

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 Building 2F, 4-3-13, Toranomom, Minato-ku, Tokyo 105-0001, Japan  
Phone: +81-3-6402-6200  
<https://www.rois.ac.jp/en/>



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



## A base for promoting data science

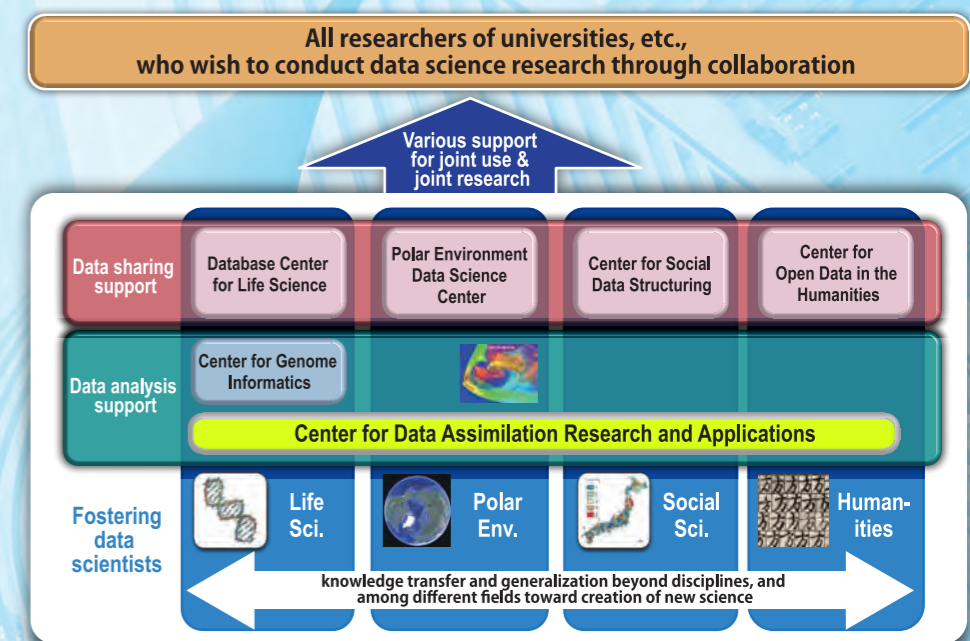
The Joint Support-Center for Data Science Research (ROIS-DS) is a joint-use, collaborative research center for the advancement of interdisciplinary data science to solve scientific and social challenges through advanced analysis of big data at a national scale. It was established by the Research Organization of Information and Systems (ROIS) in 2016 to strengthen collaboration and cooperation among universities and other institutions 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.



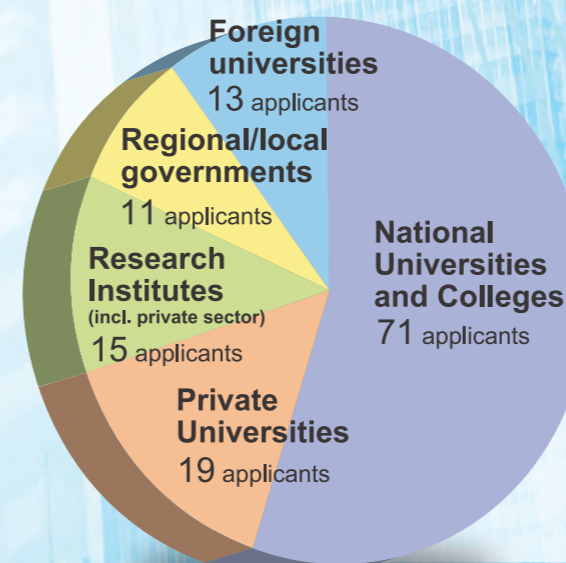
Data Science Building (Tachikawa campus)

## 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 2018, these programs covered 39 topics, 30 for "The Joint Research Program" and 9 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 2018 - Breakdown of participant affiliations -

Applications to the ROIS-DS-JOINT program came from national and private universities, research institutes, technical colleges, local governments, and foreign universities (71 organizations and 129 individuals, excluding ROIS-DS researchers).

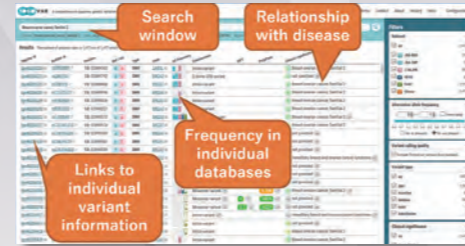


## Database Center for Life Science (DBCLS)

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 create an integrative environment that allows users to access databases distributed across the Internet. In 2018, for example, we released TogoVar, a database that integrates information on individual variants in genomic sequences collected from the Japanese population with other information such as known variants and diseases associated with those variants. TogoVar has been jointly developed with the National Bioscience Database Center, JST.
- ▶ 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 130 participants from 15 countries gathered in Matsue, Japan for BioHackathon 2018, which was our 11th international workshop.



TogoVar



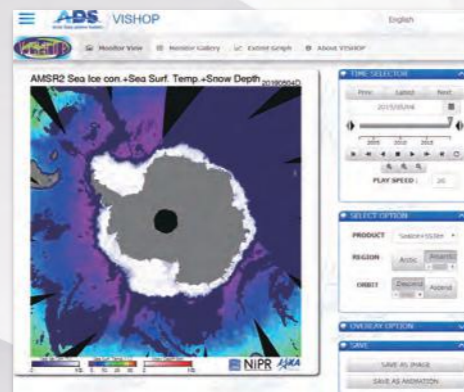
Group photo from the BioHackathon 2018 workshop  
\* Organized in collaboration with the National Bioscience Database Center of Japan Science and Technology Agency

## 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 2018:

- ▶ We developed the AADS system that extends the capabilities of the ADS system, which is widely used for archiving and publishing Arctic data, to deal with data from the both the Arctic and Antarctic polar regions. We also created and released databases that provide data from Earth observation satellites and rock sample data for the Antarctic region.
- ▶ We strengthened the meta-database system of the inter-university collaborative project IUGONET, which allows easy, cross-sectional search and analysis of various data in the field of upper atmospheric science that are held by various Japanese institutions. We also promoted cooperation with various research and observation projects in Japan.



ADS system



IUGONET

## Center for Social Data Structuring (CSDS)

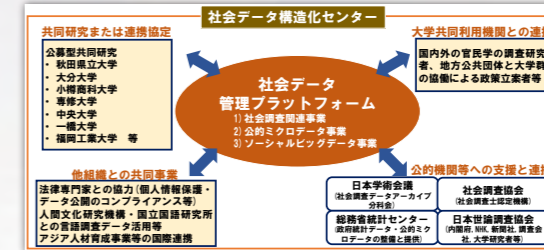
Structured and organized social data helps understand the complexities of contemporary society. Moreover, it also finds utility in solving various problems in the fields of local environment, security, and economy. CSDS was set up with the objective to create a social-data management platform in collaboration 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 from 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.

### Highlights of 2018:

- ▶ We publish data from surveys of Japanese attitudes, such as the Cross-National Comparative Survey on National Character.
- ▶ We operate the Onsite Data Analysis Room within the ROIS-DS, where users can access publicly available individual-level microdata collected by national and local governments.
- ▶ We opened a website for the CSDS that serves as an interface to a social-data management platform that the Center aims to establish. We also hosted "the Third Symposium on Structuring Human and Social Data."



International Comparative Survey of Asia Pacific Values



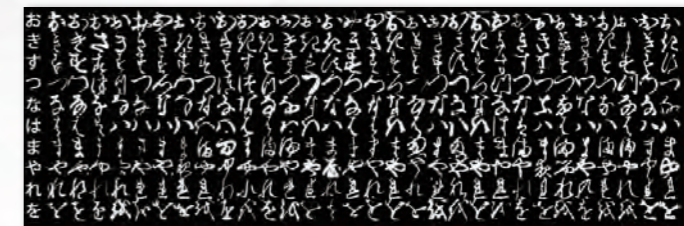
Social Data Management Platform

## 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 2018:

- ▶ We expanded the "Dataset of Pre-Modern Japanese Text," reaching approximately 3,000 volumes (including books such as Manyoshu) and also expanded the "Kuzushiji Dataset," reaching approximately 680,000 characters.
- ▶ We completed the release version of the "IIIF Curation Platform" which enables curation activities based on the International Image Interoperability Framework (IIIF), a protocol for image delivery used internationally by leading museums and libraries.
- ▶ We released new datasets and archives including the "Kuzushiji-MNIST Dataset" for machine learning (AI) on kuzushiji and "North China Railway Archive," which contains photographs for publicity activities during the Second Sino-Japanese War.



Kuzushiji-MNIST Dataset for machine learning (AI) of kuzushiji



North China Railway Archive (an archive of photographs for publicity activities during the Second Sino-Japanese War)

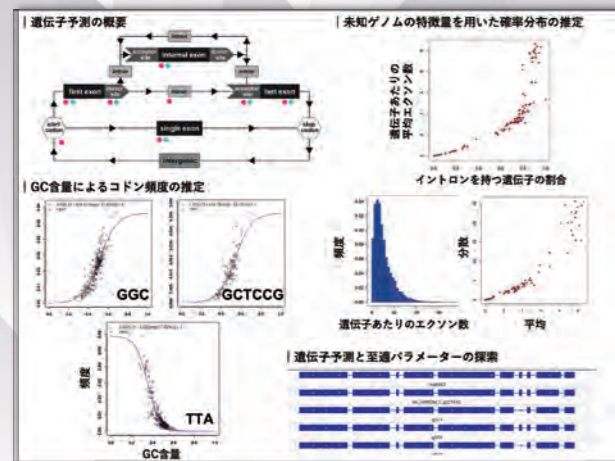


## Center for Genome Informatics (CGI)

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)

### Highlights of 2018:

- ▶ We provided analysis support for 18 genome-related research projects studying various organisms (11 species including animals, plants, fungi, and prokaryotes).
- ▶ We created a genome browser for searching and viewing information on the organism being analyzed.
- ▶ We created a genomics analysis pipeline and developed new analytical methods (gene prediction, detection of repeated sequences, etc.).



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



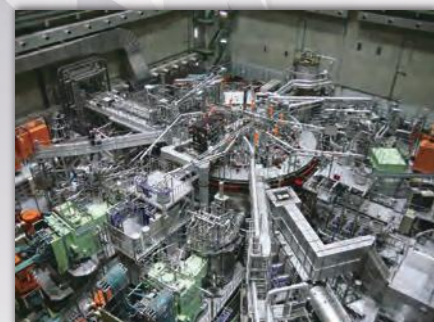
Genome browser: enabling easy access to various genome-related contents.

## 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 2018:

- ▶ We provided advice and technical guidance on implementing a data assimilation method for a heat transport simulation model for high-temperature plasma in the Large Helical Device (LHD) of the National Institute for Fusion Science.
- ▶ We provided advice and technical guidance on methods for statistically analyzing and making inferences about urban agglomeration based on urban economic models.
- ▶ We organized a seminar entitled "Time Series Analysis with the R Package KFAS: Hands-on Practice" at a conference of the Japanese Association of Risk, Insurance, and Pensions, in which the participants received a lecture on time series analysis and analyzed data using the personal computers they had brought.



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 Geoscientists 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/>



PEDSC Website  
<http://pedsc.rois.ac.jp/>



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



CODH Website  
<http://codh.rois.ac.jp/>



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



CARA Website  
<http://daweib.ism.ac.jp/cara/>

DBCLS Services list page	<a href="https://dbcls.rois.ac.jp/services.html">https://dbcls.rois.ac.jp/services.html</a>	On-site analysis room	<a href="http://ds.rois.ac.jp/center3_micro/">http://ds.rois.ac.jp/center3_micro/</a>
TogoVar	<a href="https://togovar.biosciencedbc.jp/">https://togovar.biosciencedbc.jp/</a>	Public statistics microdata research consortium	<a href="http://ds.rois.ac.jp/center3_micro/moc/">http://ds.rois.ac.jp/center3_micro/moc/</a>
BioHackathon Website	<a href="http://www.biohackathon.org/">http://www.biohackathon.org/</a>	Coelacanth genome browser	<a href="http://coelacanth.nig.ac.jp/">http://coelacanth.nig.ac.jp/</a>
Arctic Data archive System (ADS)	<a href="https://ads.nipr.ac.jp/">https://ads.nipr.ac.jp/</a>	MetaGeneAnnotator (Prokaryotic gene finder)	<a href="http://metagene.nig.ac.jp/">http://metagene.nig.ac.jp/</a>
IUGONET	<a href="http://www.iugonet.org/">http://www.iugonet.org/</a>	Platanus (Genome assembler)	<a href="http://platanus.bio.titech.ac.jp/">http://platanus.bio.titech.ac.jp/</a>
Polar data catalogue "Science database"	<a href="https://scibase.nipr.ac.jp/">https://scibase.nipr.ac.jp/</a>	P3 (Python parallelized particle filter library)	<a href="http://daweib.ism.ac.jp/support/software/P-cubed/P-cubed.html">http://daweib.ism.ac.jp/support/software/P-cubed/P-cubed.html</a>
Asia Pacific Values Survey	<a href="https://www.ism.ac.jp/~yoshino/ap2/index_e.html">https://www.ism.ac.jp/~yoshino/ap2/index_e.html</a>		
Cross-National Comparative Survey on National Character	<a href="https://www.ism.ac.jp/~yoshino/index_e.html">https://www.ism.ac.jp/~yoshino/index_e.html</a>		