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)

Inter-University Research Institute Corporation
Research Organization of Information and Systems
Joint Support-Center for Data Science Research(ROIS-DS)

Data Science Building, 10-3 Midori-cho, Tachikawa, Tokyo 190-0014, Japan
Phone: +81-42-512-9254
https://ds.rois.ac.jp/en/

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

Hulic Kamiyacho Bldg. 2F, 4-3-13, Toranomon, Minato-ku, Tokyo 105-0001, Japan
Phone: +81-3-6402-6200
https://www.rois.ac.jp/en/

brochure 202010

ROIS-DS: an interdisciplinary, joint-use, collaborative research center supporting data-driven researches

Director
Asao Fujiyama

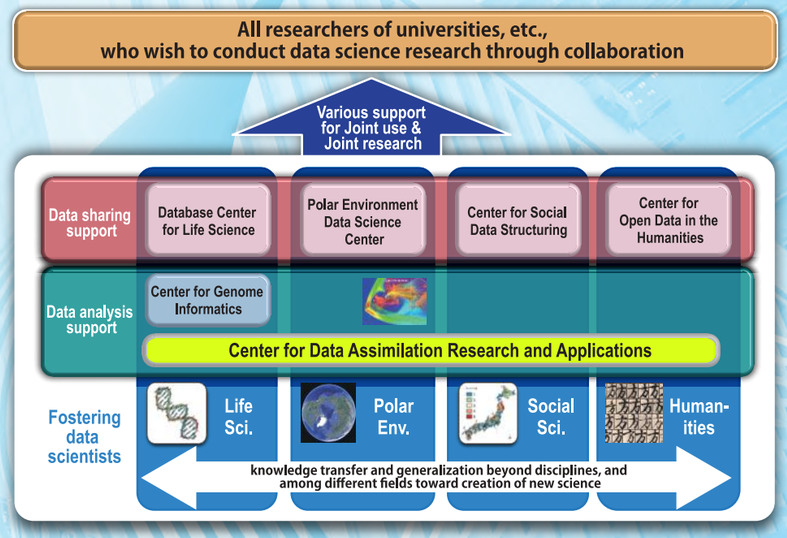
A base for promoting data science

The Joint Support-Center for Data Science Research (ROIS-DS) is a joint-use, collaborative research center for the advancement of interdisciplinary data science to solve scientific and social challenges through advanced analysis of big data at a national scale. It was established by the Research Organization of Information and Systems (ROIS) in 2016 to strengthen collaboration and cooperation among universities and other institution under the slogan data science (data-driven research). The ROIS-DS consists of a total of six centers as of November 2019 (the Database Center for Life Science, the Polar Environment Data Science Center, the Center for Social Data Structuring, the Center for Open Data in the Humanities, the Center for Genome Informatics, and the Center for Data Assimilation Research and Applications) and contributes to strengthening the research capability of universities and other institutions. The ROIS-DS handles an extremely broad range of data: large-scale data related to biological information such as genomes and genetics, observational data such as atmospheric radar data, classical documents, and micro-data from social surveys as well as official statistics. The ROIS-DS therefore cooperates with the other four institutes under the ROIS umbrella, the National Institute of Polar Research, the National Institute of Informatics, the Institute of Statistical Mathematics, and the National Institute of Genetics, as well as other institutes associated with the Inter-University Research Institute Corporations.

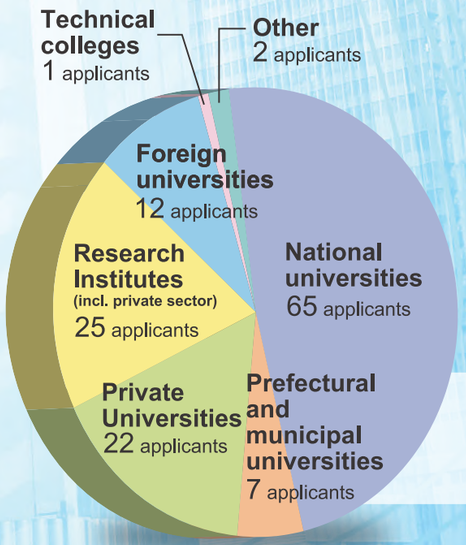
Data Science Building (Tachikawa campus)

Supporting research through data sharing, data analysis support, and human resource development

The six centers of the ROIS-DS facilitate research across a broad range of diverse domains spanning life sciences, polar environment, and statistical mathematics as well as the humanities and social sciences. The centers conduct research and provide support to activities centered around the three pillars of data science — data sharing, data analysis, and human resource development. The facilities work to strengthen the research capabilities of universities while actively cooperating with communities from various fields of academia and society in order to make meaningful contributions to academic development and social innovation.



Collaborative research



The data-science related collaboration program "ROIS-DS-JOINT" is being carried out at the ROIS-DS. This program comprises "The Joint Research Program," which conducts collaborative research by utilizing the expertise and resources of each ROIS-DS research center. The program also comprises "The Joint Research Meeting Program," which conducts research exchanges, seminars, and more at each center; in 2019, these programs covered 43 topics, 37 for “The Joint Research Program” and 6 topics for "The Joint Research Meeting Program.” In addition to "ROIS-DS-JOINT," each ROIS-DS research center offers consultation services, striving to provide new opportunities for collaboration and joint research as well as support for researchers nationwide. Details concerning the ROIS-DS's activities are announced on the organization’s website, by research coordinators, and at symposia.

ROIS-DS 2019 - **Breakdown of participant affiliations** -
Applications to the ROIS-DS-JOINT program came from national and private universities, research institutes, technical colleges, and foreign universities (75 organizations and 134 individuals, excluding ROIS-DS researchers).

Database Center for Life Science (DBCLS)

The aim of DBCLS is the promotion of open science in the field of life sciences. The center conducts research and development in the field of database integration, which is necessary for the unified use of the diverse and rapidly expanding databases created and maintained by national universities, research institutes, and the like. We focus on developing the technology necessary for integration, including terminology and its classification system (ontology) as well as the development of linked data based on standardization of the data description method. In addition, we collaborate with experts from database development organizations around the world, hold international workshops such as the annual BioHackathon, and lead the development and standardization of integration technologies. (Director: Yuji Kohara)

Main Activities of the Center

► We promote the use of the Resource Description Framework (RDF) for life science databases to realize an environment that allows users to access databases distributed across the Internet in an integrated manner. We have been continuously supporting the conversion of existing data into RDF and accepting RDF data from various organizations to the NBDC RDF Portal. As an application of the RDF portal, a single website was opened at the National Institute of Technology and Evaluation (NITE) in FY 2019, which enabled the integrated search of the microbiological RDF data provided by the NITE and RIKEN. We also improved the user interface for TogoVar, a database released in 2018 that contains information on individual variants in genomic sequences collected from the Japanese population and related diseases, so that users can have easy access to more information.

► In addition to organizing the international developer workshop BioHackathon, we promote standardization of databases through events such as the domestic version of BioHackathon, SPARQLthon (held monthly) for RDF database developments, and BLAH (Biomedical Linked Annotation Hackathon) for text mining from literature. Approximately 120 participants from 13 countries gathered in Fukuoka, Japan for BioHackathon 2019, which was our 12th international workshop.

TogoVar

Group photo from the BioHackathon 2019 workshop
＊Organized in collaboration with the National Bioscience Database Center of Japan Science and Technology Agency

Polar Environment Data Science Center (PEDSC)

This center aims to help create new data-driven polar science and contribute to research on the global environment. We therefore promote the publication and shared use of valuable data acquired through surveys and observation activities in the Arctic and Antarctic regions, and facilitate the promotion of data science in the field of the global environment study. A wide variety of research and observation data has been gathered across various disciplines through the use of surveys and observation activities; however, owing to differences in the extent of database creation for each dataset as well as their publication, PEDSC has sought to create databases and archives of actual data as well as a unified database for the meta-information (metadata) for various data such as location information and attribute information. (Director: Akira Kadokura)

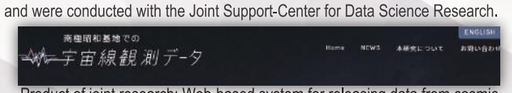
Highlights of 2019:

► We continued to register metadata, release data, and add DOIs to data through the Arctic and Antarctic Data archive System (AADS), Science Databases, and the IUGONET system.

Arctic and Antactic Data archive System （AADS）

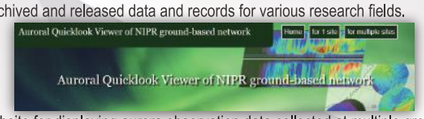
Science Database

► We supported data releases for joint research projects that were publicly solicited by and were conducted with the Joint Support-Center for Data Science Research.



Product of joint research: Web-based system for releasing data from cosmic ray observations at Syowa Station

► We archived and released data and records for various research fields.



Creation of a website for displaying aurora observation data collected at multiple ground stations

Creation of a digital archive system for historical records on the polar regions

Launch of a website that allows the general public to search for data on polar regions

Center for Social Data Structuring (CSDS)

Structured and organized social data helps understand the complexities of contemporary society. Moreover, it also finds utility in solving various problems in the fields of local environment, security, and economy. CSDS was set up with the objective to create a social data management platform in collaboration with the Science Council of Japan and the Ministry of Internal Affairs and Communications Statistics Bureau, as well as domestic and foreign survey organizations and research institutes. Specifically, we work on maintaining and improving the National Character Survey (Social Survey data), "Microdata of official statistics" that can be used onsite, as well as "Social Big Data" which shows human social behavior in real time. The activities at CSDS support the progress of humanities, social sciences, and evidence-based policy making. (Director: Tadahiko Maeda)

Highlights of 2019:

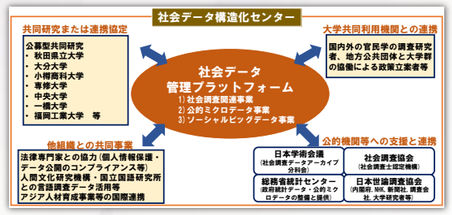
► We publish detailed cross tabulations, such as those from the East Asian Value Survey. We collaboratively use individual-level data from domestic surveys such as the Japanese National Character Survey in joint studies and publish papers and other research results.

► We operate the Onsite Data Analysis Room within the ROIS-DS, where users can access publicly available microdata collected by national and local governments.

► We trialed a social data management platform, which the Center aims to establish, refined its functions, and accumulated information to be disseminated through the platform.

► We organized a tutorial seminar in which a lawyer discussed from a legal standpoint the issues surrounding compliance at various stages, from the collection to the release of social survey data.

International Comparative Survey of Asia Pacific Values



Social Data Management Platform

Center for Open Data in the Humanities (CODH)

The Center of Open Data in the Humanities (CODH) aims to promote research and support activities based on openness and shared use of data in the field of humanities. The center conducts an analysis of characters and contents of data emerging from the massive digitization of Japanese culture, such as kuzushiji, which is pre-modern Japanese text from the Edo period and printed books since the Meiji period, using the latest technologies from informatics and statistics. Open data is essential to the development of data science in the field of humanities; however, its progress has been limited. Therefore, the center promotes openness by providing information infrastructure for sharing humanities data with the world, while collaborating with citizens, industries, and researchers from across disciplines. (Director: Asanobu Kitamoto)

Highlights of 2019:

► We improved a deep learning-based object detection algorithm to develop the KuroNet kuzushiji recognition service, which converts all the old Japanese cursive script within one page image into modern Japanese characters in about one second.

► We expanded the collection of facial expressions for Japanese picture scrolls and picture books. We used it not only for collaborative research between art history and computer science, but also for novel applications such as training data for machine learning as well as entertainment purposes.

► We held an international machine learning competition (Kaggle kuzushiji recognition) in which more than 300 participants competed for the best accuracy of their kuzushiji recognition algorithms. In addition, we expanded the existing datasets and released new datasets, such as the kuzushiji dataset, which was expanded to more than one million characters after the competition.

An example of kuzushiji recognition by the KuroNet service Original image (left; from Hyakunin Isshu Manyokan archived in the National Institute of Japanese Literature) and an image with recognized characters (right)

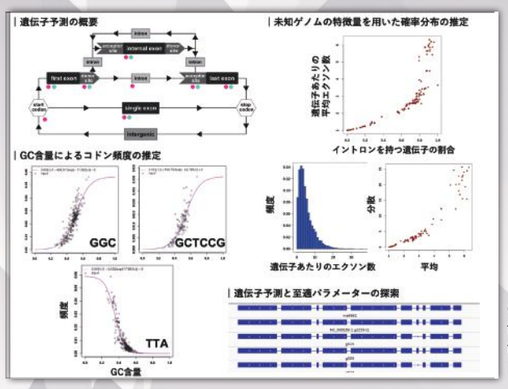
Facial expressions from Daikokumai (archived in the National Institute of Japanese Literature) selected from the collection of facial expressions. Facial images in the collection are searchable for a list view.

Center for Genome Informatics (CGI)

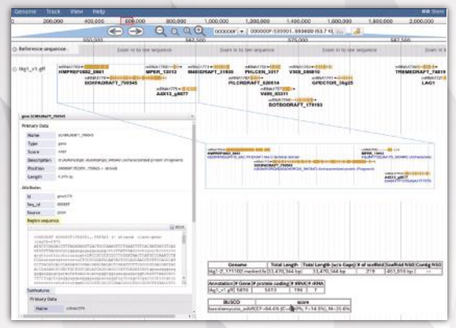
Decoding the information recorded in genome DNA is regarded as the starting point for biological research. However, genome data obtained using the latest technology is a dataset consisting of hundreds of millions of short sequences, each of which is approximately 300 letters, and the human genome in its entirety has 3 billion characters. Therefore, a structured data-driven approach to genome DNA analysis is critical to further biological research. The Center for Genome Informatics (CGI) develops and provides cutting-edge bioinformatics technology to extract new findings that lead to the development of researches across various disciplines such as in biology, medicine, and environmental science. The center supports researchers through analysis consultations and collaborative research activities. (Director: Hideki Noguchi)

Main Activities of the Center

- We support genome analysis such as the de novo genome sequencing and re-sequencing for various species (animals, plants, fungi, and prokaryotes).
- We are developing various genome analysis pipelines and genome browsers, as well as novel analysis methods such as a gene finding method and RNA-seq assembler.



Gene finding from unknown fungal genomes based on the statistical features of known fungal genes.



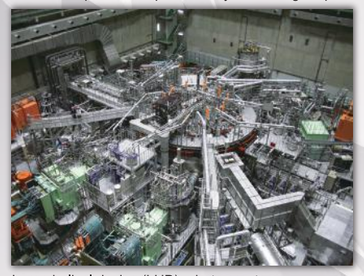
Genome browser:
enabling easy access to various genome-related contents.

Center for Data Assimilation Research and Applications (CARA)

In this center, efforts are devoted to the research and development of various methods of statistical mathematics that broaden the possibilities of simulation, such as "data assimilation," which integrates data and simulation, and "statistical emulators," which imitate simulation using statistical methods. In addition, we provide technical know-how for various technologies that unite simulation and statistical mathematics and support users experiencing issues pertaining the use of simulation in academia and the industry. (Director: Genta Ueno)

Highlights of 2019:

- We provided advice and technical consulting to introduce a time series model (a nonstationary renewal process) into earthquake forecasting.
- We provided advice and technical guidance on the implementation of a data assimilation method for simulations of liquid film flows.
- We succeeded in developing a data assimilation system for a large helical device (LHD) and gave a presentation as an invitee at a conference of the Japan Society of Plasma Science and Nuclear Fusion Research.
- We developed a computationally efficient geospatial modeling approach and released it in a package for R (statistical software).



Large helical device (LHD); photo courtesy of the National Institute for Fusion Science.

We presented this at our booth at an academic conference and gave advice to visitors about how to analyze their data.

Coordination of Research Activities

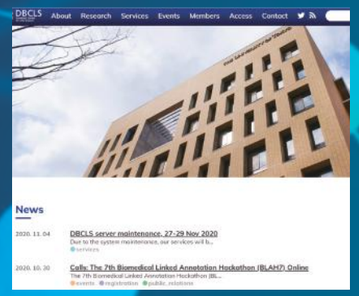
Research coordinators from the ROIS-DS lead activities such as conducting public relations at academic conferences, responding to inquiries, and supporting the initiation of joint research. We have supported research in a wide range of fields through exhibition booths at more than 30 academic conferences in biology, medicine, pharmacy, engineering, agronomy, environmental studies, earth and planetary science, statistics, and financial engineering.

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

Hands-On (Interactive workshops)

We periodically host various hands-on workshops including integrated database workshops (hosted by the JST NBDC, with the DBCLS as a co-host), RDF seminars (hosted by the DBCLS), CODH tutorials (hosted by the CODH) and hands-on sessions on data assimilation (hosted by CARA; see the photo). In addition, we conduct IUGONET seminars (hosted by the PEDSC) that include data comparison programs held both in Japan and overseas.

List of center website URLs



DBCLS Website
https://dbcls.rois.ac.jp/index-en.html

PEDSC Website
http://pedsc.rois.ac.jp/en/

CSDS Website
http://csds.rois.ac.jp/

CODH Website
http://codh.rois.ac.jp/index.html.en

CGI Website
https://genome-info.nig.ac.jp/

CARA Website
http://daweb.ism.ac.jp/cara/en/index.html

DBCLS services list page	https://dbcls.rois.ac.jp/services-en.html
TogoVar	https://togovar.biosciencedbc.jp/
BioHackathon Website	http://www.biohackathon.org/
Arctic Data archive System (ADS)	https://ads.nipr.ac.jp/
IUGONET	http://www.iugonet.org/index.jsp?lang=en
Polar data catalogue “Science database”	https://scidbase.nipr.ac.jp/?ml_lang=en
International comparative awareness survey	http://www.ism.ac.jp/~yoshino/
International comparative survey on Asia Pacific values	http://www.ism.ac.jp/~yoshino/ap2/index.html

On-site analysis room	http://ds.rois.ac.jp/center3_micro/
Public statistics microdata research consortium	http://ds.rois.ac.jp/center3_micro/moc/
Coelacanth genome browser	http://coelacanth.nig.ac.jp/
MetaGeneAnnotator (Metagenome annotator)	http://metagene.nig.ac.jp/
Platanus (Genome assembler)	http://platanus.bio.titech.ac.jp/
P3 (Python parallelized particle filter library)	http://daweb.ism.ac.jp/support/software/P-cubed/P-cubed.html