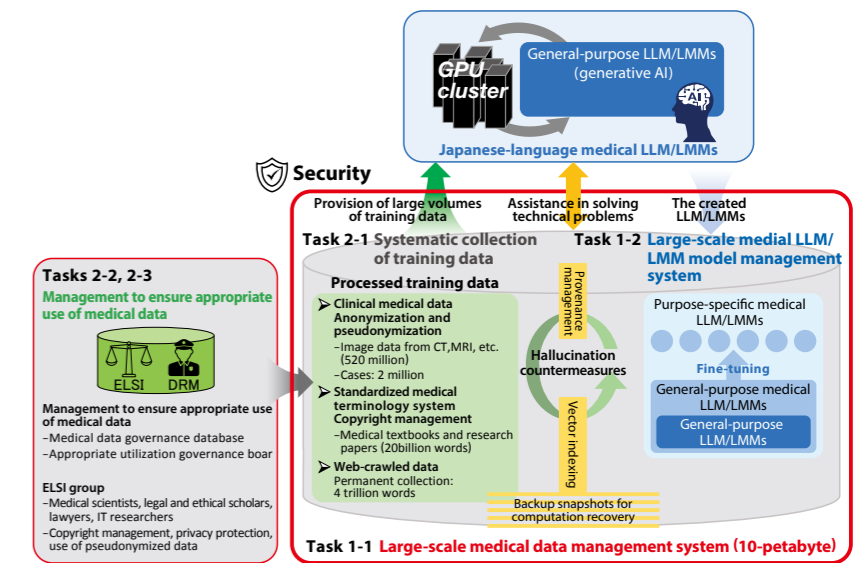


Figure 1. Medical data management system that promotes R&D on medical LLM/LMMs.



Location: National Institute of Informatics 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
 URL: https://ds.rois.ac.jp/en_center7/



Location: Huli Kamiyacho Building 2F 4-3-13, Toranomon, Minato-ku, Tokyo 105-0001 Japan
 Research Organization of Information and Systems
 URL: <https://ds.rois.ac.jp/center8/> (Japanese website)



(Japanese website)

The Collaboration Program “ROIS-DS-JOINT”

ROIS-DS invites applications every year for “ROIS-DS-JOINT,” an open joint research program on data science, to offer researchers throughout the country opportunities for research sharing and joint research. There are two types of programs offered: Joint Research to be conducted with researchers and resources from each center affiliated with ROIS-DS, and Joint Research Meetings held mainly at each center, including research exchanges, workshops, and so on.

The Joint Research Program

This program targets research related to specific data science issues conducted by small research teams affiliated with ROIS-DS or research teams utilizing ROIS-DS resources.

- The Principal Investigator must be a researcher from an educational or research institution outside of our organization (including librarians, museum curators, and researchers).
- Joint research budget: up to JPY 1,000,000.
- The budget can be used for expenses such as travel costs, service fees, and supplies necessary for conducting the research (but excludes items that institutions are generally expected to provide).

The Joint Research Meeting Program

This program targets small-scale meetings held to advance joint research that are convened for the purposes of discussion, data science-related research exchange, workshops, and so on.

- The Principal Organizer must be affiliated with an educational or research institution outside of our organization (those whose primary duties do not involve research, development, or investigation are also eligible to apply).
- Joint research budget: up to JPY 500,000.
- The budget can be used to cover travel expenses for participants and costs related to organizing the meetings.

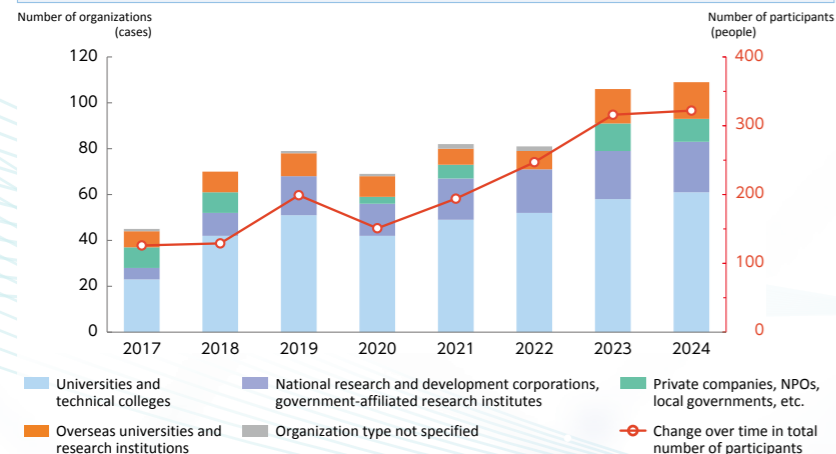
Notes for Applicants

- Applicants should **discuss and agree upon with ROIS-DS staff prior to application.**
- **Joint research expenses will be managed by ROIS-DS.** These funds will not be directly allocated to the Principal Organizer.
- Because the mission of ROIS-DS is to facilitate data sharing and analysis, projects that solely involve data collection (including surveys) are not eligible for this program.
- For more detailed information on the application process, please visit ROIS-DS website: <https://ds.rois.ac.jp/en/>.

Trends in the Number of Participating Institutions and Collaborators

Both the number of collaborators and the number of participating institutions have exhibited steady growth since the launch of the program in 2017. In FY2023, the number of participating institutions exceeded 100. The program has received research proposals from applicants affiliated with a wide range of institutions, including universities, technical colleges, and national research institutes. We have also seen a steady increase in the number of applicants affiliated with overseas universities and research institutions, particularly in Asian countries.

Trends in the number of institutions and collaborators participating in “ROIS-DS-JOINT” open joint research program (general collaborative research programs)



Recent participation in collaborative research (General collaborative research)

	2022	2023	2024
Universities and technical colleges	53	58	61
National research and development corporations, government-affiliated research institutes	19	21	22
Private companies, NPOs, local governments, etc.	4	12	10
Overseas universities and research institutions	8	15	16

DS Promotion

ROIS-DS Workshop & International Symposium

ROIS-DS Workshops serve as a platform for reporting on center activities and the outcomes of open-call joint research. The 4th Workshop was held online in FY2024 (videos of the presentations are available on ROIS-DS website). Part 1 features the keynote speech and activity reports from the centers, while Part 2 consists of poster presentations using video chat tools, which led to lively discussions on the accomplishments of joint research. The International Symposium on Data Science (DSWS-2023)—Building an Open-Data Collaborative Network in the Asia-Oceania Area was held in December 2023.



Coordination of research support

The research coordinators of ROIS-DS play a central role in public relations activities at conferences, responding to inquiries, and providing support for initiating joint research. They support the promotion of data science in a wide range of research fields, including biology, medicine, pharmacy, engineering, agriculture, environmental studies, earth and planetary sciences, statistics, and financial engineering.

The Molecular Biology Society of Japan, The Japanese Society of Human Genetics, The Japan Society for Bioscience, Biotechnology, and Agrochemistry, Society of Evolutionary Studies, Japan, The Pharmaceutical Society of Japan, The Ecological Society of Japan, Japanese Cancer Association, The Society for Biotechnology, Japan, Japan Geoscience Union, Institute of Actuaries of Japan, The Japanese Association of Risk, Insurance and Pensions, and so on.

Past achievements can be found at the following → <https://ds.rois.ac.jp/information/exhibitions/>

Information Dissemination

Hands-on Workshops

We organize various hands-on workshops, including integrated database workshops: AJACS (by JST and co-organized by DBCLS), RDF Workshop (by DBCLS), CODH Tutorial (by CODH), and Data Assimilation Hands-on (by CARA, see photo). IUGONET training sessions (by PEDSC), which include a program on data comparison, are also held in Japan and abroad.



Publications

We have published the booklet “ROIS-DS: Progress Toward Data Science,” which chronicles the history of the DS Center, as well as a collection of DS Center lectures entitled “The Future of Data Science.”

- ROIS-DS Steps Toward Data Science
- ROIS-DS Brochure
- ROIS-DS Lecture Series “The Future of Data Science”



Use of Facilities

On-site Facilities

ROIS-DS offers, together with the Center for Social Data Structuring (CSDS), an on-site research facility, equipped with an advanced security environment that includes access to control systems and surveillance cameras, thereby ensuring the safe use of survey questionnaire information and confidential data.



Opening hours: Weekdays 10 am – 5 pm

Excluding Saturdays, Sundays, national holidays, December 29 to January 3, and holidays for which a booking calendar has announced a holiday notice.

*Open by agreement on holidays and after hours

Location : Data Science Building 2F 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-0014 Japan

Email address : tokumeid(at)ism.ac.jp