International Workshop on Data Science

- Present & Future of Open Data & Open Science –

PROGRAMME and ABSTRACTS

Mishima Citizens Cultural Hall &
Joint Support-Center for Data Science Research,
Mishima, Shizuoka, Japan

12–15 November 2018

Table of Contents :

International Workshop on Data Science
- Present & Future of Open Data & Open Science –

Mishima Citizens Cultural Hall &
Joint Support-Center for Data Science Research, Mishima, Shizuoka, Japan
12–15 November 2018

Scope of Workshop :

The Workshop will focus on recent topics of interest in the field of scientific data, which are attributed to play a crucial role in accelerating "Open Science" and "Open Data" globally. Contributions from all scientific disciplines are welcome, including life and bio-science, social and human science, as well as polar science. Inter-disciplinary orientated topics on data management are especially encouraged.

A wide range of presentations will be given on topics of effective scientific data management spanning across the entire spectrum of data management - planning and policy, submission of primary and metadata, data sharing for facilitation of inter-disciplinary science, long-term preservation and stewardship with global and social perspectives.

Topics on industry-academia collaboration, education and capability building on data sciences, promoting "Open Science" via feedback to the public and archiving are also encouraged. Contributors will report on successes and challenges recently encountered, best practices and experiences learned and what is yet be done to ensure that we leave a data legacy. Fruitful discussions on data legacy and historical data issues for all branches of science are expected to give a new proxy for addressing data management issues and to achieve inter-disciplinary science linkages.

It is expected that this workshop will lead to mutual understanding of various aspects of data by different stakeholders and it will open new paths for pursuing activities in different fields of science. The activities are expected to play a central role in the promotion of inter-disciplinary sciences and new collaborative research paths based on multi-disciplinary data and directly contribute to global data activities based on the facilities provided by the "Joint Support-Center for Data Science Research (DS)" of the "Research Organization of Information and Systems (ROIS)".

Session Themes :

- **International data activity:** Various data-related aspects of accreditation schemes and their benefits, positives and negatives of current approaches of individual international initiatives, centres and networks, related data management planning, data policy, etc.
- **National data activity:** Various data-related aspects of accreditation schemes and their

benefits, positives and negatives of current approaches of individual national projects, centres and regional networks, related data management planning, data policy, etc.

- **Current status of data science:** Current status and on-going progress in the field of data science and related applications for individual disciplines and cross-disciplinary synergies. This includes a wide range of topics - databases, data systems, metadata schemes, vocabularies, ontologies, knowledge management, cloud computing, security, storage, repository practises and standards etc.

- **Current status of Inter-disciplinary science:** Current status and progress in relating to inter-disciplinary research activities; data sharing, real-time data handling and manipulation, virtual observatories, information and communications technology infrastructure protocols and architectures, sustainability and governance models.

- **Industry-academia collaboration, education and capability building:** Best practise on industry-academia collaboration, education and capability building in data science, data-driven knowledge transfer, data publication and journals, scientific awards and recognition schemes.

- **Legacy data, historical data, future on data science:** All aspects of data use evolution, legacy data, historical data and the potentials for enhancing scientific and non-scientific research developments through data sharing, citation and publication across disciplines.

Schedule Summary :

Monday 12 November 2018; Registration, Public Lecture, Icebreaker Party
Tuesday 13 November 2018; Workshop (day 1), Reception
Wednesday 14 November 2018; Workshop (day 2), Tour to NIG and DBCLS, Banquet
Thursday 15 November 2018; Workshop (day 3)

Local Organization Committee (LOC) : (* Chair)

Masanori Arita (National Institute of Genetics, ROIS)
Tomoya Baba (Data Science Promotion Section, DS, ROIS)
Susumu Goto (Database Center for Life Science, DS, ROIS)
Rue Ikeya (University Research Administrator Station, ROIS)
Akira Kadokura (Polar Environment Data Science Center, DS, ROIS)
*Masaki Kanao (Polar Environment Data Science Center, DS, ROIS)
Naoko Kato (Center for Social Data Structuring, DS, ROIS)
Asanobu Kitamoto (Center for Open Data in the Humanities, DS, ROIS)
Tadahiko Maeda (Center for Social Data Structuring, DS, ROIS)
Mari Minowa (Database Center for Life Science, DS, ROIS)
Shinya Nakano (The Institute of Statistical Mathematics, ROIS)
Takeru Nakazato (Database Center for Life Science, DS, ROIS)
Koji Nishimura (Polar Environment Data Science Center, DS, ROIS)
Hideki Noguchi (Center for Genome Informatics, DS, ROIS)

Akihiko Nomizu (Data Science Promotion Section, DS, ROIS)
Yoshimasa Tanaka (Polar Environment Data Science Center, DS, ROIS)
Hironori Yabuki (Polar Environment Data Science Center, DS, ROIS)

Advisory Committee (AC): (* Chair)

Phillippa Bricher (Australian Antarctic Division)
Taco De Bruin (Royal Netherlands Institute for Sea Research)
Shannon Christoffersen (University of Calgary)
Hiroyuki Enomoto (National Institute of Polar Research, ROIS)
*Asao Fujiyama (Joint Support-Center for Data Science Research, ROIS)
Øystein Godøy (Norwegian Meteorological Institute)
Kazuhiro Hayashi (National Institute of Science and Technology Policy)
Heidi J. Imker (Illinois University)
Toshihiko Iyemori (Kyoto University)
Yuji Kohara (Database Center for Life Science, DS, ROIS)
Ellsworth LeDrew (University of Waterloo)
Kassim S. Mwitondi (Sheffield Hallam University)
Yasuhiro Murayama (National Institute of Information and Communications Technology)
Tsuneo Odate (National Institute of Polar Research, ROIS)
Mark Parsons (Rensselaer Polytechnic Institute)
Peter Pulsifer (University of Colorado)
Hideaki Takeda (National Institute of Informatics, ROIS)
Seiji Tsuboi (Japan Agency for Marine-Earth Science and Technology)
Anton Van de Putte (Royal Belgian Institute for Natural Science)
Ryozo Yoshino (Center for Social Data Structuring, DS, ROIS)

Abbreviation :

DS: Joint Support-Center for Data Science Research
DBCLS: Database Center for Life Science
DDBJ : DNA Data Bank of Japan
NIG: National Institute of Genetics
ROIS: Research Organization of Information and Systems

Website :

https://ds.rois.ac.jp/article/dsws_2018/

Contact Address :

data.ws.loc-2018 @ nipr.ac.jp

Programme Summary :

Monday 12 November 2018	
15:30–17:30	Public Lecture (for general public, in Japanese)
16:00–19:30	Registration
18:30–19:30	Icebreaker party @ Mishima Citizens Cultural Hall

Tuesday 13 November 2018	
09:05–09:30	Opening Remarks
09:30–12:30	Session A: International and National data activity
12:30–14:00	Group Photo & Lunch
14:00–17:20	Session B1: Data science and Inter-disciplinary science
17:20–18:20	Poster Session
18:30–20:00	Reception @ Mishima Shoukou-Kaigi-Sho

Wednesday 14 November 2018	
09:05–11:20	Session B2: Data science and Inter-disciplinary science
11:20–13:20	Session C: Legacy data, Historical data, Industry-academia collaboration
13:20–14:50	Lunch & Poster
14:50–17:50	Visit to NIG & DBCLS
18:00–19:30	Banquet @ NIG Lecture Hall

Thursday 15 November 2018	
09:05–12:30	Session D: Education and capability building
12:30–14:00	Lunch & Poster
14:00–17:20	Session E: Future on data science
17:20–17:30	Closing Remarks

大学共同利用機関法人 情報・システム研究機構 Research Organization of Information and Systems

大学共同利用機関法人 情報・システム研究機構 データサイエンス共同利用基盤施設 Joint Support-Center for Data Science Research (DS)

DDBJ DNA Data Bank of Japan

For Participants :

Language :

The official conference language is English. No simultaneous interpretation service will be provided.

Session D (Education and capability building, 15 November) will be presented in Japanese, using slides with English text.

Name Badge :

Participants' name badges will be provided at the registration desk of the workshop. All participants are required to wear the badge throughout the workshop.

Stickers for side events will be pasted to the badges as soon as the payments have been made.

Group Photo:

After the morning session on 13 November (1st day, Session A), all participants are invited to take a "group photo" of the conference. The place is planned in front of the entrance of " Mishima Citizens Cultural Hall ".

WiFi service:

At the venue of "Mishima Citizens Cultural Hall", WiFi service is available. Detail information about SSID & Password will be given at the registration desk.

For Presenters :

Oral Presentation :

General presenters are allocated 20 minutes including questions and discussion time.
Keynote speakers are allocated 30 minutes including questions and discussion time.

A Window lap top PC is available for presentations, but presenters can use their own lap tops if they so wish.

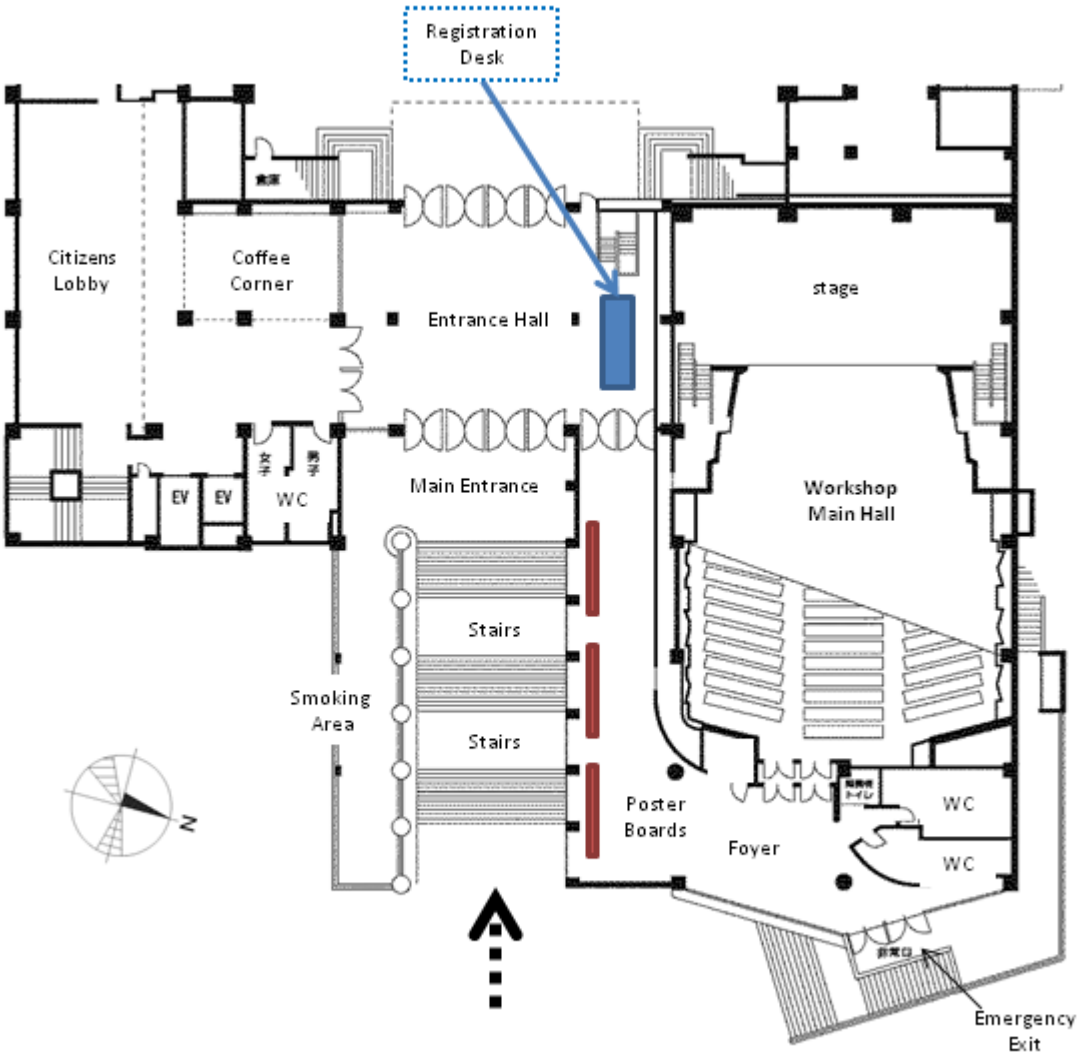
When using the PC of the conference room, please bring presentation file (ppt, pdf) on USB to the registration desk prior to the beginning of the session.

Poster Presentation :

Please prepare a poster of the maximum size within the posting board of 1200 mm (width) x 1800 mm (height).

All posters can be posted during three days on 13-15 November 2018, in front of the workshop main hall (1F, Mishima Citizens Cultural Hall).

Core Times for Poster presentations are allocated as follows;
Tuesday 13 November 17:20–18:20, & Wednesday 14 November 13:50-14:50
(for Sessions A, B, C and E)
Thursday 15 November 13:00-14:00 (for Session D)



Mishima Citizens Cultural Hall ("Mishimashimin Bunka Kaikan") 1F, Scale: 1/500

Registration :

Registration Desk :

Registration desk will open at the following date & time.
On-site registration can be acceptable for each day in front of the workshop main hall (1F, Mishima Citizens Cultural Hall).
No registration fee is required to attend the conference.

- Monday 12 November 2018 16:00-19:30
- Tuesday 13 November 2018 08:45-18:00
- Wednesday 14 November 2018 08:45-14:30
- Thursday 15 November 2018 08:45-16:00

Pre-Conference Registration :

Those who are intending to attend the workshop prior to the conference date can be made from the following URL;

https://docs.google.com/forms/d/e/1FAIpQLSc6kNOrCAvXMX0DVA42eFDF67IBpDkzTvfK8AyQ--tLrRzmWw/viewform

The Organizing Committee add to the participant list & prepare name badges before the conference.

Venue & Access :

Conference Venue :

Mishima Citizens Cultural Hall (“Mishimashimin Bunka Kaikan”)
Ichiban-cho 20-5, Mishima, Shizuoka, Japan
http://mishima-youyouhall.com/access/

It takes 3 min. southward walk from Mishima JR station to the Citizens Cultural Hall.
The Citizens Cultural Hall is surrounded westward by Mishima Municipal Park (“Rakujyuen”)
http://www.city.mishima.shizuoka.jp/rakujyu/index.html

Travel Information :

From Narita Airport to Mishima JR station:
https://www.nig.ac.jp/nig/pdf/access/Narita-NIG.pdf

From Haneda Airport to Mishima JR station:
https://www.nig.ac.jp/nig/pdf/access/Haneda-NIG.pdf

It takes about 2 hours from both the airports to Mishima.

Accommodation :

Lunch Service:

There are many restaurants near the Mishima JR station and the Citizens Cultural Hall. Lunch will not be served at the conference, delegates are advised to use the restaurants within the vicinity of Mishima JR station and the Citizens Cultural Hall.

Hotel Accommodation :

As there are several business hotels around Mishima central city area, participants are required to make their own reservations for accommodation.

The LOC is planning to reserve several rooms at discount rates for invited speakers and foreign participants in “Mishima Plaza Hotel” located within 6 minutes Walk south of the Mishima Citizens Cultural Hall.
https://www.mishimaph.co.jp/

Social Events :

There will be several social events as follows (public lecture, icebreaker party, banquet, reception, tour to NIG & DBCLS, etc.).

- **Public Lecture** (12 Nov. 15:30–17:30). The lecture is planned for the general public and it will be delivered in Japanese at the Mishima Citizens Cultural Hall.
  https://ds.rois.ac.jp/post-2635/

- **Icebreaker party** (12 Nov. 18:30–19:30). This will take place at the Mishima Citizens Cultural Hall, at the cost of 1,000 JPY per delegate. Drinks and light snacks will be served.

- **Reception** (13 Nov. 18:30–20:00). This will take place at the "Mishima Shoukou-Kaigi-Sho (1F TMO Hall)", at the cost of 3,000 JPY per delegate. The TMO Hall locates in front of Mishima Citizens Cultural Hall, at the opposite side of their facing road.

- **Tour to NIG & DBCLS** (14 Nov. afternoon 14:50–17:50). Details are demonstrated in separated item as below and related map of the NIG campus.

- **Banquet** (14 Nov. 18:00–19:30). To take place at the Lecture Hall (2F of NIG restaurant), delegates will be required to pay 3,000 JPY.

All payments for attending the foregoing events must be made at the registration desk of the conference & at the desks for individual events.

More detailed information will be announced at the conference venue.

Tour to NIG & DBCLS :

Tour to the National Institute of Genetics (NIG)
& Database Center for Life Science (DBCLS)

(14 November 2018, afternoon 14:50–17:50)

TIME TABLE :

14:50 - **Participants** are expected to gather at the main entrance of the Mishima Citizens Cultural Hall before this time. Please choose the transportation route either using A) public bus transportation or B) shuttle bus service of NIG.

15:00 - Take a **public bus transportation ("City bus")** to NIG at the bus stop 1 min. southward of the Mishima Citizens Cultural Hall (250 JPY for each half way ride). It takes about 15 min. to NIG. The delegates who are not in time to take the public bus, please use a Taxi to go to NIG.

- 15:30 **Arrive at the entrance of NIG**. Depending on the number of people, participants will be divided into one or two group(s) & migrate the following five "SPOT" places inside the campus.

15:30 - **SPOT 1:** General introduction of NIG at the "NIG Museum Exhibition Hall" besides the entrance (by Project Associate Prof. Mitsuhiko Kurusu, Office for Research Development).

16:00 - **SPOT 2:** Introduction of Advanced Genomics Center at the Computer Building (West-1F) (by Prof. Ken Kurokawa, Director of Advanced Genomics Center).

16:30 - **SPOT 3:** Introduction of Center for Genome Informatics & DBCLS at the Computer Building (West-3F) (by Prof. Hideki Noguchi, Director of Center for Genome Informatics; & Dr. Takeru Nakazato of DBCLS).

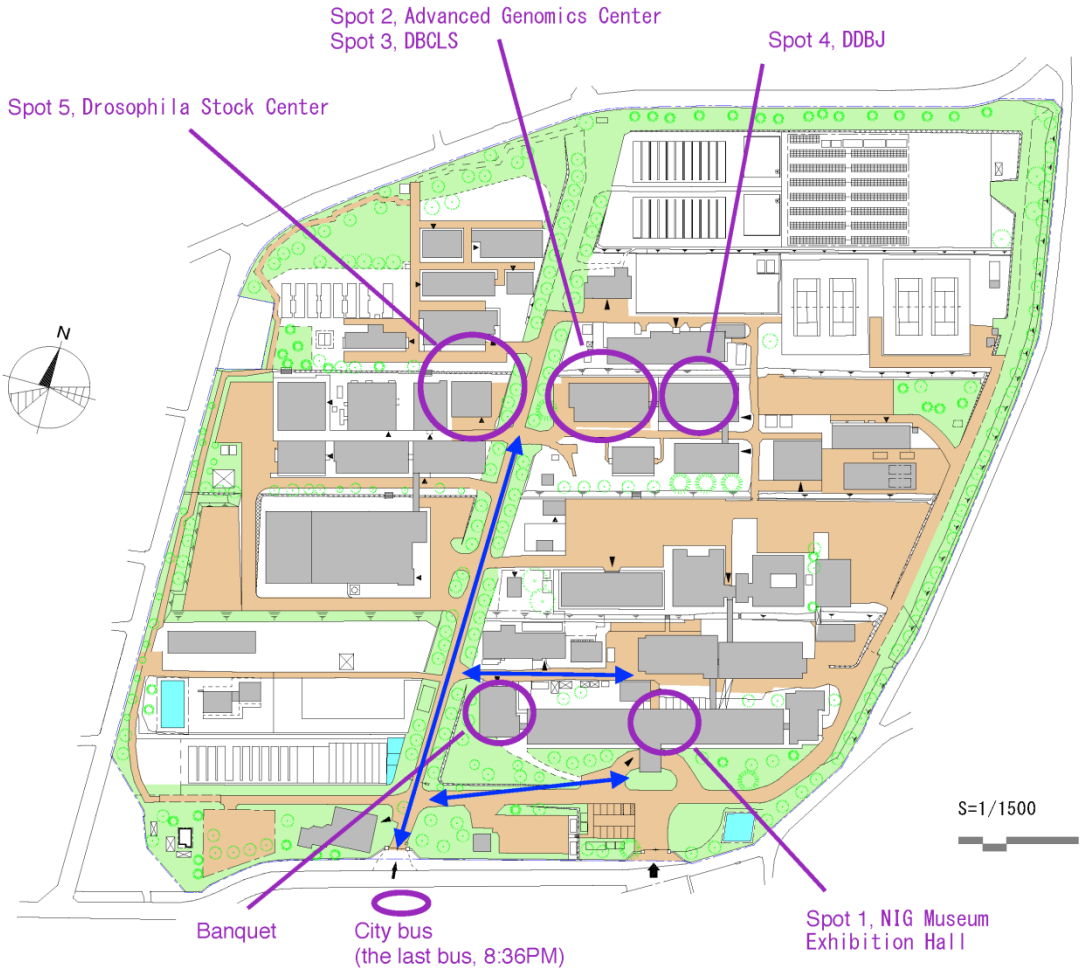
17:00 - **SPOT 4:** Introduction of the DNA Data Bank of Japan (DDBJ) at the Computer Building (East-4F) (by Prof. Masanori Arita, Director of DDBJ center).

17:30 - **SPOT 5:** Introduction of Drosophila Stock Center (by Prof. Kuniaki Saito, Director of Invertebrate Genetics Laboratory).

18:00 - **Banquet**: take place at the Lecture Hall (2F of NIG restaurant), delegates will be required to pay 3,000 JPY.

17:46, or 19:06, or 20:36 - Take a **public bus transportation ("City bus")** to the southern area near Mishima JR Station, from the road just outside NIG (250 JPY for each half way ride, It takes about 15 min.).

Detail timetable from NIG to JR Mishima Station can be checked in ;
https://www.nig.ac.jp/nig/pdf/access/busE-NIG.pdf



International Workshop on Data Science

- Present & Future of Open Data & Open Science –

PROGRAMME

Mishima Citizens Cultural Hall &
Joint Support-Center for Data Science Research,
Mishima, Shizuoka, Japan

12–15 November 2018

PROGRAMME

International Workshop on Data Science
- Present & Future of Open Data & Open Science –

12 – 15 November 2018

Mishima Citizens Cultural Hall & Joint Support-Center for Data Science Research, Mishima, Shizuoka, Japan

Monday 12 November 2018
15:30–17:30 Public Lecture (for general public, in Japanese)
16:00–19:30 Registration @ Mishima Citizens Cultural Hall
18:30–19:30 **Icebreaker party @ Mishima Citizens Cultural Hall** (Facilitator: Tomoya Baba) • **Welcome Remark: Takuji Nakamura** (Director-General, National Institute of Polar Research, ROIS)

Tuesday 13 November 2018
09:05–09:30 **Opening Remarks** (Chair: Masaki Kanao) • **Asao Fujiyama** (Director-General, Joint Support-Center for Data Science Research (DS), ROIS) (15') • **Masaki Kanao** (LOC Chair), agenda of workshop & practical information (10')
09:30–12:30 **Session A: International and National data activity** (Session chair: Akira Kadokura, Seiji Tsuboi) **Keynote 1:** *The International Seismological Centre (ISC): 50-Year Miracle* **Dmitry A. Storchak** (International Seismological Centre) (30') • *Data Management at Korea Polar Data Center, Korea Polar Research Institute* **Dongchan Joo** (Korea Polar Research Institute) (20') • *Activities of World Data Center for Geomagnetism, Kyoto* **Satoshi Taguchi** (Data Analysis Center for Geomagnetism and Space Magnetism, Kyoto University) (20') • **Coffee Break** (20') **Keynote 2:** *World Data Center for Microorganisms: The global cooperation on microbial big data* **Juncai Ma** (Chinese Academy of Sciences) (30') • *DIAS Platform Contributing to Open Science in Earth Environmental Informatics* **Masaki Yasukawa** (Earth Observation Data Integration and Fusion Research Initiative, the University of Tokyo) (20') • *Data Management at Polar Research Institute of China* **Lizong Wu** (Polar Research Institute of China) (20') • *Activities of Polar Environment Data Science Center* ***A*kira Kadokura** (Polar Environment Data Science Center, DS, ROIS) (20') • **Discussion**

12:30–14:00 **Group Photo & Lunch**

14:00–17:20 **Session B1: Data science and Inter-disciplinary science**

(Session chair: Susumu Goto, Takashi Watanabe)

Keynote 1: *Amenability of the United Nation's Sustainable Development Goals to Big Data Modelling*

Kassim S. Mwitondi (Sheffield Hallam University) (30')

- *Life Science Database Integration and Its Application for Medical Science*
  Susumu Goto (Database Center for Life Science, DS, ROIS) (20')
- *Current Situation of Data Archiving for Japanese Official Statistics*
  Shinsuke Ito (Chuo University) (20')
- *Importance of semantics and rated challenges in science and humanities according to the Humboldtian ideal*
  Bernd Ritschel (Kyoto University) (20')
- **Coffee Break** (20')

Keynote 2: *SPEDAS (Space Physics Environment Data Analysis Software): Multi-mission Heliophysics Data Management, Analysis, Visualization, and Collaboration*

Jim Lewis (Space Sciences Laboratory, University of California) (30')

- *Application of deep learning and large scale simulation to the Earth science*
  Seiji Tsuboi (Japan Agency for Marine-Earth Science and Technology) (20')
- *Domestic and international activities of DOI-minting to solar- terrestrial physics data and their citation in publication*
  Masahito Nose (Nagoya University) (20')
- *Multidisciplinary Study of the Earth's Environment in 18th-19th Centuries - A Trial to find an Approach to the Open Data and Open Science*
  Takashi Watanabe (International Program Office, ICSU World Data System) (20')
- **Discussion**

17:20–18:20 **Poster Session**

18:30–20:00 **Reception @ Mishima Shoukou-Kaigi-Sho** (Facilitator: Masaki Kanao)

- **Welcome Remark: Ryoichi Fujii** (President, Research Organization of Information and Systems (ROIS))

Wednesday 14 November 2018

09:05–11:20 **Session B2: Data science and Inter-disciplinary science**

(Session chair: Hideaki Takeda, Masaki Kanao)

Keynote 3: *Data Archives for Social Science in Korea, Challenges and Opportunities*

Hearan Koo (Seoul National University) (30')

- *Exploring public attitudes toward scientific research with visitor surveys and nationally representative surveys*
  Naoko Kato-Nitta (Center for Social Data Structuring, DS, ROIS) (20')

Keynote 4: *Age-Period-Cohort Analysis of Data Obtained from Repeated Social Surveys Such as the Surveys on the Japanese National Character*

Takashi Nakamura (Center for Social Data Structuring, DS, ROIS) (30')

- *Extended Cell Suppression Problem Towards Better Data*
  Kazuhiro Minami (The Institute of Statistical Mathematics, ROIS) (20')
- *Estimation of age-specific reporting ratio of sentinel influenza surveillance using seroprevalence data*
  Masaya Saito (The Institute of Statistical Mathematics, ROIS) (20')

- **Coffee Break** (15')

11:20–13:20 **Session C: Legacy data, Historical data, Industry-academia collaboration**
(Session chair: Koji Nishimura, Naoko Kato-Nitta)

Keynote 1: *The ISC-GEM Global Earthquake Catalogue: Making Good Use of Historical Observations*
Dmitry A. Storchak (International Seismological Centre) (30')

- *Reframing Scholarly Communication by Persistent Identifier*
  Hideaki Takeda (National Institute of Informatics, ROIS) (20')

Keynote 2: *Resource and Environment Scientific Data Sharing and Disaster Risk Reduction Knowledge Service*
Juanle Wang (Chinese Academy of Sciences) (30')

- *Data Treatment Strategies and Some Science Projects of PANSY Rader*
  Koji Nishimura (Polar Environment Data Science Center, DS, ROIS) (20')
- *Research Data Management in Industry-Academia Collaboration*
  Yuko Toda and Hodaka Nakanishi (Teikyo University) (20')
- **Discussion**

13:20–14:50 **Lunch & Poster**

14:50–17:50 **Visit to NIG & DBCLS**

18:00–19:30 **Banquet @ NIG Lecture Hall** (Facilitator: Hideki Noguchi)

- **Welcome Remark: Isao Katsura** (Director-General, National Institute of Genetics, ROIS)

Thursday 15 November 2018

09:05–12:30 **Session D: Education and capability building** (presented by Japanese, using English slides)
(Session chair: Masanori Arita, Yoshimasa Tanaka)

Keynote 1: *Health - Prediction of Infectious Disease System Dynamics using Machine Learning and Mathematics*
Shailza Singh (National Centre for Cell Science, India) (30')

- *Open genome analysis in the post-genomic era*
  Masanori Arita (National Institute of Genetics, ROIS) (20')
- *Education Programs for AI/IoT/BigData Skills Development at the Medical and Drug Discovery Data Science Consortium*
  Eli Kaminuma and Hiroshi Tanaka (Tokyo Medical and Dental University) (20')
- *Kyoto University Academic Data Innovation Unit - bottom-up promotion for various research data infrastructure*
  Takaaki Aoki and Shoji Kajita (Kyoto University) (20')

- **Coffee Break** (15')

Keynote 2: *From Human Genome Project to Genome Cohort Study*

Atsushi Shimizu (Iwate Tohoku Medical Megabank) (30')

- *Genetic relationship between tea plant (Camellia sinensis) and ornamental camellia (C. japonica)*
  Kazumi Furukawa (National Institute of Technology, Numazu College) (20')
- *Research outline with regional theme: Examples using machine learning, geographic information system, and social big data*
  Shizuo Suzuki (National Institute of Technology, Numazu College) (20')
- *Useful Tools for Education and Capacity Building about Solar Terrestrial Physics Study*
  Yoshimasa Tanaka (Polar Environment Data Science Center, DS, ROIS) (20')

- **Panel discussion** (10')

12:30–14:00 **Lunch & Poster**

14:00–17:20 **Session E: Future on data science**

(Session chair: Asanobu Kitamoto, Hironori Yabuki)

Keynote 1: *Embedding Machine Learning in the Data Life Cycle - An Example from Minerals Exploration*
Jens Klump (Mineral Resources, CSIRO in Australia) (30')

- *Toward universal information access on the digital object cloud*
  Kei Kurakawa (National Institute of Informatics, ROIS) (20')
- *Development of Japanese Research Data Discovery Service*
  Fumihiro Kato (National Institute of Informatics, ROIS) (20')
- *AI and Data Science - Machine Learning as Digital Catalyst for Data Curation*
  Asanobu Kitamoto (Center for Open Data in the Humanities, DS, ROIS) (20')

- **Coffee Break** (20')

Keynote 2: *Open Research Data Progress in Korea*
Myung-Seok Choi (Korea Institute of Science and Technology Information) (30')

- *The Cooking Recipes without Border Data set: FAIR challenges*
  Frederic Andres (National Institute of Informatics, ROIS) (20')
- *Arctic Data Archive System (ADS)*
  Hironori Yabuki (Polar Environment Data Science Center, DS, ROIS) (20')
- *Perspective of Open Science in Japan driven by Integrated Innovation Strategy*
  Kazuhiro Hayashi (National Institute of Science and Technology Policy) (20')

- **Discussion**

17:20–17:30 **Closing Remarks** (LOC Chair)

Poster Presentation @ outside the Conference Hall

Core Time 1; Tuesday 13 November 17:20–18:20, Wednesday 14 November 13:50-14:50

- *P-1: Multivariate analysis of the occupations of rental rooms by using the housing information website data*
  Hayafumi Watanabe (Center for Social Data Structuring, DS, ROIS)
- *P-2: Useful Tools for Education and Capacity Building about Solar Terrestrial Physics Study*
  Yoshimasa Tanaka (Polar Environment Data Science Center, DS, ROIS)
- *P-3: Arctic Data Archive System (ADS)*
  Hironori Yabuki (Polar Environment Data Science Center, DS, ROIS)
- *P-4: Data Processing and Archive System for the Antarctic PANSY Rader*
  Koji Nishimura (Polar Environment Data Science Center, DS, ROIS)
- *P-5: "Polar Data Journal" : A new data publishing platform for polar science*
  ***A*kira Kadokura** (Polar Environment Data Science Center, DS, ROIS)
- *P-6: A decade of history for polar data management in Japan*
  Masaki Kanao (Polar Environment Data Science Center, DS, ROIS)
- *P-7: Data and metadata sharing among AFoPS countries*
  Masaki Kanao (Polar Environment Data Science Center, DS, ROIS)
- *P-8: Development of Metadata, Conversion and Archiving of the Time Series Data of the Completed Censuses and Surveys of the BBS*
  Chandra Shekhar Roy (Bangladesh Bureau of Statistics)
- *P-9: An attempt for the thermal transport modelling of fusion plasmas based on the statistical approach*
  Masayuki Yokoyama (National Institute for Fusion Science)

Core Time 2; Thursday 15 November 13:00-14:00 (for Session D)

- *P-10: A Python library for parallelised particle filters*
  Shinya Nakano (The Institute of Statistical Mathematics, ROIS)
- *P-11: A Study on Speaker Discrimination Using Machine Learning*
  Wataru Murata and ManYong Jeong (National Institute of Technology, Numazu College)
- *P-12: A Study on Speaker Discrimination by Deep Learning*
  Shinpei Mine and ManYong Jeong (National Institute of Technology, Numazu College)
- *P-13: A Study on Tapping Diagnosis of Structures Using Machine Learning*
  Ken Higuchi and ManYong Jeong (National Institute of Technology, Numazu College)
- *P-14: Development of Traffic Flow Data Measurement System by Drive Recorder*
  Hironori Kanazawa and ManYong Jeong (National Institute of Technology, Numazu College)
- *P-15: Optimization of Pipe Support Arrangement by Machine Learning*
  Hisashi Sakashita and ManYong Jeong (National Institute of Technology, Numazu College)
- *P-16: Genomic analysis in somatic embryogenesis of tea plant(Camellia sinensis)*
  Kaito Yamaki and Kazumi Furukawa (National Institute of Technology, Numazu College)
- *P-17: Preliminary estimation of the aesthetic value of cultural ecosystem services by mapping geotagged photos from social media data around Mt. Fuji and Izu Peninsula*
  Naoto Unno and Shizuo Suzuki (National Institute of Technology, Numazu College)
- *P-18: Preliminary classification of Japanese cedar forests with aerial photographs using machine learning approach*
  Ochiai Ryosuke and Shizuo Suzuki (National Institute of Technology, Numazu College)
- *P-19: Preliminary estimation of tea leaf quality using remote sensing techniques in Numazu city*

Tomoaki Okatsu and Shizuo Suzuki (National Institute of Technology, Numazu College)

- *P-20: Preliminary application of results in Japanese historical character recognition by machine learning to locally historical documents*
  Masuhiro Yamaji and Shizuo Suzuki (National Institute of Technology, Numazu College)

International Workshop on Data Science

- Present & Future of Open Data & Open Science –

ABSTRACTS

Mishima Citizens Cultural Hall &
Joint Support-Center for Data Science Research,
Mishima, Shizuoka, Japan

12–15 November 2018

The Cooking Recipes without Border Data set: FAIR challenges

***Frederic Andres*[1]*, *Laurent D'Orazio*[2], *Johyn Papin*[3], *William I. Grosky* [4]**

[1*] *NII, Chiyoda-ku, Tokyo, Japan*
[2] *IRISA, Lannion, France*
[3] *University of Rennes, Lannion, France*
[4] *University of Michigan-Dearborn, Michigan, USA*
Email: andres@nii.ac.jp

Summary. The following is a set of guidelines for preparing an extended abstract submission for the "IOne of the grand challenges of Open Data and Open Science is to facilitate knowledge discovery and sharing by collaboration between human and machines. We review how we apply the standard FAIR reference architecture for Big Data Governance and Metadata Management to the CRWB1 dataset that is scalable. We review how the dataset supports the Findability, Accessibility, Interoperability, and Reusability without worrying about the ingestion of new data source and different structures. One challenge is to enable data integration/mashup among heterogeneous food/cooking recipes dataset and make data discoverable, accessible, and usable through a machine readable and actionable standard data infrastructure.

Keywords. Food, Cooking recipes, FAIR data model, Collective Intelligence.

1. Introduction

Food science is the study of the physical, biological, and chemical makeup of food; and the concepts underlying food processing. Recipes are perhaps the simplest examples of the connection between food science and cooking: Cooking is chemistry and Physics as there is a fusion of Cooking and *Science*, ingredients and measurements, instructions and written documentation, all designed to lead anyone to a specific, repeatable outcome that someone else has also perfected.

2. CRWB Data Set

The CRWB[1][2] data Set is a huge linked open data collection [2] of 60 years collection of cooking recipes and a CRWBvoc set of ontologies including an ingredients ontology, recipe taxonomy, a cooking process-centric ontology, and a food tasting ontology. This allows accessibility to other researchers who want to investigate in several fields (dish recipe generation, cooking execution plan optimization, recipe DNA coding, recipe similarity), find correlations between cooking recipe, nutrition issues, and food tasting, or look into whatever topic they are interested in.

3. A FAIR Data Set

The FAIR Model [1] is a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable.

The CRWB dataset is Findable: (1) its metadata and data have been assigned a globally unique and eternally persistent identifier; (2) it is described with rich metadata; (3) its metadata are indexed in a searchable service; and (4) its metadata specify the CRWB identifier.

The CRWB dataset is Accessible: (1) its metadata and data are retrievable by their identifier using a standardized CRWB protocol. (2) this CRWB protocol is open, free, and universally implementable; (3) this CRWB protocol (e.g. [3]) allows for an authentication and authorization procedure, where necessary and (4) its metadata are accessible, even when the data are no longer available.

[1] https://github.com/fredericandres/CRWB-Research-Group/wiki/1)-CRWB-RSbench-Introduction
[2] Cooking Recipes Without Border Dataset

The CRWB dataset is Interoperable: (1) the CRWB (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation (see the CRWB data structure in Fig 1); (2) the CRWB (meta)data use the CRWBvoc ontology that follow FAIR principles; and (3) the CRWB metadata and data include qualified references to other (meta)data.

The CRWB dataset is Re-usable: (1) the CRWB metadata and data have a plurality of accurate and relevant attributes; (2) (meta)data are released with a clear and accessible data usage license; (3) (meta)data are associated with their provenance; and (4) (meta)data meet food/cooking recipe domain-relevant community standards.

4. Big Data Governance Mgt Challenges

Publicly available cooking recipe data and their associated nutritional and health-related information are generated from multiple organizations, each with their own methodology of data collection and dissemination. Mechanisms and rights for accessing enhanced cooking recipe data from such diversified organizations can be challenging.

5. Big Data Metadata Mgt Challenges

Cooking recipe collections as raw data or as linked open data have different metadata structures from different data hubs. Promoting interoperable semantic meaning and syntactic structures of these data fields can be challenging. The CRWB data will help to ease the Big Data Mashup including IoT mashup tools as the integration of heterogeneous cooking recipes and applications from multiple sources including IoT for health care and other research purposes.

6. Big Data Mashup Challenges

The CRWB data will help to ease the Big Data Mashup including IoT mashup tools as the integration of heterogeneous cooking recipes and applications from multiple sources including IoT for health care and other research purpose.

7. Conclusions

The CRWB dataset follows the FAIR model for open data science and is used in the benchmarking initiative of CRWB2 research group dedicated to the evaluation of new algorithms and technologies for cooking recipe recommendation, access, and exploration. It offers tasks that are related to human and social aspects of cooking recipe and healthy eating decision-making to the research community.

Acknowledgments. We thank the NII for its support to the CRWB project.

References

1. Wilkinson, M. D., et al., The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 15 March 2016, doi:10.1038/sdata.2016.18, retrieved online on 21 June 2016
2. Andres, F., The CRWB RSbench: Towards a Cooking Recipe Benchmark Initiative. Data Engineering Meets Intelligent Food and Cooking Recipe Workshop (DECOR 2018), ICDE 2018, CNAM, Paris, France; 04/2018, DOI:10.1109/ICDEW.2018.00032, 2018
3. Papin, J., Andres, F., D'Orazio, L., A Method to build a Geolocalized Food Price Time Series Knowledge Base analyzable by Everyone. LADaS - Latin American Data Science Workshop 2018, VLDB workshop 2018., Rio de Janeiro, Brazil; 08/2018

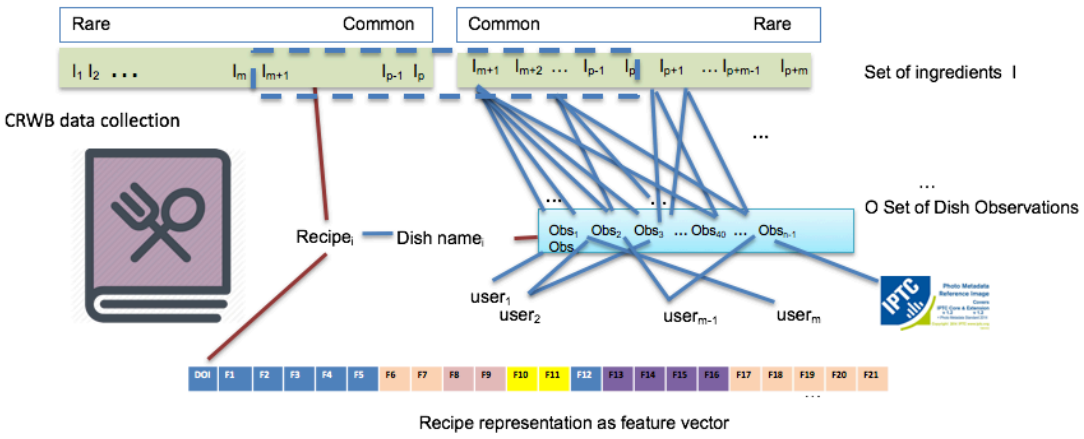


Fig 1: Example of the CRWB data structure

Kyoto University Academic Data Innovation Unit - bottom-up promotion for various research data infrastructure

Takaaki Aoki [1]*, ***Shoji Kajita*** [1]

[1]* *Kyoto University, Sakyo Kyoto 606-8501, Japan*
Email: aoki.takaaki.6v@kyoto-u.ac.jp

Summary. Kyoto University has established the "Academic Data Innovation Unit" in November 2017. The goal of this unit (known as "Kudzu Unit") is the bottom-up promotion of open data science, which is initiated in various scientific fields. In this paper, the background, motivation and current activities of the Kudzu unit are reviewed.

Keywords. Interdisciplinary collaboration, research data management, academic research data.

1. Research data openness promoted by the Japanese Government

From the view of Japanese academia, the openness of research data has been discussed based on two aspects. One is the strengthening of research fairness or integrity. Around 2014, Japanese academia was shocked by high-profile incidents of scientific misconduct. Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT) revised the old previous guideline against misconduct and issued "Guidelines for Responding to Misconduct in Research" [1], which mandates that researchers should preserve research data for a certain period and disclose it upon request. The second aspect is the promotion of open science. The term "open research data" firstly showed in A joint statement by the G8 Science Ministers in 2013. Then, the Japanese politics and business establishments issued several guidelines and statements to promote open science and initiate innovation [2]. Especially in June 2018, Japanese cabinet office issued a guideline for Japanese research and development agencies to establish data policy to lead open science activities in each scientific field [3].

2. Open research activities at Kyoto University

Kyoto University (KU) is the second largest national university and has been leading the openness of academic knowledge. For example, the KU library is the first one that adopted the open access policy on research papers. A large number of departments and research institutes have a vast amount of research databases accessible to the public. Based on the context of research integrity, KU has developed rules and organized systems to preserve research data [4] to prove to the integrity of research publications. These actions for enforcement of research integrity proceeded in a top-down scheme. On the other hand, it is considered that policy and system development for open research data at university initiated in a bottom-up project. The academic research data at KU are very diverse, which means that the development of comprehensive rules or systems covers all research filed is not possible.

3. The academic data innovation unit

The academic data innovation unit (known as "Kudzu Unit") is organized in November 2017 under the "Center for the Promotion of Interdisciplinary Education and Research" of KU. The mission of the unit is a preliminary study of policy and system development for university-wide research data management, which includes the acquisition, preservation, analyze, sharing,

and publication of research data. Unit members consist of various stakeholders around research data, which includes not only the faculties or researchers engaged with the academic databases but including the staffs of the university library, museum, research administration office, and IT management division.

The main activity of the Kudzu unit is to organize workshops annually. "The 5th open science data workshop" is held, on March 2018. The previous series of the workshop has been arranged by Kyoto University Geomagnetism Center, which is pioneering open data science activities in geoscience. Kudzu unit took over the management and framework of the workshop, which gathers the reports on the current progress of activities of open data science in various fields.

The next workshop will be scheduled in October 2018, which focuses on research data management (RDM) maturity assessment. The RDM rubric has been proposed by the California Digital Library [5]. It defines the 6 phases in research data lifecycle and 4 degrees of maturity for each phase. The participants of the workshops, mainly from KU, will evaluate the status of their research data management rules. The output of the workshop is expected that the participants will learn the standard method of RDM and perspective view on it, while the organizer will make a comprehensive map of open data science activities in KU, which will be referenced by making the roadmap of university-wide open science initiatives.

4. Conclusion

The Kudzu unit is established in order to understand the current status and develop the future plan of research data management in the university. The work of the unit is expected to foster appropriate research data management manner applicated to both research integrity and open science promotion.

References

1. Ministry of Education, Culture, Sports, Science and Technology. http://www.mext.go.jp/a_menu/jinzai/fusei/1359618.htm
2. National Institute for Informatics, https://rcos.nii.ac.jp/en/openscience/internal/
3. Cabinet Office, Government of Japan, http://www8.cao.go.jp/cstp/stsonota/datapolicy/datapolicy.html
4. Kyoto University, http://www.kyoto-u.ac.jp/en/research/ethic/research_guide/documents/research-integrity-regulations201503.pdf [10-Apr-2017]
5. California Digital Library , https://uc3.cdlib.org/2018/01/11/support-your-data/

Open genome analysis in the post-genomic era

Masanori Arita [1*, 2]

[1*] *National Institute of Genetics, Yata 1111, Mishima, Shizuoka, 411-8540, Japan*
[2] *RIKEN Centre for Sustainable Resource Science, Suehiro 1-7-22, Tsurumi, Kanagawa, 230-0045, Japan*
Email: arita@nig.ac.jp

Summary. The cost of DNA sequencing is ever decreasing; non-human sequences will soon become open commodity data like weather information, whereas human sequences become access-restricted, proprietary data like personal photographs. In this setting, the reference frames of representative genomes become more important to map and understand (possibly) streaming big data. To support the requirement of highly computational, interdisciplinary bioscience, we need to design a tiered and federated data warehousing scheme that offers reliable genome references and study contents.

Keywords. DDBJ, genome, warehousing.

1. Size of genome data in 2025

The advancement of DNA sequencing technologies surpasses that of computer chips and storages; the cost of sequencing has drastically dropped in the last decade (i.e. next-generation sequencing or NGS), and the latest instrument, Illumina NovaSeq 6000, can output 6 Tera bases within 2 days [1]. Such data cannot be transferred to or accumulated in a single, centralized warehouse, such as the International Nucleotide Sequence Database Collaboration (INSDC) between National Center for Biotechnology Information (NCBI in the United States), European Bioinformatics Institute (EBI), and DNA Data Bank of Japan (DDBJ). Indeed, Illumina Inc. extrapolates that the current sequencing ability already surpasses 35 Peta bases per year, and required storage size reaches Exabyte scale by 2025 [2,3]. To support broadband access to such data, the cloud-computing system seems to be the only practical solution.

2. Roles of centralized repositories

At least for 10 years, INSDC will remain as the unique data registrar/coordinator; physical repositories, however, become inevitably federated, e.g., with BGI-Online and sequencer vendors [4]. The responsibility of data storage and distribution will be also delegated to partner repositories, and the main focus of the registrar becomes the coordination (or standardization) of analysis frameworks. The integration of biological domain expertise is the most expensive part, and for this reason, commercial sectors expect governments or national sectors to solve this problem. In exchange, software platforms or predictions will be provided by data science companies such as Google and Amazon [5], targeting academic researchers as their major customers.

3. Importance of reference frames

Human genomes are actively analysed in various ways such as RNA-Seq, Methyl-Seq, or ChIP-Seq. All these technologies presuppose the availability of reference genomes. Analysis of streaming data for virus sensors or environmental monitoring also requires reliable references. Such templates should be open, citable, and versioned through a public community, such as INSDC.

4. Education is the crucial part

To fully exploit the fruits of new era, students should learn computer science in addition to modern biology. Early exposure is the key just as

younger kids can learn natural languages faster and better. At DDBJ we provide online coursework for different levels to support active, voluntary learners.

Acknowledgments. This work is supported by AMED-CREST (JP18gm0910005, JP18gm1010006), and the National Bioscience Database Center of JST.

References

1. Illumina Inc. NovaSeq System (data sheet), https://www.illumina.com/systems/sequencing-platforms/novaseq.html [accessed on: Oct 2018]
2. Stephens, ZD, Lee, SY, Faghri, F *et al.*, Big Data: Astronomical or Genomical? *PLOS Biology*, 13(7), e1002195
3. Schatz MC, Biological data sciences in genome research. *Genome Res* 25, 1417-1422, 2015
4. BGI Online for Big Data https://www.bgionline.com/;
   Basespace Sequence Hub by Illumina Inc. https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub.html [accessed on: Oct 2018]
5. Google Genomics in Google Cloud https://cloud.google.com/genomics/;
   Amazon Web Service (AWS) Genomics https://aws.amazon.com/jp/health/genomics/ [accessed on: Oct 2018]

Open Research Data Progress in Korea

Myung-Seok Choi [1]*, ***Sanghwan Lee***[1]

[1]* *Korea Institute of Science and Technology Information, 245 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea*
Email: mschoi@kisti.re.kr

Summary. As the advent of the Digital Era and data-driven research paradigm, Open Science, based on openness and sharing of research results, has emerged as a global agenda for scientific research. National policies for sharing and re-use of research data from publicly-funded research are in effect globally. In Korea, it is just getting started to build policies and infrastructure for open research data. In this paper, we investigate the current status of research data sharing and utilization in Korea. And then we briefly introduce a recent national strategic plan for open research data and related pilot projects.

Keywords. Open Science, Open Research Data, Research Data Management, Data Management Plan.

1. Introduction

With the advance of digital technology, such as advanced research equipment, sensors, and data processing technology, large-scale research data has been explosively produced, and a data-centric fourth research paradigm has emerged[1]. Moreover, R&D trends that emphasizes sharing and convergence are expanding to Open Science paradigm for easy access and utilization of research results from publicly-funded research projects. OECD has adopted Open Science as one of major policy agendas, and many countries are promoting policies for the spread of open science and focusing on developing infrastructures[2-5].

Open Science movement is expanding from open access of research publications to open research data which allows easy access and reuse of research data produced in the process of scientific research[2]. It is acting as the catalyst for data-driven collaborative research for global problem solving. In many countries including USA, UK and Australia, open research data policies including data management plan(DMP) have been implemented to create value through systematic management, easy access and reuse of research data from publicly-funded research projects [4 -5].

Openness is the basic premise for self-correction, the most important feature of scientific research, to work and is also the essential mechanism to deter, detect and stamp out bad science as well as to enable scientific discovery by assuring transparency in scientific research[3]. Recently, there are growing concerns over the reproducibility crisis in scientific research[6]. In order to verify the reproducibility of the research, it is necessary to disclose the data produced and used in the research. In this paper, we introduce the current status and future direction of open research data in Korea.

2. Status and Issues of Open Research Data in Korea

In Korea, national R&D information has been integrated and managed through NTIS(http://www.ntis.go.kr). While there is a sharing and utilization system of public government data and a limited range of research data, the legal basis and related infrastructures are still insufficient for accepting diverse research data as the first-class outcome of national R&D projects and sharing and utilizing them at national level.

Various types and sizes of reusable research data are being produced from national R&D projects. However, these data are stored and managed at the individual researcher or laboratory level, and partly shared by personal request. This is because data construction itself is

not recognized as major research achievements, and additional work related to data management and quality·responsibility issues are recognized as a burden to most researchers.

3. National Strategic Plan for Open Research Data

As data-driven R&D trends increase demand for research data utilization, Ministry of Science and ICT of Korea established a research data sharing·utilization strategy in January 2018 in order to construct a national system in which researchers can easily share and utilize research data accumulated through national R&D projects.

(Vision) Knowledge-assetization·big-dataization of research data

(Goal) Enabling national research data sharing and utilization by revising legal system, establishing hierarchical research data management system and infrastructures, and fostering data science experts

(Key tasks)

- Establishment of the research data management system and encouragement of data sharing communities
- Development of a national research data platform
- Support of training data and computing utilization experts
- Establishment of legal basis for nation-wide research data management · sharing · utilization
- Promotion of industrial utilization of research data and job creation

Currently action plans for pursuing the strategy are being developed and several pilot projects are underway this year.

- Organizing a research data network for collaboration and consensus building between various stakeholders
- Revising legal system for mandating data management plan
- Developing a pilot system of national research data platform
- Performing pilot projects for disciplinary data centers such as bioinformatics, materials, artificial intelligence, and large research equipment.
- Developing a data repository system for research institutes

4. Conclusions

Promoting sharing and utilization of research data from publicly funded research is becoming a global agenda to prepare for a data-driven research paradigm and is also an essential element for improving transparency and productivity of research. Open research data is accompanied by transforming research culture, so it should be approached from a long-term perspective. First of all, national data infrastructures as well as legal basis should be preceded. Then, fostering disciplinary data centers can let data utilization increase. At last, it should be able to spread to individual researchers' data sharing culture. Furthermore, close collaboration between various stakeholders such as data producers, data scientists, researchers, and funders is essential for open research data.

References

1. Hey, T., TanSley, S., Tolle, K., The Fourth Paradigm: Data-Intensive Scientific Discovery. *Microsoft Research,* 2009
2. OECD, Making Open Science a Reality. *OECD Science, Technology and Industry Policy Papers,* 25, 2015
3. The Royal Society, Science as an open enterprise, *The Royal Society Science Policy Centre Report,* 2012
4. Holdren, J., Increasing Access to the Results of Federally Funded Scientific Research. *White House OSTP Memorandum,* 2013
5. Open Research Data Forum, Concordat on Open Research Data. *Open Research Data Forum Report,* 2016
6. Baker, M., Is There a Reproducibility Crisis?. Nature, 533(7604), 452-454, 2016

Genetic relationship between tea plant (*Camellia sinensis*) and ornamental camellia (*C. japonica*)

Kazumi Furukawa[1]*

[1]* *National Institute of Technology, Numazu College, Ooka 3600 Numazu, Shizuoka, 410-8501, Japan*
Email: furukawa@numazu-ct.ac.jp

Summary. For the breeding of tea plant (*Camellia sinensis, 2n=30*), we investigated the chromosome structure and partial sequence of the tea caffeine synthase (*TCS1*) gene among the *Camellia* species used as the genetic resources of tea. A fluorescence *in situ* hybridization assay showed that the shapes of the chromosomes having 5S rDNA were similar between *C. sinensis* and *C. japonica* (2n=30). Moreover, bivalents were observed in the interspecific hybrid of the two *Camellia* species. Thus, because these species have homoeologous chromosomes, we could infer the extent of their genetic relationship. In the interspecific hybrid, a strongly methylated cytosine was detected in one chromosome. This allowed us to consider changes in epigenetic traits during the formation of hybrids. In addition, the diversity of the caffeine synthetic ability was found unique to the tea plants of the *Camellia* genus. As a result, an in/del of 83 bp was detected in the intron of the partial sequence of *TCS1*, and several mutations were detected in *C. sinensis*. In *C. japonica*, caffeine was detected in extremely small amounts, and it was thought that such a difference in the arrangement of nucleotides causes different traits. These mutations were considered important for interspecific hybridization in *Camellia* species.

Keywords: tea plant, *Camellia sinensis*, *Camellia japonica*, chromosome analysis, 5-methylated cytosine.

1. Introduction

The leaves of *Camellia sinensis* have long been used as tea throughout the world. There are several closely related species in the genus *Camellia*[1]. Ornamental camellia (*C. japonica*) is one of the genetic resources for tea breeding since it has cold tolerance and caffeine-less trait. Both *C. sinensis* and *C. japonica* have the same number of chromosomes (2n=30), and the interspecific hybridization between them is possible at low rates. To elucidate the efficiency of hybrid formation between these species, we investigated the morphological comparison of chromosomes and the diversity of the tea caffeine synthase (*TCS1*) genes in *Camellia* species.

2. Materials

A Japanese green tea cultivar Sayamakaori was used as the tea plant material in this study. The ornamental camellia interspecific hybrids Robiraki (*C. japonica* ×*C. sinensis*) and Cha chukanbohon nou 1gou (*C. sinensis* ×*C. japonica*) were used as comparison plants.

3. Chromosome analysis

A 5S rDNA probe labelled with fluorescein isothiocyanate (FITC) or indocarbocyanine (CY3) was hybridized to the chromosomes and detected by their fluorescence intensities[2]. Two signals were detected on each chromosome of *C. sinensis* and *C. japonica* (Fig. 1 top and middle). In *C. sasanqua*, several signals were detected (Fig. 1).

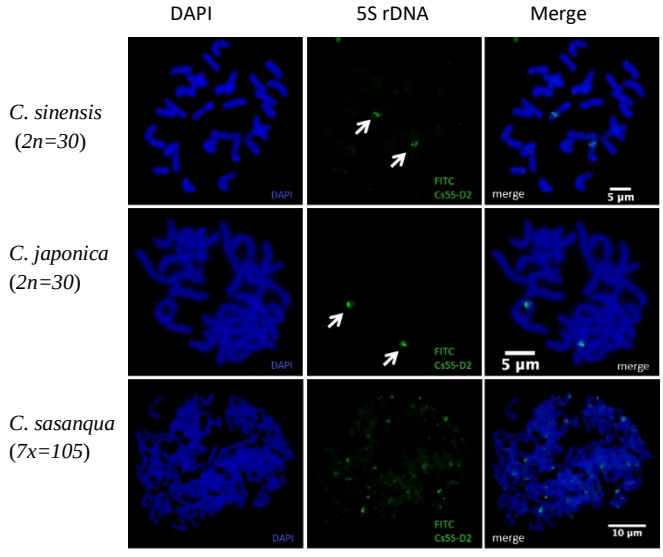


Fig. 1 Fluorescence *in situ* hybridization images of the 5S rDNA of the mitotic chromosomes of three *Camellia* species. Arrows indicate the 5S rDNA signals.

The lengths of chromosomes and the density of 4′,6-diamidino-2-phenylindole (DAPI) along their lengths were measured in *C. sinensis* and *C. japonica* using the chromosome image analysing system IV (CHIAS IV; http://www2.kobe-u.ac.jp/~ohmido/index03.htm). The lengths and signal positions of 5S rDNA were found similar in both *C. sinensis* and *C. japonica* (Fig. 2).

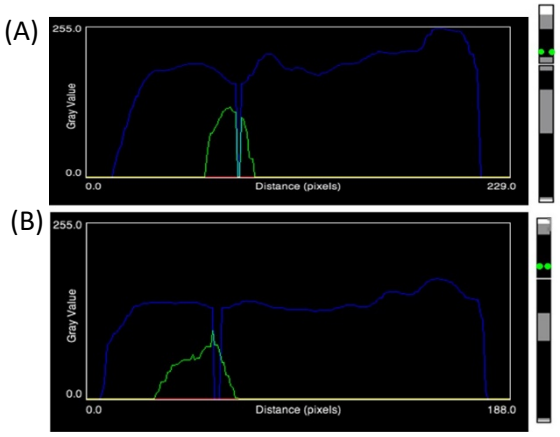


Fig. 2 DAPI density on the distance along the chromosome with 5S rDNA signal. The vertical axis represents the gray value, and the horizontal axis represents the number of pixels along the chromosome. The figures on the right side show the ideograms of the chromosome with the 5S rDNA. The green circle indicates the 5S rDNA loci. (A): *C. sinensis* (B): *C. japonica*

4. TCS1 partial sequencing

A polymerase chain reaction (PCR) was performed using the primer pair 5′-tcttcaa aggcctgtcg tct-3′ and 5′-tccccttgtttaatgccaag-3′, based on the TCS1 mRNA sequence of tea (Accession AB031280). An 83-bp in/del sequence was detected in the clones (Fig. 2). After cloning, the sequence alignment roughly divided the *Camellia* species into two groups—tea plant and other Camellia species (Fig. 3). Moreover, several polymorphisms were detected in the tea cultivar Sayamakaori (Fig. 2(B)).

5. 5-methyl cytosine

In the interspecific hybrid cultivar Cha chukanbohon nou 1gou, one chromosome showed a strong FITC signal upon immunostaining (Fig. 4). No 5-methyl cytosine signal was detected on the chromosomes of the parental plants. We need further investigation on this issue, but it will be interesting to identify uniparental chromosomes showing epigenetic regulation.

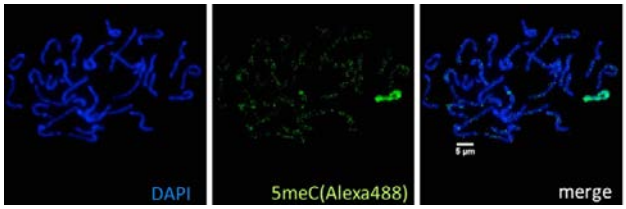


Fig. 4 Anti-5 methyl cytosine antibody signal in the interspecific hybrid cultivar Cha chukanbohon nou 1gou.

6. Conclusions

The quantitative analysis of a chromosome pair with 5S rDNA indicated homology in the chromosomes of *C. sinensis* and *C. japonica*. In the tea caffeine synthase gene, the two species differed in the presence of an 83-bp in/del in the partial sequence of the TSC1 intron. In the interspecific hybrid, 5-methyl cytosine was detected in one chromosome pair. These results could be used to understand the progeny of cross breeding between *Camellia* species and their evolution.

Acknowledgements. I thank Sakurako Ogawa (National Institute of Technology, Numazu College Advanced Course), Shohei Sugiyama (Graduate School of Agricultural Science, Kobe University), and Fuuka Kitamura (Hiroshima university) for their continuous efforts towards this work.

References

1. Ackerman, W. L., *Technical Bulletin*, 1427, US Government Print Office, 1971
2. Furukawa, K, Sugiyama, S., Ohta, T., Ohmido, N., *Chromosome Science*, 20, 9-15, 2017

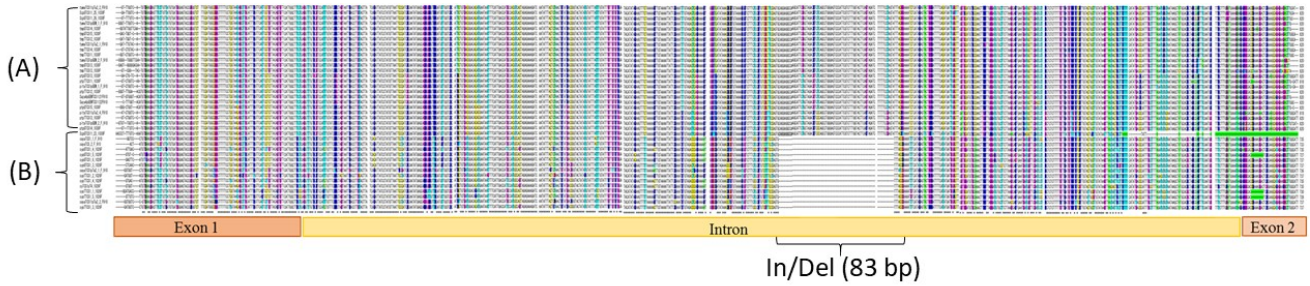


Fig. 3 Partial sequence alignment of *TCS1* cDNA in *Camellia* species. (A): *C. japonica 1*, *2*, and *C. sasanqua* (B): *C.*

Life Science Database Integration and Its Application for Medical Science

Susumu Goto[1*]

[1*] *Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 178-4-4 Wakashiba, Kashiwa, Chiba, 277-0871, Japan*
Email: goto@dbcls.rois.ac.jp

Summary. Huge numbers of life science databases have been developed and are available as open data. They can, therefore, contribute to the acceleration of open science, but they are quite heterogenous and their integrated use are still difficult. Database Center for Life Science (DBCLS) develops basic technologies for their integration based on semantic web technology and provides a portal site for data represented in the resource description framework (RDF) model, middleware for efficient access to the data, hackathons for the RDF-based database developers and application programmers, and many other tools including databases and web interfaces for medical science applications such as a genome variation database and phenotype-disease association tool for differential diagnosis.

Keywords. Resource Description Framework, SPARQL, genomic variations, diagnosis assistant, multi-omics.

1. Introduction

More than 1,700 databases have been developed in the life science research field and are available on the web [1]. They include data from various research fields. In fact, the 2018 Nucleic Acids Research database issue grouped the databases into eight broad research categories, (i) nucleic acid sequence and structure, transcriptional regulation; (ii) protein sequence and structure; (iii) metabolic and signaling pathways, enzymes and networks; (iv) genomics of viruses, bacteria, protozoa and fungi; (v) genomics of human and model organisms plus comparative genomics; (vi) human genomic variation, diseases and drugs; (vii) plants and (viii) other topics, such as proteomics databases.

Database Center for Life Science (DBCLS) [2] has been developing basic technologies for integrating those heterogeneous molecular biology databases by using the semantic web technology. Last year, we introduced the activities of DBCLS focusing on the development of middleware tools for accessing and utilizing data modelled by Resource Description Framework (RDF), and the coordination of BioHackathon and SPARQLthon, forums for database, ontology and program developers [3]. This year, we present updates of the basic technology developments in DBCLS and applications of our resources and tools to medical research.

2. Developments in DBCLS: Updates

DBCLS develops basic technologies and application programs to integrate and maintain life science databases and to enhance their usability. The tools and activities of DBCLS are grouped into the following four categories depending on the target users.

Application program development for researchers using web interface: Collaborating with DNA Data Bank of Japan (DDBJ) Center in National Institute of Genetics, we are developing integrated search tools for genome and expression data without knowing detailed data structure to support various levels of users for accessing data, because the query language SPARQL for accessing RDF databases are difficult for initial users. These are further applied for the integrated tools for medical research that are described in detail in the next section. Several RDF data can be searched via

natural language processing interface implemented in LODQA with an update for speech recognition.

Efficient access and maintenance for the integrated database: To maximize the utilities of RDF data, we have developed tools for accessing SPARQL endpoints (RDF database servers), checking their status, building SPARQL queries, and summarizing the query results. Currently, we are applying a tool for retrieving distributed RDF data to a protein sequence multiple alignment program.

Basic technology development for database integration: DBCLS supports database developers to convert their data into RDF formatted data, most of which are accessible via NBDC RDF portal [4] that is now consisting of ~47 billion triples from 21 data sets of various types of omics and other data. DBCLS also develops tools and dictionaries for knowledge extraction from texts in biomedical literature.

Standardization for database integration: We work intra- and internationally with research organizations that develop and maintain various databases and have contributed to the development of international guidelines and principles for database integration using semantic web technology. Hackathon series including BioHackathon, SPARQLthon, RDF summit and Biomedical Linked Annotation Hackathon are regularly held as forums of those collaboration.

3. Application for Medical Science

In addition to the basic technology developments, DBCLS has started application developments by collaborating with National Bioscience Database Center (NBDC) aiming at three main targets: medical science, useful substance production and breeding research. As a first step, TogoVar and PubCaseFinder have been developed for medical science applications.

TogoVar. Genome sequence variations among individuals in the Japanese population have been collected and organized so that each variation can be referenced with related links such as disease information [5].

PubCaseFinder. A differential diagnosis support system based on phenotypic similarity has been developed for rare diseases by integrating disease-phenotype databases and text mining approach [6].

4. Conclusions

As many life science data including omics data are open data and DBCLS has been focusing on such data for the integration using the semantic web technology, the databases and tools we provide are suitable for open science. In addition, open-close approach would be necessary for the data such as individual genome and personal data from cohort studies.

Acknowledgments. The author acknowledges all the members of DBCLS and NBDC for their contribution in developing tools, services, ontologies and databases, and coordinating hackathon and lecture series. The activity of DBCLS is funded, in part, by JST NBDC and ROIS International Network Formation and Data Science Education programs.

References

1. Rigden, D. J., Fernández, X. M., The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.*, 46, D1–D7, 2018
2. DBCLS homepage, http://dbcls.rois.ac.jp/ [accessed on: Oct 2018]
3. Goto, S., Life Science Database Integration Based on Semantic Web Technology. *International Workshop on Sharing, Citation and Publication of Scientific Data across Disciplines*, 21-22, 2017
4. NBDC RDF Portal, https://integbio.jp/rdf/ [accessed on: Oct 2018]
5. TogoVar, https://togovar.lifesciencedb.jp/ [accessed on: Oct 2018]
6. Fujiwara, T., et al., PubCaseFinder: A Case-Report-Based, Phenotype-Driven Differential-Diagnosis System for Rare Diseases. *Am. J. Hum. Genet.,* 103, 389-399, 2018

Perspective of Open Science in Japan driven by Integrated Innovation Strategy

Kazuhiro Hayashi [1]*

[1]* *National Institute of Science and Technology Policy, Japan*
Email: khayashi@nistep.go.jp

Summary. Open Science movement is transforming Science itself, and also transforming Industry and Society, and "Science and Society". Under this background, Open Science Policy has been developed and implemented in Japan. Open Science is placed as a main function in "Integrated Innovation Strategy" released on June 2018. In the strategy, Open Science is one of three pillars of "Source of Innovation" as well as "Social Data" and "Official Administrative Data." These are expected to contribute to Society 5.0 which is Japan's vision for the future in the 5th Science and Technology Basic Plan. This new strategy aims to enhance Data Science in new fields such as AI, Brain, and Research Data Management is getting more crucial not only for Science but also for Industry and Society. Citizen Science and its expansion towards co-creative research would also emerge as a new trend of public engagement and new framework of society-related Science.

Keywords. Open Science, Integrated Innovation Strategy, CSTI, Data Science, Citizen Science.

1. Introduction

Open Science is one of important emerging issues around Science, Technology and Innovation Policy all over the world. In policy context, Open Science is manly targeting to make research outputs by publicly-funding as open as possible and as closed as necessary to share, re-use them for innovation. [1] Open Science movement is now transforming Science itself beyond just changing the framework of scholarly publishing [2], and also transforming Industry and Society as well as "Science and Society". [3] Under this background, Open Science Policy has been developed and implemented in Japan.

This paper discusses perspective of Open Science in Japan with introduction of the latest STI Policy called "Integrated Innovation Policy" and its core module.

2. Open Science as a Source of Innovation in Integrated Innovation Strategy

"Integrated Innovation Strategy" (IIS) [4] was released by the Council for Science, Technology and Innovation (CSTI) on June 2018. After checking progress of the 5th Science and Technology Basic Plan [5] and Comprehensive Strategy on Science, Technology and Innovation 2017 [6], IIS is for setting agendas with targets to accomplish them. During consideration of designing IIS, various emerging topics were discussed at CSTI and Open Science was an issue which was not so emerging but promising issue with long-term consideration.

In the strategy, Open Science is one of three pillars of "Source of Innovation" (Knowledge Base) as well as "Social Data" and "Official Administrative Data." These three are expected to contribute to develop Society 5.0 which is Japan's vision for the future of super smart society in the 5th Science and Technology Basic Plan.

With concrete targets with figures, Open Science policy is described to enhance research data platform (data repository), making data policy of governmental research institutions, advocacy and education, and developing a survey for monitoring Open Science as well as

promoting Data Management Plan by funding agency and promoting Institutional Repository.

Open Science is now recognized as a driver of innovation to transform Science towards Society 5.0.

3. Perspective of Open Science and related activity

This new strategy aims to enhance Data Science in emerging fields such as AI, Brain Science and Material Informatics. For example. the Cabinet Office started an expert meeting to discuss details for AI. [7]

For those new emerging field and even for the conventional science field, Research Data Management (RDM) is getting more crucial. Since IIS is seeking multi-dimensional exploitation of research data, RDM is not only for Science but also for Industry and Society, implementing interoperable connection among them.

Open Science is also transforming "Science Society." RDM is effective to keep transparency of Science itself. More proactively Citizen Science is changing to exploit potential of Open Science variously. Citizen is getting involved efficiently in various research stage such as co-design and co-production exploiting ICT. Some scientists earn their research budget from citizen directly by crowd-funding, not by indirect subsidies of public funding agency coming through as tax. It would be towards co-creative research and it would also emerging as a new trend of public engagement and new framework of society-related Science.

4. Conclusions

Open Science movement is affecting a whole social system to transform society itself to adapt to the web-native society (Society 5.0). Each researcher, PI, and research institute will have to recognize it and associate with it according to the affinity of their research to current movement, which would change their Science itself as Open Science movement aims.

References

1. Open Science, European Commission, https://ec.europa.eu/research/openscience/, accessed on: Oct 2018
2. Hayashi, K., Revolution of process on publishing and sharing towards Open Science enhanced by openness of scholarly communication. *STI Horizon,* 3, 35-39, 2017 (in Japanese) , http://doi.org/10.15108/stih.00092, accessed on: Oct 2018
3. G7 SCIENCE MINISTERS' COMMUNIQUÉ, 2017 http://www.g7italy.it/sites/default/files/documents/G7%20Science%20Communiqu%C3%A9.pdf , accessed on: Oct 2018
4. Integrated Innovation Strategy (in Japanese), http://www8.cao.go.jp/cstp/tougosenryaku/, accessed on: Oct 2018
5. A Promoting Open Science in Japan "Opening up a new era for the advancement of science" Executive Summary Report by the Expert Panel on Open Science, based on Global Perspectives Cabinet Office, Government of Japan, 2015, http://www8.cao.go.jp/cstp/sonota/openscience/150330_openscience_summary_en.pdf , accessed on: Oct 2018
6. Comprehensive Strategy on Science , Technology and Innovation for 2017, http://www8.cao.go.jp/cstp/english/doc/2017stistrategy_main.pdf, accessed on: Oct 2018
7. CSTI Expert Meeting on AI (in Japanese) , https://www.kantei.go.jp/jp/singi/ai_senryaku/index.html, accessed on: Oct 2018

Current Situation of Data Archiving for Japanese Official Statistics

Shinsuke Ito[1*]

[1*] *Faculty of Economics, Chuo University, 742-1 Higashinakano, Hachioji-shi, Tokyo, 192-0393 Japan*
Email: ssitoh@tamacc.chuo-u.ac.jp

Summary. In Japan, several types of official data are released under the Statistics Act including statistical tables, original microdata, anonymized microdata, tailor-made tabulation and open data. Seven types of Anonymized microdata from Japanese official statistics including the Population Census are publicly released. Anonymized microdata from the 2000 and 2005 Census are currently available. Various disclosure limitation methods including non-perturbative methods (sampling, recoding etc.) and perturbative methods such as data swapping are applied to create anonymized census microdata. This paper describes the methods used for data archiving for official statistics in Japan, with a focus on both the current situation and future outlook.

Keywords. Data Archiving, Official microdata, Statistics Act.

1. Introduction

In Japan, several types of official data are released across a variety of formats and based on data confidentiality and user needs. These data include statistical tables, original microdata (individual data), anonymized microdata, tailor-made tabulation and open data. The data is released under the Statistics Act [1].

2. Archiving for Official Statistical Data

In Europe and North America, several types of official microdata from the Population Census are publicly available. In the United States, Public Use Microdata Sample has been available since the 1960 Census [2]. In the United Kingdom, several types of Anonymized microdata such as Sample of Anonymised Records (SAR) are available.

At present, seven types of Anonymized microdata from Japanese official statistics including the Population Census are publicly released in Japan. Anonymized microdata from the 2000 and 2005 Census is also available. Various disclosure limitation methods such as sampling, recoding, top (bottom) coding and data deletion are applied prior to release. Data swapping is applied as an additional perturbative method to create Anonymized Census microdata.

In recent years, 'Statistical Reform' has attracted attention in Japan. Statistical reform focuses on improving the usability of official statistical data as well as using big data and administrative data to foster evidence-based policy making by the Japanese national and regional governments.

3. Conclusion

The methods used for data archiving for official statistics in Japan are largely based on those developed in Europe and North America. In order to promote the secondary use of official statistics in Japan, further research into data sharing methods for official statistical data should be pursued.

References

1. Ito, S., Data sharing for official statistics: Current situation and future outlook in Japan. *Journal of Information Processing and Management*, 58, 836-843, 2016 (in Japanese)
2. Zayatz, L., Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. *Journal of Official Statistics*, 23, 253-265, 2007

Data management at Korea Polar Data Center, Korea Polar Research Institute

***Dongchan Joo*[1]*, *Hyun-cheol Kim*[1], *Goheung Kim*[1]**

[1]* *Korea Polar Data Center, Korea Polar Research Institute, Incheon, Korea South*
Email: dc.joo@kopri.re.kr

Summary. The Korea Polar Data Center (KPDC) is in charge of collecting, managing, and sharing data acquired through researches conducted by the Korea Polar Research Institute (KOPRI). The KPDC has large volumes of research data of which most of them are shared for scientific and non-commercial purposes. In recent years, data management has become increasingly important due to the surge of new data from more active surveys than the past.

Keywords. Polar Data, Open Data, King Sejong Station, Jang Bogo Station, Dasan Station, IBRV ARON.

1. Introduction

The KOPRI opened its first Antarctic research station, the Antarctic King Sejong Station, in 1988. Subsequently, the KOPRI established the Arctic Dasan station in 2002 and IceBreaker Research Vessel ARAON in 2009. In 2011, the KOPRI established its second Antarctic research station, the Jang Bogo Station, located in the mainland of the Antarctica.

Based on above infrastructures in Arctic and Antarctic areas, researches on various fields such as oceanography, biology, climate change, geophysics, remote sensing, and space weather have actively been carried out and have generated huge volumes of research data. In 2003, the KPDC was established to follow demands for more systematic management of data, complying with the Antarctic Treaty Article III (1) (c) "*Scientific Observations and results from Antarctica shall be exchanged and made freely available*", and to participate in such international joint effort for Antarctic research.

The KPDC is designed to play as a hub of the Korean polar research data from both Arctic and Antarctic. All data acquired from polar researches are registered at KPDC. Anyone can retrieve metadata and can request to use the original data if necessary. Recently, a project to upgrade KPDC system has started for better data management and user-friendly system by applying GIS, data visualization, and statistical techniques.

2. Past

The KPDC is an organization dedicated for managing different types of data acquired during scientific researches that South Korea carries out in Antarctic and Arctic regions. South Korea, as an Antarctic Treaty Consultative Party (ATCP) and an accredited member of the Scientific Committee on Antarctic Research (SCAR), established the center in 2003 as a part of its effort to joint international Antarctic research.

In 2010, South Korea formulated a "Master-plan for polar data center of South Korea" to build its own polar data center. Since then, ongoing efforts had been made to construct solid data management infrastructures by implementing polar data management policies, developing and operating a metadata management system, and establishing the data management system for each research area and digital data management system.

In 2012, the KOPRI started providing training courses and raising awareness to researchers who are participating in polar studies in order increase efficiency of data management. In addition, it established the schemes for data storage and back-up for acquired digital data to be safely managed and permanently stored in its own

system. Furthermore, it has contributed to polar researches of the international society by sharing the internally obtained polar data and experiences.

In 2017, it reorganized the overall data sharing system to improve the accessibility and convenience for those who use the polar data, along with the efforts for advancing the related systems supported by the specialized IT consulting.

3. Present

The KPDC is developing a new system to provide comprehensive and various polar information so that researchers can have easy access to polar data by efficient management and distribution of the data. It also aims to expand its infrastructure for safe management and storage of polar data, as well as to upgrade, on a continuous basis, the systematic tools for the comprehensive management of polar data obtained from diverse research fields of polar regions. This new system will be open in December, 2018 (Figure 1).

4. Future

Until now, the KPDC has focused its efforts on data collection and management. In the future, the KPDC is going to start a new challenge, big-data science, which is beyond data management.

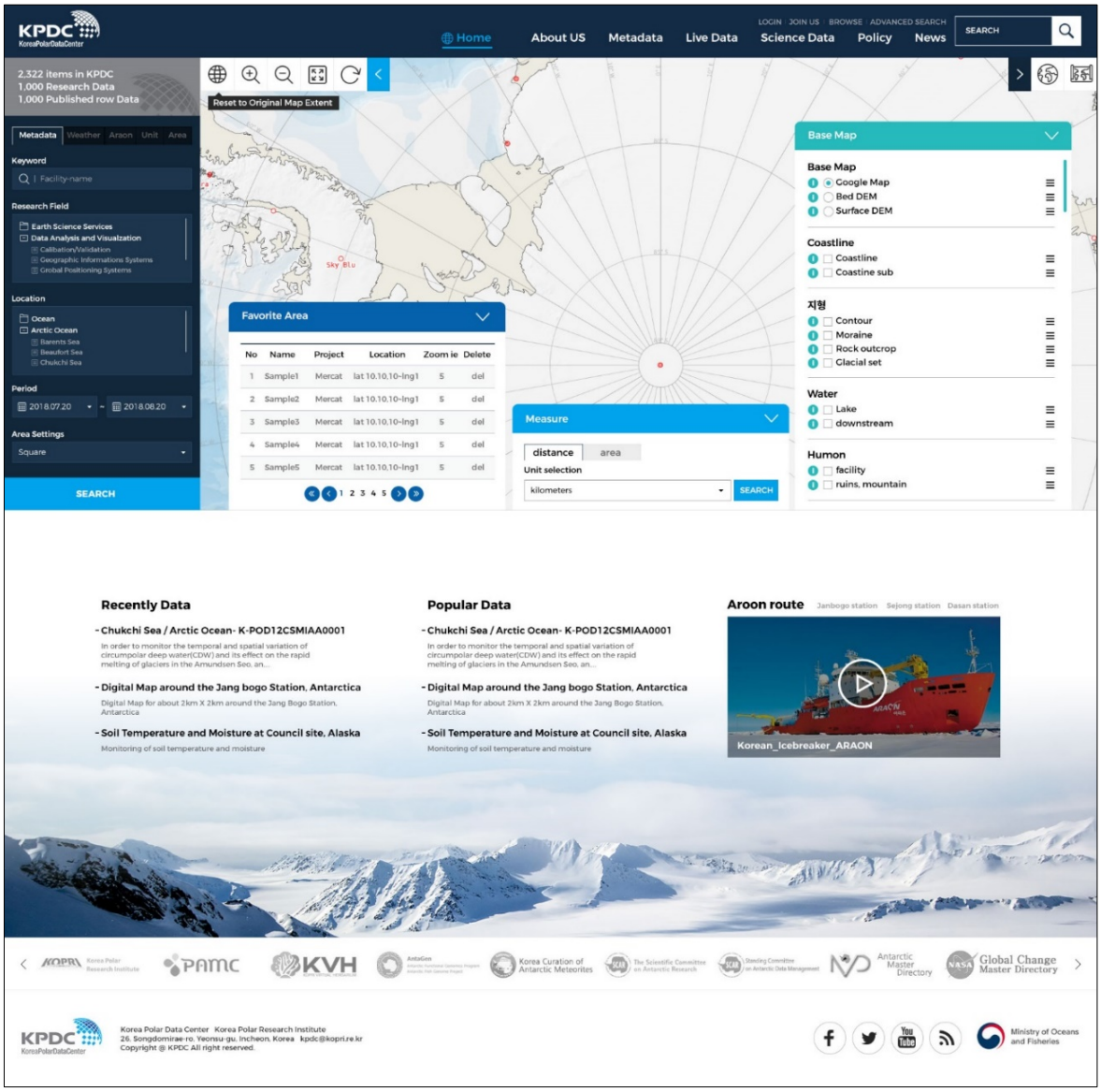


Figure 1. Main screen of the KPDC web site under development (URL: https://kpdc.kopri.re.kr).

Activities of Polar Environment Data Science Center

Akira Kadokura[1]*

[1]* *Polar Environment Data Science Center, DS, ROIS, 10-3, Midoricho, Tachikawa, Tokyo 190-8518, Japan*
Email: kadokura@nipr.ac.jp

Summary. Activities of the Polar Environment Data Science Center (PEDSC) are introduced. PEDSC has been established in the Joint Support-Center for Data Science Research (DS) of the Research Organization of Information and Systems (ROIS) in 2017. Purpose of the PEDSC is to promote collaboration with the data obtained by the research activities in the polar region, and to play a key role of the data activity in polar science to contribute to the global environment research.

Keywords. polar science, data science, open data, data publishing, data journal.

1. Introduction

Activity of the PEDSC is closely related with the research and observation activities of the National Institute of Polar Research (NIPR). Main purpose of the PEDSC is to promote the opening and utilization of the various data stored in NIPR in collaboration with universities and other institutions in Japan and foreign countries. In FY2017, we defined seven categories and made a yearly plan based on those categories to be carried out during the whole 5 years of the DS project period until FY2021.

2. Data

Data to be handled by the PEDSC are obtained in both Antarctic and Arctic regions mainly by the four research groups in NIPR; Space and upper atmospheric sciences group, Meteorology and glaciology group, Geoscience group, and Bioscience group. Various data in various research fields have been obtained so far, e.g. aurora and upper atmosphere, meteorology and marine science, snow and ice, geology, geomorphology, seismology, gravity, and biology. Those data are stored and archived in various forms, and basically classified in two categories, time series data and sample data. The former data are sampled at a fixed interval and recorded continuously during some observation period. The latter data are obtained at some specific date at some specific locations as a sampled material, e.g. rock, meteorite, ice core, sea water, air, etc. For such sample data, both their catalogues and analysis data are created and stored.

3. Database system

Each data mentioned in the Section 2 are stored and archived in each database in each research field by each research group in NIPR in each form and style. There are also following general database systems for the data in NIPR, which can handle various data in various research fields.

3.1 Science database

Science database (https://scidbase.nipr.ac.jp/) is a metadata database for all the data in all the research fields of polar science, and has a close relationship with international data activities, e.g. NASA Global Change Master Directory (GCMD), Standing Committee on Antarctic Data Management (SCADM) under the Scientific Committee on Antarctic Research (SCAR), etc.

3.2 Arctic Data archive System

Arctic Data archive System (ADS) (https://ads.nipr.ac.jp/) is a metadata and actual data database system mainly for Arctic projects such as GRENE (Green Network of Excellence) and ArCS (Arctic Challenge for Sustainability). ADS is also equipped with online visualization and analysis tools, and is used for collaboration with

Japanese and international communities for Arctic research

3.3 IUGONET system

IUGONET (Inter-university Upper atmosphere Global Observation NETwork) system (http://www.iugonet.org/) is a metadata database system which is developed in an inter-university project among NIPR and 4 universities for upper atmospheric science research. IUGONET is also equipped with display and analysis software tools (UDAS: iUgonet Data Analysis Software) for actual data, and is close relationship with international data activities such as SPEDAS (Space Physics Environment Data Analysis Software) and SPASE (Space Physics Archive Search and Extract) groups. Workshop and school for the IUGONET system are regularly held for students and researchers in Japan and abroad.

4. Current problems

Current needs and problems on the data and database systems in NIPR are; (1) A synthetic database system to cover, search and utilize all the data and database in all the research fields is not constructed and needed to understand the polar science activities as a whole. (2) Status of the processing, archiving and opening of the data is in wide variety for each data and database, depending on status of the resources of manpower, hardware and software.

5. Activity plan of PEDSC

Current targets of the PEDSC are defined in the following seven categories; (1) To construct a synthetic database for all the research fields of polar science. (2) To make the existing database systems (Science Database, ADS, IUGONET) upgraded and interoperable with each other. (3) To promote archiving, opening, and sharing of the time series digital data in each research field. (4) To promote archiving, opening, and sharing of the sample data in each research field. (5) To promote the data publication through the "Polar Data Journal ", data journal of NIPR. (6) To promote collaboration with universities and other institutions in Japan and international communities. (7) To promote data science using the database and database system.

Staff of the PEDSC in 2017-2018 JFY consists of a manager with one associate professor, three specially-appointed associate professors, and two office assistants.

6. Activities in FY2017-2018

In FY2017, all the members of the PEDSC moved to the new building "Data Science building" from the main NIPR building, and made a future activity plan during FY2017-2021. As for each category, (1) System design of the unified database system has been fixed in FY2018. (2) Several Antarctic data were processed by ADS. System design to develop ADS to AADS (Arctic & Antarctic Data archive System) has been carried out. (3) Data management and publishing system for the PANSY data has been constructed. (4) Database system for the Antarctic rock sample was constructed. Some of the ice core data were registered in the ADS system. (5) 2 data papers were published and 5 papers are now under review for the Polar Data Journal. (6) ROIS-DS International workshop on data science (IWDS) was held in FY2017 organized mainly by the PEDSC, and the second IWDS will be held in November, FY2018 (this time). (7) 3 and 7 proposals were accepted as PEDSC associated ROIS-DS collaboration program in FY2017 and FY2018, respectively.

7. Summary

PEDSC, established in FY2017, is now carrying out various activities along a yearly activity plan during FY2017-2021 in the seven categories specified to promote the archiving, opening and sharing of the polar science data in Japan.

"Polar Data Journal": A new data publishing platform for polar science

Akira Kadokura [1]*, ***Yasuyuki Minamiyama*** [2], ***Masaki Kanao*** [1], ***Takeshi Terui*** [2], ***Hironori Yabuki*** [1], ***Kazutsuna Yamaji*** [3]

[1]* *Polar Environment Data Science Center, DS, ROIS, 10-3, Midoricho, Tachikawa, Tokyo 190-8518, Japan*
[2]*National Institute of Polar Research, 10-3, Midoricho, Tachikawa, Tokyo 190-8518, Japan*
[3] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*
Email: kadokura@nipr.ac.jp

Summary. Polar Data Journal (PDJ) is a free-access, peer-reviewed online journal of the National Institute of Polar Research for publishing original research data/dataset, covering broad range of research disciplines including Arctic, Antarctic, or other polar regions. PDJ was newly born in January, 2017, and the first data paper of PDJ was published in October, 2017. So far, two papers were published, and currently five papers are under review. A special issue is planned to invite more submissions to raise the value of the PDJ as the first data journal in polar science community.

Keywords. data journal, data publication, open data, polar science, DOI.

1. Aim and scope of Polar Data Journal

Polar Data Journal (PDJ) has been launched on 19 January, 2017 by the National Institute of Polar Research (NIPR) with a help of the JAIRO Cloud online repository system (https://community.repo.nii.ac.jp/) of the National Institute of Informatics (NII). PDJ is a free-access, peer-reviewed and online journal. It is dedicated for publishing original research data/dataset, furthering the reuse of high-quality data and the benefit to polar sciences. PDJ aims to cover broad range of research disciplines involving Arctic, Antarctic, or other polar regions, especially earth and life sciences. PDJ primarily publishes data papers, which provide detailed descriptions of research data/dataset (e.g. Methods, Data Records, Technical validation). PDJ does not require any new scientific findings, so PDJ also welcomes submissions describing past valuable data/dataset which has not been published yet. PDJ requires submitted papers to be passed our peer-review process. PDJ also requires that authors should deposit their data/dataset to trustworthy data repository before submitting their manuscripts. PDJ guarantees data authenticity by publishing all review reports together with accepted manuscripts at the same time.

2. History of data publication of NIPR

Since Antarctic station Syowa had been established in January, 1957, "Antarctic Record" has been launched in December, 1957 as a journal for recording the Antarctic activities. "JARE Data Reports" is a journal for annual report of the scientific data obtained in each Japanese Antarctic Research Expedition (JARE), and has been launched on August, 1968. In December, 1996, "NIPR Arctic Data Reports" has been launched as a data journal for the scientific activity in Arctic region. Consideration of launching the PDJ started in June, 2015. Many aspects of the PDJ has been considered (e.g. Submission guideline, selection of platform, development environment, policy, review process, journal name, board organization, etc.) before launching on January 19, 2017. In future, both "JARE DATA Reports" and "NIPR Arctic Data Reports" will be combined and moved to PDJ.

3. Board organization of PDJ

Editorial board of PDJ consists of 12 board members, including two foreign members, three non-NIPR domestic members, and seven NIPR members. Advisory board of PDJ consists of 10 board members, including eight foreign members and two non-NIPR domestic members.

4. Review process of PDJ

Review process of PDJ is as follows: 1. Authors submit their manuscript by using an online Editorial Manager (EM) system, and they also upload their original data to an appropriate repository; 2. EM sends a receive information to editorial office (Office); 3. Office confirms the original data on the repository; 4. Office copies the original data to the JAIRO Cloud; 5. EM sends a receive information to editorial board (Editor); 6. Editor confirms the original data; 7. Editor invites two referees (Referee); 8. Referee receives the manuscript; 9. Referee reviews both manuscript and original data; 10. Referee sends reply comment to EM; 11. EM sends the reply comment to Editor; 12. Editor sends feedback to EM; 13. Authors receives the feedback from EM; 14. After accepted, Office copies the original data at accepted timing to the JAIRO Cloud; 15. Office requests the manager of the data repository to publish a Data DOI to the accepted data; 16. EM sends a proofreading comment to Office; 17. Office sends publish report of all review processes; 18. EM publish manuscript with data DOI using the JAIRO Cloud.

5. Appropriate repository for PDJ

As the appropriate repository for uploading the data for PDJ, following criteria are required: 1. The data/dataset in the repository must have a persistent identifier (such as DOI); 2. The data/dataset in the repository must be available free of charge and without any barriers except for a standard registration to get a login free of charge; 3. Anyone must be free to copy, distribute, transmit, and adapt the datasets as long as they give credit to the original authors.

6. Submission guidelines

Structure of the data paper submitted to PDJ should consist of the following items: title, authors, affiliations, abstract, background and summary, location (or observation), methods, .data records, technical validation, usage notes (optional), acknowledgements, author contributions, competing interests, figures (optional), tables, references, data citations.

7. Current and future of publication

So far, 8 submissions were received. 2 were published, 1 is on the editor's decision after reviewing, 4 are under review, and 1 was rejected. To enhance and invite more submission, a special issue is now planned. Recruit and support activities for inviting submissions are also necessary for the data obtained in the JARE, arctic projects, and in the international programs of SCAR (Scientific Committee on Antarctic Research) or IASC (International Arctic Science Committee), etc.

8. Summary

PDJ is a new data publishing platform for polar science, born in January, 2017. It is important to invite more submissions of data papers to raise the value of the PDJ as the first data journal in the polar science community.

Education Programs for AI/IoT/BigData Skills Development at the Medical and Drug Discovery Data Science Consortium

Eli Kaminuma[1]*, ***Hiroshi Tanaka*** [2]

[1]* *Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo, Tokyo, 113-8510, Japan*
Email: ekmds@tmd.ac.jp

Summary. The number of healthcare data science studies is increasing nowadays, however there are few data scientists familiar to life science domain in Japan. To solve this problem, we established the medical and drug discovery data science consortium in 2017, for assisting consortium participants to increase their AI/IoT/BigData skills in healthcare domain. The mission of the consortium is (1) enhancing education programs for doctoral students and industrial careers and (2) conducting open innovation meetings to exchange advanced progress of healthcare data science.

Keywords. Data Science Consortium, Medical Care, Drug Discovery, AI/IoT/BigData, Open Innovation.

1. Introduction

In healthcare domain, education programs to develop AI/IoT/BigData skills at training courses are not popular in Japan. Thus, human resources with novel data science skills such as artificial intelligence and big data are few in Japanese healthcare research community. To resolve this problem, we established the Medical & Drug discovery Data Science (MD-DS) Consortium [1]. The consortium aims to develop education programs to acquire the ability to combine knowledge of medical and drug discovery with novel data science skills. In this presentation, our education programs for healthcare data science are introduced.

2. Medical & Drug Discovery Data Science Consortium

In 2017, the Japanese government MEXT started data science education grants named by Doctoral program for Data-Related InnoVation Expert (D-DRIVE) [2]. The D-DRIVE program focuses on AI/IoT/BigData skills development for the fourth industrial revolution. Our Tokyo Medical and Dental University (TMDU) was adopted as a D-DRIVE grant program of the first call. Thus we established the Medical & Drug discovery Data Science consortium, which is characterized as follows.

- **One of the 2017-18 D-DRIVE grant programs**
  In 2017 D-DRIVE grants, the Japanese govenment MEXT selected four institutes of TMDU, University of Electro-Communications, Waseda University, and Osaka University. Then Hokkaido University was selected for 2018 D-DRIVE program.
- **Characteristic features of MD-DS consortium among D-DRIVE members**
  Of five data science consortiums, our MD-DS consortium represents characteristic feature of *healthcare* domain such as medical care and drug discovery. The data science consortium of Hokkaido university is also an organization concerning to *infrastructure* domain.

The D-DRIVE grant program seems to refer to the INSIGHT Data Science Fellows Program [3] of US government to bridge the gap between academia and a career in data science.

3. Education Programs for

AI/IoT/BD Skills Development at the MD-DS Consortium

In 2018, the MD-DS consortium called respective 30 participants of doctoral students and industrial careers. As the result, fifty-three participants were recruited from 22 companies and 8 universities/institutes.

The curriculum of the MD-DS consortium provides two main training components of "Medical Big Data" and "AI Drug Discovery". Moreover, basic training components are constituted by "Data Science Basics", "Genomic Science", "Computational Programming", "Life Science Ethics". The total number of consortium classes is around 100. Then programming practices supported statistical analysis and machine learning with R and Python languages.

Most classes are provided between July to December, and then visiting training programs are started at October. Visiting training practices of data science are formatted to internship for doctoral students and programming practices at major research institutes for industrial careers. Table 1 presents the list of research institutes for visiting training programs.

All classes and visiting training programs are reviewed by questionnaire survey to participants and thus the survey results contribute improvement of the quality of education programs.

4. Conclusions

We introduced the outline of the Medical & Drug discovery Data Science consortium of a Japanese data-related educational grant program. The MD-DS consortium holds open innovation meetings four times a year in addition to developing human resource education programs. Then, promoting open innovation is also a valuable target of the MD-DS consortium.

Acknowledgments. We are thankful to Takeshi Hase, Katsuyuki Takeuchi, Yuko Igarashi, Shuhei Mitome, Akiko Miyamoto, Akihiro Sekiguchi, Hiromi Shimoda as staffs of the medical and drug discovery data science consortium. This study was partially supported by a Japanese government MEXT grant of the Doctral program for Data-Related InnoVation Expert.

References

1. Medical & Drug discovery Data Science Consortium, http://md-dsc.com/ [accessed on: Oct 2018]
2. Japanese government MEXT D-DRIVE grant, http://www.mext.go.jp/a_menu/jinzai/data/index.htm [accessed: Oct 2018]
3. INSIGHT data science fellows program, https://www.insightdatascience.com/ [accessed: Oct 2018]

Table 1. List of research institutes to provide visiting training programs for industrial careers.

Research Institutes	Training Programs	Institute Location
National Center for Global Health and Medicine, National Center for Neurology and Psychiatry, Japanese Foundation for Cancer Research, TMDU	EMR/EHR Data Mining, Cancer Image Analysis, Deep Learning	Tokyo
Tohoku Medical Megabank Organization	Precision Medicine	Miyagi
Keio University	Metabolomics	Yamagata

A decade of history for polar data management in Japan

Masaki Kanao[1]*

[1]* *Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8518, Japan*
Email: kanao@nipr.ac.jp

Summary. Diverse data accumulated by many science disciplines make up the most significant legacy of the International Polar Year (IPY2007-2008). The Polar Data Center (PDC) of the National Institute of Polar Research (NIPR), followed by the Polar Environment Data Science Center (PEDSC) of the Joint Support-Center for Data Science Research (DS) have responsibility to manage these polar data in Japan as a National Antarctic Data Center (NADC). During the IPY, a significant number of multi-disciplinary metadata records were compiled from IPY- endorsed projects. A tight collaboration has been established between the Global Change Master Directory (GCMD), the Polar Information Commons (PIC), and the newly established World Data System (WDS). In this presentation, a decade of history of polar data management is demonstrated. Linkages of data sharing among Asian Forum for Polar Sciences (AFoPS) countries, moreover, should be promoted by the involved countries in near future.

Keywords. International Polar Year, data management, Polar Environmental Data Science Center, National Antarctic Data Center, Asian Forum for Polar Sciences.

1. Introduction

At the 22nd Antarctic Treaty Consultative Meeting (ATCM) in 1998, affiliate countries were obliged to ensure that scientific data collected from Antarctic programs could be freely exchanged and used. Following Article No.III.1.c of the Antarctic Treaty, each country is required to establish NADC and to properly disclose the data collected from involved scientists. The PDC/PEDSC has performed the function of a NADC for Japan and established a data policy in February 2007, based on the requirements of the Standing Committee on Antarctic Data Management (SCADM) of the Scientific Committee on Antarctic Research (SCAR). This contributed to the subsequent SCAR Data and Information Management Strategy [1].

Dedicated data services have been conducted by PDC/PEDSC as a member of NADC under SCAR. Several different aspects of scientific data collected in polar region have great significance for global environmental research. To construct an effective framework for long-term strategy of the polar data, data must be made available promptly by Internet technologies such a repository network service. In addition to activities in polar science communities of SCAR and the International Arctic Science Committee (IASC), tighter linkages expected to be established with other cross-cutting science bodies under ICSU, such as CODATA, and WDS. Linkages among these data-management bodies need to be strengthened in the post IPY era.

2. Data Management

The International Polar Year (IPY 2007-2008) was the world's most diverse science program. It was conducted during the 50th anniversary of the International Geophysical Year (IGY 1957-1958). The IPY greatly enhanced the exchange of ideas across nations and scientific disciplines to unveil the status and changes of planet Earth as viewed from polar region. The interdisciplinary exchange helped us understand and addressed grand challenges such as rapid environmental change and its impact on society. Eventually, Japanese researchers participated to 63 projects endorsed by the IPY Joint Committee. The huge amount of

data accumulating during IPY should be the most important legacy if it is well preserved and utilized [2].

The science database provided by PDC/PEDSC has a tight connection with AMD in GCMD. In addition to the IPY-related data, data from Japanese national and other international projects had been compiled. In totally, 300 metadata were compiled in Japanese Antarctic portal in GCMD. PDC/PEDSC stores its metadata in their original format, but this includes the main items listed in GCMD Directory Interchange Format (DIF). There are tight cross-linkages in corresponding metadata in AMD. Metadata collected by IPY projects for Japan have also been accumulated in an IPY portal of GCMD. More than 250 metadata from Japan were stored in the IPY portal [3, 4]. This constitutes a significant proportion of all IPY metadata to GCMD.

3. Data Legacy

SCADM has been strongly connected with IPY data-management activities (IPY Data and Information Service: IPY-DIS). IPY data policy emphasized a need to make data available on the "shortest feasible timescale." In accordance with IPY data policy, IPY-DIS recommended that data be formally cited when used, and the IPY Data Committee has developed initial guidelines for how data should be cited. The guidelines harmonized different approaches, and they adopted by many data centers relating polar area. After the end of IPY, a new initiative, the Polar Information Commons (PIC), began as a framework for open and long-term stewardship of polar data and information [2]. The PIC serves as an open, virtual repository for vital scientific data and information and provides a shared, community-based cyber-infrastructure fostering innovation and improved scientific understanding while encouraging participation in research, education, planning, and management in polar region. PIC developed specialized tools that produce a small, machine-readable "badge" that is attached to the data. However, the badge requested data users to adhere to basic ethical norms of data use including proper citation. This service was coupled with a cloud-based repository that may not have a suitable archive elsewhere. NIPR/DS have made contributions to PIC, both by attaching badges and registration in the repository. As of October 2017, Japan had contributed more than 50 data sets to the PIC.

Polar data have great relevance for modern, global environmental research well beyond the polar region. It is critical to explore the cloud approaches such as the PIC to develop an effective framework for open and long-term stewardship of polar data. The status of data-management before and after the IPY in Japan was introduced in this report. Several different aspects of the scientific data collected in the polar region have great advantage for global environmental research as well as in future. Linkages of data sharing among Asian Forum on Polar Science (AFoPS) countries, moreover, should be promoted by the involved countries in near future.

Acknowledgments. The authors express their appreciation to all collaborators of the IPY activities. They also acknowledge the members of SCADM and IPY Data committee for their efforts to adhere to data-management issues.

References

1. Finney, K., SCAR Data and Information Management Strategy 2009-2013. *SCAR Ad-hoc Group on Data Management*, 34, 2009
2. Parsons, M. A., Godoy, Ø., LeDrew, E., de Bruin, T., Danis, B., Tomlinson, S., Carlson, D., A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science. *Journal of Information Science*, 1-21, 2011
3. Kanao, M., Kadokura, A., Okada, M., Yamnouchi, T., Shiraishi, K., Sato, N., Parsons, M. A., THE STATE OF IPY DATA MANAGEMENT: THE JAPANESE CONTRIBUTION AND LEGACY. *Data Science Journal,* 12, WDS124-WDS128, 2013
4. Kanao, M., Okada, M., Friddell, J. and Kadokura, A., Science Metadata Management, Interoperability and Data Citations of the National Institute of Polar Research, Japan. *Data Science Journal*, 17, 1–6, 2018

Data and metadata sharing among AFoPS countries

Masaki Kanao[1]*

[1]* *Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8518, Japan*
Email: kanao@nipr.ac.jp

Summary. The Polar Environmental Data Science Center (PEDSC) of the Joint Support-Center for Data Science Research (DS), the Research Organization of Information and Systems (ROIS) has a responsibility to manage the data for Japan as a National Antarctic Data Center (NADC) during the last few decades. At the International Polar Year (IPY2007-2008), a significant number of multi-disciplinary metadata/data have been compiled mainly from IPY- endorsed projects. These collected metadata have a tight collaboration with the Global Change Master Directory (GCMD), the Polar Information Commons (PIC), as well as several centers belonging to the World Data System (WDS). Among activities in polar communities of the Scientific Committee on Antarctic Research (SCAR) and the International Arctic Science Committee (IASC), tighter linkages of data sharing within the Asian Forum for Polar Sciences (AFoPS) countries should be promoted by the involved Asian countries.

Keywords. Polar Environmental Data Science Center, data sharing , polar communities, AFoPS.

1. Introduction

Diverse data accumulated by many science disciplines make up the most significant legacy of the International Polar Year (IPY2007-2008). The Polar Data Center (PDC) of the National Institute of Polar Research (NIPR), followed by the Polar Environment Data Science Center (PEDSC) of the Joint Support-Center for Data Science Research (DS) have responsibility to manage these polar data in Japan as a National Antarctic Data Center (NADC). During the IPY, for instance, a significant number of multidisciplinary metadata records were compiled from IPY- endorsed projects. A tight collaboration has been established between the Global Change Master Directory (GCMD), the Polar Information Commons (PIC), and newly established World Data System (WDS). In this presentation, a history of data management at NIPR is summarized, focusing on the era after the IPY.

The PDC/PEDSC has been performed the function of a NADC for Japan and established a data policy in February 2007, based on the requirements of the Standing Committee on Antarctic Data Management (SCADM) of the Scientific Committee on Antarctic Research (SCAR). This contributed to the subsequent SCAR Data and Information Management Strategy. Several different aspects of scientific data collected in the polar region have great significance for global environmental research in this century. To construct an effective framework for long-term strategy of the polar data, data must be made available promptly and new Internet technologies such a repository network service must be employed. In addition to the activities in polar science communities of SCAR and the International Arctic Science Committee (IASC), tighter linkages must be established with other cross-cutting science bodies under ICSU, such as CODATA, and WDS.

2. AFoPS and perspective

The Asian Forum for Polar Sciences (AFoPS) is a non-governmental organization established in 2004 to encourage and facilitate cooperation for the advance of polar sciences among countries in the Asian region. The Forum consists of its six members, i.e, the national polar research institutions representing China, Japan, South Korea, India, Malaysia and Thailand. AFoPS also has four observer countries: Indonesia, Philippines, Sri Lanka and Vietnam, respectively.

The objectives of AFoPS has been recognized as

the value of scientific research in bi-polar regions for the benefit of human activities, recognizing the importance of international cooperation in the polar regions and the need to work closely with other national operators, as well as aiming to serve the common interests in both polar sciences and logistics. Member countries will work together for the tasks as follows; provide a foundation for cooperative research activities; present Asian achievements toward international polar communities; encourage more Asian countries' involvements in polar sciences.

Major Activities of AFoPS include; provide a forum to seek a common view on polar affairs among member countries; develop and support cooperative programs on polar activities (i.e., joint science projects, logistic cooperation and personnel exchange program between polar expeditions and institutes, etc.); convene joint symposia and workshops for sharing scientific results, information and experience joint symposium; conference activities within AFoPS working groups (WGs); support non member countries to develop their national polar programs; invite scientists to field expeditions and institutes; invite scientists to AFoPS meetings; provide personnel training and cooperation in outreach activities; produce joint publications on polar sciences.

Among data activities in polar communities of SCAR and the International Arctic Science Committee (IASC), moreover, tighter linkages of data sharing among AFoPS countries should be promoted in near future by the involved members of Asian countries. Establishing a new portal inside GCMD, for example, the simple and first step for data sharing and inter-operability. Detail discussion will be expected in this conference of DSWS-2018.

Acknowledgments. The authors would like to express their appreciation to many collaborators involving polar data communities, in particular to the member of PEDSC, data committee members on SCAR, IASC, WDS, CODATA and IPY.

Development of Japanese Research Data Discovery Service

Fumihiro Kato[1]*, ***Teruhito Kanazawa***[1], ***Kei Kurakawa***[1], ***Ikki Ohmukai***[1]

[1]**National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, 101-8430, Japan*
Email: fumi@nii.ac.jp

Summary. Finding and reusing research data are a crucial part of Open Science. CiNii Research is a discovery service for finding scholarly resources including research publications and datasets produced by research projects in Japan. It aggregates a variety of metadata related to Japanese research activities. Sources of metadata contain NII scholarly information services, institutional repositories, persistent identifiers and other scholarly services. It extracts research entities from aggregated metadata to construct a large-scale scholarly knowledge graph that illustrates relationships among research entities. It helps researchers not only to find research datasets and relevant articles used for their daily research activities but also to get an idea of a research theme to start their project. This presentation introduces development progress of CiNii Research.

Keywords. Open Science, Research Data, Discovery Service, Knowledge Graph.

1. Introduction

National Institute of Informatics (NII) hosts scholarly information services for researchers. CiNii[1] is a discovery service by NII for Japanese research literatures such as articles, books, and dissertations. It harvests and integrates metadata of publications from institutional repositories, the National Diet Library, academic societies and other scholarly databases in Japan. As sharing, finding and reusing research data are one of key concepts of Open Science, NII has launched a CiNii Research project to enhance CiNii to support research data as a first-class citizen since 2017.

CiNii Research aims to enable search and discovery of publications and datasets produced by research projects in Japan. CiNii Research consists of three components. The first one is to aggregate metadata related to research projects in Japan. The second one is to extract research entities and their relationships from aggregated metadata to construct a large-scale scholarly knowledge graph. The last one is to provide a discovery service for research entities by indexing nodes of the knowledge graph. This presentation reports on the progress of the development.

[1] https://ci.nii.ac.jp/en

2. Aggregation

The first component is to aggregate metadata of research entities related to research projects in Japan. NII has already collaborated with Japanese universities and institutions to collect metadata behind NII scholarly information services. IRDB[2] is a national aggregator of institutional repositories. It includes about 2.9 million records from 685 repositories as of Sep. 2018. And the number of datasets is about 55 thousand records (2.5% of total). As CiNii uses metadata collected by IRDB, we will update IRDB with JPCOAR Schema 1.0[3] which is the latest metadata schema for Japanese institutional repositories to support new features like identifiers and open access policies.

Another aggregator is KAKEN[4] that collects and hosts result reports of Grants-in-Aid for Scientific Research (KAKENHI) which is one of the major research funds by the Government of Japan. NII begins to collect persistent identifiers like DOI. Japan Link Center (JaLC)[5] is the DOI registry agency in Japan and NII is one of board members of JaLC. Hence JaLC will be our primary

[2] http://irdb.nii.ac.jp/analysis/index_e.php
[3] https://schema.irdb.nii.ac.jp/en
[4] https://kaken.nii.ac.jp/en/
[5] https://japanlinkcenter.org/top/english.html

DOI source as Japanese repositories use JaLC to assign DOIs to research entities including research datasets.

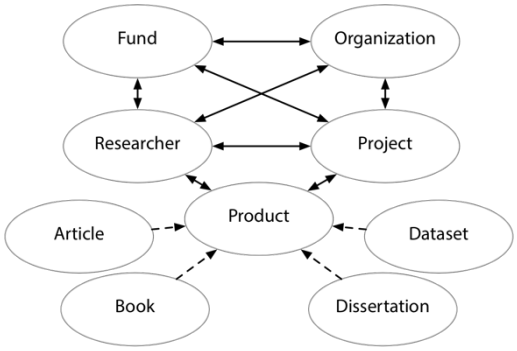


Figure 1. Entity Types

3. Knowledge Graph

Constructing a knowledge graph of research entities is an essential part of a modern discovery service as links between research entities help to find further related research entities. Figure 1 describes targeted entity types including Product, Researcher, Project, Organization and Fund. Product is defined as a superset of Article, Book, Dissertation and Dataset. CiNii Research currently focuses on Product, Researcher and Project to extract research entities from metadata and identify them with persistent identifiers and name disambiguation techniques.

Acquiring links between identified entities is the hardest part of creating a knowledge graph as explicit links in metadata are rather a few. Our challenge is to extract links from scholarly services to integrate into the knowledge graph. For instance, KAKEN has reports, products and researchers of projects so that the system can obtain links among them. The main issue of this challenge is that each scholarly service is independent and only a part of national researcher identifiers is shared. Therefore, linking the same entities among services is important.

Metadata including identifiers of research entities and links between research entities will increase in future as NII has been developing a new version of institutional repository system called WEKO3 to implement the JPCOAR Schema. The current WEKO is used by about 500 Japanese universities and librarians to input relevant identifiers in their public repositories. They will help us to grow and refine our knowledge graph.

Knowledge graph is also important for a global collaboration with other discovery services. Scholix [1] provides an interoperability framework for exchanging links between scholarly literatures and datasets. OpenAIRE provides LOD [2] to share their links. Our links should be also shared with international activities in future.

4. Discovery Service

CiNii Research provides a simple input form to search by keywords. A user can select a type described in the section 3 from tabs before searching words. If a user selects the “dataset” tab, search results are filtered only for datasets. CiNii Research does not support a facet search that typical discovery services implement for their search results to keep the results as much as simple currently.

The discovery service will connect to our research data management platform called GakuNin RDM [3] to import specific research data after finding it from search results.

5. Conclusions

CiNii Research is a discovery service for Japanese scholarly resources based on knowledge graph currently under development. It will help researchers to find research datasets and other resources related to research activities in Japan.

References

1. Burton, A., Koers, H., Manghi, P., et al., The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib,* 23, 1, 2017
2. Alexiou, G., Vahdatai, S., Lange, C., Papastefanatos, G., Lohmann, S., OpenAIRE LOD Services: Scholarly Communication Data as Linked Data. *Save-SD 2016, Lecture Notes on Computer Science,* 9792, 45-50, 2016
3. Komiyama, Y., Yamaji, K., Nationwide Research Data Management service of Japan in the Open Science Era. *In: 6th International Congress on Advanced Applied Informatics,* Hamamatsu, Japan, 129-133, 2017

Exploring public attitudes toward scientific research with visitor surveys and nationally representative surveys

Naoko Kato-Nitta[1,2]*, ***Tadahiko Maeda***[1,2]

[1]* *Center for Social Data Structuring, Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS), 10-3 Midori-cho, Tachikawa, 190-8562, Japan*
[2] *Department of Statistical Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, 190-8562, Japan*
Email: naokokn@ism.ac.jp

Summary. Despite the promotion of public engagement in science, there has been little empirical research on the attitudinal characteristics toward scientific research of visitors in science events and the extent to which such individuals are distinctive from the general population. We statistically explored this topic by contrasting samples from visitor surveys and archived nationally representative surveys. The results showed that the visitors believed more in the value of scientific research than the general public, but there was no difference regarding assessment of the levels of national science or of the national economy. Furthermore, visitors were more actively engaged than the general public not only in scientific activities such as going to science museums and attending science cafes, but also in literary and artistic activities such as going to art museums and reading novels or history books. This research should be a good practice for utilizing archived social survey data for promoting public communication of science.

Keywords. science communication, archived social survey data, visitor survey, the Japanese National Character Survey, scientific outreach activities.

1. Introduction

Scientific research bolsters national competitiveness and economies and often results in industrial innovations that provide solutions to the environmental challenges such as climate change. Yet, not all scientists know the better ways to convey their scientific findings to the lay public. In such circumstances, the governments in many countries including Japan promote the scientific contributions of scientists to the public through social activities such as the national institutes' open houses by disseminating scientific information to visitors [1].

Better understanding their visitors to such scientific outreach events and how 'the lay public' reacts for their scientific findings should be one of the great interests for scientists and science communicators; however, there is little data showing how visitors to scientific outreach events are distinct from the general public. We elucidated socio-cultural and attitudinal distinctiveness of visitors in science communication activities by statistically contrasting them to the respondents of nationally representative surveys [2]. Further, we explored determinants of visitors' exhibit viewing behaviours by utilizing different methods of measurement.

2. Methods

The statistical analyses used data from the following three surveys:

Survey 1: a 2012 visitor survey at the Institute for Molecular Science (IMS),

Survey 2: the 2013 Japanese National Character Survey,

Survey 3: a 2014 web-based Internet survey of Japanese citizens.

Statistical comparisons utilizing the above data were made based on the following hypotheses:

H1: Visitors (to scientific outreach activities) participate in science-related activities more often than the general public.

H2: Visitors participate artistic and literary-related activities more often than the public.

H3: Visitors show more favourable attitudes toward the value of science than the public.

H4: Visitors' assessments of the level of Japanese science, art, or economy are not different from those of the public.

The above comparisons include analyses of estimates adjusted for the distributions of the attribute variables of age, gender, and education with direct method of standardization.

For the above survey 1, we further assessed visitors' exhibit viewing behaviours of the total viewing time and the total number of exhibits viewed, and explored the determinants of those behaviours with regression analysis.

3. Results

The visitors believed more in the value of scientific research than the general public, but there was no difference regarding assessment of the levels of national science (Table 1) or of the national economy. To sum up the statistical tests, all the above hypotheses were basically supported.

The results of regression analyses showed that people who have higher scientific cultural capital viewed more exhibits and stayed longer at the events.

4. Conclusions

Scientific experts tend to think the people with whom they engage in dialogue during scientific outreach activities are the 'general public. However, our findings suggest that there are not only socio-cultural and attitudinal differences between visitors to their exhibits and the general public but also behavioural differences between highly engaged visitors and other visitors.

The results provide essential information for scientists involved in the institutional communication of science by providing insight into potential or non-attending visitors. We would like to emphasize that the empirical comparison between the general public and the visitors of the scientific institution was made possible with the archived social survey data that are representative of Japanese population.

Acknowledgments. This study was conducted under the ISM Cooperative Research Program (2016 ISM.CRP2- 2031, 2015 ISM.CRP 2-2039) and KAKENHI (15H03424). The data for Survey 3 used the data developed for KAKENHI (24380118).

References

1. Kato-Nitta, N., The influence of cultural capital on consumption of scientific culture. *Public Understanding of Science*, 22.3, 321-334, 2013
2. Kato-Nitta, N., Maeda, T., Iwahashi, K., Tachikawa, M., Understanding the public, the visitors, and the participants in science communication activities. *Public Understanding of Science*, published online before print, doi.org/10.1177/0963662517723258, 2017

Table 1. Differences in attitudes toward science

Items/Categories	Survey 2 (n=1585/1572)		Survey1							
			Crude (n=299 to 302)				Adjusted(n=261 to 292)			
	%	SE	%	SE	Diff test's Z[*4]	Chi-squared Test	%	SE	Diff test's Z[*4]	Chi-squared Test
Item 1.Science improves daily life?										
1.A lot	38.9%	1.224%	70.9%	2.658%	10.942	277.002	61.9%	4.402%	5.037	85.883
2.A little bit	45.7%	1.251%	23.6%	2.486%	-7.945	(df=3)	31.2%	4.318%	-3.229	(df=3)
3.Not at all	10.4%	0.767%	1.0%	0.590%	-9.695	<.001	1.3%	1.027%	-7.073	<.001
4.Other/Don't know	5.0%	0.547%	4.5%	1.207%	-0.402		5.6%	2.225%	0.248	
Item 2. Level of S & T in Japan										
1.Very high	34.7%	1.201%	35.7%	2.809%	0.329	2.624	38.5%	4.528%	0.810	3.070
2.Fairly high	52.2%	1.260%	49.1%	2.931%	-0.967	(df=3)	46.1%	4.550%	-1.299	(df=3)
3.Low(fairly/very)	7.3%	0.657%	9.6%	1.729%	1.247	0.453	9.6%	2.650%	0.853	0.381
4.Other/Don't know	5.7%	0.586%	5.5%	1.336%	-0.156		5.7%	2.332%	0.003	

AI and Data Science - Machine Learning as Digital Catalyst for Data Curation

Asanobu Kitamoto [1*, 2*]

[1*] *Center for Open Data in the Humanities (CODH), Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS)*
[2*] *National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan*
Email: kitamoto@nii.ac.jp

Summary. Data is a valuable asset, but how can we characterize its value? We propose four types of values, namely intrinsic value, basic value, added value and persistent value, and discuss how machine learning can contribute to increasing basic and added values. In the examples of automatic tagging for image collection, we realized that the role of machine learning-based approach is to give tags of general nouns, and provides a good starting point for human-based higher quality tags of proper nouns. We finally propose the "digital catalyst" metaphor to emphasize that machine learning can help to lower the barrier in the middle of findability workflow and reduce energy or motivation to start the tedious work of metadata annotation.

Keywords. Machine learning, FAIR data, data curation, image tagging, digital catalyst.

1. Conceptualizing the Value of Data

In the age of data-driven science and open science, data is a valuable asset. But what is the best way to characterize the value of data and how can we increase them? Here we introduce our conceptualization of the value of data into four types; namely intrinsic value, basic value, added value and persistent value.

1. *Intrinsic value* refers to the value of raw data, such as how rare or how relevant it is to solve a research question. This is especially important for scientists and scholars from the viewpoint of new findings and discovery, but its usage heavily depends on the expertise.
2. *Basic value* refers to the organization of data, typically with good metadata and standard format. This is especially important for librarians from the viewpoint of FAIRness to make data findable, accessible, interoperable and reusable.
3. *Added value* refers to the integration of data, typically with new combination and presentation. This is important for curators from the viewpoint of creativity to illustrate new values created from combination, or to allow serendipity for unexpected discovery.
4. *Persistent value* refers to the temporal dynamics of values projected in the long term. This is especially important for archivists from the viewpoint of selection and preservation of data with proper planning of infrastructure and organization.

Those types of values can be roughly mapped to existing jobs, but digital transformation requires re-definition of roles to take advantage of new methodology. In particular, this paper focuses on machine learning in terms of its contribution to increasing basic value and added value.

2. FAIR Data and Findability

In our definition, basic value is tightly connected to the concept of FAIR, and in the context of technical innovation, findability is the focus of attention. Findability can be increased by human labor such as annotating high quality metadata to the data, but this human labor is exactly the reason that prevents experts from doing it. How can technological innovation such as artificial intelligence (AI) increase the findability of data?

Knowledge representation is the first approach. For example, Google launched "Google

Dataset Search" in September 2018 to help find the dataset on the web. This service asks people to assign tags defined in schema.org, which is an interoperable schema across disciplines. Giving a tag requires human labor, but it is much simpler than writing a full description, and is still an effective solution for increasing findability.

Machine learning is the second approach. Machine learning also requires human labor for creating training dataset, but it is done in advance, and it could be done by others. Although machine learning's final goal is generalization to unseen data, the current technology works better on data similar with the training data.

3. Tagging for General or Proper Nouns

Machine learning has potential to be applied to any numerical data, but the most attractive result is currently obtained for general image data. To study the effectiveness of workflow to increase findability, we prepared four types of image collections, namely (1) ethnological field work photographs, (2) archaeological field work photographs, (3) disaster recovery photographs and (4) historical photographs. We then applied automatic image tagging [1] and had a brief interview with experts who created image collections to compare the current result and their expectation. Among other lessons, we focus in this paper on the role of automatic tagging.

We observed that automatic tagging is more effective on the general noun than the proper noun. For example, automatic tagging can give a "person" tag, while human tagging can go further to give person's name as a tag. Typically, it is considered that a tag of proper nouns, or entity names, is more valuable than general nouns, but the story is not that simple.

Ethnological photograph is one example that general noun works effective. In their field work, they record anything they think relevant, and later try to organize them to find characteristics of the local culture. In this case, a simple timeline organization is not useful because the subject of photographs is mixed in time and space. On the other hand, general noun tags are useful for roughly grouping photographs by subjects, which is useful for comparing multiple photographs.

Disaster photograph has similar characteristics, because disaster is also a multi-faceted event. We also realized that there is serendipity in automatic tagging. A tag given for a very different object is useful for grouping some types of objects that share the same appearance.

General noun has less value in archaeological photograph, because a proper noun, or an entity, such as ruins is the typical target of photographs. Its space-time variability is also smaller than other collections.

Finally, the collection of historical photographs is a promising target, because it captures people's daily lives more than specific objects. What kind of subjects are more often taken than others? This kind of research question could be quickly answered by automatic tagging approach.

4. Machine Learning as Digital Catalyst

The preliminary study suggests that machine learning is not about the automation of findability workflow, but works as "digital catalyst" to help human work. The term catalyst is inspired from chemistry as a substance to lower the barrier in the reaction process and require less energy to achieve the same result.

In a similar sense, the role of digital catalyst is to reduce high motivation energy requested for experts to complete findability workflow, such as metadata annotation. Basic tagging results at least provides a good starting point for higher quality metadata given by humans. In this sense, machine learning can be considered as digital catalyst to reduce the burden of human labor and allow humans to focus on higher quality work.

Acknowledgments. Photograph collections were provided from the following collaborators: Dr. Taku Iida in National Museum of Ethnology, Dr.

Yoko Nishimura in Toyo University, Ms. Hinako Suzuki in National Research Institute for Earth Science and Disaster Resilience, and Dr. Toshihiko Kishi and his colleagues in Kyoto University.

References

1. Krasin , I., et al. , OpenImages: A public dataset for large-scale multi-label and multi-class image classification, https://github.com/openimages, 2017

Embedding Machine Learning in the Data Life Cycle - An Example from Minerals Exploration

Jens Klump[1]*, ***Jess C. Robertson***[1], ***Alistair White***[1], ***Francky Fouedjio-Kameny***[1,2], ***Ryan R.P. Noble***[1], ***Nathan Reid***[1], ***Maciej Golebiewsky***[3], ***Ryan Fraser***[1]

[1]* *CSIRO Mineral Resources, 26 Dick Perry Avenue, Kensington, WA, 6151, Australia*
[2] *Stanford University, 450 Serra Mall, Stanford, CA, 94305, USA*
[3] *CSIRO Information Management & Technology, 36 Gardiner Road, Clayton, VIC, 3168, Australia*
Email: jens.klump@csiro.au

Summary. Machine Learning (ML), as a data science tool, has the potential to greatly accelerate the interpretation of data. However, data analysis and the use of ML tools are often applied only in the late stages of the data lifecycle. We used ML for identification of geochemical anomalies as part of our fieldwork. To make a decision about which areas to target for further exploration, companies must first characterise the 'background' chemistry to be able to robustly identify these anomalies. However, current workflows from field sample collection to lab analysis, to the production of background geochemical maps may take months. By chaining together a suite of novel hardware, portable characterisation and machine learning workflows we have reduced this initial part of the data lifecycle to 24 hours.

Keywords. machine learning, spatial data, mobile application, minerals exploration, geochemistry.

1. Introduction

The identification of chemical anomalies in potentially thick cover sequences has led to the discovery of large mineral deposits at depth. To make a decision about which areas to target for further exploration, companies must first characterise the 'background' chemistry to be able to robustly identify these anomalies. However, current workflows from field sample collection to lab analysis, to the production of background geochemical maps take months. By chaining together a suite of novel hardware, portable characterisation and machine learning workflows we have reduced the turnaround time for preliminary geochemical data acquisition and analysis to 24 hours. Our toolchain included:

- Digital metadata capture using tablets and a mobile server running a custom FAIMS sampling app,
- International Geo Sample Numbers (IGSN) for sample tracking and management,
- Field-based sample preparation coupled with optimised pXRF, and
- A machine learning system deployed in the field lab to generate updated maps, predict new measurements before they were taken, and suggest locations for new samples.

We demonstrate these efficiency gains in a geochemical survey situated in the remote Nullarbor Plain in South Australia. Using these tools and helicopter-based sampling we were able to collect and analyse samples and conduct targeted infill sampling to generate regional geochemical and proxy-mineralogical maps of soil and rock over an area of nearly 4000 km^2 within seven days, carried out the next day at the field camp.

2. FAIMS mobile app and IGSN

FAIMS Mobile is a framework for low-cost, customisable mobile and web applications to collect data during field research in the geosciences and other domains. FAIMS Mobile can integrate many different data types. Interfaces, data structure, automation, validation, and data export interfaces can be customised to

accommodate different research approaches and methods. All features work offline, with data synchronising among multiple devices opportunistically when a connection to a server is available [1].

IGSN is a persistent, globally unique, web-resolvable identifier for physical samples [2]. In this project, we used it to identify and track samples from the sampling event in the field and their analysis in the field lab, to their subsequent transfer to storage at our home institution.

The FAIMS mobile app was used with pre-printed sample labels. The contextual information available allowed us to populate almost all elements of the sample description automatically, resulting in significant time savings and improved data quality. We reduced field sampling times from 20 minutes to 5-6 minutes per site for five different sampling media.

3. Geochemical field laboratory

Samples were prepared for analysis at a provisional laboratory set up in the fieldwork area. For our preliminary analysis, soil samples were dried, ground to a fine powder, and pressed into pellets. The pellets were analysed using a portable X-ray fluorescence spectrometer (pXRF). Streamlining the workflow reduced analytical preparation and processing time to four minutes per sample.

Since not all elements of interest can be directly detected with this method we used other chemical elements present in the soil to act as proxies for the elements of interest. XRF analyses are influenced by the structure of the analysed material. We used a machine learning algorithm to correct the pXRF measurements.

4. Data analysis

A probabilistic model used the relationship between geospatial data and new pXRF data collected on-the-fly. The results demonstrate a better understanding of the regional geochemistry, quantitative assessment of variation and directed the infill sampling. These workflows will enable the minerals exploration industry to conduct field sampling campaigns more efficiently while at the same reducing the risk of missing the next major discovery by following a statistically robust sampling strategy [3].

5. Conclusions

In this field sampling campaign, we combined mobile data capture with on-the-fly data analysis into an uninterrupted processing chain. This machine learning enhanced workflow lead to significant savings in time and cost for conducting the sampling campaign and resulted in improved survey results based on a statistically robust sampling strategy.

Acknowledgments. The Coompana field campaign was supported by the Geological Survey of South Australia, IMDEX Ltd., and Aerotech Australia. We thank the traditional owners for their permission to conduct this survey on their land.

References

1. Ballsun-Stanton, B., Ross, S. A., Sobotkova, A., Crook, P., FAIMS Mobile: Flexible, open-source software for field research. *SoftwareX,* 7, 47–52, 2018
2. Devaraju, A., Klump, J., Tey, V., Fraser, R., Cox, S. J. D., Wyborn, L. A. I., A Digital Repository for Physical Samples: Concepts, Solutions and Management. *In Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (Eds.), Research and Advanced Technology for Digital Libraries (TPDL 2017), Cham, Switzerland, Springer International Publishing,* 10450, 74–85, 2017
3. Robertson, J. C., Cole, D., White, A. J. R., Fouedjio, F., Ballsun-Stanton, B., Noble, R. R. P., et al., Accelerating minerals exploration with in-field characterisation, sample tracking and active machine learning. *In Geophysical Research Abstracts,* 20, EGU2018-4169, 2018

Data Archives for Social Science in Korea, Challenges and Opportunities

Hearan Koo[1]*

[1]* *Korea Social Science Data Archive, Seoul National University Asia Center, Room 250, Gwanak-ro 1, Gwanak-gu, Seoul, 08826, Republic of Korea*
Email: hrkoo@snu.ac.kr

Summary. Open science has gained considerable attention in scientific communities in the past few years. One of the key elements in open science is open access to research data. While it is not a new idea, the recent development in open data movement presents many challenges and opportunities for social science data archives. This presentation addresses several issues on the sustainability of social science data archives in Korea. It consists of two parts; the first part introduces social science data archives and their development in Korea, particularly focused on KOSSDA (Korea Social Science Data Archive) and discusses the challenges and opportunities faced by KOSSDA in the wave of open science movement. The second part talks about the recent development of a social science data platform within Asian contexts, called NASSDA (Network of Asian Social Science Data Archives). It concludes with few remarks on future roles and position of data archives in open science movement.

Keywords. research data, social science, Korea, Network of Asian Social Science Data Archives (NASSDA).

1. Introduction

Open science has gained considerable attention in scientific communities in the past several years. While it means different things to different stakeholders, open access to research data is one of the key issues in open science agenda. In social science disciplines, it is not a new idea. For a long time, research data in social science has been opened and shared with broad research communities. Data archives have played a critical role in providing digital platform for researchers to open and share their own data.

The recent surge in open science, however, presents many challenges and opportunities for the existing data archives. This presentation addresses several issues on the sustainability of data archives in Korea.

2. Social Science Data Archives in Korea

Compared to Europe and US, social science data archives emerged relatively late in Korea. KOSSDA (Korea Social Science Data Archive), one of the leading data archives in Korea, started as a non-profit social science library in 1983. It transformed to an integrated platform for quantitative and qualitative research data and literature in 2006 and provides integrated online service of data and literature.

One of the reasons for late development is data culture. Data culture in Korean academic community was not favourable for data archives. Researchers did not consider data produced by public funds as a public good and were hesitant to share their data to re-use by others.

There has been a significant progress in data publishing, since KOSSDA has established. Many government-funded research institutions and other data-oriented organizations built their own repositories. Open Data Portal and Korean Research Memory, managed by government, began to service data online. However, there is little reward to be given to data sharing and data publishing does not gain the status of a scientific publication yet.

Meanwhile, the demand for open data is on the rise and open data policy evolved rapidly.

3. Challenges and Opportunities

The recent development in open science and open data movement presents the number of challenges and opportunities for data archives

3.1 Self-publishing vs. professional archiving

The current open data movement facilitates self-publishing of research data. A lot of open data platform such as Dataverse provide automated tools and services for self-archiving which have long been provided by professional archivists.

Self-publishing enables quick and easy publish without expert help and allows full control over researchers' own data. On the other hand, professional archiving guarantees consistent and secured archiving and assures quality of published data. Although different ways of archiving have their own merits, researchers can choose various data publishing options and it becomes a potential threat to the job of professional archivists.

3.2 General repository vs. domain repository

There are various kinds of repositories (CESSDA, 2017). General purpose repository piles up of numbers of data from different scientific disciplines. It accepts a wide range of data types and is suitable for cross-disciplinary data. Domain specific repository, on the contrary, tends to accept limited types of data and to focus more on specialized data management.

It is reported that using domain-specific repository for data sharing is not a preferred option in social science disciplines. Data archives need to devise more incentives that induce researchers to choose data archives for their open platform.

3.3 Repository vs. curation

According to Wiley's survey in 2016, the primary motivation behind data sharing is to increase the impact and visibility of one's research (CWTS, 2017). In this respect, data curation is essential.

Although repository offers search, navigate, and visualize functionalities, data curation in general repositories is limited. Data archives has invested a lot of efforts to curate and promote the re-use of data. Making published data searchable, visible, and re-usable is crucial for the future of data archives.

4. International networking

As part of extending data curation services, international networking is recognized for its importance. KOSSDA becomes aware global nature of research data and the necessity of coordination of data services in an international context.

KOSSDA has been actively involved in the discussion of founding NASSDA (Network of Asian Social Science Archives) with representative archives of three other countries – SSJDA in Japan, CNSDA in China, and SRDA in Taiwan.

NASSDA was born in 2016. It aims to build a sustainable infrastructure for social science research in Asian regions, which is intended to be an open platform of sharing and distributing social science research data.

5. Conclusions

Social science data archives are facing critical challenges in the wave of open science and open data movement. But challenges can become opportunities. As open science continues to evolve, it will restructure data ecosystem in Korea. Researchers and funders will acknowledge the values of data sharing and cooperate to implement appropriate open data policy. The successful introduction and enforcement of a data sharing policy will support the job of data management and curation. More innovative efforts of data archives are needed to maintain a stable position in data ecosystem.

References

1. CESSDA Training Working Group, CESSDA Data Management Expert Guide, Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/DMGuide, 2017
2. Centre for Science and Technology Studies (CWTS), Open data: The Researcher perspective, Elsevier, 2017

Toward universal information access on the digital object cloud

Kei Kurakawa[1*]

[1*] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan*
Email: kurakawa@nii.ac.jp

Summary. Universal information access on the digital network has been a long-lasting objective since the dawn of the Internet age. One of the technologies to aim at the universal information access is Digital Object Architecture proposed by Kahn and Wilensky in late 1980s. RDA (research data alliance) as a professional data community founded in 2013 re-focuses on this technology and intends continuous development and adaptation for scientific data. The paper describes the technological core information management and access scheme discussed in RDA leading to Digital Object Cloud as a promising next generation data world.

Keywords. Universal information access, digital object cloud, persistent identifiers (PID), PID centric data management and access.

1. Introduction

Universal information access on the digital network has been a long-lasting objective since the dawn of the Internet age [1]. In 1945, Vannevar Bush proposed the first idea of document management system. In 1960s, Ted Nelson introduced topics of hypertext, and Doug Engelbart implemented the ideas on working systems. The World Wide Web was invented by Tim Berners-Lee in late 1980s as covering these ideas of knowledge representation on the Internet. Digital Object Architecture [2] was proposed by Kahn and Wilensky in late 1980s to manage and access the digital objects on the Internet. These technological inventions continue to evolve into state-of-the-art information access.

A professional data community, RDA (research data alliance) was established in 2013 and launched biannual plenary meetings to develop community consensus and technological specifications on the professional data management and access among multi-disciplinary fields. In the professional data community of scientific experimental research, such as geoscience, life science, material science, experimental fields of natural language, several premises to manage and process the domain specific data exist in the heart of the community as follows. They are as follows.

- Computational data scheme should not be complicated, ever-lasting, and independent on computer technology changes.
- Data format and data attributes are complicated at some professional levels.
- Only the professionals can deal with data processing and management.
- Of course, the professionals have good knowledge of the domain.

In the context of the community understanding of the data management, the RDA mainly focuses on the Digital Object Architecture and enhance it to adopt for each disciplinary data management so as to share the data for everyone.

This paper illustrates the idea of the enhancement of the Digital Object Architecture discussed in the RDA technical community among which information scientists and engineers believe that Digital Object Architecture is the best way to universal information access for those professional data communities.

2. Digital Object Cloud

Digital Object Cloud (DOC) is the universal information access platform vision on the basis of

Digital Object Architecture. The following subsections explain elements of the component.

2.1 Digital Object Architecture

Digital Object Architecture [2] consists of the following elements; Digital Object (DO), Handle (a unique, persistent identifier (PID)), Handle System, DO repositories, DO registries. DO is the central object that is any unit of information represented in digital form, may be structured as a digital object within the Internet. The structure of a DO, including metadata, is machine and platform independent. Every DO has a unique persistent identifier or a Handle. The Handle Server is the resolution system that maps handles to state information of the digital object which includes location and others. DO can be stored in DO repositories, and DO registries let DOs to be federated among DO repositories.

2.2 Virtual layer for information access

DOC exists on the basis of the Digital Object Architecture [3]. It consists of the layered architecture, which are, from top to bottom, network of digital objects, network of internet devices, and computing and storage. The network of digital objects is virtually layered on the network of internet devices with its computing and storage. Users intend to get access digital objects by PID on the virtual layer.

2.3 PID centric approach to data management and access

PID is a key to get access to the digital object on the DOC. We assume that millions of DOs are spreading out on the DOC. To automate every process for the DOs, such as collecting, evaluating, pre-processing, calculating, publishing DOs, we need to know data types of DO and relationship attributes among DOs. Data type registries stores such the data types of DOs [4]. A local Handle server pushes relationship attributes among DOs, we call it kernel information, in response to the client which requests the state information of DO by a PID [5].

For automating data processing, data typing is a necessary task. To proceed this task, a data provider has two steps [6]. In the first step, it prepares data and its documents about data types. In the second step, it provides data with data types. Needless to say, the last step is inevitable for the automation.

2.4 Data discovery

Automatic data processing and data discovery are different tasks. Metadata to achieve those tasks are separately managed. Dataset search engine such as Google dataset search (GOODS) considers organizing datasets from search engine perspectives finally to provide catalogue metadata to users [7,8].

3. Conclusions and Future Work

Universal information access is a general requirement for all people who imagine the digital world of the Internet. RDA deal with technologies for this, especially Digital Object Architecture. This paper illustrated recent technical developments emerged in the RDA. We hopefully continue to develop this kind of technologies and adapt it for real professional data cases.

Acknowledgments. The author is thankful for RDA Kernel Information WG discussions.

References

1. Denning, P. J., Kahn, R. E., The long quest for universal information access. *Communications of the ACM*, *53*(12), 34. DOI: 10.1145/1859204.1859218, 2010
2. Kahn, R. E., Wilensky, R., A framework for distributed digital object services. *International Journal on Digital Libraries 6*, 2, DOI: 10.1007/s00799-005-0128-x. (First made available on the Internet in 1995 and reprinted in 2006 as part of a collection of seminal papers on digital libraries), 2006
3. Lannom, L., Wittenburg, P., Global Digital Object Cloud (DOC) - A Guiding Vision. http://hdl.handle.net/11304/a8877a1a-9010-428f-b2ce-5863cec4aff3, 2016
4. Broeder, D., Lannom, L., Data Type Registries: A Research Data Alliance Working Group. *D-Lib Magazine*, *20*(1/2), DOI: 10.1045/january2014-broeder, 2014
5. Weigel, T., Plale, B., Parsons, M., Zhou, G., Luo, Y., Schwardmann, U., Quick, R., Hellström, M., Kurakawa, K., RDA Recommendation on PID

Kernel Information. https://www.rd-alliance.org/sites/default/files/RDA%20Recommendation%20on%20PID%20Kernel%20Information.pdf, 2018

6. Kurakawa, K., Sekiya, T., Baba, Y., Making data typing efforts or automatically detecting data types for automatic data processing? *Research Data Alliance 11th Plenary Meeting*, Berlin, Germany, 2018.03.21-23. https://www.rd-alliance.org/sites/default/files/rda11_poster_20180321_kurakawa.pdf, 2018
7. Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., Whang, S. E., Goods: Organizing Google's Datasets. *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, 795–806, DOI: 10.1145/2882903.2903730, 2016
8. Google Dataset Search, http://g.co/datasetsearch

SPEDAS (Space Physics Environment Data Analysis Software): Multi-mission Heliophysics Data Management, Analysis, Visualization, and Collaboration

James W. Lewis[1]*

[1]* *University of California, Berkeley, Space Sciences Laboratory, 7 Gauss Way, Berkeley, California, 94720, USA*
Email: jwl@ssl.berkeley.edu

Summary. SPEDAS (Space Physics Environment Data Analysis Software) is a software framework and collection of tools developed to support heliophysics research. Its core features include the ability to download data from remote servers as needed, management of a local cache of downloaded data, a rich collection of analysis and modelling tools, and extensive support for time-series plots and other types of data visualization. In addition to the core, general-purpose tools, SPEDAS includes a plug-in architecture to permit rapid development and deployment of modules to support new missions, data types, and visualizations. By adopting an open-source development philosophy, adhering to recognized interfaces and metadata standards where possible, and implementing a robust quality assurance program, SPEDAS has emerged as a key enabling technology of the multi-mission Heliospheric/Geospace System Observatory (H/GSO) concept.

Keywords. Heliophysics Software Analysis Visualization Collaboration.

1. SPEDAS development history and motivation

In the early 1990s, researchers at the UC Berkeley Space Sciences Lab (SSL) [1] developed a set of software tools, written in the IDL programming language, to support the multi-instrument data analysis needs of the FAST satellite mission. [2] These tools (notably, the TPLOT package for plotting time-series data), were shared informally among researchers in the field, and continued to evolve.

In the early stages of the THEMIS (Time History of Events and Macroscale Interactions during Substorms) mission, the science team adopted this collection of tools as the preferred data analysis environment for both the space-based segment, and a complementary ground-based segment (a network of magnetometers and all-sky imagers). The THEMIS version of these tools was released as TDAS (THEMIS Data Analysis Software).

As researchers gained familiarity with TDAS, other projects with similar data analysis needs expressed interest in modifying TDAS to suit their needs. That effort eventually led to the initial release of SPEDAS, where the emphasis of the software was on the general-purpose, multi-mission capabilities, and TDAS became one of several plug-in modules that leveraged the general-purpose SPEDAS framework to meet the needs of specific missions.

2. SPEDAS capabilities

SPEDAS provides many of the tools and services needed to accomplish a typical workflow in heliophysics research:

- Automatically download data from remote servers and manage a local data cache
- Perform coordinate transformations and other analysis operations
- Interface with GEOPACK magnetic field modelling library
- Use HAPI (Heliophysics Application Programming Interface) or NASA's SPDF (Space Physics Data Facility) [3] to locate

and download data from a large selection of archived data sets

- Generate publication-ready plots using the command-line TPLOT package, or the SPEDAS GUI (Graphical User Interface)

3. SPEDAS development and distribution strategy

SPEDAS is under nearly continuous development, and enhancements and bug fixes are checked into the source code repository on a daily basis. A nightly snapshot of the source code repository is made freely available for download from the SPEDAS web site [5], for users who want access to the latest features.

On a semi-regular basis (1-2 times per year), formal SPEDAS releases are built, thoroughly tested by the SPEDAS core development team, and made available to the community.

To make SPEDAS available to the broadest possible audience, for every formal release, the SPEDAS team builds a set of VM (Virtual Machine) executables for Windows, Linux, and Mac, which are free to download and do not require an IDL license. Although the executable releases do not support access to the IDL command line, most of the key SPEDAS features are exposed via the GUI, so the free executable release is a viable option for many researchers and students.

4. Conclusions

SPEDAS enables the loading, analysis, and visualization of data sets from many different sources and missions, in a single, unified environment. Its ease of use, open approach to software development and distribution, and the development team's commitment to provide a quality software product to the widest possible audience, have led to its adoption by several other projects. The focus on multi-mission support has provided fertile ground for collaboration between missions, and achieving the goals of the Heliophysics/Geospace System Observatory.

Acknowledgments. The author is grateful to Dr. Yoshimasa Tanaka, Dr. Masaki Kanao, and the local organising committee for travel support and the opportunity to participate in this conference. This work was supported by NASA contracts NAS5-02099 and NNG17PZ01C.

References

1. University of California, Space Sciences Lab, http://www.ssl.berkeley.edu/ [September 2018]
2. McFadden, J. P., Ergun, R., Carlson, C. et al., Science Operations and Data Handling for the FAST Satellite, *Space Science Reviews*, 98,169-196, 2001
3. Space Physics Data Facility, https://spdf.gsfc.nasa.gov/ [September 2018]
4. Faden, J. B., Weigel, R. S., Merka, J. et al. , Autoplot: a browser for scientific data on the web, *Earth Science Informatics*, 3, 41-49, 2010
5. SPEDAS, http:/www.spedas.org [September 2018]

World Data Center for Microorganisms: The global cooperation on microbial big data

Juncai Ma [1*]

[1*] *Institute of Microbiology, Chinese Academy of Sciences,NO.1 Beichen West Road, Chaoyang District, Beijing 100101, China*
Email: ma@im.ac.cn

Summary. Microbial resources are one of the most important natural resources in the world, which is the scientific basis to support the development of biotechnology and life sciences. WFCC-MIRCEN World Data Centre of Microorganisms (WDCM), which once hosted in Australia and Japan and then moved to China in 2010, plays a crucial role in providing a database of microorganisms, analysis of the function and establishing a platform of international communication. WDCM launches the international project Global Catalogue of Microorganisms (GCM) to construct a data management system and a global catalogue to help organize, unveil and explore the data resources of culture collection worldwide. Now 112 international culture collections from 43 countries and regions joined GCM. GCM provides a comprehensive view on the microbiological material made accessible online by public collections, and the function of Analyzer of Bio-resources Citation. Future developments such as "BIG DATA" technology including semantic web or linked data will allow the system to provide more flexible data integration broader data sources. Linking WDCM strain data to broader data sets such as environmental, chemistry and research literature can add value to data mining and targeting microorganisms as potential sources of new drugs or industrial products. Cooperation with other organizations and institutions promoted broad utilization of WDCM data platform. The WDCM is exploring collaboration with the World Health Organization (WHO) for the establishment of a database allowing influenza virus information integration. Moreover, the WDCM database is able to provide services for the implementation of the TRUST code of conduct to allow culture collections to comply with the Convention on Biological Diversity and Nagoya protocol for Access and Benefit Sharing. The unique strain identifier available through the WDCM will and the further utilization of information extracted by ABC implements key provisions of the Nagoya Protocol and provides required transparency, legal certainty while lowering transaction costs and reducing administrative and governance burdens. The WDCM will work with Research Infrastructures, Publishers, Research funders, Data holders and individual collections and scientists to ensure data interoperability and provide the environment for enhanced tools for research and development. WDCM is prone to evolve and continue. WDCM now started the development of the data platform to microbiome. Recently, WDCM announced the launching of Global Microbial Type Strain Genome and Microbiome Sequencing Project in the 7th WDCM Symposium, marking the GCM project has begun to enter a new stage (GCM 2.0). Focused on exploring the genomic information of microorganisms, this project has planned to sequence all uncovered prokaryotic type strains together with select eukaryotic type strains, construct a database for genomics data sharing, and also provide online data mining environment. Working groups responsible for selecting bacteria and fungal strains, drafting SOP, managing intellectual property right and legal issues and constructing database have already embarked on the pioneer stage of GCM 2.0. The project will establish a cooperation network for type strain sequencing and functional mining, and complete genome sequencing of over 10000 species of microbial type strains in five years.

Keywords. WDCM database, GCM project, microbiome, BIG DATA, TRUST code.

Extended Cell Suppression Problem Towards Better Data

Kazuhiro Minami[1]*, ***Yutaka Abe***[2]

[1]* *The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo, 190-8562, Japan*
[2] *National Statistics Center, 19-1 Wakamatsu-cho, Shinjuku-Ku, Tokyo, 162-8668, Japan*
Email: kminami@ism.ac.jp

Summary. The cell suppression problem (CSP) for tabular data has been studied in the research community of official statistics for many years. In CSP, we usually suppress regular internal cells to marginal cells because information on marginal sums is more likely to be available elsewhere. We, however, find a class of tables with no cell suppression pattern for regular cells that protects sensitive cell values properly. Such a situation occurs when a small value of a marginal sum restricts the upper bound of each cell contributing to that sum narrowing the ranges of their possible values. To study positive aspects of suppressing marginal cells in CSP, we compare the approach of suppressing both marginal cells and regular cells with that of suppressing only regular cells after preprocessing unsafe rows and columns. Our initial experimental results show that suppressing marginal cells is an effective way of reducing information loss in terms of the number of suppressed cells.

Keywords. cell suppression problem, data utility, statistical disclosure control.

1. Introduction

The cell suppression problem (CSP) [1] for tabular data has been studied for many years. The goal of CSP is to find the minimum set of secondary suppressed cells (i.e., a suppression pattern) that ensures sufficient uncertainty on the value every sensitive cell under the presence of linear relationships concerning marginal sums.

In CSP, we usually suppress regular internal cells to marginal cells because information on marginal sums is more likely to be available elsewhere. However, when we apply a SDC tool for CSP [2] to tabular data produced from the official survey ``National Survey of Family Income and Expenditure,'' in Japan [3], we find that, if we suppress only regular internal cells in a table, there exist a class of tables with no feasible suppression pattern for a safe table. Such a situation occurs when a marginal sum restricts the upper bound of each cell contributing to that sum narrowing the ranges of their possible values. smaller than a given threshold value.

We consider two approaches to address this issue. One is to perform preprocessing on an input table to remove ``unsafe'' rows and columns, which we cannot ensure enough uncertainty of values for primary suppressed cells. After the preprocessing, we perform secondary cell suppression on regular internal cells. The other approach is to include marginal cells as possible targets for suppression.

In this paper, we experimentally evaluate the two approaches using tabular data produced from the official survey ``National Survey of Family Income and Expenditure [3]. Our experimental results show that to suppress marginal cells not only solves the issue of unsafe tables and but also reduces information loss of the original table effectively.

2. Background

Figure 1 shows an overview of the CSP algorithm. The algorithm takes an original table and security parameters as inputs and outputs a suppressed table and its corresponding suppression pattern, which is a binary matrix that specifies the positions of suppressed cells. There are two security parameters in the input. The unit frequency threshold t is used for primary suppression.

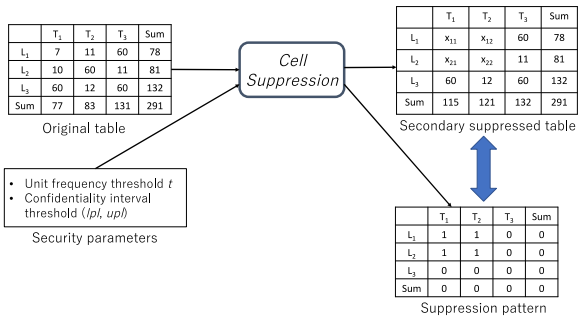


Figure 1: An algorithm for cell suppression problem (CSP)

The algorithm primarily suppresses cells whose value is smaller than t. The second parameter is a confidentiality interval threshold, which is a pair of two values lpl and upl, which ensures that each primary suppressed cell has a sufficient range of possible values as shown in Figure 2.

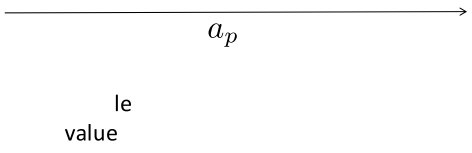


Figure 2: Confidentiality interval

3. Issue of no feasible solution

There are two cases where there is no feasible suppression pattern for tabular data. The first case is that a marginal sum s is smaller than an upper bound threshold upl. Then, the upper bound of every cell belonging to that marginal is smaller than upl, violating the condition in Figure 2. The second case occurs when a difference between the cell value a_p of a primary suppressed cell and its marginal sum a_{sum} is smaller than upl. If this is the case, $\overline{a_p} = a_{sum} < a_p + upl$ holds again violating the safety condition.

4. Experimental results

We evaluate two approaches to address the issue in Section 3. One is to perform preprocessing on an input table to remove unsafe rows and columns and then perform secondary suppression on internal cells. The other is to suppress marginal cells as well as internal cells. We use a set of frequency tables produced from the public survey data of National Survey of Family Income and Expenditure [3] in 2014. The sample size of the survey is 51,768 households, and their geographical areas in Japan are divided into ten regions. We prepare frequency tables of the households by crossing geographical regions and exact age of a head of household in several different ranges.

The bar graph in Figure 3 compares the total number of suppressed cells with the two methods. This results shows that the method of solving the extended CSP performs better than the preprocessing method in terms of data utility because the numbers of suppressed cells are smaller than those of the preprocessing method by 6% to 20% respectively.

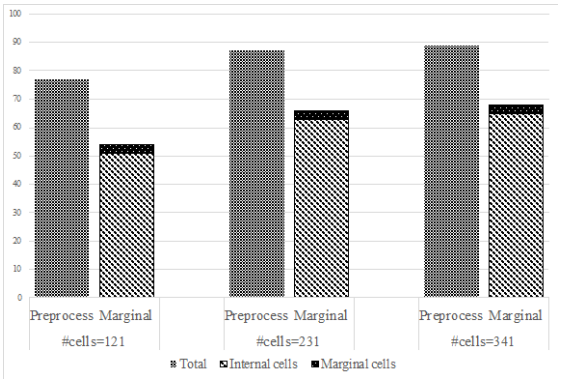


Figure 3: Comparisons of suppressed cells in total. The bars for the new method consists of the counts of internal cells and marginal cells with different colors.

References

1. Castro, J., Recent advances in optimization techniques for statistical tabular data protection. *European Journal of Operational Research,* 216, 257–269, 2012
2. Minami, K., Abe, Y., Statistical disclosure control for tabular data in r. *Romanian Statistical Review,* 4, 67–76, 2017
3. National survey of family income and expenditure, http://www.stat.go.jp/english/data/zensho/index.html

Amenability of the United Nation's Sustainable Development Goals to Big Data Modelling

Kassim S. Mwitondi[1]*, ***Barnabas N. Gatsheni***[2] ***and Isaac Munyakazi***[3]

[1]**Sheffield Hallam University, Faculty of Science, Technology and Arts; Sheffield S1 2NU. United Kingdom*
[2]*University of Johannesburg, Dept of Applied Information Systems, PO Box 524 Auckland Park 2006 RSA*
[3]*Ministry of Education, KG 7 Avenue, Kigali, Republic of Rwanda*
Email: k.mwitondi@shu.ac.uk

Summary. As the world gets increasingly entrenched into the information age, winners and losers are defined by levels of connectivity and utilisation of the potential knowledge that flows along shared networks. This paper examines the challenges and opportunities the developing world faces in attaining the United Nation's Sustainable Development Goals (SDGs). It explores the potential for converting SDG-embedded attributes into measurable data attributes amenable to Big Data modelling for socio-economic transformation. Assuming accessibility of locally authoritative data, interdisciplinary skills, infrastructure and relevant policies, each SDG is viewed as a data source and user node with measurable attributes. Attainment of the goals depends much on the strategies each country puts in place for harnessing and modelling the highly dynamic interactions of the data attributes associated with SDGs. This approach leads to the widely published, but probably still lacking ground level practical realisation, concepts of Big Data and Data Science. Based on theoretical implementations of unsupervised and supervised modelling, the paper proposes a Development Science Framework (DSF) for integrating and modelling data attributes harnessed from SDGs. The power of the proposed framework derives from the interaction between the inner layer - i.e., formalisation of the nature of relationships among the highly correlated, multivariate attributes across the set of nodes and the outer layer, a governing shell at each node. We show that sustainable attainment of the SDGs is a coherent phenomenon that is independent of the individual goals and heavily depends on these relationships and the utilisation of the knowledge from the integrated modelling environment. Real socio-economic data are used to validate the mechanics of an implementing algorithm in addressing data modelling inconsistencies and variability. Illustrations focus on interdisciplinary shared approaches to identifying what works, via which we expound the complexity of SDGs data and grey literature.

Keywords. *Big Data Modelling, Data Science, Development Science, Sustainable Development Goals, Value and Measurable Objectives.*

1. Background

In 2015 the United Nations member states signed up to 17 SDGs spanning across various aspects of life. Each goal was defined with measurable aims for improving human quality of life, via targets and indicators set to be achieved by 2030 [1]. The UN Statistics Division is responsible for collecting and cataloguing data from member states for monitoring progress-which is quite a big ask, given the dynamics and disparate standards and policies across member states [2]. Addressing these challenges, requires adaptive capacities in developing relevant tools and skills and promoting interdisciplinary collaborative work.

Based on theoretical implementations of unsupervised and supervised modelling, the paper proposes a Development Science Framework (DSF) for integrating and modelling data attributes harnessed from the SDGs. It highlights challenges and opportunities for exploiting inherent knowledge in data attributes as described in [3]. The DSF presents a general framework inspired by data- based considerations of SDGs attainment in a triangular

relationship of authoritative data availability, interdisciplinarity and development. Its main idea derives from viewing each SDGs in Figure 1 as a data source/user node with multi-faceted attributes and dynamics to be harnessed, stored and modelled in co-ordinated strategies with other nodes, and leading to attainment of Vision, Values and Measurable Objectives (VVMO).

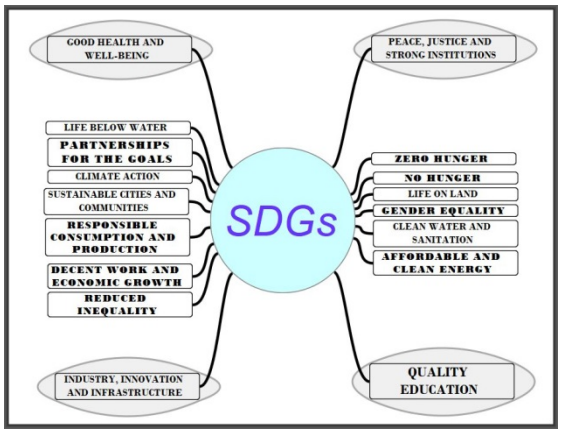


Figure 1: Sustainable Development Goals

This framework aligns with the widely published, but probably still lacking ground-level practical realisation, concepts of Big Data and Data Science. Thus, in carrying out the analyses, we follow the scientific approach proposed in [4], attributing science to a system of postulates and statements and sequentially testing them against actual experience via observations and experiments. The proposed framework traverses across the SDGs, providing spatio-temporal comparative examples.

1.1 Motivation

The link between computational intelligence and development is discussed in [5-7]. We set off from the reasoning that variations in data and modelling outcomes arise from estimation of sources of knowledge, hence the need for identifying the key attributes; understanding how they interact and appreciating the impact they have on societies. These requirements are fundamental in the modern era in which the potentials to harness and make use of data are increasingly outpacing capacities.

1.2 Research Question & Objectives

The paper seeks to answer the following question. How can developing nations convert the surrounding large volumes of data to useful knowledge? We set the following objectives.

1. General: Knowledge–based layout for SDGs.
   1.1. To propose a wide range of development data sources, including grey literature.
   1.2. To determine the natural path of data flows and how they can be followed.
   1.3. To establish how multiple interacting agents can identify the key attributes, understand how they interact and appreciate their impact on societies.
2. General: Knowledge Extraction from Data.
   2.1. To highlight theoretical and practical justifications for development based on inherent information in SDGs attributes.
   2.2. To outline the requirements for creating a Development Science Framework index.
   2.3. To propose an indexation prototype using SDGs' descriptive/inferential parameters.

Fulfilling the foregoing objectives hinges on knowledge management (KM)–a process described in [8] as requiring not only technical but also cultural adaptation. In other words, KM involves, *inter–alia,* associating each SDG note in Figure 1with people, time, locations, resources, organizations and conditions–forming a highly complex source of knowledge, often referred to as entity extraction.

The SDGs constitute a complex multivariate dataset with highly interacting variables. For example, the 4th SDG, Quality Education, focuses on access, inclusivity, universality as well as women and girls, among others. The goal also identifies Sub-Saharan Africa, a region also characterised by a high percentage of young population and high birth rates, as most struggling to access education. Many more attributes can be listed within this node and equally many can be directly or indirectly associated with attributes in the remaining 16 nodes. It is in this context that we envision SDGs as a Big Data problem and, consequently, expect countries and institutions to identify and manage

SDG-node related attributes for knowledge extraction purposes.

The paper is organised as follows. Section 2 outlines the methods–DSF structure, data sources and implementation strategy. Section 3 provides illustrations based on real data and discusses issues relating to data randomness and variation. Finally, concluding remarks are made in Section 4.

2. Methods

The SDGs in Figure 1 entail multiple targets and indicators and, as such, they naturally lead to critical questions and issues. Some of raised issues relate to the role of KM in Governments [9]. Thus, the DSF consists of multiple layers of highly disparate, and often fragmented, data that lead to objective identification, monitoring and adjustment of the Values and Measurable Objectives (VVMO).

2.1 Data Sources

Data sources, rated by comparative amenability to their SDGs Big Data Modelling, are in Table 1.

Source\Repository	Online Access	Rating
UN	http://data.un.org/	Average
UNSD	https://unstats.un.org/sdgs/indicators/database/	Good
WHO	http://www.who.int/healthinfo/statistics/en/	Average
FAO	http://www.fao.org/faostat/en/	Average
ILO	http://www.ilo.org/global/statistics-and-databases/lang–en/index.htm	Average
NISR	http://www.statistics.gov.rw/statistical-publications/subjects	Very Good
STATS SA	http://www.statssa.gov.za/	Very Good

Table 1: Amenability of SDGs Big Data Modelling

2.2 Implementation Strategy

The strategy is to treat SDGs-related data attributes as forming two strongly correlated layers-the *outer* and *inner* layer (Figure 2). The fabric on which the SDGs nodes operate epitomises *governance* and forms the *outer* layer-a voluminous and high–dimensional data node, largely a function of the prevailing internal and external political spectra. It plays an extremely vital role in attaining the VVMO. Its core function is to clear perturbations, crises or disagreements–i.e., VVMO support through the prevailing political spectrum.

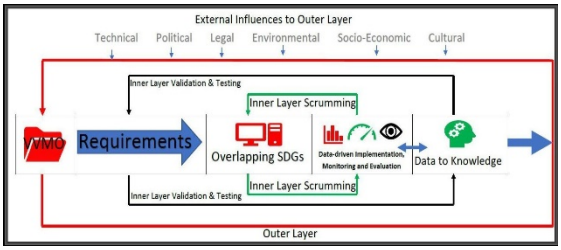


Figure 2: Layered structure for SDGs Big Data Modelling

The Inner Layer is multifarious and highly complex–it comprises of multivariate data attributes from all the SDGs, interactions, functional and hidden relationships that must be uncovered, monitored and appropriately adopted as inputs in enhancing the SDGs. The two sub–structures operate on the global fabric, influenced by several external factors as illustrated in Figure 2.

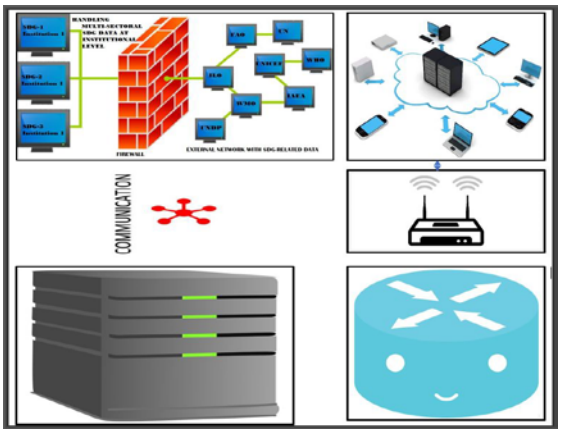


Figure 3: A graphical illustration of SDGs data sharing

Figure 3 illustrates a typical SDGs data communication mode required for implementing the structure in Figure 2, bearing in mind the ratings in Table 1. Together they define two main components of DSF–design and implementation. Both design and implementation require as inputs, authoritative data, tools and techniques, skills and resources–all governed by policies and other non–inner layer factors. It is imperative to recognise these as part of the data attributes to be modelled–not least because they are all embedded in one or more SDGs. We set off from the premises that development is a function of successful SDGs–a reasonable assumption for our research question. Hence, to determine the main drivers of SDGs, we focus on the SDGs-related

data tasks, tools, techniques and features in Table 2.

Data Task & Purpose	Tools	Feature
Data Creation, Access	Data Sources & Ontologies	Data Sharing
Development Index	Based on SDG Parameters	Dynamics
Data Visualisation	Exploratory, Pattern Recognition, Image & Shape Analysis	Adaptability, Dynamics
Data Clustering	Unsupervised Modelling: PCA, SOMs, K-Means, EM algorithm families, MCMC and all derivatives	Adaptability, Dynamics, Cohesion
Establishing confidence levels of association among data attributes	Association Rules, Probability	Confidence Levels, Priors, Linking Priors to Posteriors, Incorporating New Priors (knowledge)
Predicting, allocating new cases into known categories or those from newly uncovered structures	Simple regression, multivariate analyses, support vector machines, neural networks, decision trees, logistic regression	Model complexity, over-fitting, robustness, training & validation

Table 2: Selected examples of SDGs Big Data modelling tasks, tools, techniques and measurability features

It is also reasonable to assume that both the Gross National Product per Capita (GNPPC) and the Human Development Index (HDI) in a country reflect the country's productivity, which, on the other hand, is a function of the 4th SDG, Quality Education and the 9th, Industry, Innovation and Infrastructure. Clearly, the two SDGs are also highly correlated and so we need to pay attention to multicollinearity in analysing data from them.

2.2.1 Indexation and Visualisation

Given relevant data, we can extract related parameters and use them as proxies for the level of development in numerical or graphical formats. For instance, if we denote SDG–relevant variables by

$$\Omega = \{x_{ij}\}\, i = 1,2,\dots,n \;\&\; j = 1,2,\dots p \qquad (1)$$

where $n \gg p$, then, the parameters associated with Ω can either be computed from data samples–if distributional priors and densities are known, or they can be estimated from data. Without loss of generality, assumed that any $x_{ij}^{*} \in \Omega; \mathrm{i} > j \geq 2$ are associated with the parameters $\theta = \{\hat{\mu}_i, \hat{\sigma}_i, \hat{\rho}_{ij}\}$.

One common example is the Human Development Index (HDI) in its computational variants [10], a composite function of Life Expectancy (LE), Educational Attainment (EA) and Standard of Living (SL). One variant fixes the annual maximum & minimum values of each indicator and deduces the Achievement Level (AL)

$$AL_{LE} = \frac{\mu_{LE} - \min LE}{\max LE - \min LE} \Rightarrow HDI = \frac{AL_{LE} + AL_{EA} + AL_{SL}}{3} \qquad (2)$$

Due to the complexity of indexation, a study aimed at improving an SDG phenomenon may consider variables from sources as diverse as brainstorming, pilot projects, past experiences or lessons from elsewhere, most of which may be highly correlated. Table 3 presents few selected driver–variable examples for SDG #4 and SDG #9 to illustrate the level of complexity involved.

Phenomenon	Variables	Nature of Variable	SDG Source and Scope
Quality Education	Quality of Teaching	Composite	GDP, R&D, Teaching Colleges
	Teacher\Student ratio	Simple	Population, Teachers Output
	Knowledge Transfer	Composite	Industrial Partnership, Private Sector
	Education Index	Composite	Economic Development
	Vocational Training	Composite	Number of VTs
Productivity\Innovation	Investment	Composite	Infrastructure, wages, Training, FDI
	Technology	Composite	ICT, Knowledge Management
	Skills	Composite	Education Index, Vocational Training
	Markets	Composite	Exports, Imports, Legislations

Table 3: Selected driver–variables for SDG #4 and SDG #9

2.2.2 Unsupervised & Supervised Modelling

To optimally partition Ω we must identify k vectors $x_1^{*}, x_{2,\dots,}^{*} x_k^{*} \in \mathbb{R}^n$ that solve the function

$$\min_{x_1, x_{2,\dots,} x_k \in \mathbb{R}^n} f(x_1,\, x_{2,\dots,} x_k) = \sum_{j=1}^{p} \mathcal{D}_j\,(x_1,\, x_{2,\dots,} x_k) \qquad (3)$$

and if we let an indicator variable $z_{i=1,2,\dots,n}$, denote group membership, we can search for optimal values through iterative smoothing of the random variable $x|z = k$ for which we compute the parameters θ.

3. Conclusions

Prior to looking into causes and modes of improvement for any process, we must have a good understanding of experiences from the past–immediate and distant. The iterative nature of Figure 2 implies that to generate descriptive or inferential parameters, improvement requires that we determine what previously worked well and what did not. Answers to such questions lead to further questions such as to how previous strategies can be improved upon. The answer to this question can only be obtained if it is known as to why there were failures, if any. Finally, and most importantly, answers to all the foregoing questions are data–driven and, quite often, they lie uncollected in fragmented data files, project reports and grey literature. A quick inspection of selected open data sources and repositories in

Table 1 shows that data–driven SDG implementation is more likely to succeed if driven by data that are locally harnessed and utilised. Thus, while many multilateral institutions host global data on their servers, optimum benefits to local communities can only be achieved if locally authoritative data sources and repositories are used.

Limitations on modelling and domain knowledge skills are common, and one practical example of how this could be addressed is for Governments, the UN and other multilateral organisations to support national level development Data Science Centres by running regular free webinars hosted by interdisciplinary teams.

Acknowledgments. We are grateful to the National Institute of Polar Research for inviting us to this workshop and funding the participation of one of the authors. We are also thankful to our colleagues at Sheffield Hallam University, University of Johannesburg and the Ministry of Education in Rwanda for allowing us time to concentrate on this work.

References

1. SDGI. Sust. Dev. Goals Indicators, 2017
2. SDGs. Sustainable Dev. Goals, 2015
3. Berman, J., Principles of Big Data: Preparing, Sharing & Analyzing Complex Inform. Morgan Kaufmann, 2013
4. Popper, K., Objective Knowledge: An Evolutionary Approach,1972
5. Dean, T., Kanazawa, K., A model for Reason. *About Persist. & Causat. Comput. Intel.,* 5, 2, 142-150, 1989
6. Mwitondi, K., The Role of Comput. Intel. *In Dev. Countries; The First Seminar on Comput. Intel. for Societal Dev. in Dev. Countries, Sheffield,* 17th February 2017
7. Mwitondi, K., Big Data Challenges and Opportunities in the Dev. World. The 4th Intern. Conf. on Big Data Analy. & Data Mining: Future Techn. for Knowledge Discov. *Data,* 2017
8. Parsons, M., Godøy, Ø, LeDrew, E., de Bruin, T. Danis, B., Tomlinson, S., Carlson, D. A., Concept. Fram. for Manag. Very Diverse Data for Complex, Interdisciplinary Science. *Journ. of Inform. Science,* 37, 555-569, 2011
9. Liebowitz, J., Will Knowledge Manag. Work in the Govern.? *Elect. Govern. an Intern. Journ.* , 1, 1-7, 2004
10. Jahan, S., Jespersen, E., Human Development for Everyone. Human Development Report, UNDP , 2016

Age-Period-Cohort Analysis of Data Obtained from Repeated Social Surveys such as the Surveys on the Japanese National Character

Takashi Nakamura[1]*

[1]* *Center for Social Data Structuring, The Joint Support-Center for Data Science Research, The Research Organization of Information and Systems (ROIS), 10-3 Midori-cho, Tachikawa, Tokyo, 190-0014, Japan*
Email: nakamura@ism.ac.jp

Summary. Cohort analysis is a method of estimating the age, period and cohort effects from a set of repeated social survey data. However, this technique confronts an identification problem. The author has successfully overcome this problem by introducing a Bayesian model with a gradually changing parameter assumption and a procedure for selecting the optimal model based on the Akaike Bayesian information criterion (ABIC). This keynote presents the Bayesian logit age-period-cohort model and its application to the Japanese national character study data.

Keywords. Repeated cross-sectional survey, Identification problem, Bayesian logit model, ABIC.

1. Introduction

Cohort analysis is a method of estimating the age, period (calendar year), and cohort (birth time) effects from a set of repeated social survey data, classified by age group and survey period. This method is widely employed in various fields [1]. However, it confronts an identification problem, wherein the age, period, and cohort effects cannot be uniquely decomposed without some prior information. To overcome this problem, [2-4] introduced a Bayesian model with a gradually changing parameter assumption and a procedure for selecting the optimal model based on the Akaike Bayesian information criterion (ABIC) [5].

2. Bayesian logit age-period-cohort model for a standard cohort table

2.1 Standard cohort table

A table of observed proportions, classified by age group and survey period, is called a standard cohort table. In this table, the range (in years) covered by each age group equals the interval (in years) between successive survey periods. In such a table, a cohort travels over the cells of the table in the lower-right direction.

2.2 Logit age-period-cohort model

Let π_{ij} denote the response probability of a certain category of a question for the i^{th} age group in the j^{th} survey period. The logit age-period-cohort (APC) model decomposes the logit η_{ij} of π_{ij} into

$$\eta_{ij} \equiv \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta^G + \beta_i^A + \beta_j^P + \beta_k^C, \quad (1)$$
$$i = 1, \dots, I;\ j = 1, \dots, J,$$

where β^G is the grand mean and β_i^A, β_j^P and β_k^C are the effects of the i^{th} age, the j^{th} period and the k^{th} cohort, respectively. Further, I, J and K are the numbers of the age groups, survey periods and cohorts, respectively. Note that the relationships, $k = k(i,j) = j - i + I$ and $K = I + J - 1$ hold for an $I \times J$ standard cohort table.

Model (1) can be rewritten in terms of vectors and matrices as

$$\boldsymbol{\eta} = \beta^G \mathbf{1} + \boldsymbol{X}_A \boldsymbol{\beta}^A + \boldsymbol{X}_P \boldsymbol{\beta}^P + \boldsymbol{X}_C \boldsymbol{\beta}^C, \quad (2)$$

where $\boldsymbol{X}_A$, $\boldsymbol{X}_P$ and $\boldsymbol{X}_C$ are appropriate design matrices that express the relationship between the cells in a cohort table and the three respective factors. The parameter vectors $\boldsymbol{\beta}^A$, $\boldsymbol{\beta}^P$ and $\boldsymbol{\beta}^C$ in

Eq. (2) are subject to the following zero-sum constraints:

$$\mathbf{1}'\boldsymbol{\beta}^A = \mathbf{1}'\boldsymbol{\beta}^P = \mathbf{1}'\boldsymbol{\beta}^C = 0. \quad (3)$$

2.3 Log likelihood function

Let m_{ij} denote the sample size and y_{ij} the observed response count for the (i,j) cell. Assuming a product binomial distribution for $\boldsymbol{y}$, the kernel of the log likelihood function of model (2) with the constraints (3) is given by

$$\log f(\boldsymbol{y}|\boldsymbol{\beta}) = \boldsymbol{y}' \log \boldsymbol{\pi} + (\boldsymbol{m} - \boldsymbol{y})' \log(\mathbf{1} - \boldsymbol{\pi}),$$

where $\boldsymbol{\beta} = [\beta^G, (\boldsymbol{\beta}^A)', (\boldsymbol{\beta}^P)', (\boldsymbol{\beta}^C)']'$.

2.4 Identification problem

For an arbitrary real number Δ, let

$$\gamma_i^A = \beta_i^A + \{i - (I+1)/2\}\Delta,$$
$$\gamma_j^P = \beta_j^P - \{j - (J+1)/2\}\Delta,$$
$$\gamma_k^C = \beta_k^C + \{k - (K+1)/2\}\Delta.$$

Then,

$$\beta^G + \gamma_i^A + \gamma_j^P + \gamma_k^C = \beta^G + \beta_i^A + \beta_j^P + \beta_k^C = \eta_{ij},$$

can be demonstrated, which shows that model (1) has an infinite number of decompositions that provide the same η's. This is the identification problem in cohort analysis.

Consider the sum of the squares of the first-order differences of the effect parameters including

$$\sum_{i=1}^{I-1}(\gamma_i^A - \gamma_{i+1}^A)^2 = \sum_{i=1}^{I-1}(\beta_i^A - \beta_{i+1}^A)^2 + 2(\beta_1^A - \beta_I^A)\Delta + (I-1)\Delta^2;$$

this suggests that minimising the weighted sum of these differences is key to attaining a parsimonious decomposition and overcoming the identification problem.

2.5 Gradually Changing Parameter Assumption

To overcome the identification problem, an assumption, suggested in the previous subsection, is made, wherein successive parameters gradually change. This can be realised by minimising the following sum of squares,

$$\frac{1}{\sigma_A^2}\sum_{i=1}^{I-1}(\delta_i^A)^2 + \frac{1}{\sigma_P^2}\sum_{j=1}^{J-1}\left(\delta_j^P\right)^2 + \frac{1}{\sigma_C^2}\sum_{k=1}^{K-1}(\delta_k^C)^2, \quad (4)$$

where $\delta_i^A = \beta_i^A - \beta_{i+1}^A$ and σ_A^2, σ_P^2 and σ_C^2 are called hyperparameters.

Letting the difference vectors including $\boldsymbol{\delta}^A = \boldsymbol{D}_I\boldsymbol{\beta}^A$, $\boldsymbol{\delta}_* = [(\boldsymbol{\delta}^A)', (\boldsymbol{\delta}^P)', (\boldsymbol{\delta}^C)']'$ and the hyper-parameter vector $\boldsymbol{\sigma} = [\sigma_A^2, \sigma_P^2, \sigma_C^2]'$, a prior distribution,

$$\pi(\boldsymbol{\delta}_*|\boldsymbol{\sigma}) = (2\pi)^{-\frac{M}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{\delta}_*'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_*\right),$$

is obtained from Eq. (4) where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_A^2, \dots, \sigma_A^2, \sigma_P^2, \dots, \sigma_P^2, \sigma_C^2, \dots, \sigma_C^2)$, and $M = I + J + K - 3$. Now, model (2) can be rewritten as

$$\boldsymbol{\eta} = \delta^G\mathbf{1} + \boldsymbol{X}_A\boldsymbol{D}_I^-\boldsymbol{\delta}^A + \boldsymbol{X}_P\boldsymbol{D}_J^-\boldsymbol{\delta}^P + \boldsymbol{X}_C\boldsymbol{D}_K^-\boldsymbol{\delta}^C.$$

2.6 ABIC and MAP estimate

To estimate the parameter vector $\boldsymbol{\delta} = [\delta^G, (\boldsymbol{\delta}_*)']'$ (or $\boldsymbol{\beta}$), the hyperparameters are first determined by minimising ABIC, which is considered to be a function of $\boldsymbol{\sigma}$, defined by

$$\mathrm{ABIC}(\boldsymbol{\sigma}) = -2\log\int f(\boldsymbol{y}|\boldsymbol{\delta})\,\pi(\boldsymbol{\delta}_*|\boldsymbol{\sigma})\mathrm{d}\boldsymbol{\delta}_* + 2h,$$

where h is the number of free parameters. Once $\widehat{\boldsymbol{\sigma}}$ is obtained, the maximisation problem

$$\max_{\boldsymbol{\delta}} \log f(\boldsymbol{y}|\boldsymbol{\delta})\,\pi(\boldsymbol{\delta}_*|\widehat{\boldsymbol{\sigma}})$$

can be solved to obtain the maximum a posteriori (MAP) estimate $\widehat{\boldsymbol{\delta}}$.

Candidate models such as the G model with only the grand mean, A, P, C, AP, AC, PC, and APC models can be considered and compared. For example, the C model is expressed by $\boldsymbol{\eta} = \beta^G\mathbf{1} + \boldsymbol{X}_C\boldsymbol{\beta}^C$.

3. Application

An application of the above mentioned Bayesian cohort model to the Japanese national character study data is shown.

References

1. Mason, M. M., Fienberg, S. E., (Eds.), *Cohort Analysis in Social Research, Beyond the Identification Problem*. Springer-Verlag, New York, 1985
2. Nakamura, T., A Bayesian cohort model for standard cohort table analysis. *Proc. Inst. Statist. Math.*, 29, 77-97, 1982 (in Japanese)
3. Nakamura, T., Bayesian cohort models for general cohort table analyses. *Annals Inst. Statist. Math.*, 32, 353-370, 1986
4. Nakamura, T., Reconsideration of a Bayesian age-period-cohort model with age-by-period interaction effects. *Proc. Inst. Statist. Math.*, 53, 103-131, 2005 (in Japanese)
5. Akaike, H., Likelihood and the Bayes Procedure. In: Bernardo, J. M., et al. (Eds.), *Bayesian Statistics*, University Press, Valencia. 143-166, 1980

A Python library for parallelised particle filters

Shin'ya Nakano[1]*, ***Yuya Ariyoshi***[1, 2], ***and Tomoyuki Higuchi***[1]

[1]* *The Institute of Statistical Mathematics, ROIS, 10-3 Midori-cho, Tachikawa, 190-8562, Japan*
[2] *Now at Nippon Bunri University, 1727 Ichigi, Oita 870-0397, Japan*
Email: shiny@ism.ac.jp

Summary. Particle filters are state-estimation techniques based on Monte Carlo approximation that use a large number of particles. Particle filters are applicable even to nonlinear or non-Gaussian problems. They are thus used for a variety of purposes. However, particle filter methods take a prohibitive amount of computational time for high-dimensional problems when a huge number of particles required. Parallel computing is an effective way to reduce the computational time, but the particle filtering algorithm is not easy to parallelise. We are developing a Python library named P^3 (Python library for parallelized particle filter), of the particle filtering algorithms with high parallel efficiency. In this paper, we explain an overview of the design and characteristics of P^3.

Keywords. Particle filter, Sequential Bayesian estimation, Parallel computing.

1. Introduction

Particle-filters (PF) [1, 2] are state estimation techniques applicable to non-linear and non-Gaussian state-space models and they used for a wide variety of purposes, including nonlinear timeseries analysis, tracking of targets through moving images, and data assimilation. PF techniques, which are based on the philosophy of Monte Carlo methods, use large numbers of particles to represent the probability distribution of state variables and perform calculations via sequential Bayesian estimation. The accuracy with which the probability distribution is represented increases with the number of particles used, so in PF methods, using the huge number of particles is desirable. In particular, as the number of state variables to be estimated simultaneously increases, the well-known "curse of dimensionality" becomes increasingly prominent, and the number of particles needed to represent the probability distribution grows exponentially. Because the computational cost increases as the number of particles grows, the computation time is a major issue in practical PF implementations.

One approach for achieving high-speed inference using large numbers of particles is to use parallel computation. However, attempts to parallelise the standard filtering procedure used in PF methods not only result in extremely complicated programmes but also fail to achieve significant parallelisation efficiency. A number of methods for improving the efficiency of parallel implementations have been proposed [3-5], but the problem of programming complexity remains unresolved.

The code that we present here, P^3 (Python library for parallelised particle filter), is implemented as a Python library to facilitate the use of PF methods with high parallel efficiency. Although Python is an interpreted programming language, an abundance of powerful packages are available, including high-performance numerical programming libraries such as numpy and scipy, libraries for computational statistics and machine learning, and parallel computing tools such as mpi4py. Furthermore, the use of Python seems to have made significant inroads in many fields of scientific and technical computing. Existing Python libraries implementing sequential Bayesian estimation methods applicable to general state-space models include filterpy [6]. However, P^3 was designed with the specific intention of using parallel computation to handle

nonlinear, non-Gaussian state-space models of relatively large sizes.

2. Overview of P^3

P^3 offers two types of methods to improve the ease of parallelising the resampling step. One is hierarchical resampling [4] and the other is alternately lattice-pattern switching (ALPS) [5, 7]. For linear state-space models, to which Kalman filtering is applicable, specifying matrices such as the state transition matrix and observation matrix suffices to define the model. However, for the general nonlinear state-space models handled by PF methods, the way in which the problem is defined varies from problem to problem. For example, data assimilation problems make use of partial differential equations of various forms, and non-Gaussian probability distributions can be involved. To ensure that P^3 can be applied to a wide variety of cases, the code is designed to allow users some amount of freedom in defining the state-space model. To allow the computation of these equations to be carried out by P^3 code modules, the user must describe state vectors and state-space models in predefined formats (class State and class System, respectively) defined in system.py. The functions needed to execute PF calculations are provided as member functions of class Filter; once class State and class System have been defined, the user can invoke functions in class Filter to solve state-estimation problems.

Many existing computational libraries for sequential Bayesian estimation have used implementations in which the state vector is organised in the form of a single array. However, in high-dimensional state-space models, it is not uncommon for the various state variables to have different meanings. In such cases, the range of possibilities addressable by a computational code is enlarged by allowing state variables with different meanings to be treated as variables of different names. In particular, for data assimilation problems in which system models are constructed from simulation models, the variable names used in the original simulation code are available, and it is convenient to use these when writing code for system models as well. For this reason, P^3 does not store state vectors as arrays but rather as instances of a class (structure) called class State. Detailed explanation on P^3 is given in our recent paper [8].

Acknowledgments. The development of P^3 was supported by a Grant-in-Aid for Scientific Research (B) (project number 26280010) from the Japan Society for the Promotion of Science.

References

1. Gordon, N. J., Salmond, D. J., Smith, A. F. M., Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F*, 140, 107–113, 1993
2. Kitagawa, G., Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comp. Graph. Statist.*, 5, 1–25, 1996
3. Bolić, M., Djurić, P. M., Hong, S., Resampling algorithms and architectures for distributed particle filters. *IEEE Trans. Signal Process.*, 53, 2442–2450, 2005
4. Nakano, S., Population-based quasi-Bayesian algorithm for high-dimensional sequential problems and hierarchization of it for distributed computing environments. in *Proceedings of 2010 IEEE World Congress on Computational Intelligence*, 2010
5. Nakano, S., Higuchi, T., A dynamic grouping strategy for implementation of the particle filter on a massively parallel computer. in *Proceedings of 13th International Conference on Information Fusion*, 2010
6. Labbel, R., Kalman and Bayesian Filters in *Python*, (http://filterpy.readthedocs.io), 2015
7. Nakano, S., Higuchi, T., Weight adjustment of the particle filter on distributed computing system. in *Proceedings of 15th International Conference on Information Fusion*, 2480–2485, 2012
8. Nakano, S., Ariyoshi, Y., Higuchi, T., P^3: Python Parallelized Particle filter library. *Proceedings of the Institute of Statistical Mathematics*, (under review)

Data Treatment Strategies and Some Science Projects of PANSY Radar

Koji Nishimura[1, 2*]

[1*] *Polar Environment Data Science Center, Midoricho 10-3, Tachikawa, 1908518 Japan*
[2] *National Institute of Polar Research, Midoricho 10-3, Tachikawa, 1908518 Japan*
Email: knish@nipr.ac.jp

Summary. The Project of Antarctic Syowa Station MST/IS Radar (PANSY) has started full time full power operation of the radar from 2015 in Syowa station (69S, 39E). In this workshop, we paresent a talk about our current strategies about data archive, transport and publication, and some related science projects.

Keywords. PANSY Radar, data transport, data archive, data publication.

1. Introduction

PANSY Radar [1] is capable of measuring the neutral atmosphere from 1.5 to about 100 km in altitude, and scatter from the ionosphere. Syowa Station, in which the radar is installed, is isolated in data communication, which depends on a geostation-ary satellite channel, as well as in physical access solely supported by a dedicated icebreaker ship. Radar is a type of instrument that potentially generates large amount of data and the limitation of the data transport is a persistent issue to be dealt with to maximize the potential of the instrument. In this presentation, we talk about data management and transport strategies, and science projects that depend on the use of massive amount of data.

2. Data Transport & Archive

Our radar is continuously in operation and yielding data of roughly 3 MB/min (~4 GB/day) as time series. In order to reduce the occupancy in the communication bandwidth, however, we reduce the time resolution of data to about 400 kB/min and transfer realtime via Intelsat to Japan. The full resolution original data are manually carried to Japan once a year by the ship. Until the transport, for a risk of accidental losses, the data are archive with the distributed storage system specially developed for PANSY, which is installed in physically separate buildings in the station.

3. *Some Future Projects*

We conduct some science projects that take advantage of massive amount of data.

1. Multiple channel Imaging Observations

Using multiple subarrays as independent receivers, it is expected that we could obtain high resolution spatial image of the scatters from the mesosphere caused by a several hypothetical mechanisms.

2. Low Altitude Observations:

The radar has a lack of sensitivity at the boundary layer from the ground to 1,500 m because of the blanking time for receiver protection. We are planning to employ a small set of auxiliary antennas solely dedicated for reception to cover this height range. In order to obtain accurate estimate of wind velocities from the equipment, however, we need to solve a complicated inversion problem using quite a lot of data real time.

3. 4D Turbulence Observations.

At present, estimating the strength of the atmospheric turbulence from echoes requires some specific spatio-temporal model that has not been approved by observations. We have developed a theory directly to observe the space-time spectrum of the turbulence using massive multi-beam multi-channel radar observations.

References

1. Sato, K., Tsutsumi, M., Sato, T., Nakamura, T., Saito, A., Tomikawa, Y., Nishimura, K., Kohma, M., Yamagishi, H., Yamanouchi, T., Program of the Antarctic Syowa MST/IS radar (PANSY). *J. Atmospheric Sol.-Terr. Phys.,* 118, 2-15, 2014

Domestic and international activities of DOI-minting to solar-terrestrial physics data and their citation in publication

Masahito Nosé [1]*, ***Yasuhiro Murayama*** [2], ***Takenari, Kinoshita*** [3], ***Yukinobu Koyama*** [4], ***Michi Nishioka*** [5], ***Mamoru Ishii*** [5], ***Manabu Kunitake*** [2], ***Koji Imai*** [2], ***Toshihiko Iyemori*** [6], ***and Takashi Watanabe*** [2]

[1]* *Institute for Space-Earth Environmental Research, Nagoya University, Nagoya, Japan*
[2] *Strategic Program Produce Office, National Institute for Information Communications Technology, Tokyo, Japan*
[3] *Japan Agency for Marine-Earth Science and Technology, Kanagawa, Japan*
[4] *Oita National College of Technology, Oita, Japan*
[5] *World Data Center for Ionosphere and Space Weather, National Institute for Information Communications Technology, Tokyo, Japan*
[6] *World Data Center for Geomagnetism, Kyoto, Kyoto University, Kyoto, Japan*
Email: nose.masahito@isee.nagoya-u.ac.jp

Summary. Data-DOI, data publication, and data citation will promote "Open Science". Recognizing their importance, solar-terrestrial physics (STP) data centers in Japan have been working to mint DOI to their database since August 2013. We participated from October 2014 in a 1-year pilot program for DOI-minting to science data launched by Japan Link Center, which is one of the DOI registration agencies. In the pilot program, a procedure of the DOI-minting for STP data was established. As a result of close collaboration with Japan Link Center, the first case of data-DOI in Japan (doi:10.17591/55838dbd6c0ad) was created in June 2015. The first case of data citation in Japan was also made. As of September 2018, there are 18 data-DOIs for the STP data. The effort of the DOI-minting to scientific data will be continued for database in our data centers and those in other fields of Earth Science.

Keywords. data DOI, data publication, data citation, World Data System, solar terrestrial physics.

1. Introduction

The DOI system was originally developed by publishers and introduced as a common identifier for publication in late 1990s. Now, more than 5000 publishers participate in the DOI system. DOI is applicable not only for usual publication articles but also for any objects such as a piece of online content (e.g., PDF files, movie files, etc.) or a physical asset (e.g., DVD, an item of equipment, rock samples, etc.). Therefore, it can be mint to research data or database.

Minting DOI to scientific database and data citation with DOI provide much benefit to both researchers and data providers. (1) Researcher can more easily locate the data used in the paper, obtain necessary information of the data (i.e., metadata), and validate the findings of the paper. (2) Data providers can gain professional recognition and rewards for their labors to publish and manage data set in the same way as for traditional publications. Figure 1 explains a benefit of DOI-minting and data publication for data providers. First, data providers publish their data with DOI. This activity is called "data publication". When researchers or data users use the data in their researches and publish a paper, the data can be cited in that journal papers with the DOI (i.e., "data citation"). Then, as usual publication, it becomes possible to count the usage of data or to measure the importance of the data, which is called "data citation metrics". According to the results of metrics, the data providers can gain "rewards" for their published data. This gives motivation or incentive of data publication to the data providers, resulting in formation of the positive cycle.

Figure 1. Benefit of DOI-minting for data providers. Positive cycle is formed with data publication, data citation, data citation metics, and rewards for data publication.

2. DOI-minting to solar-terrestrial physics data

Recognizing the importance of data citation, solar-terrestrial physics data centers in Japan started discussion to mint DOI to their own database in August 2013. In DOI-minting, we need a contact to a relevant "Registration Agency" which is qualified by International DOI foundation (https://www.doi.org/registration_agencies.html). There are currently 10 registration agencies, some of which are Crossref and DataCite. Each registration agency has its own field of specialty. After survey and discussion, we find that Japan Link Center (JaLC) is a proper agency to register DOIs for our case, because JaLC aims at public information services to promote science and technology in Japan and it handles scientific and academic metadata and content from holders nationwide, including national institutes and universities.

As shown in Figure 2, we develop a web-based system to register metadata with JaLC and to create landing pages of data. This system is shared among the solar-terrestrial physics data centers. JaLC assigns the DOI prefix to each data center. The data centers submit metadata of dataset, to which they want to mint DOI, in addition to information of DOI suffix. Then the registration server sends a query of mapping between DOI and URL of the landing page to JaLC, resulting in completion of the DOI-mining. The system can handle version of the landing pages when the data are updated.

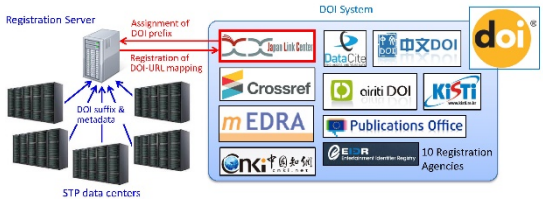


Figure 2. DOI system and 10 registration agencies. Japan Link Center (JaLC) is chosen for a registration agency in our practice. We develpoed a sever that works solar-terrestrial physics data centers and JaLC to mint DOI to the solar-terrestiral physics database.

JaLC started a 1-year pilot program to mint DOI to the database from October 2014. As a result of close collaboration with JaLC, we successfully created the first case of data-DOI in Japan (doi:10.17591/55838dbd6c0ad) in June 2015 for mesospheric wind velocity data (30min. mean) observed with MF radar at Poker Flat, Alaska. As of September 2018, there are 18 data-DOIs for the STP data in Japan.

3. Conclusions

Since we have more data in our data center, we will continue this activity. We will also share our experience and system with others who are interested in DOI-minting to their database.

In the field of geophysics (including solar terrestrial physics), interests to the DOI-minting are rapidly growing. At the next International Union Geodesy and Geophysics General Assemblies that will be held in Montreal, Canada, in July 2019, a joint session entitled "Geoscience data licensing, producing, publication and citation" is planned. This session solicits contributions presenting actual practices and future plans of data licensing, producing, publication, and citation of scientific data, and possible related topics. The international effort will be continued for DOI-minting to scientific data in solar terrestrial physics.

Importance of semantics and related challenges in science and humanities according to the Humboldtian ideal

Bernd Ritschel [1*]

[1*] *Graduate School of Science, Kyoto University, Oiwake-cho, Kitashirakawa, Sakyo-ku, Kyoto, 606-8502, Japan (until 2016)*
Email: berndritschel@yahoo.de

Summary. The Humbolditan education ideal was created by the brothers Alexander and Wilhelm from Humboldt at the beginning of the 19th century. Both together formed the Prussian concept of a humanistic and holistic education and science system, combining both humanities and natural sciences. The philosopher Wilhelm emphasizes language skills, linguistics and semantics as main part of human and societal culture and therefore as a main pillar for furthering universal education and leading to a high level of general knowledge. His universal concept of humanistic education combines in an ideal way humanities, natural sciences and art. Another pillar of his approach is academic freedom in education, science and research for students, teachers and scientists. Freedom especially means the independence from dogma, authority, tradition and politics. This also implies the non-interference and non-intervention of public and private funders with topics in education and science. Alexander and Wilhelm can be appreciated as the masterminds of a modern global Open Science approach, emphasizing a world-wide network of science and research, including a cross-domain inquiry and exchange of scientific data and results. While the success of these concepts in Germany had led many countries, such as USA and Japan to integrate Humboldtian ideas into their education systems, the situation has shifted dramatically. Nowadays the trend of market-oriented education and research in Germany as well as the influence of political correctness and gender mainstreaming is increasing, and is largely ignoring the Humboldtian humanistic education and science ideal. There is also a tendency that research funding increasingly comes from private entities and therefore research activities are especially influenced by private economic interests. The process of neglecting the Humboldtian ideal is accompanied by the change of semantics and the reinerpretation of well known concepts in science and society. Instead of a revival of Humboldt's concepts, higher education systems and standards are mainly driven by economic concerns and neoliberal ideas of the institutions. The new premise for education, science and research, inevitably leads to a conflict between humanistic and market-driven approaches to higher education. There are serious doubts about the long-term sustainability of this new model. The focusing on the market as main indicator for higher education, as well as scientific-technological and social progress, will narrow the human mind and intellectual power, and possibly even set back or stop further societal developments. Acontemporary humanistic and ethics-driven Humboldtian approach should also contain substantial education about recent and modern history. This knowledge is necessary for a comprehensive impact assessment of science and research activities for society and beyond, e.g. in nuclear sciences, but also technological developments, e.g. artificial intelligence and autonomous systems.

Keywords. Humbodlitan Ideal, Political Correctness, Gender Mainstreaming, Semantics, Concepts.

References

1. Ritschel, B., The importance of the Humboldtian ideal for education, science and research at present, http://wdc.kugi.kyoto-u.ac.jp/wdc/news/1603.pdf

Development of Metadata, Conversion and Archiving of the Time Series Data of the Completed Censuses and Surveys of the BBS

Chandra Shekhar Roy[1]*

[1]* *Senior Maintenance Engineer-IT, Bangladesh Bureau of Statistics, Statistics & Informatics Division, Ministry of Planning, E27/A, Agargaon, Dhaka-1207, Bangladesh*
Email: csroy.sme@bbs.gov.bd

Summary. Bangladesh Bureau of Statistics (BBS) has vowed to convert Census/Surveys micro data for next generation technology format use processed in IBM proprietary technology. After Bangladesh's independence, there is a rich repository of statistical information in IBM 360 to ES/9000 mainframe 8600 tapes, dating back to late 1970s and early 2000s time. The overarching objective is to strengthen the prevailing national statistical archiving system. BBS will be making available of this large volume of converted data to the citizens of Bangladesh (if needed world community) so that academic and scholarly debates can take place taking cognizance of historical data. This is a big step towards dissemination of Big Data covering Bangladesh's economic developmental trend since its birth in 1971. By revisiting time series data, it is hoped that well-informed and meticulous policies can be designed and formulated in the future. Most fundamentally, availability and easy accessibility to such a large volume of Big Data will inspire reassessing economic theories and indicators of development informing Bangladesh's position in global rankings like SDG.

Keywords. Mainframe, Historical data, IBM, Archiving, SDG.

1. Introduction

The Bangladesh Bureau of Statistics (BBS) is the government's apex body providing technical and administrative guidance in matters of all official statistics in Bangladesh. BBS is the national statistical organization (NSO) responsible for collecting, compiling and disseminating statistical data of all the sectors attributable to the Bangladesh economy. The underlying function is to meet data-needs of diverse users and stakeholders, e.g. national level planners, development partners and other agencies. Presently, the Bureau of Statistics possesses a large volume of statistical information dating back to the time of Bangladesh's independence since 1971.

However, much of these data sets were stored on computer media that is no longer in common use, deterring accessibility to such historical data. In 1973, BBS installed the IBM System 360, followed by IBM System 4341 in 1981 and IBM ES/9000 a year later. Following emergence of disk-system, these mainframe tapes, as a storage tool, faced serious criticism with their maintenance costs in Bangladesh. With the onset of personal computers and a proliferation of statistical software in the 1990s, the IBM ES/9000 was discontinued from 1999 onwards.

Fortunately a huge quantity of 9-track tape was used to preserve the micro data. So there is a rich repository of statistical information (census, survey and geo-code) in these mainframe tapes, dating back to late 1970s and early 2000s. The Government of Bangladesh has initiated a project to convert all the time series statistical microdata set in a re-useable next generation technology format. In this initiative, a legacy data recovery lab will also be set up in the BBS.

2. Towards Big Data

The overarching objective of this step is to strengthen the prevailing national statistical archiving system. The BBS will be making available of this large volume of converted data to the users in Bangladesh and for the world community, so that academic and scholarly

debates can take place for the cognizance of historical data.

Data recovery and conversion from mainframe EBCDIC format into ASCII is a big step towards dissemination of Big Data covering trend of economic development in Bangladesh since her birth in 1971. By revisiting historical trends, it is hoped that well-informed and meticulous policies can be designed and formulated in the future. Most fundamentally, availability and easy accessibility to such a large volume of Data will inspire revisiting economic arena and indicators of development informing Bangladesh's position in global rankings.

3. Outcome of the Project:

-Supply older/historical data to the researcher's, stakeholders along with Metadata;
-This information will be helpful to review the history of progress of the country;
-Opportunities for sharing historical data both local and foreign countries;
-Legacy data set, will be helpful in big data preparation;
-Time series data, to predict future values based on previously observed data.
Data Conversion project activities starts with Planning, leading to Analysis and Design, progressing to Conversion of Data, finishing in Metadata - where converted data loaded in the Time series database.
In this stage BBS, in conjunction with IT expert & Statistician, prepares a plan, which is also intended to shorten the duration of the Conversion along with metadata process and also reduces business impact and risk.

4. Execution of Data Conversion

In our dram project 60% percent works are Machine intervention i.e. re-engineering where a group of IBM mainframe compatible hardware (9-track tape drive, Spool tape cleaner) need to be used. More than one scientific oven also needs to be use to remove fungus and stickiness of 9-track spool tapes.

i. Exclusive data conversion software needs to be use to convert data from EBCDIC to ASCII format. a) First of all this software will capture the data from the magnetic and will create an Image file, So that physical tapes no more needed in future. b) Another beauty of this software is It will create ASCII text from the captured image files. Finally, we will get the numerals data i.e. 0 to 9. Statistically we called it data records. All these records are meaningless unless we shape it with proper layout/dictionary.

ii. Rest 40% of our project work is Human intervention. Within the 8600 tapes there is programming tapes, data tapes and geo-code tapes. Most of the programs were written under COBOL language. In the programming and Geo-code tapes the output come numerals with characters. There is a challenge if programming tapes fails to read or convert corresponding data tapes, we need third party statisticians or SME assistance. Defining and implementing data quality standards to ensure consistency across the different databases.

a) Their duty is to create metadata of the converted ASCII file data. In this case, they need hard copy of survey/census data collection questionnaire/schedule. It can be found in the publication of that particular census/survey.

b) Data Profiling and Cleansing: Ensure that proper data profiling and data cleansing procedures are in place so that the original data is of high quality. This helps to smoothen out the subsequent data conversion procedures.
In light of data users view metadata should be very constrictive and under a database also, we called it Metabase (metadata + database) where each and every instructions need to be mentioned. Following data conversion, ensure that the duplicate master data has eliminated, reducing the risk of incorrect transactions and unreliable reports. The project should satisfy all principles of data management and data governance. The project adherence to FAIR principles: The DCMPT philosophy has always been consistent with the principles of Findability, Accessibility, Interoperability, and Re-usability.

iii. The final goal of our project is to create or develop time series data both census and surveys from year 1972 to 2021. Data conversion is thus, a task critical from both business and technical perspectives.

Acknowledgments. The authors express their appreciation to all collaborators of the project

activities. They also acknowledge the members of DCMPT and Metadata formulation committee for their efforts to adhere to Historical data conversion issues.

References

1. Open Government Data Strategy Bangladesh Data for all, policy guidelines
2. John, W. C., Van Bogart, National Media Laboratory, USA

Estimation of age-specific reporting ratio of sentinel influenza surveillance using seroprevalence data

Masaya M. Saito[1]*, ***Hiroshi Nishiura***[2]

[1]**The Institute of Statistical Mathematics, Midori-cho 10-3, Tachikawa, Tokyo, 190-8562, Japan*
[2] *Graduate School of Medicine, Hokkaido University, Kita 15, Nishi 7, Kita-ku, Sapporo, 060-8638, Japan*
Email: saitohm@ism.ac.jp

Summary. Using serosuveillance data during 2009/10 swine flu pandemic in Japan, we have assessed the what proportion of infected people are reported in the sentinel surveillance, which is open to public via IDWR. After confirming these proportions do not change very much, we estimate incidence rate in seasons 2010/11 and 15/16. Our method, in principle, tries to count people who have been infected over the season in concern regardless their severity, while conventional methods capture patients who have developed enough symptoms so as to seek a medical institution. Therefore, the ratio of two may be interpreted as the crude ratio between symptomatic and symptomatic plus asymptomatic. In average, the ratio is about three and greater than unity even with the lowest estimates.

Keywords. Seasonal influenza, Disesase burden, Serosurveillance, Sentinel observation of influenza.

1. Introduction

In order to assess diseases burden, continuous monitoring of its cumulative incidence over year is necessary. However, as for influenza, the cases are tracked only in sentinel medical institutions (SMIs), and we need to somehow estimate the cumulative incidence from such a sample. A conventional method, called *multiplier method*, assumes SMIs would be sampled randomly from all the institutes in Japan [1]. However, as the proposers themself recognized, this assumption may not be justified since, in fact, SMIs are recruited on a voluntary basis. In order to overcome this limitation, utilization of whole the medical claims linked to influenza patients [2] or reweighting each SMI w.r.t. the number of its outpatients at a certain duration [3] have been considered.

We consider to employ seroprevalence gain as an additional information. Serosurveillance is held every year before a major epidemic wave comes (specifically July to September). Season 2009/10 was induced by pandemic virus 2009pdm and hence vaccination administrated in 2008 is expected to yield very limited gain in seroprevalence. Therefore, the difference of the seroprevalence between 2010 and 2009 can be interpreted as just the incidence rate during season 2009/10. Under this assumption, Mizumoto el al. [4] estimated age-specific incidence rates and concluded that the reported proportion is much higher in children while very low in elderly people. In this study, we try to examine whether the estimated proportion reported is applicable to other season to estimate incidence rate by applying the same procedure to successive seasons and examining whether proportion reported is near constant.

2. Method

We use three data sets for (i) age-specific sentinel influenza cases, (ii) serological prevalence, and and (iii) virological type-specific patients.

Assuming that sampling of serosurveillance participants and eventually (via any clinical and statistical processes) recorded SMI cases are governed by a single probability parameter each. The governing parameters are the infected incidence (denoted as $q(a,v,y)$ with age-group a, virological type v, and year y, or just q) and the

reported proportion as SMI cases (denoted as $p(a,v,y)$ or just p). As a summary, the likelihood function is a product of three binomials:

$$P = {}_M C_m \{p(q_2 - q_1)\}^m \{1 - p(q_2 - q_1)\}^{M-m} \times \prod_{y=1,2} {}_{N_i} C_{n_i} q_i^{n_i} (1 - q_i)^{N_i - n_i},$$

where M is Japanese population (in an age group), m is the number of SMI cases, N_1, N_2 are the number of sero-surveillance participants in the 1st and 2nd years, and n_1, n_2 are the corresponding seropositives. The goal is to estimate seroprevalence gain $q_2 - q_1$ and reported proportion p. An estimate and an approximated confidence interval of each probability parameter are provided by a conventional Gaussian approximation: The estimates are given simply by $\hat{q}_i = n/N_i$ $(i = 1,2)$ and $\hat{p}_i = m/\{(\hat{q}_2 - \hat{q}_1)M_i\}$.

3. Results

In 2009/10 season, children and young adults are intensively suffered from the pandemic flu: ages 10-14 exhibited the maximal incidence, 62% (95%CI: 57-66), which is reported in Mizumoto et al. [4]. In seasons 2011/12 and 15/16, the incidence of this subtype took roughly constant low values (~10%) over age groups, which is comparable to other competitive strains.

The reported proportion assessed as the ratio between expected cases from the seroprevalence elevation and the sentinel observation is summarized in Fig 1. The estimate based on 2009pdm and AH3 in seasons 10/11 and 15/16 are identical to that based on the 2009 pandemic within their uncertainties, though which are large. Therefore the reported proportion estimated from the 2009 pandemic data is applicable to other seasons to assess the incidence rate.

4. Conclusions

We have assessed what proportion of people who have been infected in a single flu epidemic season. Comparing such a assessment w.r.t the 2009 pandemic data against those w.r.t. 2010/11 and 15/16 seasons, we conclude that this ratio is almost time invariant. The incidence rate in several seasons estimated using this fact will be introduced in the presentation.

Acknowledgments. This work has been supported by JSPS KAKENHI Grant Number 18K11541.

Disclaimer. This article is a memo-random on a on-going research and any implication present here is not authorized by the peer review process.

References

1. Hashimoto, S., Murakami Y., Taniguchi K., et al., Annual incidence rate of infectious diseases estimated from sentinel surveillance data in Japan. *J. Epidemiol.,* 13(3), 136-41, 2003
2. Nakamura, Y., Sugawara, T., Kawanohara, H., et al., Evaluation of estimated number of influenza patients from national sentinel surveillance using the national database of electronic medical claims. *Jpn. J. Infect. Dis.,* 68(1), 27-29, 2015
3. Kawado, M., Hashimoto, S., Ohta, A., et al., Improvement of Influenza Incidence Estimation Using Auxiliary Information in Sentinel Surveillance in Japan. *The Open Infectious Diseases Journal,* 10, 29-36, 2018
4. Mizumoto, K., Yamamoto, Y., Nishiura, H., Computational and Mathematical Methods in Medicine. Article ID 637064, 1-8, 2013

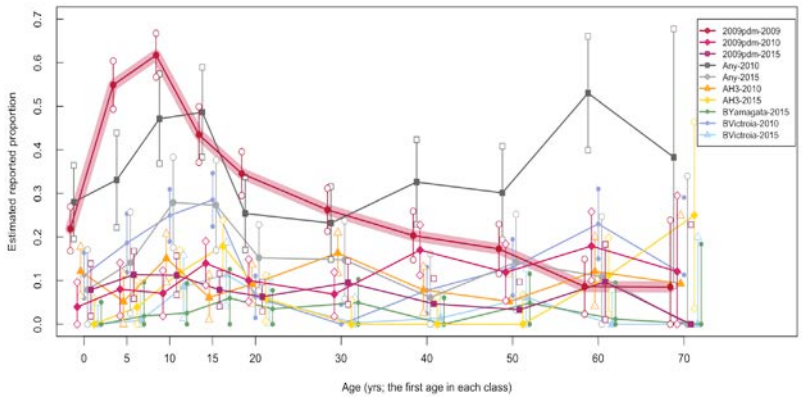


Figure 1. Estimated reported proportion based on various subtypes and different seasons.

From Human Genome Project to Genome Cohort Study

Atsushi Shimizu[1]*

[1]* *Division of Biomedical Information Analysis, Iwate Tohoku Medical Megabank Organization*
Iwate Medical University, 2-1-1, Nishitokuta, Yahaba-cho, Shiwa-gun, Iwate, 028-3694, Japan
Email: ashimizu@iwate-med.ac.jp

Summary. In 2003 when the Human Genome Project was completed, we finally obtained the blueprint of ourselves. At first, the function of genome sequence that accounted for 98% of the outside of the gene region was almost unknown. However, it was found that the human genome included several control regions and non-coding genes by the huge improvement of molecular biology research and bioinformatics analysis. Because of the advent of the next generation of the sequencer in the late 2000s, the personal genome sequencing progressed rapidly and we became able to understand the humankind as a species by studying individuals. Today, large-scale genome cohort studies are ongoing all over the world to investigate the onset of diseases caused by the interaction between genes and environments. In this lecture, we will present an overview ranging from the Human Genome Project to the genome cohort study.

Keywords. Human Genome Project, Genome Cohort Study, Bioinformatics.

Health - Prediction of Infectious Disease System Dynamics using Machine Learning and Mathematics

Shailza Singh[1]*

[1]* *National Centre for Cell Science, NCCS Complex, Ganeshkhind, SP Pune University Campus, Pune-411007, India*
Email: singhs@nccs.res.in

Summary. In India, there is a global organization called as Data Science for India (DSI), dedicated to educating and empowering high school students in India with a strong analytical foundation and critical tools for data science. Targeted bootcamps workshops are held in different thematic areas concerning data science viz., designing the DSI Data Literacy curriculum, creating the DSIx datascience bootcamp program and improving the DSI exploratory data science curriculum. India has seen more than 400% rise in demand for data scientists, as per recent 2018 data index reports. Use of big data architecture is helping to manage the expeditious data growth in health care industry. Here, I would like to discuss an empirical study performed on an infectious disease, leishmaniasis through machine learning and synthetic biology.

Keywords. leishmaniasis, machine learning, systems biology, synthetic biology, genome.

1. Introduction

Leishmaniasis is an endemic parasitic disease, predominantly found in the poor locality of Africa, Asia and Latin America. It is associated with malnutrition, weak immune system of people and their housing locality. At present, it is diagnosed by microscopic identification, molecular and biochemical characterisation or serum analysis for parasitic compounds. In this study, we suggest a new approach for diagnosing Leishmaniasis using cognitive computing and synthetic biology.

2. Machine Learning

Two different datasets of Leishmaniasis, Genetic data and Image data, were collected from databases and processed. The algorithm for training and developing a model, based on the data were prepared and coded using python. These algorithms and their corresponding datasets were integrated using Tensor flow data frame. An Artificial Neural Network trained model with multi-layer perceptron was developed as a diagnostic system for Leishmaniasis. It was developed using convolution neural network for image data and recurrent neural network for genetic data. The cognitive models of the trained network were interpreted using the maps and mathematical formula of the influencing parameters. The credit of the systems was measured using the accuracy, loss and error of the systems. These integrated systems of the leishmaniasis dataset and neural network acted to be the good choice for diagnosis with higher accuracy and lower error. Through this approach, all records of the data were effectively incorporated into the system. The experimental results of mean square error and regression analysis of genetic data, shows good prediction accuracy and it could be a better solution for diagnosing Leishmaniasis in future.

The model created through this approach for the collected and processed data of human Leishmaniasis genetic data has a good fit of accuracy for the diagnosing process. It showed 85.71% of accuracy, with just 1.94 as a value of loss function. This makes it to be a good model for the prediction of Leishmaniasis using its genetic dataset.

3. Systems dynamics through Synthetic biology

Quality health care can be maintained by descriptive, predictive and prescriptive big data analytical techniques and systems approaches being rightly called as predictive, preventive, participative and personalized medicine, I, next

talk about the work which revolves around modulating signaling in *Leishmania* infected macrophages. Till date, leishmaniasis is treated with traditional pharmaceutical approaches. Newer approaches to modify the host action against the parasite would be a better option to resolve the *Leishmanial* parasite. The strategies adopted helped devise an immuno-modulatory synthetic signaling circuit targeting CD14-TNF and EGFR pathways in Leishmaniasis. When macrophages are infected with *Leishmania,* multiple signaling inputs arrive simultaneously and/or sequentially, that are subsequently integrated for an anti-inflammatory pathophysiology in leishmaniasis. Novel biological devices may be constructed, that revert these signals for a pro-inflammatory leishmanicidal response to curb infection. CD14, TNF and EGFR pathways has been considered for dynamical system theory analysis, whose parameters were fitted within the known biological limits for a desired and predictable system behavior (BIOMD0000000477). Network metrices indicate that the signaling cascade is a real world network with crosstalk points MKK1/2, MKK3/6, MKK4/7 and IKK–NFkB, central to modulations for different phenotypic responses. Furthermore, the downstream gene network of CD14-TNF-EGFR pathway was reconstructed. Network analysis showed that NFkB links the signaling and gene network, used as a point of intervention through a synthetic circuit embedded within the negative autoregulatory feedback loop. A chimeric protein kinase C (PKC) was incorporated in the synthetic circuit, under the transcriptional regulation of Lac repressor and IPTG, as an inducer. The chimeric PKC via IKKb phosphorylation activates NFkB, and modulates the gene expression from an anti-inflammatory to a pro-inflammatory phenotype in *in vitro L. major* infected macrophage model.

4. Conclusion

How to tame big data with systems biology and systems pharmacology is the order of the day. Here, I have provided two examples as to how this can be achieved, but in addition the stages for many potential projects are set which can explore different aspects of this intriguing puzzle. In addition, the framework discussed here provides a template for the inclusion of the newer data that will be gathered in the coming years for different types of infectious disease.

References

1. Mandlik, V., Shinde, S., Chaudhary, A., Singh, S., Biological network modeling identifies IPCS in Leishmania as a therapeutic target. *Integrative biology : quantitative biosciences from nano to macro,* 4, 1130-1142, doi:10.1039/c2ib20037f, 2012
2 Mandlik, S. S. a. V., Structure based investigation on the binding interaction of transport proteins in leishmaniasis: insights from molecular simulation. *Molecular BioSystems*, doi:10.1039/C4MB00713A, 2015
3 Orgnization, W. H., Global Health Observatory (GHO) data, 2016. *Leishmaniasis : Situation and trends*, 2017
4 Mandlik, V., S. P., Bopanna, R., Basu, S., Singh, S., Biological Activity of Coumarin Derivatives as Anti-Leishmanial Agents. *PLOS ONE,* 11, 1-15, doi:10.1371/journal.pone.0164585, 2016
5 Milsee Mol, M. S. P., Singh, S., Immune signal transduction in leishmaniasis from natural to artificial system: Role of feedback loop insertion. *Biochimica et Biophysica Acta*, doi:10.1016/j.bbagen.2013.08.018, 2013
6 Milsee Mol, D. K. a. S. S., Nano-synthetic devices in leishmaniasis: a bioinformatics approach. *Frontiers in Immunology,* 6, 1-6, doi:10.3389/fimmu.2015.00323, 2015
7 Prevention, U. C. f. D. C. a., Diagnosis of Leishmaniasis. *CDC's Division of Parasitic Diseases and Malaria*, 2016
8 Alaa S. AlAgha, H. F., Bassam H. Hammo, Ala' M. Al-Zoubi, Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine. *Artificial Intelligence In Medicine*, doi:10.1016/j.artmed.2018.04.009, 2018
9 Shaikhina, T., N. A. K., Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine,* 75, 51-63, doi:10.1016/j.artmed.2016.12.003, 2017
10 Khosla, E., D. R., Sharma, R. P., Nyakotey, S., RNNs-RT: Flood based Prediction of Human

and Animal deaths in Bihar using Recurrent Neural Networks and Regression Techniques. *Procedia - Computer Science,* 132, 486–497, 2018

11 Quan Doa, T. C. S., Chaudri, J., Classification of Asthma Severity and Medication Using TensorFlow and Multilevel Databases. *Procedia - Computer Science,* 113, 344–351, 2017

12 Pizzi, N., L.-P. C., Mansfield, J., Jackson, M., Halliday, W. C., et al., Neural network classification of infrared spectra of control and Alzheimer's diseased tissue. *Artificial Intelligence In Medicine,* 7, 67-79, 1995

13. Sacks, D., Noben-Trauth, N., The immunology of susceptibility and resistance to Leishmania major in mice. *Nat. Rev. Immunol.,* 2, 845–858, 2002

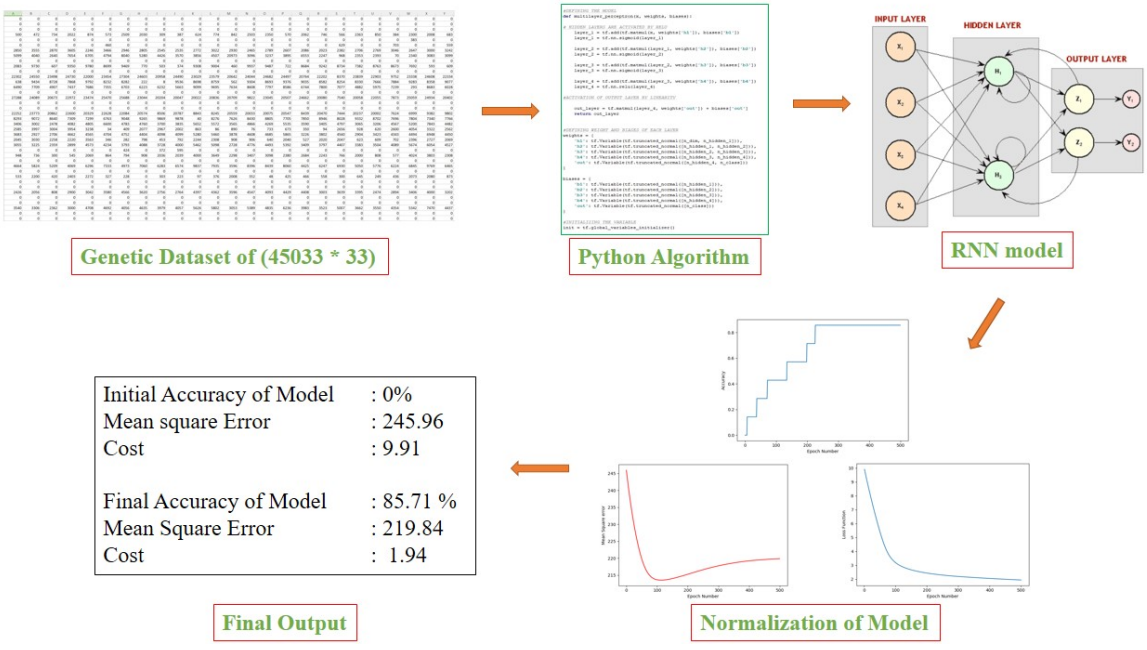
14. Liese, J., Schleicher, U., Bogdan, C., The innate immune response against Leishmania parasites. *Immunobiology,* 213, 377–387, 2008

15. Ropert, C., Closel, M., Chaves, A. C., Gazzinelli, R. T., Inhibition of a p38/stress-activated protein kinase-2-dependent phosphatase restores function of IL1 receptor-associate kinase-1 and reverses Toll-like receptor 2- and 4-dependent tolerance of macrophages. *J. Immunol.,* 171, 1456–1465, 2003

16. Little, J. W., The SOS regulation system. In Regulation of Gene Expression in Escherichia coli. in: (eds) Lin, E. C. C., Simon L. A., R.G. Landes Company, Austin, TX, USA, 453 – 479, 1996

17. Rosenfeld, N., Elowitz, M. B, Alon, U., Negative autoregulation speeds the response times of transcription networks. Journal of molecular biology, 323(5), 785-793, 2002

18. Becskei, A., Serrano, L., Engineering stability in gene networks by autoregulation. *Nature,* 405(6786), 590-593, 2000



Git-Hub Link:
https://github.com/shailzasingh/Machine-Learning-code-for-analyzing-genetic-dataset-in-Leishmaniasis

The International Seismological Centre (ISC): 50-Year Miracle

Dmitry A. Storchak[1*]

[1*] *International Seismological Centre, Pipers Lane, Thatcham, Berkshire, Rg19 4NS, U.K.*
Email: Dmitry@isc.ac.uk

Summary. For over 50 years, the International Seismological Centre (ISC) has been fulfilling its unique mission of producing the most definitive, complete and accurate long-term account of seismicity of the Earth. This work is based on a free and open exchange of data with several hundreds of seismological institutions, networks and observatories in over hundred countries around the world. The result – the Seismological Bulletin of the ISC and several associated datasets are freely available to all and routinely used by the Geoscience research community and referenced in several hundreds of scientific articles each year.

Keywords. Earthquake, bulletin, hypocentre, magnitude, mechanism.

1. Introduction

The International Seismological Centre (ISC) was founded in 1964 and follows the work of the International Seismological Summaries. It is based in the United Kingdom and charged with collection, integration and re-evaluation of seismological bulletins of earthquakes and other seismic events produced by many hundreds of geoscience institutions, seismic networks and individual observatories worldwide.

2. The ISC Mission and Operating Principles

The original data were always collected on a free data sharing principle, where each local data reporter was supplying a bulletin of earthquakes and other seismic events in their region in exchange for the global bulletin from the ISC that contains data of all reporters. In its work, the ISC acts as a truly international organization, not being responsible to any particular government and operating on a non-profit basis.

The operations of the ISC are only possible due to long-term support of many tens of Member-Institutions and commercial Sponsors. There never was a solid commitment from any Member to sustain the financial support of the ISC for any considerable length of time. Nevertheless, the quality and free availability of the earthquake related products that the ISC was publishing have successfully secured such support for over half of a century.

There is a clear realization that the long-term operations conducted by the ISC could not be sustained on commercial principles due to a very large overall cost of running seismic networks worldwide and routine processing of their data. It is only by a mutual consent of mostly academic and public institutions that the operations of the ISC could have been sustained for over 50 years.

3. Data Products of the ISC

In this presentation we describe the main data products of the ISC and their usefulness in various fields of Geoscience research:

- The ISC Bulletin (1964-2018)
- The International Seismograph Station Registry (1904-2018)
- The IASPEI Reference Event List (1959-2014)
- The ISC-GEM Global Instrumental Earthquake Catalogue (1904-2014)
- The ISC-EHB Bulletin (1964-2015)
- The ISC Event Bibliography (1904-2018)
- The International Seismological Contacts database.

4. Conclusions

Based on the widest possible long-term data exchange with seismic networks worldwide, the ISC was able to produce the seismological datasets that greatly contributed to the advancement of Geoscience. Free and open exchange of data has been a key to the ISC success.

Acknowledgments. We are grateful for the financial support of 65 current Member-Institutions in 48 countries as well as 16 public and commercial Sponsors. Among them we acknowledge the NSF Award 1417970. We also acknowledge additional project support received from CTBTO, USGS (Award G18AP00035), FM Global, Lighthill Risk Network, Willis Towers Watson, BGR, Reftek, GeoSIG and Guralp.

The ISC-GEM Global Earthquake Catalogue: Making Good Use of Historical Observations

Dmitry A. Storchak[1]* ***and Domenico Di Giacomo***[1]

[1]* *International Seismological Centre, Pipers Lane, Thatcham, Berkshire, Rg19 4NS, U.K.*
Email: Dmitry@isc.ac.uk

Summary. The ISC-GEM Global Instrumental Earthquake Catalogue is a key product of the International Seismological Centre (ISC). It is designed for use by the scientific community as well as by commercial companies for evaluation of seismic hazard and risk on a regional and global scale. The catalogue covers the entire period of instrumental recordings in seismology and contains the parameters of moderate to large earthquakes since 1904. It is important to note that major earthquake parameters in the catalogue (epicentre, depth and magnitude) were determined using the same procedure. Thus, usage of historical paper-based global earthquake summaries and individual station bulletins was highly important to enlarge the overall period for which the original seismogram measurements are available in digital form, ready for further interpretation.

Keywords. Global, earthquake, location, magnitude, homogeneity.

1. Introduction

The ISC-GEM Global Instrumental Earthquake Catalogue (last Version 5.1) is a key product of the International Seismological Centre (ISC). It was specifically designed and further developed as an extension of the general ISC Bulletin to satisfy the needs of the seismic hazard and risk research community.

The ISC-GEM catalogue was built using funding contributed by both the ISC Member-Institutions and commercial companies interested in the development of this product.

Much of the work on the early instrumental period (before 1964) has required collection, scanning and further processing of paper-based documents, some of which are available only in a few copies distributed around the world. This international collaboration was key to the overall success of the catalogue.

2. Working with historical documents

The majority of the original routine observations made by observatories around the world since 1964 have been captured by the standard Seismological Bulletin of the ISC which is available in digital form. Nevertheless gradual development of seismological observational standards during the 20th century has led to a gradual adoption of these standards by observatories worldwide. Hence, even for earthquakes since 1964, some of the key data, such as amplitudes and periods of surface waves had to be recovered from various, often non-digital sources.

For the years before 1964, the observational data were available mostly from paper-based sources that were only partly digitised at the start of our project – the historical global earthquake summaries (such as the ISS, ISA, Gutenberg Notepads etc) as well as a variety of individual seismological observatory bulletins that contained key information allowing to constrain earthquake size.

A gigantic effort was made not only to bring these historical data into the digital era, but also to process them and finally to make them available to the entire community for alternative uses in other projects.

3. Working with commercial companies

During the work on the ISC-GEM catalogue we received almost universal encouragement from both academic research and industries such as earthquake insurance, re-insurance and seismic engineering. Several commercial companies were willing to provide limited funding for our work, yet many chose to limit their support to mere encouragement.

One of the major stumbling points was the open availability of the ISC-GEM catalogue at all times, which was dictated by the majority of the ISC Members that are public institutions themselves. Many commercial companies were unwilling to contribute funds towards the catalogue development because of the product's open availability to those competitors who chose not to contribute towards the project.

Nevertheless, it was possible to convince some of the companies to contribute even though they had a very limited lead-time on using the product ahead of their competitors. Their support was often motivated by good will and the general good benefits that the ISC-GEM catalogue work brings to humankind.

4. Conclusions

The original development and further advancement of the ISC-GEM catalogue involved a lot of work with historical paper-based materials. Not only have these materials been put to good use for more accurate assessment of seismic hazard and risk worldwide, they have also been made available to a wider community for use in other fields of research. We observed an expected collision of the free data availability principle and the size of the commercial funding for our research. It was therefore very rewarding to see that some commercial entities chose to support our project despite no obvious commercial advantages over their competitors.

Acknowledgments. We are grateful for the financial support of 65 current Member-Institutions in 48 countries as well as 16 public and commercial Sponsors. In particular, the ISC-GEM work currently benefits from additional grants received from NSF, USGS, FM Global, Lighthill Risk Network, Willis Towers Watson and BGR, We specifically acknowledge the NSF Award 1417970 and USGS Award G18AP00035.

Research outline with regional theme: Examples using machine learning, geographic information system, and social big data

Shizuo Suzuki [1*]

[1*] *Department of Electronic Control System Engineering, National Institute of Technology, Numazu College, Ooka Numazu, Shizuoka 410-8501 Japan*
Email: shizuo.suzuki@numazu-ct.ac.jp

Summary. Three examples of research with regional theme were presented in this workshop. An image classification by aerial photographs was conducted to identify bamboo forests automatically. The two representatively coastal forests dominated by black pine in Shizuoka Prefecture were investigated to show changes in forest area during about 30 years. Main tourist hot spots around Mt. Fuji were analysed to demonstrate the potential of photo-sharing sites. These three results are considered to contribute forest monitoring and maintenance, and mapping of tourism potential.

Keywords. Aerial photograph, black pine, Flick, Random Forests, World cultural heritage.

1. Introduction

National Institute of Technology, Numazu College is promoting regionally collaborative activities among industry, academia and government with the aim of improving functions of education and research, and revitalizing the regional economy. Students as well as teachers of the college conduct joint research with staffs of companies, municipalities, research institutes and universities through graduation research and internship.

Three examples of research with regional theme were presented in this workshop. First, an example using machine learning. Bamboo forests were preliminary investigated by aerial photographs and estimated by machine learning in the eastern area of Shizuoka Prefecture. Next, an example using geographic information system. Changes in area of two coastal forests (one is Miho-Matsubara which is a component of Fujisan World Cultural Heritage site and the other is Senbon-Matsubara which is one of geo-site in Izu Peninsula UNESCO Global Geopark) were investigated by aerial photographs. Finally, an example using social big data. Mapping of tourism potential was conducted using social data in Flickr around Mt. Fuji.

2. A example using machine learning

This research shows an image classification by aerial photographs to identify bamboo forests automatically. The survey covered areas including the foot of Mt. Ashitaka in Numazu City. Random Forests which is a type of machine learning was used for the image classification. The classification accuracy was compared among RGB (red-green-blue) color space, HSV (hue saturation value) color space, and SLIC (simple linear iterative clustering) processing. Additionally, the classification was conducted at canopy level of bamboo forests. The model by the HSV color system combined with the texture classification showed the highest accuracy. These results suggest that the proposed method is applicable to the forest monitoring and the planning as a useful tool.

3. A example using geographic information system

Miho-no-Matsubara and Senbon-Matsubara are the representatively coastal forests dominated by black pine (*Pinus thunbergii*) in Shizuoka Prefecture. This research used the aerial photographs from 1974 to recent years to quantify the area change of coastal forest.

Percentages of both black pine forest area increased about 20% from 1976 to 1988 and decrease about 11% to 16% from 1988 to recent year. These results suggest that the increasing and decreasing in coastal forest area are due to enhanced growth of black pine itself as well as invasion of broad-leaved tree species after seawall construction and logging after pine wilt disease, respectively. These results are expected to contribute to the future maintenance of the two coastal forests.

4. A example using social big data

Social big data show great promise for geographical research, especially in the field of tourism geography. Visualizing the location information of photographs taken by tourists is a promising method to measure tourist activity. Photo-sharing sites are important sources for gathering location histories of tourists. Internet photo-sharing sites such as Flickr offer geotagged photographs, from which geographical information is retrievable by databases using the application programming interface (API) of such sites. The aims of this study were to identify and analyze the main tourist hot spots around Mt. Fuji and demonstrate the potential of photo-sharing sites. Geotagged photographs on Flickr were differentiated according to whether they had been taken by Japanese or foreigner. The study used the API from Flickr to obtain the latitude and longitude of the geo-tagged photographs, and then used geographic information system (GIS) to examine spatial patterns with areas of touristic value around Mt. Fuji. The spatial distribution patterns were analyzed using spatial statistical techniques in the GIS. The results revealed differences in main hot spots between Japanese and foreigner. Japanese posted photos taken in various places. In contrast, foreigners posted photos taken only in areas around the main tourist attractions on the north-eastern side of Mt. Fuji. This study demonstrated the value of photo-sharing site information as an approach to map tourism potential.

5. The other examples presented by poster

Preliminary investigations are presented by poster session on 1) aesthetic value of cultural ecosystem services by mapping geotagged photos from social media data around Mt. Fuji and Izu Peninsula, 2) classification of Japanese cedar forests with aerial photographs using machine learning approach, 3) tea leaf quality using remote sensing techniques in Numazu city, and 4) results in Japanese historical character recognition by machine learning to locally historical documents.

Acknowledgments. The three research examples were conducted in collaboration with Naoya Suwa, Koki Shimizu, Raiki Shida, Takahiro Katsumata, Kichi Kato, Ryodai Kato, Akari Sakai, Yoshitaka Suzuki, Taisuke Yasuda, and Yusuke Suzuki.

Activities of World Data Center for Geomagnetism, Kyoto

Satoshi Taguchi[1,2]*, ***Masahiko Takeda***[1], ***Hiroaki Toh***[1], ***Toshihiko Iyemori***[1]

[1]* *Data Analysis Center for Geomagnetism and Space Magnetism, Graduate School of Science, Kyoto University, Kyoto, 606-8502, Japan*
[2] *Department of Geophysics, Graduate School of Science, Kyoto University, Kyoto, 606-8502, Japan*
Email: taguchi@kugi.kyoto-u.ac.jp

Summary. The World Data Center (WDC) for Geomagnetism, Kyoto, provides geomagnetic field data supplied from a worldwide network of magnetic observatories, and geomagnetic indices derived in this data center to the community. In early 2011, we launched the near-real-time index data service. Immediately afterwards, the number of requests to the data server doubled, and has reached as many as 2,000,000 requests per month since then. The launch of the near-real-time index service is the result of a successful international collaborative project among six organizations in the USA, Russia, and Japan including this WDC in the early 2000s. This practice, which has moved the research community towards open science, teaches us that an effective approach for advancing WDC-based open science is to conduct collaborative projects based on needs of users together with data providers and external organizations that recognize the value of satisfying user needs.

Keywords. World Data Center, near-real-time data service, collaborative project, open science.

1. Introduction

The World Data Center (WDC) for Geomagnetism, Kyoto, is a regular member of the International Council for Science/World Data System. The center provides geomagnetic field data supplied from a worldwide network of magnetic observatories and geomagnetic indices to researchers specializing in solar terrestrial physics and geomagnetism, graduate students majoring in these academic disciplines, and citizens with an interest in those research fields.

The geomagnetic indices derived by the WDC are the Auroral Electrojet (AE) index [1], Disturbance Storm-Time (Dst) index [2], and ASY/SYM index [3]. These indices provide information on how certain regions of the near-Earth space are electromagnetically disturbed, which is often called Space Weather. The AE and Dst indices have long histories, and both were endorsed by the International Association of Geomagnetism and Aeronomy in 1969.

2. Users in the World

Figure 1 shows the percentage of the number of accesses to the WDC data server by country. The WDC data server is most often accessed by domestic (Japan) servers, followed by those in Russia, USA, and Germany.

Figure 2 shows the number of monthly requests to the data server. The recent average frequency is approximately 2,000,000 times per month. The frequency increased sharply in early 2011, which was exactly the time when the WDC started providing near-real-time information on the AE and Dst indices. i.e., the launch of the "Real-Time Space Weather Monitoring" system. The launch of this system undoubtedly satisfied the needs of users.

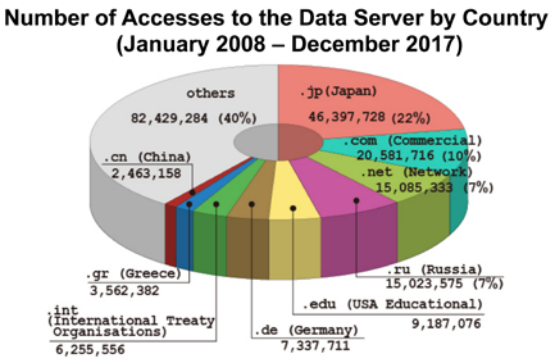


Figure 1. Number of accesses to the WDC data server by country from January 2008 – December 2017.

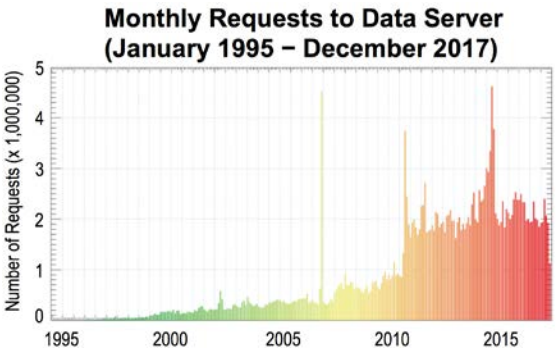


Figure 2. Monthly requests to the data server from January 1995–December 2017.

3. Near-Real-Time Index Service

3.1 International collaboration in early 2000s

The launch of the near-real-time index service is the result of a successful international collaborative project between six organizations in the USA, Russia, and Japan in the early 2000s for the purpose of upgrading magnetometer and installing real-time data transmission instruments in observatories in Siberia, Russia. Four out of the six organizations, i.e. the Johns Hopkins University Applied Physics Laboratory (USA); Geophysical Institute, University of Alaska (USA); National Institute of Information and Communications Technology (Japan); and Institute for Dynamics of Geosphere (Russia) are neither a data center for geomagnetism, nor a data provider (i.e. an organization operating an observatory) but are research institutes that recognize the value of satisfying user needs.

3.2 Open science accelerated by near-real-time service

The near-real-time index service needs further improvement. Users also want the WDC to launch the digital information service for the near-real-time AE index, which is not yet included in the current service. To provide this service, highly advanced methods need to be introduced for prompt data-quality checks.

This WDC has experienced successful collaborations with other external organizations, which eventually moved the research community toward open science. This practice teaches us that an effective approach for advancing WDC-based open science is to conduct collaborative projects based on needs of users together with data providers and external organizations that recognize the value of satisfying user needs, which is summarized in Figure 3.

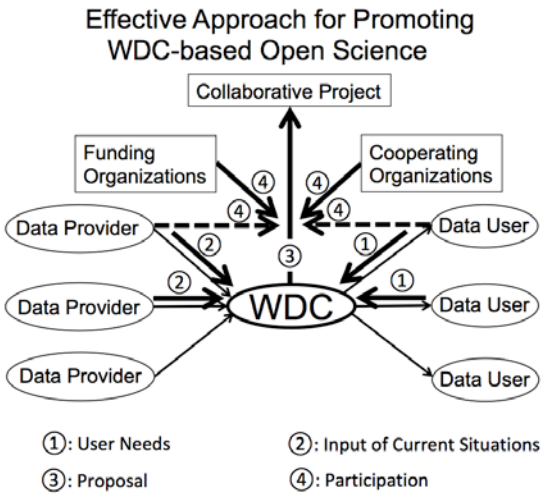


Figure 3. An effective approach for promoting WDC-based open science.

4. Conclusions

The activities of the WDC for Geomagnetism, Kyoto have moved the research community towards open science, and will do so for years to come. An effective approach for advancing WDC-based open science is to conduct collaborative projects based on the needs of users together with data providers and external organizations that recognize the value of satisfying user needs.

References

1. World Data Center for Geomagnetism, Kyoto, Nose, M., Iyemori, T., Sugiura, M., Kamei, T., Geomagnetic AE index, doi:10.17593/15031–54800, 2015
2. World Data Center for Geomagnetism, Kyoto, Nose, M., Iyemori, T., Sugiura, M., Kamei, T., Geomagnetic Dst index, doi:10.17593/14515–74000, 2015
3. Iyemori T., Rao, D. R. K., Decay of the Dst field of geomagnetic disturbance after substorm onset and its implication to storm-substorm relation. *Ann. Geophys.*, 14, 608–618, doi: 10.1007/s00585-996-0608-3, 1996

Reframing Scholarly Communication by Persistent Identifier

Hideaki Takeda[1]*

[1]* *National Institute of Informatics, 2-1-2, Hitotshubashi, Chiyoda-ku, Tokyo 101-8430, Japan*
Email: takeda@nii.ac.jp
ORCID: 0000-0002-2909-7163

Summary. Open Science is a new form of science with the infosphere created by the Internet. Before the Internet era, scholarly communication is relatively well organized such as journal publications by scientists and scholars in universities and research institutes. But there become more complicated and difficult with such open infosphere. Persistent Identifiers give filtering and anchoring scientific information among the vast infosphere. The system of persistent identifier has started by DOI, and now other PIDs are following DOI. By assigning PIDs to every piece of scholar communication, scientific activities will be accessible, findable, and preservable.

Keywords. Open Science, Persistent Identifier, DOI, ORCID.

1. Introduction

We, scientists and scholars, believe that science is open. It simply means that there are no barriers between science and people. But it is not true, or has not been true until now. In fact science is not open to public, even science is not open to each other. Science in different disciplines are not open to each other.

There are invisible barriers between science and people, i.e., accessibility, understandability and professionality. Fortunately or unfortunately digital technology, in particular, the Internet technology is breaking these barriers step by step. At first, the barrier by accessibility is broken. Before the Internet era, information on science is published by books and journals and communicated in public and private scholar meetings. Except private meetings, they are open in principle. But in fact, these books and journals are only available at academic libraries and central libraries so that ordinal people are not easy to reach them.

The Internet technology, in particular, Web technology created the infosphere where everyone in everywhere can create, share and use information. So can scientific information.

Then the barrier by understandability is going to be broken now. Information on science is usually very difficult for people because languages are quite different from ordinal languages. Artificial Intelligence is now supporting people by their inference facility with knowledge like ontologies.

The barriers by professionality means that reliability is ensured only in the professional scientists and scholars. Non professorial people are limited to participate into scientific activity.
It is also going to be broken. The structure of professionality has been supported by social and economic system such as authority and budget. The Internet flattened the social structure and created new economic system such as crowdsourcing.

So now science is REALY going open. Here I focus on the first aspect, i.e., accessibility, because it is now changing our business.

2. Digital Object Identifier (DOI)

STM publishing, in particular, scholar journals are the important role in scholar communication. Journals are media not only for communicating but also delivering and archiving information. When the publishers were changing their journals from paper to digital, they noticed that the Internet itself if not so suitable for delivering and archiving information, because the Internet

mainly focused on the current information. But scientific information has longer life.

Digital Object Identifier (DOI) was created to save this situation. DOI mechanism is very simple, i.e., resolving DOI names to actual URLs where information is available. While URLs tend to change rapidly, DOIs are fixed even original URLs are changed. We can expect longer life of access information just like catalogues in libraries.

DOI is not just a technical system rather it's value is that it is socially supported. The DOI Foundation and the contracted registration agencies (RA) keep the system work.

3. The Role of Persistent Identifiers (PIDs) in Scholarly Communication

The success of DOI tells us new insights about scholarly communication in the Internet Era. The infosphere is the promising space for scholarly communication for its great flexibility. New types of communication can easily be created on the Internet. Sharing data and process is a good example. But the flexibility also causes the problems in scholarly communication. Information is easily lost. Judging reliability of information is not easy. DOI and its metadata can successfully anchor the information on the Internet with the traditional manner in science.

We have usually trusted papers and books that can be accessible via libraries. In the Internet Era, infosphere is almost infinite, but we just use papers with DOI just like papers in libraries.

So this anchoring function of DOI is important for science to live with the infosphere by the Internet. It is the role of Persistent Identifier (PID) in science.

4. The Information Flow in Science with PID

Now various PIDs are following DOI. ORCID and ISNI are identifiers to people. In particular, ORCID is getting common for scientists and scholars because many publishers use ORCID to identifier authors of papers. Open Funder Registry [1] is started to identify funders supporting the research work. Identifiers for affiliations will start soon[2].

We will have more IDs in science soon. Figure 1 shows how information flow in science will be with PIDs. By assigning PIDs to every piece of scholar communication, scientific activities will be accessible, findable, and preservable.

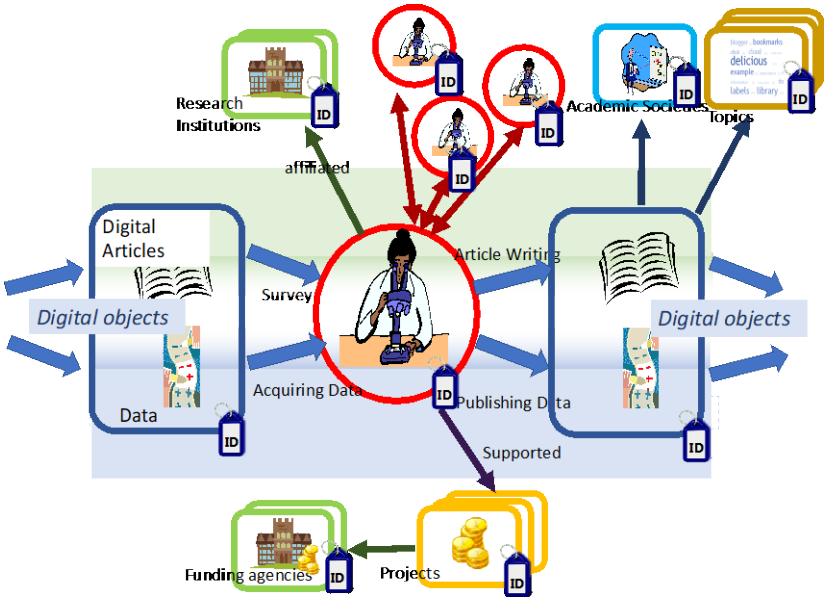


Figure1: Information Flow in Science with PIDs

[1] https://www.crossref.org/services/funder-registry/

[2] https://orcid.org/content/organization-identifier-working-group

Useful Tools for Education and Capacity Building about Solar Terrestrial Physics Study

Yoshimasa Tanaka[1*,2,3], ***Norio Umemura***[4], ***Atsuki Shinbori***[4], ***Shuji Abe***[5], ***Satoru UeNo***[6], ***Masahito Nose***[4], ***and IUGONET project team***

[1*] *Joint Support-Center for Data Science Research, Research Organization of Information and Systems, 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8518, Japan*
[2] *National Institute of Polar Research, ROIS, 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8518, Japan*
[3] *The Graduate University for Advanced Studies (SOKENDAI), Shonan Village, Hayama, Kanagawa 240-0193 Japan*
[4] *Institute for Space-Earth Environmental Research, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601 Japan*
[5] *International Center for Space Weather Science and Education, Kyushu University, Motooka, Nishi-Ku, Fukuoka 819-0395, Japan*
[6] *Kwasan and Hida Observatories, Graduate School of Science, Kyoto University, Kurabashira, Kamitakara-cho, Takayama, Gifu 506-1314, Japan*
Email: ytanaka@nipr.ac.jp

Summary. We introduce two tools for solar terrestrial physics study, developed by the Inter-university Upper atmosphere Global Observation NETwork (IUGONET) project. One is a data analysis software based on Space Physics Environment Data Analysis Software (SPEDAS); we have provided a plugin software for SPEDAS, called iUgonet Data Analysis Software (UDAS), which enables users to deal with various upper atmospheric data obtained by the IUGONET members using SPEDAS. The other is a metadata database for upper atmospheric data, called IUGONET Type-A, which is one-stop web service that allows users to search data, get the information of data, view quick-look plots of data, find scientifically interesting events, interactively plot data, and proceed to more advanced data analysis using SPEDAS. These tools may also be useful for the education about solar terrestrial physics study in universities or high schools and capacity building in emerging countries.

Keywords. IUGONET, SPEDAS, analysis software, metadata database, upper atmosphere.

1. Introduction

Inter-university Upper atmosphere Global Observation Network (IUGONET; http://www.iugonet.org) is a Japanese inter-university project, which aims at sharing upper atmospheric data, including solar and planetary data, obtained with various ground-based instruments by Japanese universities and institutes since International Geophysical Year (IGY; 1957-1958). The project started in FY2009 originally by Tohoku University, Nagoya University, Kyoto University, Kyushu University, and National Institute of Polar Research, and then some other universities and institutes joined the project. We have primarily developed two tools; one is a metadata database, called IUGONET Type-A, which allows users to cross-search various kinds of the upper atmospheric data distributed across the IUGONET members, and the other is a data analysis software that is capable of analyzing such a variety of data in an integrated fashion. In this paper, we present the characteristics of these tools.

2. IUGONET Data Analysis Software

The iUgonet Data Analysis Software (UDAS) developed by the IUGONET project is a plug-in software for Space Physics Environment Data Analysis Software (SPEDAS; http://spedas.org/wiki/) [2]. The SPEDAS is a grass-roots data analysis software for space physics community, which is written in the Interactive Data Language (IDL;

https://www.harrisgeospatial.com/SoftwareTechSoftwa/IDL.aspx). One of the important features of SPEDAS is an automatic downloading of data files via the internet. The SPEDAS was originally developed as TDAS by scientists and programmers of the UC Berkeley and UCLA and other contributors to analyze satellite- and ground-based observational data obtained by the THEMIS mission [3], and then it was extended to support various kinds of data from multiple missions. By using the UDAS plugin, users can easily load, visualize, and analyze the data released by the IUGONET members with useful functions in the SPEDAS/IDL.

Recently, we developed template routines that allow users to easily create routines for loading their own data onto SPEDAS, by modifying only about 10 lines of the templates. At present, the templates called UDAS EGG (Easy Guide to Generate your load routines) support data files in the Common Data Format (CDF) and the ascii format, and will support other formats in near future, e.g., netCDF and FITS. This tool will help to expand the base of potential users of SPEDAS.

3. IUGONET Metadata Database

Since the upper atmospheric data obtained by the IUGONET members have been archived and opened to public separately by each university or institute, it is often difficult and time-consuming for users to find, get, and analyze the data. In order to solve this problem, we have developed a metadata database for upper atmospheric data, which can search various kinds of data distributed across the IUGONET members [4]. The second version of the metadata database was named IUGONET Type-A (http://search.iugonet.org/). It provides one-stop web service, which allows users to search data, get the information of data (i.e., metadata such as contact persons, access URL, and data use policy), view quick-look plots of data, find scientifically interesting events, interactively plot data, and proceed to more advanced data analysis using SPEDAS. The IUGONET metadata format was designed based on the Space Physics Archive Search and Extract (SPASE) metadata model with some modifications for the upper atmospheric data. The IUGONET Type-A was produced mainly for researchers in the field of the solar terrestrial physics, but it may be useful for the capacity building in the emerging countries and the education in universities or high schools. We regularly have data analysis workshops several times a year in Japan and sometimes in other countries, especially in Asia and African region, to diffuse the IUGONET data and tools.

4. Conclusions

The IUGONET project has developed the metadata database and the integrated data analysis software for the upper atmospheric data. It helps to share the data with researchers in the solar terrestrial physics community, effectively analyze various kinds of data, and promote interdisciplinary study. These tools may also be useful for the capacity building in the emerging countries and the education in the universities or high schools.

Acknowledgments. The IUGONET project was supported by the Special Educational Research Budget (Research Promotion) [FY2009] and the Special Budget (Project) [FY2010-2014] from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

1. Tanaka, Y.-M., Shinbori, A., Hori, T., Koyama, Y., Abe, S., Umemura, N., Sato, Y., Yagi, M., Ueno, S., Yatagai, A., Ogawa, Y., Miyoshi, Y., Analysis software for upper atmospheric data developed by the IUGONET project and its application to polar science. *Adv. Polar Sci.,* 24, 231-240, doi: 10.3724/SP.J.1085.2013.00231, 2013
2. Angelopoulos, V., The THEMIS mission. *Space Sci. Rev.,* 141(1-4), 5-34, doi: 10.1007/s11214-008-9336-1, 2008
3. Abe, S., Umemura, N., Koyama, Y., Tanaka, Y.-M., Yagi, M., Yatagai, A., Shinbori, A., Ueno, S., Sato, Y., Kaneda, N., Progress of the IUGONET system - metadata database for upper atmosphere ground-based observation data. *Earth, Planets and Space,* 66, doi:10.1186/1880-5981-66-133, 2014

Research Data Management in Industry-Academia Collaboration

Yuko Toda[1]*, ***Hodaka Nakanishi***[2]

[1]* *Teikyo University, 2-11-1 Kaga, Itabashi-ku, Tokyo 173-8605, Japan*
Email: y.toda@med.teikyo-u.ac.jp
[2] *Teikyo University, 2-11-1 Kaga, Itabashi-ku, Tokyo 173-8605, Japan*
Email: nakanishi@med.teikyo-u.ac.jp

Summary. The data management is an important issue, but it is not well discussed when contracting collaborative research between industry and academia. Universities are recently requested to preserve research data because of the research integrity while companies tend to hold their right to the data. The description concerning research data management in the contracts on industry-academia collaboration of Teikyo University was analysed. In the contracts of collaborative research, only a few contracts prescribe data management issue. It is necessary to negotiate not only for research integrity, but also for the importance of open data and the social role of the universities considering the interests of companies. Technology transfer offices of university should consider research data management in addition to intellectual property.

Keywords. collaborative research, contract, industry-academia, attribution, research data.

1. Introduction

The handling of research results, including research data as well as intellectual property, is an important issue in collaborative research between industry and academia. Although the issues on intellectual property is often discussed, the management of research data, such as creating, accessing, using, storing, and disposing data, is not well discussed at the negotiations of contract between company and university.

In this paper, the cases of research data management in joint research contracts between Teikyo University and companies are summarized. Several points to be considered for the research data management in the contracts are proposed.

2. Background

For the treatment of intellectual property right, the Ministry of Education, Culture, Sports, Science and Technology formulates Sakura Tool which indicates eleven types of contracts from the viewpoint of the attribution of intellectual property rights in collaborative research. As the Sakura tool does not explicitly include the attribution of data, the treatment of the research data in a contract depends on the definition of "outcome" [1].

Regarding the handling of data among companies, the Ministry of Economy, Trade and Industry has formulated "contract guidelines on data usage authority" in 2017 [2]. This guideline is organized from the viewpoint of usage rights of research data. However, companies often ask for ownership of research data in the collaborative research especially in a clinical research.

In the context of research integrity, it is required to keep experiment notes, documents, numerical data, image data etc. for 10 years after the publication of the related papers [3]. To meet the requirement, research data should be stored in the university properly while companies often request the ownership of the data. It is necessary to coordinate which party should keep the research data in the collaborative research between industry and academia.

3. Research Data in Contracts

In this study, 45 contracts of collaborative research between Teikyo University and

companies (excluding cancelled cases) which were checked by Teikyo Technology Transfer Center from April 2018 to August 2018 were analysed. 29% of the contracts mentioned “data” and 71% did not (Table 1). Here, the phrase "mention data" means that the term "data" is described in the context of data management or explaining research content, and it does not mean the mere use of the term such as for explaining misconduct activities.

Table 1. Description of “data” in contracts.

Description of data	number (%)
“data” is described in the contract	13 (29%)
“data” is not described in the contract	32 (71%)
Research results don’t include data	13 <41%>
Research results may include data	19 <59%>
Total	45 (100%)

71% of total contracts did not mention data management. About 60% of the contracts without data description determine that all technological outcomes are considered to be research results and research data can therefore be included in the research results. However, about 40% of the contracts limit research results to materials, and research data is therefore out of the scope.

29% of total contracts mentioned data and determined data management such as attribution of data or permission of using the data.

As data has a significant role in a clinical research, pharmaceutical companies often request universities to make the data attribution to the company in a contract. In the past, Teikyo university allowed companies to hold the ownership of the data, but now it is necessary for universities to have the right to access the research data for 10 years from the viewpoint of research integrity.

4. Conclusions

In the contracts of the collaborative research between industry and academia, the management of research data is getting important.

For the coordination of industry and academia relationship, not only the importance of open data and the social role of the universities, but also research integrity should be considered and the interests of the companies as well. Technology transfer offices of university should now consider research data management in addition to intellectual property.

References

1. Anderson Mori & Tomotsune, *Survey on the way of handling outcomes in collaborative research in the open and closed strategy era*, 2018 (in Japanese)
2. Ministry of Economy, Trade and Industry. *Contract Guidelines on Data Utilization Right*, 2017 (in Japanese)
3. Science Council of Japan, *Answer: Improving Integrity in Scientific Research*, 2015 (in Japanese)

Application of deep learning and large scale simulation to the Earth science

Seiji Tsuboi[1]*, ***Daisuke Sugiyama***[1]

[1]* *JAMSTEC, Yokohama, Kanagawa, 236-0001, Japan*
Email: tsuboi@jamstec.go.jp

Summary. Recent progress in both data science and large scale simulation is now giving an impact on Earth science research. Particularly, application of artificial intelligence, which is realized by using convolutional neural network, to the Earth science has shown possibilities to make significant progress in various fields. Here we show a combination of deep learning approach and large scale simulation to develop new procedure to locate earthquake. We have applied our procedure to regional earthquakes in Kanto region, Japan. Although the number of earthquakes is limited, our results suggest that our approach may be applicable to locate earthquakes in various scales.

Keywords. Deep learning, Seismology, Synthetic seismograms, large scale simulation, Spectral-Element method.

1. Introduction

Traditional method of locating earthquake has been based on the arrival times of seismic waves, such as P-wave and S-wave, and one-dimensional Earth model. Here we develop different approach combining numerically computed theoretical seismograms and deep machine learning. We calculate theoretical seismograms for realistic three-dimensional Earth model by using the Spectral-Element method program SPECFEM3D (https://github.com/geodynamics/specfem3d_globe) and use these seismograms to create seismic wave propagation images at the surface of the Earth. Then we use these images as training dataset of convolutional neural network. We build neural networks for determination of hypocentral parameters by using the package TensorFlow (https://github.com/tensorflow), such as epicenter, depth, origin time and magnitude, and applied these networks to actual seismograms to examine if this procedure works to locate earthquake and determine magnitude. Although the number of earthquakes is small and the regional extent is quite limited, the results demonstrate that it is feasible to locate earthquakes by using this approach. Advantages of using this approach to locate earthquakes and determine magnitude are; accuracy of hypocenter parameters can be increased by accumulating theoretical seismograms for various earthquake location and size as learning dataset of deep machine learning; there is no need to read arrival times and amplitude of seismic waves to locate earthquakes; three dimensional Earth structure can be included without additional computational cost to locate earthquakes; seismologically rare but inevitable case, such as earthquakes which happen concurrently in different location, can be included in learning dataset. This technique may be applied to determine moment tensor solution of the earthquakes in realtime monitoring of seismic activity.

Acknowledgments. We used the seismic waveform provided by the NIED K-net and earthquake catalog provided by Japan Meteorological Agency. Computation of theoretical seismograms was conducted on the Earth Simulator and supercomputers in JAMSTEC, Japan.

Resource and Environment Scientific Data Sharing and Disaster Risk Reduction Knowledge Service

Juanle Wang[1]*

[1]* *Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A, Datun Road, Chaoyang District, Beijing, 100101, China*
Email: wangjl@igsnrr.ac.cn

Summary. The disaster risk reduction is an urgent task facing the world. Many developing countries are weak on the disaster risk reduction information processing and application capacity. Knowledge service plays important role on disaster risk reduction field in the Big Data era. A Disaster Risk Reduction Knowledge Service System (http://drr.ikcest.org) was built based on disaster information sharing and thematic knowledge service, which realizes the sharing of knowledge resources such as disaster data, map, expert, organization, document and courseware. In future, the Disaster Risk Reduction Knowledge Service should focus on the standard for disaster metadata, the integration of disaster metadata resources, developing disaster information search engine based on big data and location based disaster risk early warning knowledge service. The current global data integration and sharing trends and activities of WDS data centres in China were also introduced in this presentation.

Keywords. Resource and Environment, Data Sharing, Disaster risk reduction, disaster data, knowledge Service.

1. Introduction

Disaster data is a necessity for establishing Disaster Risk Reduction (DRR) capabilities. The integration and sharing of DRR data and the provision of knowledge services is an emerging trend in the field of DRR[1]. Knowledge services represent a fundamental transformation of knowledge resources from 'possession' to 'application'. Innovations in knowledge service theories, methods and techniques are necessary to drive the continuation and development of knowledge services[2], which are expected to deeply influence the growth of human society in the near future.

The United Nations Educational, Scientific and Cultural Organization (UNESCO) has always prioritised global cooperation in DRR by promoting DRR knowledge sharing services[3]. With the implementation of the Sendai Framework for Disaster Risk Reduction[4], UNESCO hopes to strengthen cooperation in disaster data standards, DRR education, and both national and regional disaster databases in a greater number of developing countries. In 2015, UNESCO proposed a set of requirements for the construction of a DRR knowledge service by the International Knowledge Centre for Engineering Sciences and Technology (IKCEST), which is managed by the Chinese Academy of Engineering (CAE). Based on current trends towards data integration and international collaboration, the DRR knowledge service was established to address the demand of DRR works around the world, especially on knowledge service applications for earthquake, drought and flood disasters. This knowledge service will provide a long-term source of DRR information and thematic knowledge services for international organisations, government institutions, research and educational institutions, businesses and broader society.

2. Resource and environment data sharing trends

Resource and Environmental Science is a comprehensive discipline that applies rational use of resources and environmental protection to the fields of production and environmental construction from an ecological point of view. The trends of global resource and environment data sharing are summarized as below.

(1) Scientific Data Curation achieve a high degree of consensus.

(2) Scientific data and related research enter the era of big data.

(3) Scientific data publishing model promotes scientific data sharing.

(4) Data archiving policies are developed and used in funding projects.

(5) Authoritative data centers gather global resources.

(6) Data center management models are designed and applied in varies data centers.

(7) Many research data centers are established in the world.

(8) Scientific data sharing standards and mechanisms are needed.

(9) The authentication of trustworthy repositories grows recently.

3. Technical Methodologies

The overall method of building the Disaster Risk Reduction Knowledge Service System is to take the standard formulation for disaster-related metadata as the breakthrough point to realise the collection of various knowledge resources, including a metadata-based disaster science database, disaster map resources, a disaster expert database, a disaster institution database, a disaster event database, a database of disaster open directory projects, a database of disaster-related information Web mining, a disaster literature database, a disaster popular science database, and a disaster video courseware database. In the development environment of an open platform, the system delivers visualized subject-specific knowledge services in typical disasters, including earthquakes, droughts, flooding, and freezing rain and snow, and it enables interactive application by users through the online system.

4. The Realisation of Thematic Knowledge Applications

The use of thematic knowledge services is an important approach for the organisation of services in DRR knowledge service platform. Based on the needs of its target users and featured resources, a knowledge service system constructs specialised service products by means of online and offline models, to provide valuable DRR knowledge services to users around the globe. DRR provides knowledge services regarding the platform, technology, data, education, and other aspects of current global disaster prevention and mitigation. The modes of service include the query, browsing, downloading, analysis, and visualization of various types of disaster-related knowledge and resources. The service contents cover eight aspects in three main categories: resource content (including data services, map services, institution services, expert database services, and disaster event services), resource dissemination (including video courseware training services and science popularization services), and resource and knowledge applications (e.g., Global Earthquake Daily Distribution Map Service, China and International Experience in Natural Disaster Relief, Map Visualization Services of China Historical Disasters, etc.).

5. Conclusions

Under the DRR directives of UNESCO, we have constructed a DRR knowledge service system for the sharing of disaster information and the provision of thematic knowledge services, based on the formulation of disaster metadata standards. This system has successfully actualized the integration and sharing of knowledge resources like disaster data, disaster maps, expert

opinions, institutional data, research literature, and educational materials.

Disaster Risk Reduction Knowledge Service system (DRR) can be accessed online (http://drr.ikcest.org). Till the end of June, 2018, the DRR platform has abstracted 12 thousand page views per month, about 24% from abroad and 76% from domestic. Targets for UN Sustainable Development Goals and Sendai Framework for Disaster Risk Reduction, DRR will long-term persistently provide disaster risk reduction knowledge services to international organizations, government agencies, research institutions, educational institutions, commercial institutions and the public.

Acknowledgments. Many Thanks for the guidance by the experts of the Natural Science Sector of the UNESCO Office for Disaster Risk Reduction and the Secretariat of the International Knowledge Centre for Engineering Sciences and Technology (IKCEST) of the Chinese Academy of Engineering. Thanks all the team members from Institute of Geographic Sciences and Natural Resources Research, and Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences.

References

1. Xie, Y. B., Design Knowledge Services via Internet-An Analysis on the Functions of China Knowledge Center for Engineering Sciences and Technology (CKCEST). *China Mechanical Engineering*, 28(6), 631-641, 2017 (in Chinese)
2. Wen, Y. K., Jiao, Y. Y., Constructing a scientific framework for knowledge service in semantic Web environment. *Journal of Information Resources Management*, 1, 99-104, 2011 (in Chinese)
3. UNESCO Disaster Risk Reduction, http://www.unesco.org/new/en/natural-sciences/special-themes/disaster-risk-reduction/
4. United Nations. Sendai Framework for Disaster Risk Reduction 2015-2030. United Nations Office for Disaster Risk Reduction (UNISDR), https://www.preventionweb.net/files/43291_sendaiframeworkfordrren.pdf, 2015

Multivariate analysis of the occupations of rental rooms by using the housing information website data

Hayafumi Watanabe[1]*, ***Yu Ichifuji***[2], ***Masahito Suzuki***[3], ***Satoshi Yamashita***[4]

[1]**Research Organization of Information and Systems, 10-3, Midori-ku Tachikawa-shi, Tokyo, 453-424, Japan*
[2] *Nagasaki University, 1-14, Bunkyo-machi, Nagasaki City, Nagasaki, 852-8521*
[3] *UD Asset Valuation Co., Ltd., 1-3-6, Andoujimachi,, Chuo-ku, Osaka 542-0061, Japan*
[4] *The institute of statistical mathematics, 10-3, Midori-ku Tachikawa-shi, Tokyo, 453-424, Japan*
Email: hayafumi.watanabe@gmail.com

Summary. The apartment loan is a loan for rentals such as condos and apartments. This loan is a very large loan which is the account for a percentage of more than 10 percent of the whole Japanese banks' loan. However, a risk model of the apartment loan with the appropriate accuracy has not been provided in Japan mainly due to the lack of data. Thus, in order to develop the risk model, we preliminarily analyse and compare two types of data set: the survey data which is made by the real estate appraiser and the housing information website data. As a result, it was found that (i)the web data is approximately corresponding to the survey data which was made by experts with respect to statistical properties, (ii)The AR (accuracy ratio) value of the simple multivariate regression model developed in this study which explains the transition of rooms from vacancy to occupation takes about 0.4 and (iii)this transition probability of the model is mainly explained by an age of a building.

Keywords. Web data, Apartment loan, Multivariate analysis, Real estate, Comparison between databases.

1. Introduction

The apartment loan is a loan for rentals such as condos and apartments. This loan is a very large loan which is the account for a percentage of more than 10 percent of the whole banks' loan in Japan. However, a risk model of the apartment loan with the appropriate accuracy has not been provided mainly due to the lack of database of the occupation of Japanese rental rooms. There are difficulty in making nationwide the database of occupation by investigators in a traditional manner to cost an excessive amount of money or time. Thus, in this study, we examine the possibility of construction of database of the occupation of rental rooms for the apartment loan by using the housing information website data.

2. Related work

There are few studies in which discuss the occupation of Japanese rental rooms based on the individual room data. The rare example of study of individual room occupation in Japan has been given by Kobayashi [1]. In the paper, he made the model which describe the periods of that rooms are empty by using the data of the occupation of rooms provided by three real estate agencies (673 rooms). Moreover, he simulated earnings based on results of data analysis. The main difference between Kobayashi's study and our study is that we use the web data, which are open to the public, large scale and easily scalable, but are expected to have sampling biases.

3. Data set

We use the two types of data sets for our study: (i) the survey data which is made by the real estate appraiser (Survey data) and (ii) the housing information website data (Web data) once every three months.

3.1 Survey data

Survey data is made by real estate appraisers. They investigate rooms in a certain area obtained by stratified sampling method from the room lists provided by a certain bank once every three months (4,333 rooms). In this investigation, they visit the rooms and check whether rooms are empty or not and the conditions of rooms such as cleanliness and fees within the possible.

3.2 Web data

We obtain the Web data from housing information website data by the web crawling on 8th, 18th and 28th of each month. We use the data from 11/2014 to 7/2016 and focused on corresponding area to the survey data. This data contains various information of rooms such as a distance from a station, a dimension of a room, super markets, ages of a room, a number of stories of building etc (35,807 rooms).

4. Results

In this section, we present the summary of results obtained by analysis of above-mentioned two types of data.

4.1 Sampling bias of web data

By comparison of between survey data and web data, we found that the web data has approximately corresponding statistical properties to the survey data which was made by experts and sampled randomly (stratified sampling). Accurately, however, this correspondence is valid for the statistics for buildings, but is not valid for the statistics for rooms. Figure 1 shows the example of the agreement of probability distribution between survey data and web data. Here, the black solid line indicates the cumulative probability distribution of the duration of empty (days) for Web data, the red dashed line for survey data and the green dotted line indicates the exponential distribution.

4.2 Multivariate analysis

The AR (accuracy ratio) value of the simple multivariate regression model (i.e., the logistic regression model) developed in our study which explains the transition of rooms from vacancy to occupation takes about 0.4. In addition, this transition probability of the model is mainly explained by an age of a building.

5. Conclusions

In this paper, in order to check the usability of web data for the risk model of the rental home financing, we analysed the two types of data:(i) Survey data and (ii) Web data. By comparison between two data, we found that the web data is approximately corresponding to the survey data which was made by experts with respect to statistical properties. In addition, we showed that web data has ability of prediction of occupation of rooms by using the simple multivariate regression model.

References

1. Kobayashi, S., A Study about How to Estimate Future-Cash Flow in Real Estate Evaluation with Micro-Analysis of Move-in and Move-out. *Financial Planning Kenkyu,* 16, 18–27, 2016 (in Japanese)

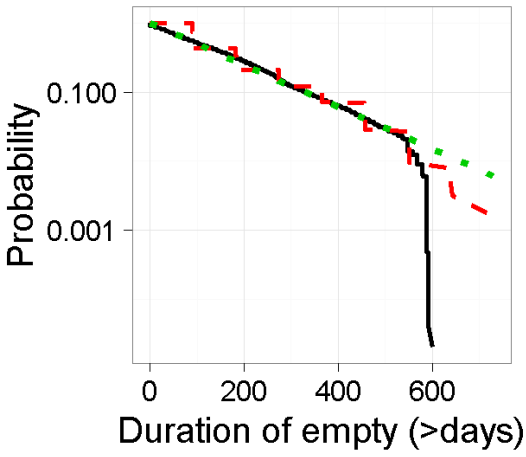


Figure 1. The cumulative probability distribution of the duration of empty (days) between survey data and web data.

Multidisciplinary Study of the Earth's Environment in 18th-19th Centuries - A Trial to find an Approach to the Open Data and Open Science

Takashi Watanabe[1*]

[1*] *World Data System International Programme Office, c/o NICT, Koganei 184-8795, Japan*
Email: takashi.watanabe@icsu-wds.org

Summary. In data-oriented activities addressing the Open Data and Open Science, to assure extensive multidisciplinary usage of data is crucial. However, there still exists many problems mainly caused by differences in "culture" of individual disciplines, particularly on Open Data issue. To specify current problems, provisional search and usage of data are conducted by accessing various kinds of data opened on the Internet to study the environmental and economic movements in the interval of 18th-19th Centuries. It is found that we need to pay effort to have ready-to-use datasets for scientists out of the field. Many old data are still remained to be digitized in machine-readable formats. Special effort will be needed also to manage data only available in publications as the forms of tables or diagrams. The ethical issue on reuse of these data is essential. Involving citizen scientists opening their data will be important also. By this trial, the study of this interval is found to be a good starting point of a potential multidisciplinary research project of the comprehensive study of the future human society under the influence of the global environmental change.

Keywords. Open Science, Open Data, Environmental Sciences, History, Economics, Climate Change.

1. Introduction

Although multidisciplinary data usage is very important in the new paradigm of science, Open Data and Open Science, this is not an easy task. For example, the majority of conventional databases have been constructed for usage mainly by related domain scientists. It is still difficult to find appropriate data for out-of-the-field researches through current search engines. Even when one succeeded to find a dataset to be used for his/her study, it will still be difficult to use them without appropriate knowledge on the data because the documentation of the data are not necessary to be prepared thinking about potential users outside the discipline.

To find current obstacles in multidisciplinary usage of data, a trial of search and analysis of data is performed on environmental and economic data covering the interval of 18th-19th Centuries. This interval was located in the recovery phase of the Little Ice Age, starting in the late Medieval era, and several important movements, e.g., the French Revolution and the Industrial Revolutions in England, were taken place under the influence of relatively cold environment and high volcanic activities. A study of this interval will be important to know the world just before the anthropological environmental changes became prominent. In this research activity, we will be able to expect to collaborate with citizen scientists because we are finding many datasets relevant to our research as the outcomes of their personal researches. From the point of view of the Open Science, collaboration with them will be highly desired.

2. Provisional Data Search and Data Analysis

A trial to search for data was performed via Internet by using currently available search engines, just to be done by a user who do not have sufficient knowledge on the data. For example, environmental data acquired by modern instrumental measurements can be obtained without important difficulty, a lot of problems exist on reconstructed data basing on old proxy records, e.g. temperature data in the interval before the era of instrumental records.

Many data of this kind are published only in the form of diagrams or tables in scientific articles as outcomes of the research works. We find similar situation for old data on, for example, market prices of wheat in old days. We will have technical and ethical issues in reuse of these data. A part of historical data, e.g. data of prices of materials at local markets and reconstructed weather data basing on old diaries, have been opened also by dedicated citizen scientists, mainly via their Web pages. In the point of view of Open Science, involving them into research activities will be important but a mechanism of quality assessment of these data will be needed.

For the present provisional study, several examples of basic data showing economic and environmental situations in 18th-19th Centuries, mainly in Europe, are shown in Fig. 1a-d. As a measure of the economic activity in Europe, The time series of the wheat price at London [1] is given in Fig. 1a, showing an interesting rise of price in the interval from about 1800 to 1820. As seen in the sunspot number data [2] given in Fig.1b, this interval was situated in the interval of relatively low solar activity (Dalton Minimum) appeared in 1790-1830. The British yearly averaged temperature [3] shown in Fig.1c shows that the climate in 18th-19th Centuries was relatively cold, presumably due to the low solar activity and a series of several intense volcanic eruptions [4], including Laki in 1783 and Tambola in 1815 (Fig. 1d). There were many reports on world-wide depressions of agricultural and commercial activities taken place in association with these big eruptions.

Although various studies on this interesting interval are still in progress by researchers of the history and the economics [5], influences of changes in natural environment are not treated as important factors because big political and economic movements were taken place in this interval particularly in Europe, e.g., the French Revolution and the Industrial Revolutions in England, etc. Detailed multidisciplinary studies of environmental and social relationship from the global point-of-view will open a new approach to get better understanding of movements of the human society.

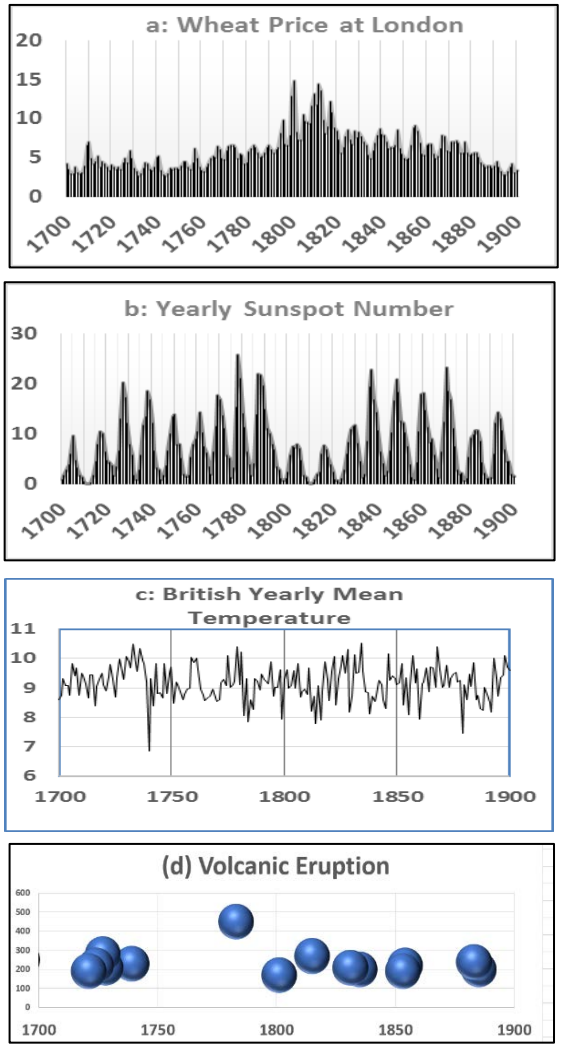
3. Concluding Remarks

It is found that we still have many technical and ethical problems to assure multidisciplinary usage of data. To resolve these problems, collaborations among scientists in various research domains will be important. To stimulate the collaboration, it will be important to have a common interest (or a driving force) by selecting an "attractive" research project to perform multidisciplinary studies. The study of environmental and social connections in 18th-19th Centuries will be a good starting point of a potential research program on the future of the Earth under the global climate change. This program will be attractive also for citizen scientists.

References

1. Clark, J., The price history of English agriculture, 1209 – 1914. http://faculty.econ.ucdavis.edu/faculty/gclark/papers/Agprice.pdf, 2003
2. World Data Centre for the production, preservation and dissemination of the international sunspot number. http://www.sidc.be/silso/datafiles, 2018
3. Met Office Hadley Centre Central England Temperature Data, https://www.metoffice.gov.uk/hadobs/hadcet/data/download.html, 2018
4. Ice Core Volcanic Eruptions Data Tables, http://chandra.harvard.edu/edu/formal/icecore/Volcanic_Eruptions_Data_Tables.pdf, 2018
5. Freeman, C., Rouca, F., As time goes by from the Industrial Revolutions to the Information Revolution. Oxford, 2001

Figure 1. Time series of economic and environmental data in 18th-19th Centuries: (a) wheat price (shilling/bushel) at London [1], (b) yearly sunspot number [2], yearly mean temperature (deg. C) in the central England [3], and the level of volcanic eruption (conductivity of the ice core, $s/cm^{-1} \times 10^{-2}$) in the world [4].



Data Management at Polar Research Institute of China

Lizong Wu[1]*, ***Beichen Zhang***[1], ***Jie Zhang***[1]

[1]* *Polar Research Institute of China, 451 Jinqiao Road, Shanghai, 200136, China*
Email: wulizong@pric.org.cn

Summary. Chinese National Arctic & Antarctic Data Center (CN-NADC) established in 2003 by the State Oceanic Administration (SOA) of China to manage and disseminate scientific data resulting from research within the China Antarctic and Arctic Scientific Expedition. CN-NADC helps fulfill China's obligations under Article (III).(1).(c) of the Antarctic Treaty which states that "Scientific observations and results from Antarctica shall be exchanged and made freely available. The CN-NADC responsibilities include management of Antarctic and Arctic scientific data for the long-term making it available in an easily accessible form, providing metadata records for all China Antarctic and Arctic scientific research data available for public searching in an effective form, and enabling China Antarctic and Arctic Expedition data is available for SCAR, IASC, SOOS, WDS and other international data systems. Since 2003, CN-NADC is a sub-platform of "National Data Sharing Infrastructure of Earth Science" supported by the Ministry of Science and Technology (MOST) and the Ministry of Finance of People's Republic of China. In 2018, CN-NADC upgrade to one of National Data Center which is a national data center system include about 20 data centers for all disciplinse and regions in China. The CN-NADC carry out various practice in repository management functions and related systems and facilitates the custodian's right to be cited as the source of published data. Data publication activities are cooperation with "Global Change Research Data Publishing and Repository". Scientists within the China Antarctic and Arctic Expedition are responsible for complying with all aspects of the China Antarctic and Arctic Program Data Policy, including submission of raw and processed data, derived products and associated metadata in an acceptable form to the Data Centre within the timelines agreed for data submission within their Data Management Plan.

Keywords. Data sharing, Data quality manage, History data rescue, Data publication.

Arctic Data Archive System (ADS)

Hironori Yabuki[1*,2], ***Takeshi Sugimura***[2], ***Takeshi Terui***[2]

[1*] *Polar Environment Data Science Center, DS, ROIS, 10-3 Modori-cho, Tachikawa, Tokyo, 190-8518, Japan*
[2] *International Arctic Environment Research Center, NIPR, ROIS, 10-3 Modori-cho, Tachikawa, Tokyo, 190-8518, Japan*
Email: Yabuki.hironori@nipr.ac.jp

Summary. Arctic Data archive System(ADS), through proceed with the visualization and the development of online analysis system of integrated big data, aiming for integrated analysis information platform, not only as a mutual distribution platform of data, we have developed a system that enables open access research data and scientific knowledge obtained in the Arctic research. Various applications and services developed by ADS should not be used only in the Arctic but should be used as a bipolar data publish platform. The ADS team is currently preparing to publish not only Arctic data but also Antarctic data.

Keywords. Arctic, Global warming, ArCS, Data Management.

1. Introduction

The easy access use is made possible from the industrial and the social public using research results(thesis and research data, etc.) using a public research fund, and a concept as open science aiming at linking it to creation of innovation by opening the new way as well as promoting a scientific technical research effectively is showing a rapid expanse to creation of worldwide. And the principle opening to the research result and data by a public research fund by GRC (Global Research Council), OECD (Organization for Economic Cooperation and Development) and G8 in 2013 etc.

Under these background, even in Arctic research, open access of a variety of variation mechanism and scientific knowledge, such as future prediction result brought about by actual grasp their environment change has been demanded.

In order to clarify the environmental variation system of complex Arctic with a variation of the time-space scale that is different consisting of air-land- marine, and human sphere, through interdisciplinary research, through interdisciplinary research, a wide variety of observational data, simulation data, satellite data, and even there is a need for the creation of A New knowledge of using the big data that integrates the research results. Also those integrated with the big data, scientific knowledge by using these, it is necessary to continue to properly publish to society.

2. Development of Arctic Data archive System(ADS)

Arctic Data archive System(ADS: https//ads.nipr.ac.jp), through proceed with the visualization and the development of online analysis system of integrated big data, aiming for integrated analysis information platform, not only as a mutual distribution platform of data, we have developed a system that enables open access research data and scientific knowledge obtained in the Arctic research.

ADS has been doing the systems development of following up to now.

- Metadata Management System by own metadata schema.(KIWA: Fig.1)
  - Metadata exchange system by using OAI-PMH, GI-cat.
  - Currently, this service is carried out in cooperation with GCW in WMO. Also this service have done coordination with GEO-Portal.

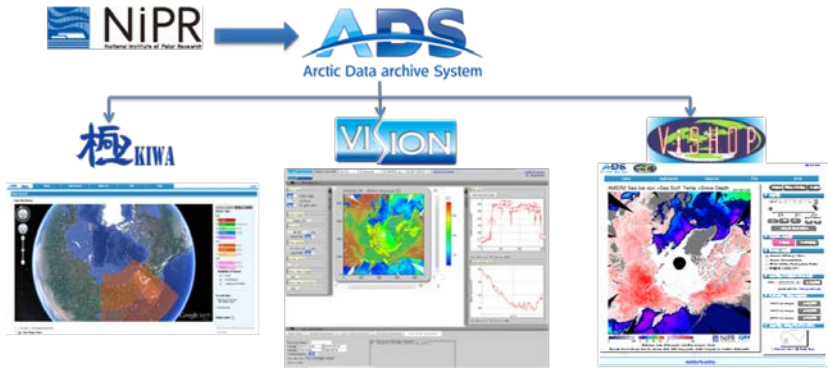


Fig.1 : Structure of ADS, Research data registration system and Metadata search service(KIWA), Online visualization application for Climate, Satellite and Simulation data(VISION) and Semi-real-time polar environ. obs. Monitor and Sea Ice prediction(VISHOP)

  - A system for space-time search using GoogleEarth collected data
  - DOI (Digital Object Identifier) registration system
- Visualization and analyzed system for the satellite data and grid data by online(VISION: Fig1)
- System to Semi-real-time polar environ. obs. monitor and sea ice prediction in the Arctic, Antarctic by using the satellite data (AMSR2) that is delivered in near-real-time from JAXA.(VISHOP:Fig1)
- System for visualizing numerical data such as time-series data(VISION-Graph)

3. Future development and challenges

ADS is not only a system that provides the data to various data users and stakeholders, in order to promote joint research and international cooperation in the Arctic region, anyone that is aimed at developing integrated analysis platform through the available Web interface. Furthermore in ADS, to developing of the information providing service of push-type in accordance with the needs of stakeholders.

By widely publish the technology developed in ADS, to promote the technology transfer system construction, to help the same technical problem solved in other areas. In ADS future, to carry out research and development the following items.

- Advancement of data and meta-data registration and retrieval system
- Promotion of data registration and data usage
- Enhanced of international cooperation of data and metadata
- Advancement of visualization, basic data analysis and the like of software and Web applications in order to provide an integrated analysis platform
- System construction of the push-type information service
- Advancement of small and medium-sized data server linkage function by ADS grid
- System Technology publishing, which is research and development in the ADS and technology transfer promotion to other systems.

4. Conclusions

The share of research data and scientific knowledge in the Arctic and non-Arctic nations, there are need for coordination of data repository and data center in a various country.

Important to drive the open-science, it is important data published and data cited, it is necessary to promote these data published and data cited. We, through the development of ADS activity, believe that can contribute to the sharing of research data and scientific knowledge in the Arctic and non-Arctic nations.

DIAS Platform Contributing to Open Science in Earth Environmental Informatics

Masaki Yasukawa[1]*

[1]* *Earth Observation Data Integration and Fusion Research Initiative, the University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan*
Email: yasukawa@iis.u-tokyo.ac.jp

Summary. Data Integration and Analysis System (DIAS) is one of data centric cloud. DIAS collects and integrates massive and diverse data such as earth observation data and socio-economic data, and various applications on DIAS platform provide beneficial information to solve societal issues such as environmental problems in Japan and overseas. This article gives the concept of DIAS at first. As a case study on DIAS's contribution to open science, I introduce a data sharing system for flood forecasting in Sri Lanka, and data collecting systems for citizen science of butterfly and oriental stork.

Keywords. DIAS, open science, data centric cloud, data sharing, citizen science.

1. Introduction

Data Integration and Analysis System (DIAS) started from 2006. The goals of DIAS are to collect and store earth observation data; to analyze such data in combination with socio-economic data, and convert data into information useful for crisis management with respect to global-scale environmental disasters, and other threats; and to make this information available within Japan and overseas. The prototype of DIAS was developed in 2010. Then, phase II of DIAS from 2011 made further advancement and expansion to apply DIAS as a social and public infrastructure. The current project of phase III has started since 2016 with the aim of its practical operation.

2. DIAS

The three systems that comprise DIAS are shown.

- Infrastructure system: integrates Earth and local observation data and numerical models, socio-economic data, and other massive datasets relating to the global environment.
- Application development: implements storage, search, analysis, visualization, and other data-related functions on the infrastructure system to provide scientific knowledge and resolve global environmental and societal issues.
- Research and development (R&D) community: DIAS provides a close R&D environment that allows domain scientists who study the environment and IT experts to promote joint research and development through cooperative work, planning, and production.

DIAS project is unique, because a community to support application development on data infrastructure is established in addition to the construction of data infrastructure.

3. DIAS Platform

DIAS provides the platform to process the global environment data. When the domain researchers install their programs in DIAS, and they use real-time data archived on DIAS, the application of real-time processing is easily realized. The applications related to an open science on the DIAS platform are introduced as follows.

3.1 Data Sharing System for Flood Forecasting in Sri Lanka

In Sri Lanka, a large flood occurred in late May 2017, and many residents were sacrificed. For the secondary disaster precaution and the basin reconstruction, a web-based data sharing system

for flood forecasting has been developed in June 2017, and the data providing has been started to the local stakeholders.

This system has data processing tools and visualization tools. The data processing tools include data collection, data correction, and real-time prediction. The visualization tools can display various real-time data such as rain gauge data, satellite data, rain map, rain forecast and flood prediction.

The local stakeholders use this system for flood monitoring. Also, this system is useful for teaching material for technology transfer. For more effective flood prediction in future, they will consider the suitable flood prediction model, and construct the new system themselves.

3.2 Data Collecting System for Citizen Science of Butterfly

The butterfly is familiar with citizen, and suitable as the index of global warming and urbanization.

A data collecting system for citizen science of butterfly in Tokyo has been constructed in 2009. The participant uses the data upload tool, and uploads the monitoring data with photograph. Data manager of butterfly expert uses the data quality control tool, and do data cleansing. After the cleansing, the data is open to the public by data visualization tool.

In 9 years, about 50,000 records were uploaded. A lot of findings such as the distribution expansion of southern butterfly, the distribution expansion of alien species, and existence of the species on the red-list were brought by the power of the citizens.

3.3 Data Collecting System for Citizen Science of Oriental Stork

The last one of wild oriental stork in Japan was died in 1971, due to farmland consolidation, disappearance of wetlands by river refurbishment, and the use of pesticides. By the activity of feral population reproduction with breeding and releasing, about 100 oriental storks inhabit satoyama landscape in Japan now.

For contributing to the monitoring activity of oriental storks, a data collecting system for citizen science of the stork has been developed by enhancing the collecting system of butterfly in April 2018. In this activity, the feature is to identify the individual number of the stork. About 1,800 records with photograph were collected from April to August 2018. The habitat map and the safety confirmation list were made from the dataset. Using the location, the behaviour and the food, the clarification of the history of life and personality will be expected in future.

4. Conclusions

DIAS is not only a data repository, but also an application platform. The field of the application on DIAS platform is wide range. Also, some of applications are contributing to open science for earth environment. Especially, data collecting systems for citizen science are the practical examples of open science. By collecting more data of butterfly and oriental stork and developing the visualization tools, a lot of new findings on ecology will be expected in future.

Acknowledgments. This study received the support of the Ministry of Education, Culture, Sports, Science and Technology study trust business "Global Environment Information Platform Development & Promotion Program (DIAS-PF)".

References

1. Kawasaki, A., Yamamoto, A., Koudelova, P., Acierto, R. A., Nemoto, T., Kitsuregawa, M., Koike, T., Data Integration and Analysis System (DIAS) Contributing to Climate Change Analysis and Disaster Risk Reduction. *Data Science Journal*, 16, 41, 1-17, 2017
2. Yasukawa, M., Ikoma, E., Nemoto, T., Rasmy, M., Tsuda, M., Ushiyama, T., Tamakawa, K., Koike, T., Kitsuregawa, M., Prototyping a Data Sharing System for Flood Forecasting: A Case Study on Sri Lanka, *3rd International Symposium on Big Data Analytics in Science and Engineering (BASE 2017)*, 2017
3. Ikimoni, http://butterfly.diasjp.net/ [accessed on: September 2018] (in Japanese)
4. Citizen Science of Oriental Stork, https://stork.diasjp.net/ [accessed on: September 2018] (in Japanese)

An attempt for the thermal transport modelling of fusion plasmas based on the statistical approach

Masayuki Yokoyama [1*, 2]

[1*] *National Institute for Fusion Science (NIFS), National Institutes of Natural Sciences (NINS)*
[2] *SOKENDAI (The Graduate University for Advanced Studies)*
both at 322-6 Oroshi, Toki, Gifu 509-5292
Email: yokoyama@nifs.ac.jp

Summary. A statistical approach has been attempted for the thermal transport modelling for plasmas in fusion experiment [1]. This approach has become possible with the analysis database accumulated by the extensive application of the integrated transport analysis suite, TASK3D-a [2], to the LHD (Large Helical Device) experiment at National Institute for Fusion Science (NIFS) [3]. After the first attempt that was published in Ref. [1], the validity check of the obtained regression expression for the ion heat diffusivity profiles has been made to be compared with several LHD experimental results. The first attempt [1] and the successive progress will be reported in the workshop, as one of possible and innovative "data-driven" approaches in fusion research.

Keywords. fusion experiment, Large Helical Device (LHD), TASK3D-a, thermal transport modelling, statistical approach.

1. Introduction

Conventionally, scaling laws for the global energy confinement time (τ_E) have been one of the approaches to systematically grasp the energy confinement property of fusion-experiment plasmas, and are also considered as one of the guidelines to design/predict future device.

On the other hand, physics-based transport models have been employed to predict the plasma performance, such as expected temperature profiles for certain operation scenarios. However, it should be recognized that a satisfactory prediction in a short time scale has not been realized.

In this poster presentation, an attempt to overcome such difficulties will be reported, based on the statistical approach utilizing analysis database created by the extensive application of TASK3D-a to LHD plasmas.

2. Regression analysis based on TASK3D-a database

Recent development of TASK3D-a, and its extensive application to a wide-ranging LHD plasmas have created the transport analysis database which includes profile information such as ion and electron temperatures (T_i and T_e), electron density (n_e), heating deposition, and ion and electron heat diffusivities (χ_i and χ_e), and others.

The accumulation of TASK3D-a analyses results has led to the attempt at deducing regression analysis for the ion heat diffusivity, χ_i, with certain parameters based on a statistical approach. Here such data are collected from so-called high-*T*i scenario [4] in LHD experiment.

On performing statistical analysis, χ_i is dimensionally normalized by Bohm diffusion coefficients, $T_i/(eB)$. Candidate predictive variables are also made into dimensionless variables, such as, normalized ion collisionality (ν_i^*), the normalized ion Larmor radius (ρ_i^*), and the temperature ratio (T_e/T_i).

Here, as a standard exercise in scaling studies, the assumed simple power-law scaling model has been transformed to the log-linear form. Multiple

ordinary least squares (OLS) regression analysis has resulted in the regression expression for $\chi_i/[T_i/(eB)]$. An important statistical measure of the quality of the model is the ratio R^2 of the variation explained by the model to the total variation. The obtained value is R^2=0.83, which is a relatively high value, indicating that the above equation reasonably reproduces the response variable. The root-mean-square-error (RMSE) value is 0.28.

The validity check of the obtained regression expression for the ion heat diffusivity profiles has been made through its implementation into the predictive TASK3D calculations to be compared with several LHD experimental results. It has indicated the promising characteristics for reproducing the ion temperature profiles with a reasonable agreement with experimental observation.

3. Conclusions

A statistical approach is proposed as the thermal transport "modelling" in fusion plasmas. Statistically-confident regression expression for the ion heat diffusivity for LHD high-T_i plasmas has been obtained, which can be directly implemented into the predictive simulation as a transport "model." The validating calculations have shown promising characteristics, which will be reported in the workshop.

Acknowledgments. The author appreciates valuable advices on statistics from Prof. K. Shimizu and Prof. Y. Iba (The Institute of Statistical Mathematics, ISM). Dr. H. Yamaguchi (National Institute for Fusion Science) is also acknowledged for his contributions on the validity check of this approach. This work has been supported by the NIFS Collaborative Research Programs, NIFS14KNTT025, NIFS14UNTT006 and NIFS18KNTT046, and by the ISM Collaborative Research Program, H30-1002.

References

1. Yokoyama, M., *Plasma and Fusion Research,* 9, 1302137, 2014
2. Yokoyama, M., et al., *Nuclear Fusion*, 57, 126016, 2017
3. Takeiri, Y., IEEE *Transaction on Plasma Sciences*, 46, 1141, 2018.
4. Nagaoka, K., et al., *Nuclear Fusion*, 55, 113020, 2015

International Workshop on Data Science

- Present & Future of Open Data & Open Science –

AUTHORS INDEX

Mishima Citizens Cultural Hall &
Joint Support-Center for Data Science Research,
Mishima, Shizuoka, Japan

12–15 November 2018

Authors Index :

(Alphabetical Order; First Author Only; Indicate ("Poster" Presentation-Number))