Support Activities

-Data Sharing Support

Databases	Research Fields, Communities	Activities
＜ Integrated Use of Life Science Database ＞ -TogoDX（Life Science Data Explorer） -TogoVar（Japanese Genetic Variation Database） -RDF Portal（Life Sciences RDF Data List） -A suite of tools to support RDF data retrieval and creation supporting the above	Life Sciences	Establish an environment for the integrated use of databases, develop essential technologies for database integration, and maintain international standards
＜ Integrated Database for Polar Science ＞ -Science Database -Arctic and Antarctic Data archive System（ADS） -Inter-university Upper atmosphere Global Observation NETwork（IUGONET）	Polar and Global Environment	Contribute to global environmental research by promoting data science and the publication and sharing of various scientific data obtained from the Arctic and Antarctic Regions
＜ Data on social conditions（social sciences）＞ -Survey on Japanese national character -International comparative study of awareness -（On-site analysis lab）	Social survey data, Microdata from public surveys, Social big data	Formation of an inter-university network infrastructure for effectively using data on social conditions（social surveys, public micro databases, social big data）and contribution to local communities.
＜ Creation of humanities（digital humanities）disciplines based on data science ＞ -Dataset of Japanese pre-modern text -Dataset of Kuzushiji in Japanese pre-modern text -Dataset of Edo Cuisine Recipes -Collection of Facial Expressions -Edo big data -Database for Sharing Information about Historical Materials	Digital Humanities, Machine learning, Art history, Palaeoclimatology, Japanese cultural studies	Develop novel humanities research in the era of open science, such as data-driven humanities, which analyses humanities materials（historical documents）using cutting-edge technologies in informatics and statistics, as well as humanities big data, which allows cross-disciplinary use of datasets built on humanities research outcomes

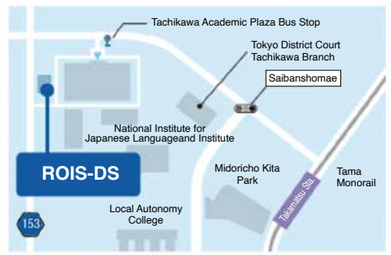
-Data Analysis Support

Analysis objects	Research Fields, Communities	Activities
＜ Genomic data analysis of various species ＞ -New Genome Determination -Genome Re-sequencing -Transcriptome Analysis -Metagenomic Analysis	Genome Biology Genome Medicine Genomic Drug Discovery	Provide informatics-based analytical support for diverse life science research based on large amounts of sequence data from next-generation DNA sequencers
＜Theoretical and applied research based on numerical simulations ＞ -Human Flow Simulation -Spacecraft Simulation -Coastal Ocean Model -Integrated Transport Simulation of Nuclear Fusion Plasmas -Magnetosphere-Ionosphere Model	Traffic Engineering Space Engineering Geophysics Nuclear Fusion Studies All other fields that use data-centered science concepts and methods	Research consultation and support and joint research utilizing methods such as data assimilation and emulators Organize hands-on workshops and experiential learning activities related to the above

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

10-3 Midori-cho, Tachikawa-shi, Tokyo 190-0014

https://ds.rois.ac.jp/en/



Contact: Data Science Promotion Section Tel: +81-42-512-9254　　email: ds_suishin@rois.ac.jp

Joint Support-Center for Data Science Research
Research Organization of Information and Systems（ROIS）

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

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

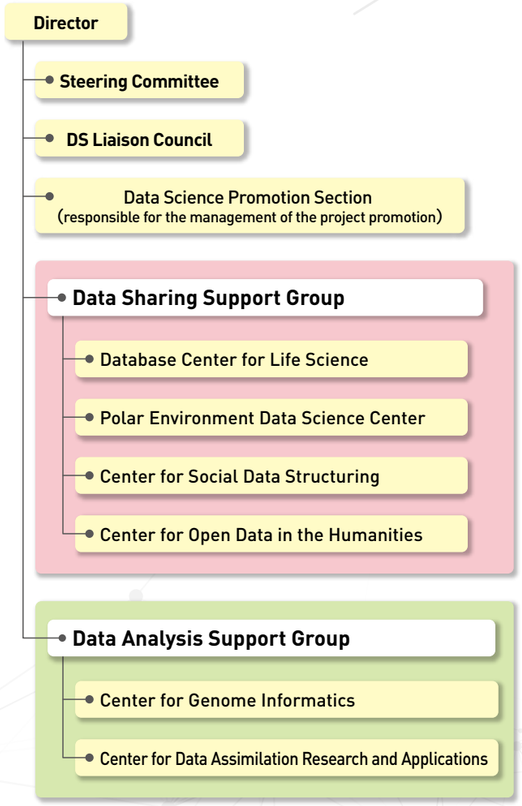
-Cooperation, Collaboration, and Development -

Interdisciplinary Resource-Sharing and Supporting Researchers with Data-driven

Center of ROIS for Promoting Data Science

Research Organization of Information and Systems "Joint Support-Center for Data Science Research（ROIS-DS）" is a resource-sharing and joint research center within ROIS that promotes data science nationwide on an integrated basis, to solve scientific and social challenges through the advanced analysis of large-scale data. Established in the fiscal year 2016 by ROIS, ROIS-DS boosts cooperation and collaboration with universities and other organizations with a focus on data science（data-driven research）. As of 2022, ROIS-DS consists of six centers, namely Database Center for Life Science（DBCLS）, Polar Environment Data Science Center（PEDSC）, Center for Social Data Structuring（CSDS）, Center for Open Data in the Humanities（CODH）, Center for Genome Informatics（CGI）and Center for Data Assimilation Research and Applications（CARA）, and contributes to enhancing the research capabilities of universities and other institutions. ROIS-DS handles an extensive range of data, from a large amount of bioinformatics data on genomes and genes to observation data obtained by atmospheric radar and pre-modern Japanese text, social surveys, and public microdata. Therefore, we are carrying out our activities in cooperation with four research institutions which are subsidiary organizations of ROIS, namely National Institute of Polar Research, National Institute of Informatics, The Institute of Statistical Mathematics, and National Institute of Genetics, as well as the other research institutions of the Inter-University Research Institute Corporation.

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



◆ Database Center for Life Science（DBCLS）◆

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

◆ Polar Environment Data Science Center（PEDSC）◆

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

◆ Center for Social Data Structuring（CSDS）◆

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

◆ Center for Open Data in the Humanities（CODH）◆

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.

◆ Center for Genome Informatics（CGI）◆

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

◆ Center for Data Assimilation Research and Applications（CARA）◆

Solving problems in various sciences and industries through the data assimilation technique

Joint Research Center for Research

Hiroyuki Araki,
Director of ROIS-DS

Outline:

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

Details of our support projects is on page 12.

Support Project（Data Analysis Support）

Genomic data analysis support project
Data fusion computing support project

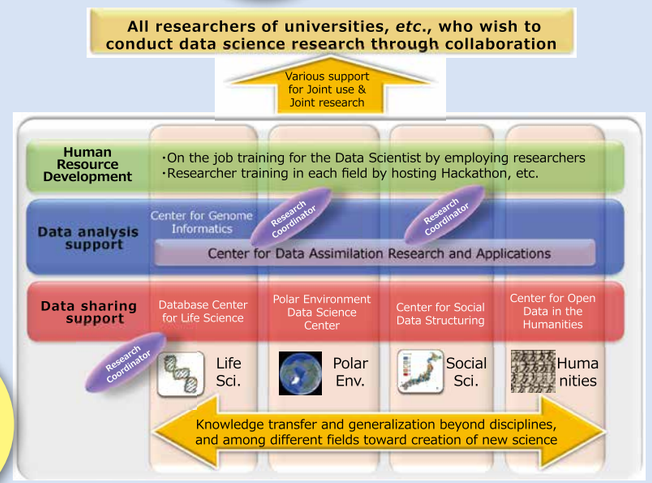
Human Resource Development（Training of Data Scientists）

Data science expert development through on-the-job training
Training of data science educators（DS instructors）

The Collaboration Program（ROIS-DS-JOINT）

Joint research program
Joint research meeting

See page 10 for more details.



Hub Activities:

Data Scientists Education and Training Hub

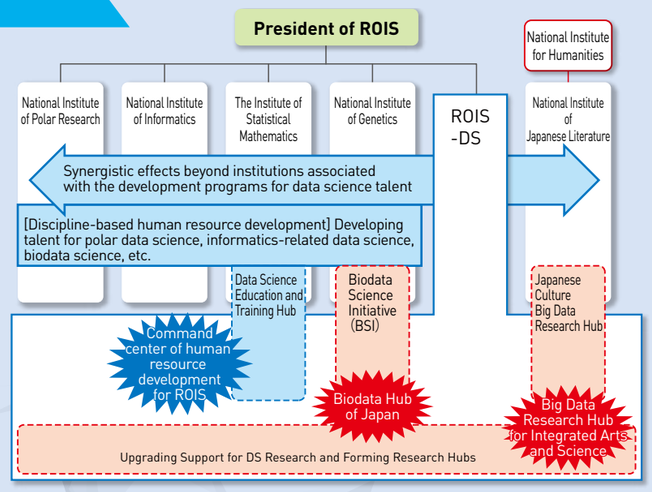
ROIS-DS and other institutions work together to train data science experts at various levels and in various fields

Biodata Science Initiative（BSI）

Support for world-class research efficiency through cooperation and integration between the ROIS-DS and National Institute of Genetics（NIG）regarding the collection, organization, standardization, etc. of data

Japanese Culture Big Data Research Hub

ROIS-DS and National Institutes for the Humanities cooperate to analyse Japanese culture using data-driven methodologies



Promotion of data science through cross-organizational reform of ROIS with ROIS-DS as a core（synergy effects through collaboration and integration）

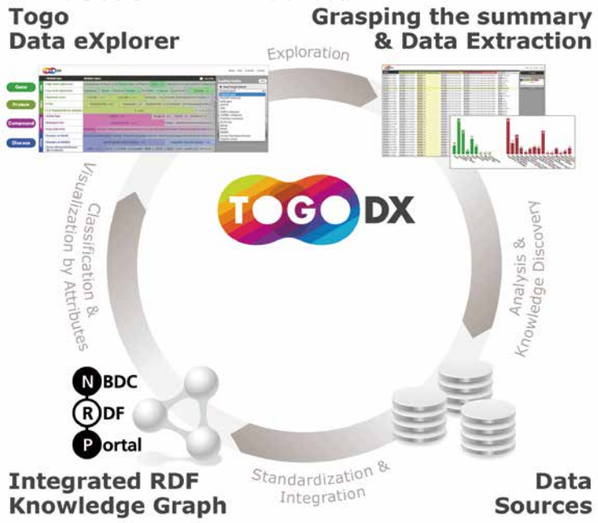
Database Center for Life Science (DBCLS)

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 owned and generated. 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. We also invite experts of database development organizations from around the world to organize annual developers' conferences such as BioHackathon, leading the international technical development and standardization for integration.

Yuji Kohara,
Director of DBCLS

▶ In order to realize the environment under which databases (DBs) scattered on the web can be utilized in an integrated manner, we, striving to build knowledge graphs of life science data, continuously support transporting various data in Resource Description Frameworks (RDF) and accept RDF data from different organizations. In 2021, we developed TogoDX, a framework for integrated exploration of information accumulated in various DBs, and made public TogoDX/Human (https://togodx.dbcls.jp/human/), an application enabling one-stop exploration of information on humans from a knowledge graph integrating national and international databases. The tool helps narrow down research targets, interpret and discuss experimental results, and apply the narrowed-down results to integrated analysis. Future development will include the enhancement of data on humans and the application of TogoDX to other species.



Data analysis platform powered by TogoDX, a framework for exploring the knowledge graph which integrates various databases

▶ Implementing the integrated use requires collaboration with national and international DB professionals. Therefore, we have organized various workshops, such as the international BioHackathon in which joint development works were intensively performed over the past decade. Despite the difficulty in holding international events where international participants have to be invited amid the COVID-19 pandemic, we continue to organize domestic BioHackathon (annually) and SPARQLthon/Togothon (monthly, two consecutive days) to promote DB standardization.

A group photo taken at the domestic BioHackathon held in September 2022. In Kochi, 79 participants worked on various development projects.

■ **Through ROIS-DS-JOINT, DBCLS supports members of any university, research institute, company, etc. who are interested in developing and applying the following technologies and related methods:**

Keywords: [Life science data integration] [Knowledge graphs] [Large-scale data analysis]

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.

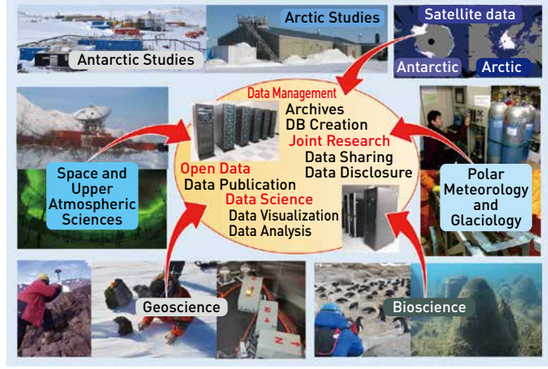
The purpose of the Polar Environment Data Science Center (PEDSC) is to promote the opening publication and sharing of various types of scientific data originally obtained by scientific observations and research activities in the Antarctic and the Arctic regions by the polar science research community in various fields led by National Institute of Polar Research (NIPR), to strengthen the collaborations between universities and other external communities, and to support the creation of further scientific outputs and further advancement of the importance of the polar science research. Internationally, PEDSC is also expected to play the key role of a national data center for polar science in Japan. The goals of PEDSC are to play a central role in the data activities in the polar science research in Japan, to create a new data-centric polar science, and to contribute to the advances of global environmental studies.

Akira Kadokura,
Director of PEDSC

▶ Data handled by PEDSC are obtained through scientific observations and research activities in the Antarctic and Arctic regions. Various observations and research activities are carried out in both polar regions mainly by National Institute of Polar Research (NIPR) in various research fields, including upper atmospheric sciences, meteorology, marine science, glaciology, geoscience, and bioscience, and a wide diversity of data are obtained in various forms, such as digital data in various recording media, or samples data obtained and stored in various containers.
The obtained data are processed and analyzed and from which, after being converted to physically meaningful data, scientific results are generated. In order to generate reliable scientific results, the reliability of the data should be ensured. For that purpose, the data should be securely stored without being lost, degraded or falsified, and be usable for everyone, the reproducibility of the same scientific results should be maintained, and so on.

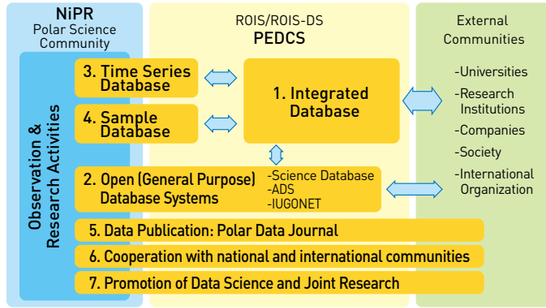
▶ On the other hand, studies such as global environmental change may produce entirely new results by simultaneously combining a wide variety of data from many different fields.
In such cases, meta-information on the location and attribute of those various data (metadata) should be treated in a unified manner. Also, it could be possible that data from a research field is applied to unrelated research fields, which leads to unpredictable new findings and new values. To this end, openness and simple accessibility of the data are important. We aim to support such activities for processing, analyzing, archiving, sharing, and opening polar science data (to promote further collaboration with external communities).

Activities of PEDSC



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

Activities of PEDSC



PEDSC bridges the various data activities between polar science and external communities.
http://pedsc.rois.ac.jp/en/activity

■ **Through ROIS-DS-JOINT, PEDSC supports members of any university, research institute, company, etc., who are interested in using the following data and database systems:**

Keywords: [Polar Science Data] [Science Database] [IUGONET]

Center for Social Data Structuring (CSDS)

To Contribute to Solving Social Issues through Maintaining and Sharing various Data Obtained from Society

The activities of the Center for Social Data Structuring (CSDS) are to maintain and make available a wide range of data obtained from the measurement of various phenomena in society—such as data obtained through social surveys targeting individuals or organizations, official microdata collected as a result of statistical surveys conducted by the public sectors such as government agencies, and social big data collected from real-time monitoring of social activities using various devices—, to put such data to a wide range of practical use, to promote empirical scholarship for the solution of various social issues, and to develop research infrastructure for the implementation of policy making based on empirical data. The Center's activities also include developing basic technologies that facilitate data maintenance and utilization.

Tadahiko Maeda,
Director of CSDS

▶ 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 dissemination of the compliance-related issues associated with the administration of the social surveys.

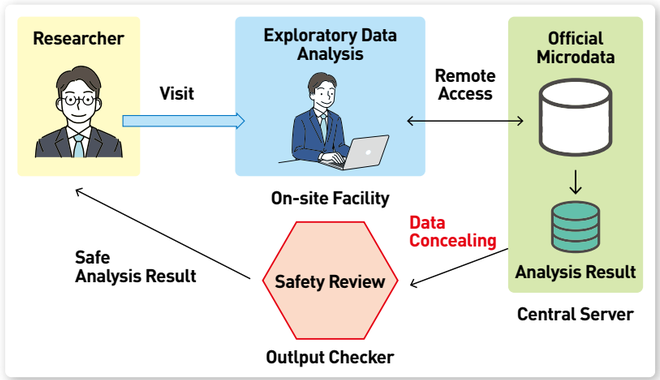
Official Microdata Project

The project is responsible for the maintenance of official microdata and the development of resource-sharing systems, R&D for online data analysis systems, operation of an on-site analysis facility, etc. 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 policy making (EBPM) using government statistics. It also functions as the secretariat of Research Consosium 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 ever-changing real-world conditions in social infrastructure such as roads and traffic.

Lifecycle of Data



On-site analysis process

■ **Through ROIS-DS-JOINT, CSDS supports members of any university, research institute, company, etc. who are interested in maintenance and resource-sharing related to the following data, and in development and research of basic technologies related to various processes which lead to resource-sharing.**

Keywords: [Social Survey Data] [Official Microdata] [Social Big Data]

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

Center for Open Data in the Humanities (CODH) is advancing its research mainly on two themes, aiming for the promotion of open science and digital innovation in the field of humanities. The first is the study of "data-driven humanities" that innovates research methods of the humanities by applying cutting-edge data-driven technologies in informatics and statistics. We create new knowledge from humanities data through new technologies such as research on AI kuzushiji recognition for pre-modern Japanese text and research on art history using the International Image Interoperability Framework (IIIF). The second is the study of "humanities big data" in which big data generated in the humanities field is used for research in other fields. We are engaged in research on historical climatology, which involves retrieving weather data from diaries in the Edo period to reconstruct the climate in the past.

Asanobu Kitamoto,
Director of CODH

miwo: App for AI Kuzushiji Recognition

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

IIIF Curation Platform

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

Collection of Facial Expressions (KaoKore)

Using the IIIF 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," 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.

Rekiske

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

■ **Through ROIS-DS-JOINT, CODH supports members of any university, research institute, company, etc. who are interested in developing and applying the following technologies and related methods:**

Keywords: [Pre-modern Japanese Text, Kuzushiji] [IIIF] [Humanities Big Data]

Center for Genome Informatics (CGI)

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

(Japanese website)

With the advancement of next-generation sequencing (NGS) technology, analyses using NGS, such as novel genome sequencing, resequencing, and transcriptome analysis, have been widely conducted in various life science research fields. However, NGS data are only fragmental data of base sequences, and the volume of such data is enormous. Therefore, in order to efficiently analyze them and obtain results that meet the purposes, it is essential to have knowledge and skills in bioinformatics in addition to knowledge of biology.
The Center for Genome Informatics (CGI) conducts research and development of information science technology for the rapid and highly accurate analysis of large amounts of genome data and supports real data analysis using cutting-edge methods and trains personnel for this purpose.

Hideki Noguchi,
Director of CGI

▶ CGI provides support to researchers of universities, research institutions, etc, for various kinds of genomic data analyses (Fig. 1).
During the six years from 2016 to 2021, we provided analysis support for total of 51 tasks at the request of 32 universities and research institutions (affiliated institutions of the representative).

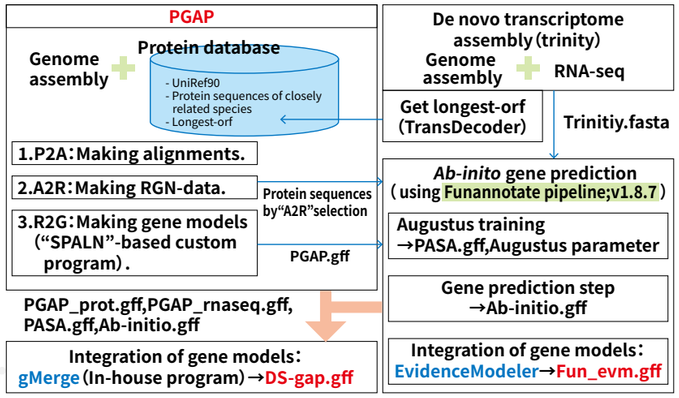
▶ The genomic data handled are mainly NGS sequence data, but the research objectives and experimental conditions vary widely depending on the study. The target organism species also vary from mammals and other vertebrates to insects, plants, fungi, and prokaryotes. Therefore, appropriate analysis methods must be selected according to genome size, structure, and evolutionary background. Taking advantage of our rich analytical experiences, CGI provides flexible and highly accurate analytical support according to research objectives.

▶ In addition, we are committed to securing an environment to provide cutting-edge analytical methods by developing, for smooth implementation of analysis support, various analysis pipelines, such as the genome annotation pipeline (Fig. 2) and the genome resequencing pipeline as well as novel analytical methods, including gene prediction methods, RNA-seq assemblers, meta-genome species classification, etc.

What We Analyze

- *De novo* genome sequencing
  - *De novo* genome assembly
    - Construct, by use of NGS data, genomes of novel organism species without reference sequence
  - Genome annotation
    - Identify and annotate gene location on genomic sequence and exon-intron structures
- Genome resequencing
  - Whole genome resequencing
    - Compare sequence reads of whole genome to reference genome and detect SNV and Structural Variation (SV)
  - Targeted genome sequencing/epigenetics analysis
    - Exom, RAD-seq, ChIP-seq, HiC-seq etc.
- Transcriptome analysis
  - Gene structure/expression analysis
    - Conduct gene structure identification and gene expression quantitative analysis through *De Novo* assemble mapping of RNA-seq
  - non-coding RNA analysis
    - RNA secondary structure prediction, targeted retrieval of miRNA, etc.
- Metagenomic analysis
  - Metagenomic assembly
    - *De novo* assemble metagenomic sequence reads
  - Species classification, gene prediction
    - Metagenomic sequence clustering, gene prediction, pathway analysis etc.

Genome data analysis performed at our center



Genome annotation pipelines

■ **Through ROIS-DS-JOINT, CGI supports members of any university, research institute, company, etc. who are interested in developing and applying the following technologies and related methods:**

Keywords: [Bioinformatics] [Next-generation Sequencing (NGS)] [Genomic DNA/RNA Data Analysis]

Center for Data Assimilation Research and Applications (CARA)

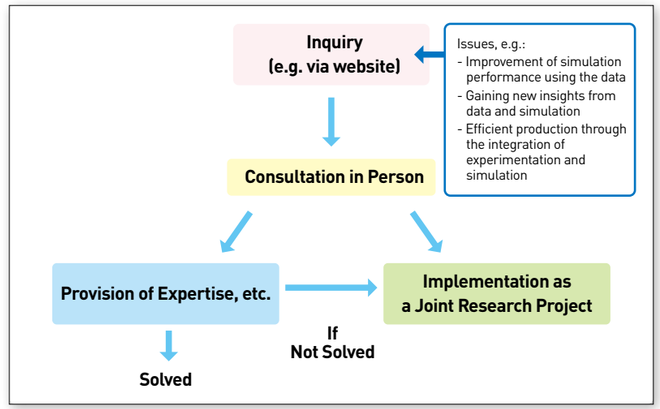
Problem-solving through the integration of simulation and observational data: Want to keep trying a simulation that doesn't work? How can observational data be used for prediction?

Data assimilation is a method of integrating observational data with numerical simulations. It allows the development of a data assimilation system, simulations that enable highly accurate predictions, and emulators, simulations that significantly reduce processing time. The center offers a consultation service for data assimilation research, providing in-person advice and technical guidance and support for problem-solving. Consultation and joint research are available for the data assimilation method based on statistical science and for the application of data assimilation. We are looking forward to your contact on how to deploy data assimilation or what to do next following data assimilation calculations.

Genta Ueno,
Director of CARA

【On-the-spot relay of consultation】
Thank you for visiting us in Tachikawa. Please give us about two hours.

I understand that you intend to deploy data assimilation. It still needs to be implemented in your field. Where did you hear the term "data assimilation?" So, you listened to my seminar. That's nice to hear. So, you have a simulation model and observational data.
All right, then please explain. Thank you. I see, thank you very much.
Now, let me explain the concept of data assimilation. I want to express all variables in your material as x and all equations as f. Is that OK? I'll write it on the whiteboard. Yes, x is a huge vector. I group variables with different time steps separately. In the same way, the data shall be taken together as y.
Now, let us get to the main issue of data assimilation. The problem is that x and y do not match, so data assimilation uses y to correct x. The difficulty with correction is that a simple replacement of value x with value y does not work. A number of parameters are needed for the correction to be carried out successfully. A simple exercise will help you understand.
Do you want a sample program? For the future, it is better to write a program from scratch. Please refer to the book I mentioned. If your simulation is in FORTRAN, it is also good to start data assimilation in FORTRAN. Good luck!



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

Photo of the case settled

■ **Through ROIS-DS-JOINT, CARA supports members of any university, research institute, company, etc. who are interested in developing and applying the following technologies and related methods:**

Keywords: [Data Assimilation] [Simulation] [Emulator]

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, etc.

The Joint Research Program

-The principal investigator must be a researcher（including librarians, curators, and researchers at a museum or other institutions）at an educational or research institution other than ROIS

-ROIS-DS will provide up to 1,000,000 JPY to each approved project as research expense, which is limited to the direct cost necessary to conduct the project such as consumables and cost of labors used at the host centers, for example.
Salary and the purchase of general equipment and consumables are not allowed.

The Joint Research Meeting Program

-The principal investigator must be a member of an educational or research institution other than ROIS
（the applicant could be someone whose primary responsibility is not research and development or surveying）
⇒ e.g. Applicant can be a private sector researcher or local government officer

-ROIS-DS will provide up to 500,000 JPY in total to each approved plan as expenses for travel and rental room fees, for example.

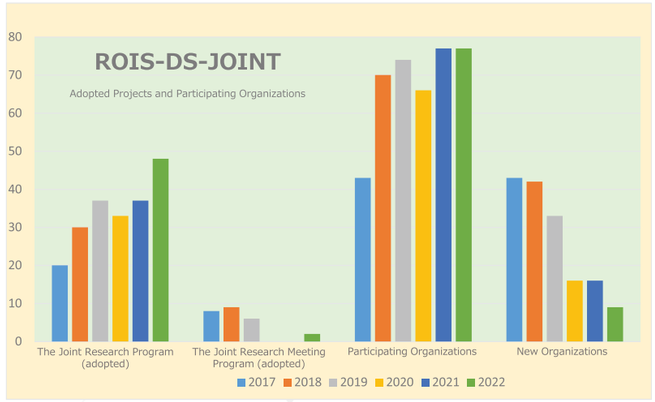
General remarks

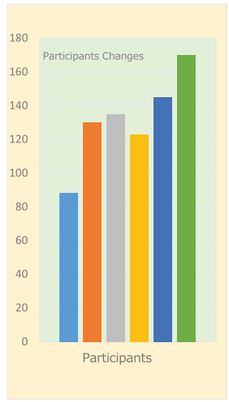
-Each proposal has to be discussed and agreed upon with the host scientist(s) prior to the submission.

-At least one of the ROIS-DS members has to be included in your proposal as the host scientist(s).

The research funds are managed by DS centers, not allocated directly to principal investigators.

Further information on the application and other instructions are provided on the ROIS-DS website（https://ds.rois.ac.jp/en_koubo/）.





Participating Universities coverage rate

（by type of affiliated universities in the Joint Research Program. Cumulative numbers as of the end of March 2022）

National Universities: 34/86 universities（39.5%）

Public Universities: 10/97 universities（10.3%）

Private Universities: 31/627 universities（4.9%）

Note: university types are based on the list of institution numbers（https://www-kaken.jsps.go.jp/kaken1/kikanList.do）.

DS Promotion

● ROIS-DS Workshop

The event is held to disseminate the outcomes of open joint research projects. The second workshop was held online on January 2022 (video of the lecture is available on the DS facility website) . Following the invited lectures in the first part and reports on the activities of each center in the second part, the third part provided an opportunity for a poster session using the video chat tool, which led to active discussions on the joint research outcomes.

Made by Spatial Chat Ltd

● Coordination of research support

The research coordinators of ROIS-DS play a central role in engaging in public relations activities at conferences, response to inquiries, support for start-up of 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, etc.

Information Dissemination

● Hands-on（Experimental Workshops）

We organize various hands-on workshops regularly nationwide and internationally, including hands-on training workshop for integrated database : AJACS (organized by JST NBDC and co-organized by DBCLS), RDF Workshop（organized by DBCLS）, CODH Tutorial (organized by CODH) and Data Assimilation hands-on (CARA, see photo).

● Publications

We published a booklet, ‘ROIS-DS: Steps towards Data Science,’ featuring the history of ROIS-DS.

■ PR brochures（in Japanese）
-ROIS-DS: Steps toward data science
-ROIS-DS brochure
-ROIS-DS Lecture Series
'The Future of Data Science’

Use of Facility

● On-site Facility

Joint Support-Center for Data Science Research（ROIS-DS）offers, with the Center for Social Data Structuring（CSDS）, an on-site research facility, equipped with an advanced security environment such as access control systems and surveillance cameras to ensure the safe use of survey questionnaire information and confidential data.

＜Opening hours＞ Weekdays 10 am – 5 pm
Excluding Saturdays, Sundays, national holidays, 29 December to 3 January, and holidays for which a booking calendar has announced a holiday notice.
*Open by agreement on holidays and after hours