PROGRAMME and ABSTRACTS

International Symposium

Global Collaboration on Data beyond Disciplines

23 – 25 September 2020

Online Conference

International Symposium
- DSWS-2020 -

“Global Collaboration on Data beyond Disciplines”

PROGRAMME and ABSTRACTS

Online Conference

23–25 September 2020

Circular

International Symposium
“Global Collaboration on Data beyond Disciplines”

23 – 25 September 2020

Online Conference

Background and Scope:

One of the important objectives of the current Open Science movement is the creation of new research fields and technologies based on the vast quantities of data now being produced from almost all scientific domains. Adherence to the FAIR (Findable, Accessible, Interoperable, and Reusable) Principles for data have been recognized as a norm for data-oriented activities in this Open Science era, and is leading to efforts to develop Open Data infrastructure such as enhanced and integrated metadata catalogues, persistent identifiers (PIDs; e.g., Digital Object Identifiers) and metadata standards for research data management, certification of data repositories addressing long-term management and stewardship of quality-assessed data, and so on. Moreover, in many academic facilities (e.g., universities), the use of persistent identifiers for people, places, and other entities is becoming best practice in preservation and provision of data produced by their research activities.

Further efforts are still needed to resolve the various challenges that currently exist for scientific research data; in particular, the sharing and reuse of such data. Although the importance of multidisciplinary data integration has been widely advocated, data reuse by scientists either within or across disciplines is still not easy from the point of view of the FAIR Principles: for example, there may be difficulties in discovering and accessing the data or insufficient information on the data to enable easy analysis. Often additional information is also needed to assist understanding when sharing research data with the general public, including policymakers. To improve the situation, it is important to stimulate collaborations among scientists from various disciplines and to establish systems that facilitate the interactions between the users and providers of research data.

On the basis of the above, the Joint Support-Center for Data Science Research (DS) of the Research Organization of Information and Systems (ROIS) and the Committee of International Collaborations on Data Science of the Science Council of Japan (SCJ) are organizing a symposium to share information on current international research data activities addressing Open Science, data-centric science, and interdisciplinary data-driven science. A key step in making science open lies in improving data quality and transparency so that researchers can easily share their work, and this necessitates a robust and global-scale infrastructure. The symposium will be a remarkable opportunity to discuss the future development of such data-oriented infrastructure in the Open Science era, and thus ensure the long-term preservation and equitable provision of quality-assessed research data. The World Data System (WDS) of the International Science Council (ISC) and its partner organization Open Researcher and Contributor ID (ORCID) are leading global initiatives that can galvanize the community to find solutions to these data-related issues, and at the same time work with the community to improve data discoverability and interoperability for effective data reuse.

The goal of this symposium is to build consensus on various aspects of research data management by all stakeholders in alignment with Open Research policies and initiatives. It will explore new paths for activities significant in promoting interdisciplinary and collaborative research and data reuse under different scientific disciplines based on evidence and feedback from data communities. Such global data activities are expected to be strengthened by the endeavours and facilities provided by DS, ROIS.

Session Themes:

Potential session topics proposed for the symposium are:

- **Opening Session**: Opening Addresses, Keynote Talks, 10 Years Event of the International Programme Office of the World Data System (WDS-IPO).

  Conveners: *Masaki Kanao, Yasuhiro Murayama, Takashi Watanabe, Rorie Edmunds

- **WDS Members' Forum 2020**: This WDS-led session is an interim, online version of the biennial WDS Members' Forum (also known as 'Data Repositories Day'), which has been postponed until 2021 such that it aligns with International Data Week. The session will be split into two parts:

  - The first part will be a Scientific Session consisting of report backs from the WDS membership on topics of interest and importance to the data repository community. All WDS Members will have the opportunity to showcase the visions and foci of their organizations for the coming five years though a series of lightning talks—either in pre-recorded videos or live—to facilitate discussions on opportunities created, challenges faced, and areas in which WDS can assist.
  - The second part will be a Plenary Session that provides a formal mechanism for the WDS Scientific Committee to consult with the community on issues of relevance to data repositories. It will share information on the current and prospective WDS activities, and ask for feedback and buy-in from the WDS membership and beyond.

  This virtual event is open to both WDS Members and general participants having an interest in the endeavours of WDS—those from outside of the WDS family will be very much welcomed. Since this meeting incorporates the biennial WDS business meeting, Representatives of WDS Member Organizations—in particular, those of WDS Regular and Network Members—are expected to participate.

  Conveners: *Rorie Edmunds, WDS-SC, WDS-IPO, WDS-ITO

- **Regional Activities on Data in the Asia & Oceania Area**: Sharing information on data-oriented collaborations on data in the Asia-Oceania area, e.g. plans of collaborating platforms have been going on in Australia and Malaysia and discussions in the series of WDS Asia and Oceania Conferences, etc. Principal Organizers are WDS and Regional Office of Asia and Pacific (ROAP).

  Conveners: *Toshihiko Iyemori, Mazlan Othman, Juanle Wang, Takashi Watanabe

- **WDS-ORCID Strategic Workshop: Adoption of PIDs in Asia–Oceania:** This jointly hosted session by ORCID (Open Researcher and Contributor ID) and the World Data System of the International Science Council (WDS) will introduce to the Asia–Oceania community Persistent Identifiers (PIDs) from both global and national perspectives, and their importance as a part of global open research infrastructure. It will also (1) raise awareness about PIDs and their use across different scientific disciplines; (2) explore practical use cases of PIDs by national data systems and data repositories, especially within the Asia–Oceania context; and (3) showcase PID standards/best practices, as well as identify the resulting opportunities and benefits.

Convener: *Chieh-Chih Estelle Cheng, David Castle, Rorie Edmunds, Aminata Garba, Masao Mori,
Karen Payne, Hideaki Takeda

- **Sharing of the COVID-19 Data**: Sharing information on data involving COVID-19 pandemics for all related branch of disciplines, including social, genome and biological aspects. As this session focus on international interest involving the pandemics, all related topics on COVID-19 data issues are welcome.

Conveners: *Elaine Faustman, Marc Nyssen, Masaki Kanao, Tadahiko Maeda, Tomoya Baba,
Mari Minowa

- **Forum of Early Career Data Scientists in the Asia & Oceania Area**: A forum of young-generation data scientists in the Asia-Oceania area to identify current problems and future plans. This part will be led by the WDS ECR Network.

Conveners: *Akira Kadokura, WDS ECR Network members, Juanle Wang, Yoshimasa Tanaka,
Takashi Watanabe

- **Promotion of Multi-Disciplinary Data Analysis**: Presentations on research projects based on multi-, and/or inter- disciplinary data use and data-led innovation of technology. Themes of space-weather data analyses are also included.

Conveners: *Takashi Watanabe, Asanobu Kitamoto, Masahito Nose, Ryuho Kataoka

Keynote Speakers:

Mark A. Parsons (Editor-in-Chief, Data Science Journal)
Mazlan Othman (Regional Office of Asia and Pacific, International Science Council)
Jens Klump (Mineral Resources, CSIRO in Australia)
Lesley Wyborn (Australian National University)
Yusuke Komiyama (National Institute of Informatics)
Devika Madalli (Indian Statistical Institute)
Masanori Arita (National Institute of Genetics)
Priyanka Pillai (Research Data Stewardship and Health Informatics)
Kassim S. Mwitondi (Sheffield Hallam University)

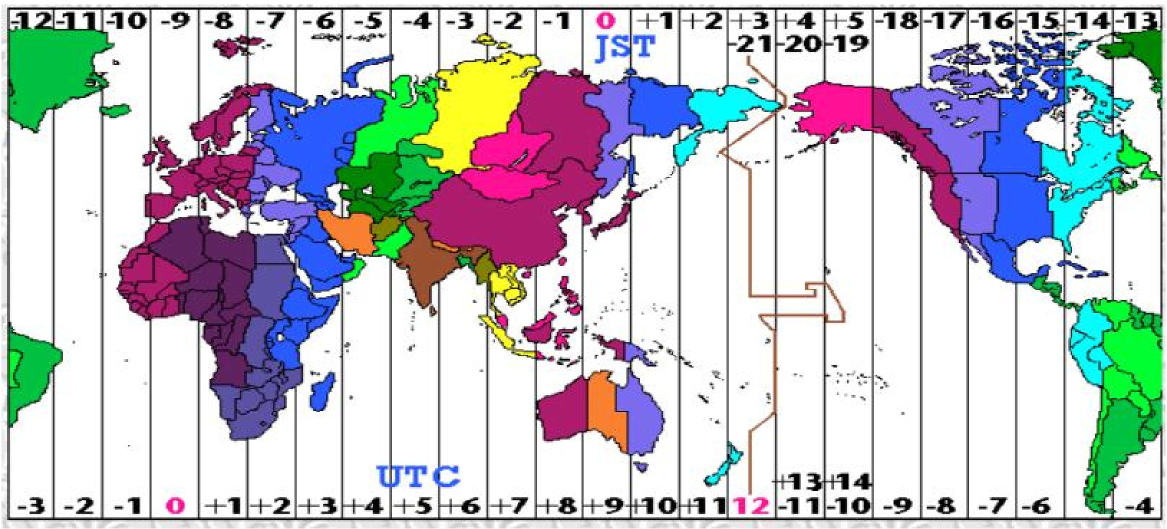
Conference Programme:

(JST=UTC+9h)

Dates	Time-slot 1	Time-slot 2	Time-slot 3
23 SEP (WED)		**Opening Addresses, Key-note Talks, 10 Years Event of WDS-IPO** **(05:00-08:00 UTC)** **(14:00-17:00 JST)**	**WDS Members' Forum 2020** **(12:00-15:00 UTC)** **(21:00-24:00 JST)**
24 SEP (THU)	**Regional Activities on Data in the Asia-Oceania Area** **(01:00-03:30 UTC)** **(10:00-12:30 JST)**	**WDS-ORCID Strategic Workshop** **(05:00-08:00 UTC)** **(14:00-17:00 JST)**	**Sharing of the COVID-19 Data** **(10:00-16:00 UTC)** **(19:00-25:00 JST)**
25 SEP (FRI)	**Forum of Early Career Data Scientists in the Asia-Oceania Area** **(01:00-03:30 UTC)** **(10:00-12:30 JST)**	**Promotion of Multi-Disciplinary Data Analysis** **Conclusive Addresses** **(05:00-08:30 UTC)** **(14:00-17:30 JST)**	

Detail Programme for all seven sessions is available from here;

Programme & Session detail: DSWS-2020 Programme & Session detail.pdf



Presentations:

Online Presentations:
General presenters are principally allocated 20 minutes including questions and discussion time.
Keynote speakers are principally allocated 30 minutes including questions and discussion time.
Several sessions prepare "lightning talk" presentations for few minutes, depending on the session content.

There is no poster presentation in this conference, therefore, those who wants to make poster presentation could submit the poster content into the Extended Abstract.

All presentations will be made via the online Meeting Application (Zoom).
Detail of the link information on the Zoom meeting room for all sessions is available by reply mail when you registered from the Pre-Registration Form
"Pre-conference registration form".

Zoom meeting room:

Audio and Video control:
On entering the Zoom meeting room for each session, all delegates will be muted, and they should turn their video mode to "OFF". While, delegates can change these settings, it is advised that they stay muted and keep their video in the OFF mode throughout the time they are listening to the presentations to avoid causing disturbance and save bandwidth. Presenters should turn on the audio and video modes only when it is their turn to present. The buttons to turn on the audio and video mode will typically be at the bottom left corner of your screen; but muting and unmuting can also be done by clicking on the dots at the top right corner of your image. Muting and unmuting can also be done via your audio headset if you are using one.

Asking questions and discussions:
During question time and discussions, delegates should only turn "ON" the "microphone" (unmute themselves), when it is clear they have been given an opportunity to talk by the session chairs. This can happen in different ways – for example, it could be a quick question to be raised in the middle of the presentation or it may follow the directions of the session chairs. Delegates can raise their hand electronically. You can raise your hand by clicking on the icon labelled "Participants" at the bottom of your screen and then click the button "Raise Hand." You can also use the “Chat Room”, at the bottom of your screen, to send text messages to everyone or to an individual participant. The chatroom will be monitored by the conveners of the session.

Sharing Screen:
Presenters will be expected to share their screens with the rest of the audience. Sharing your screen is a straightforward process. Simply select "Share Screen" at the bottom of your screen, which will display various options for you to share. You will typically want to share your Desktop, as that is what you will be looking at. Once you have sone so, select your Power Point slides or any other application you may wish to share and click "Share" at the bottom right corner. You can end screen sharing by selecting the relevant button at the top of your screen. If sharing external content with audio or video, such as YouTube clips, make sure you check the box "Share Computer Sound" at the bottom left corner of your screen.

Full Screen Mode:
Please note that you might not be able to see some of the options if you are in Full Screen mode. You can use the Escape key to leave the mode and use the double arrow sign, at the top right corner of your screen to enter full screen mode.

For those who are not familiar with screen sharing, we advise that they send their presentation files to the session conveners prior to the conference. Otherwise, you may send your presentation movie file to the conveners instead of an actual oral presentation.

Video recording:
All the sessions are planned to be recorded into the Zoom cloud. After the symposium has been completed, delegates can watch the session videos using the link information that will be provided by the Local Organizing Committee (LOC). The video watching period is set to be terminated by the end of October 2020.

Please feel free to contact the LOC and/or the session conveners, if you have any questions or issues you would like to raise.

Abstract submission:

All presenters are required to prepare an Extended Abstract to be included in the conference "Programme and Abstracts" booklet. The booklet will be published on the conference website.

The following doc.file contains a set of guidelines for preparing an Extended Abstract submission. Guidance is given on layout, text formatting, figures/tables, and references to ensure that all submissions are clear and consistent. The template is presented exactly as your extended abstract should appear, and it is highly advised to use it to prepare your submission in Word format. It is noticed that the manuscript should not exceed two pages in length, and should fit within the margins given in this template.

Abstract Format: Guidelines for Extended Abstracts - dsws.2020.doc (please download from conference website)

All the manuscript file(s) for Extended Abstracts should be emailed to [dsws2020.admn (at) gmail.com] by **31 August 2020**.

Each submission will be given a light review by a member of the LOC or AC (see below) to confirm the typesetting and formatting adhere to the template, as well as the suitability of the presentation for inclusion in the programme.

Registration:

Pre-conference registration and presentation submission can be made from here. No registration fee is required to attend the conference.
"Pre-conference registration form"

Detail Zoom meeting information for all sessions will be sent by reply mail when you registered.

For the participants who are staying in places/countries where the above Google Form cannot be accessed for any reasons, please use the "Offline pre-registration form (xls.file) ". The "Offline pre-registration form" should be emailed to [dsws2020.admn (at) gmail.com] prior to the conference.

Social Events (provisional):

The organizers are planning the following social events in conjunction with the symposium.

- **10-Year Anniversary Event for the International Programme Office of the World Data System** (23 September, 0500-0800 UTC). This will take place by online virtual meeting by Zoom.

Local Organizing Committee (LOC): (* Chair)

Tomoya Baba (Data Science Promotion Section, DS, ROIS)
Akira Kadokura (Polar Environment Data Science Center, DS, ROIS)
*Masaki Kanao (Polar Environment Data Science Center, DS, ROIS)
Mari Minowa (Data Science Promotion Section, DS, ROIS)
Yasuhiro Murayama (National Institute of Information and Communications Technology)
Koji Nishimura (Polar Environment Data Science Center, DS, ROIS)
Akihiko Nomizu (Data Science Promotion Section, DS, ROIS)
Masahito Nose (Nagoya University)
Hideaki Takeda (National Institute of Informatics, ROIS)
Yoshimasa Tanaka (Polar Environment Data Science Center, DS, ROIS)
Hironori Yabuki (Polar Environment Data Science Center, DS, ROIS)
Takashi Watanabe (WDS-IPO)

International Advisory Committee (AC): (* Chair)

Phillippa Bricher (Australian Antarctic Division)
Taco De Bruin (Royal Netherlands Institute for Sea Research, Netherland)
Chieh-Chih Estelle Cheng (ORCID)
Shannon Christoffersen (University of Calgary, Canada)
Alexander de Sherbinin (NASA SEDAC; WDS-SC)
Rorie Edmunds (WDS-IPO)
Hiroyuki Enomoto (National Institute of Polar Research, ROIS)
Elaine Faustman (University of Washington; WDS-SC)
*Asao Fujiyama (Joint Support-Center for Data Science Research, ROIS)
Susumu Goto (Database Center for Life Science, DS, ROIS)
Yuko Harayama (ORCID Board Researcher Member)
Kazuhiro Hayashi (National Institute of Science and Technology Policy)
Heidi J. Imker (Illinois University, USA)
Toshihiko Iyemori (Kyoto University; International Union of Geodesy and Geophysics; WDS-SC)
Ryuho Kataoka (National Institute of Polar Research, ROIS)
Asanobu Kitamoto (Center for Open Data in the Humanities, DS, ROIS)
Jens Klump (Mineral Resources, CSIRO in Australia)
Yuji Kohara (Database Center for Life Science, DS, ROIS)
Tadahiko Maeda (Center for Social Data Structuring, DS, ROIS)
Masao Mori (Tokyo Institute of Technology)
Hideki Noguchi (Center for Genome Informatics, DS, ROIS)
Tsuneo Odate (National Institute of Polar Research, ROIS)
Mazlan Othman (Regional Office of Asia and Pacific, International Science Council)
Mark Parsons (Rensselaer Polytechnic Institute, USA)
Peter Pulsifer (University of Colorado, USA)
Seiji Tsuboi (Japan Agency for Marine-Earth Science and Technology)
Genta Ueno (Center for Data Assimilation Research and Applications, DS, ROIS)
Juanle Wang (China Academy of Science; WDS-SC)

Organized by:

Joint Support-Center for Data Science Research (DS), Research Organization of Information and Systems (ROIS)
Committee of International Collaborations on Data Science, Science Council of Japan (SCJ)

Supported by:

National Institute of Information and Communications Technology (NICT)
Research Organization of Information and Systems (ROIS)
World Data System (WDS), International Science Council (ISC)
Open Researcher and Contributor ID (ORCID)
Science Council of Japan (SCJ)

Workshop Website:

https://ds.rois.ac.jp/article/dsws_2020/

Important Dates:

Registration & Abstract submission open: 01 April 2020
Presentation submission deadline: 15 August 2020
Extended Abstract submission deadline: 31 August 2020
Programme booklet online: 15 September 2020
Registration deadline: 23 September 2020
Symposium: 23–25 September 2020

Contact Address:

dsws.loc-2020 (at) nipr.ac.jp

International Symposium
- DSWS-2020 -

“Global Collaboration on Data beyond Disciplines”

PROGRAMME

Online Conference

23–25 September 2020

PROGRAMME

International Symposium
“Global Collaboration on Data beyond Disciplines”

23 – 25 September 2020

Online Conference

Wednesday 23 September 2020
Opening Session (05:00-08:00 UTC)

Conveners: *Masaki Kanao, Yasuhiro Murayama, Takashi Watanabe, Rorie Edmunds

05:00–05:20 **Opening Addresses** (Chair: Masaki Kanao)

- **Asao Fujiyama** (Director-General, Joint Support-Center for Data Science Research (DS), ROIS) (10’)
- **Masaki Kanao** (LOC Chair), outline of symposium & practical information (10’)

05:20–06:35 **Keynote Talks** (Chair: Takashi Watanabe)

Keynote 1: *Past, present and future of global data collaboration*
Mark A. Parsons (Editor-in-Chief, Data Science Journal) (25’)
Keynote 2: *Open Science for Asia and the Pacific*
Mazlan Othman (Regional Office of Asia and Pacific, International Science Council) (25’)
Keynote 3: *Do we all need to become data scientists?*
Jens Klump (Mineral Resources, CSIRO in Australia) (25’)

06:35–06:45 **Group Photo @ Zoom Screen & Coffee Break**

06:45–08:00 **10 Years Event of the International Programme Office of the World Data System (WDS-IPO)**
(Chair: Yasuhiro Murayama; each talk is allocated for 5’-10’)

- **Hideyuki Tokuda** (President, National Institute of Information and Communications Technology (NICT))
- **Ryoichi Fujii** (President, Research Organization of Information and Systems (ROIS))
- **Kazuhiko Takeuchi** (Vice President, Science Council of Japan)

- **Daya Reddy** (President, International Science Council)
- **Alexander de Sherbinin** (Chair, Scientific Committee of World Data System)
- **Elaine Faustman** (Vice Chair, Scientific Committee of World Data System)

- **Chieh-Chih Estelle Cheng** (Manager APAC Engagement, ORCID)
- **Rorie Edmunds** (International Program Office, World Data System)

- **Discussion**

Wednesday 23 September 2020

WDS Members' Forum 2020 (12:00-15:00 UTC)

Conveners: * Rorie Edmunds, WDS-SC, WDS-IPO, WDS-ITO

12:00–13:30 **Scientific Session – Member Lightning Talks**

(Chairs: Alex de Sherbinin, Rorie Edmunds)

- *Welcome & Introduction* (5')
  Alex de Sherbinin (Chair, WDS-SC), **Rorie Edmunds** (Executive Director, WDS)
- *Panel Discussion 1: FAIR Data and Interoperability* (15')
  Marc Nyssen (Moderator)
  Nick Thieberger (PARADISEC)
  Wendy Gross (World Data Service for Paleoclimatology)
  Michael Diepenbroek (PANGEA – Data Publisher for Earth & Environmental Science)
- *Panel Discussion 2: Automation, Scalability, and Cloud Use* (20')
  Aminata Garba (Moderator)
  Jeanne Behnke(NASA-ESDIS Project)
  Martina Stockause (DKRZ – WDC Climate)
  Jerry Carter (Incorporated Research Institutions for Seismology)
  Andrii Shelestov (Ukrainian Geospatial Data Center)
- *Panel Discussion 3: Sustainability: Making the case for domain repositories through value of information, operational use, & web services* (20')
  Robert Chen (Moderator)
  Mamoru Ishii (WDC – Ionosphere and Space Weather)
  Adam Shepherd (Biological and Chemical Oceanography Data Management Office)
  Doug Schuster (National Center for Atmospheric Research)
  Frederic Clette (WDC – Sunspot Index and Long-term Solar Observations)
- *Breakout Discussions* (20')
  Karen Payne, David Castle, Loana Popescu (Moderators)
- *Wrap-up of Scientific Session* (5')
  Alex de Sherbinin (Chair, WDS-SC)

13:30–13:50 **Coffee Break**

13:50–15:00 **Plenary Session – Community Consultation**

(Chairs: Alex de Sherbinin, Rorie Edmunds)

- *Welcome & Introduction* (5')
  Alex de Sherbinin (Chair, WDS-SC), **Rorie Edmunds** (Executive Director, WDS)
- *WDS-ITO Activities (+ Discussion)*
  Karen Payne (Associate Director, WDS-ITO) (15')
- *Data Together: WDS & Member Involvement (+ Discussion)*
  Ingrid Dillo (Vice Chair, WDS-SC) (15')
- *Candidate Membership & Mentoring (+ Discussion)*
  Wim Hugo (Director of Strategy, WDS) (15')

- *Domain vs Generalist Repositories (+ Discussion)*
  Alex de Sherbinin (Chair, WDS-SC) (15’)
- *Wrap-up of Plenary Session*
  Alex de Sherbinin (Chair, WDS-SC) (5’)

Thursday 24 September 2020

Regional Activities on Data in the Asia & Oceania Area (01:00-03:30 UTC)

Conveners: *Toshihiko Iyemori, Mazlan Othman, Juanle Wang, Takashi Watanabe

01:00–03:30 **Regional Activities on Data in the A & O Area**

(Chairs: Mazlan Othman, Toshihiko Iyemori)

- *National Cross-Domain Activities on Research Data in Australia and Connections with International Activities (Invited)*
  Lesley Wyborn (Australian National University) (20’)
- *WDS-led Activities on Data in the Asia-Oceania Area*
  Takashi Watanabe (International Program Office, World Data System) (15’)
- *Open Science and its Policy towards Open Science Paradigm*
  Kazuhiro Hayashi (National Institute of Science and Technology Policy) (15’)

- *General-purpose research-data management service for international research collaboration (Invited)*
  Yusuke Komiyama (National Institute of Informatics) (20’)

Coffee break (10')

- *Scientific Data Standard System Construction and its Application for Data Archiving in China*
  Juanle Wang (Chinese Academy of Sciences) (15’)
- *Collection, Archive and Sharing Space Weather information in NICT*
  Mamoru Ishii (National Institute of Information and Communications Technology) (15’)
- *Making historical data Re-useable: a case study of the challenges and success of Bangladesh Bureau of Statistic’s historical data conversion project*
  Chandra Shekhar Roy (Bangladesh Bureau of Statistics) (15’)

- *Data Stewardship: Indian Perspective (Invited)*
  Devika Madalli (Indian Statistical Institute) (20’)

- **Discussion / Summary**

Thursday 24 September 2020

WDS – ORCID Strategic Workshop: Adoption of PIDs in Asia-Oceania (05:00-08:00 UTC)

Conveners: *Chieh-Chih Estelle Cheng, David Castle, Rorie Edmunds, Aminata Garba, Masao Mori, Karen Payne, Hideaki Takeda

05:00–06:20 **Part 1 – PIDs: A Vital Component of Global Research Data Infrastructure**

(Chairs: Rorie Edmunds, Estelle Cheng)

- **Introduction by Chairs (5')**
- *Adopting Persistent Identifiers in Open Research Infrastructure: ORCID and beyond*
  Chieh-Chih Estelle Cheng (ORCID, APAC Engagement Manager) (20')
- *Japanese DOI RA, Japan Link Center (JaLC) and its collaboration with ORCID*
  Masashi Hara (Japan Science and Technology Agency) (20')
- *The IGSN Global Sample Number - A PID for Physical Objects*
  Jens Klump (Mineral Resources, CSIRO in Australia) (20')
- **Discussion & Wrap Up (15')**

06:20–06:40 **Coffee Break**

06:40–08:00 **Part 2 – Practical Use Cases: From National PID Systems to Sector Applications**

(Chairs: Rorie Edmunds, Estelle Cheng)

- **Introduction by Chairs (5')**
- *The roles of PIDs in Scientific Data Management in China*
  Xiaoyan Hu (National Space Science Center, CAS, Space Science Big Data Technology Laboratory) (15')
- *Korea Research Data Management Platform and PID*
  Sa-kwang Song (Korea Institute of Science and Technology Information) (15')
- *DataCite Commons – A new service to explore the PID Graph*
  Martin Fenner (DataCite, Germany) (15')
- *The Power of PIDs at the Cambridge Crystallographic Data Centre*
  Ian Bruno (Cambridge Crystallographic Data Centre) (15')
- **Discussion & Wrap Up (15')**

Thursday 24 September 202

Sharing of the COVID-19 Data (10:00-16:00 UTC)

Conveners: *Elaine Faustman, Marc Nyssen, Masaki Kanao,
Tadahiko Maeda, Tomoya Baba, Mari Minowa

10:00–12:00 **Sharing of the COVID-19 Data - 1**

(Chair: Tomoya Baba, Tadahiko Maeda)

Keynote 1: *Open Access and Data Sharing of Nucleotide Sequence Data*
Masanori Arita (National Institute of Genetics) (25')

- *Time-series analysis of directional sequence changes in SARS-CoV-2 genomes and an unsupervised*

explainable AI for studying corona virus genomes
Toshimichi Ikemura (Nagahama Institute of Bio-Science and technology) (20’)

- *Susceptibility to COVID-19 infection: Insights from mathematical modelling*
  Ryosuke Omori (Research Center for Zoonosis Control, Hokkaido University) (20’)
- *Wastewater-based epidemiology for COVID-19: Perspectives for environmental surveillance of SARS-CoV-2 in wastewater*
  Ryo Honda (Faculty of Geosciences and Civil Engineering, Kanazawa University) (20’)
- *Spatial analysis of COVID-19 spread using compositionally-warped Gaussian process*
  Daisuke Murakami (Center for Data Assimilation Research and Applications, DS, ROIS) (20’)

- **(Discussion)**

- **Coffee Break**

12:00–14:00 **Sharing of the COVID-19 Data - 2**
(Chair: Marc Nyssen, Mari Minowa)

Keynote 2: *Sharing data for a coordinated response during COVID-19 pandemic*
Priyanka Pillai (Research Data Stewardship and Health Informatics) (25’)

- *Sharing literature annotation regarding COVID-19 through PubAnnotation*
  Jin-Dong Kim (Database Center for Life Science, DS, ROIS) (20’)
- *COVID-19 Data and Knowledge Hub*
  Liu Chuang (Global Change Research Data Publishing & Repository and GCdataPR -WDS) (20’)
- *COVID-19 Data Sharing Initiative by WDCM and NMDC*
  Linhuan Wu (Institute of Microbiology, Chinese Academy of Sciences) (20’)
- *Machine Learning based Peptide Therapeutics for Covid-19*
  Shailza Singh (National Center for Cell Science, India) (20’)

- **(Discussion)**

- **Coffee Break**

14:00–16:00 **Sharing of the COVID-19 Data - 3**
(Chair: Elaine Faustman, Masaki Kanao)

Keynote 3: *Monitoring Sustainable Development Goals Amidst COVID-19 Through Big Data, Deep Learning and Interdisciplinarity*
Kassim S. Mwitondi (Sheffield Hallam University) (25’)

- *The RDA Community Response to COVID-19*
  Mark Leggott (Research Data Canada) (20’)
- *VODAN-in-a-box: a FAIR toolkit for improving reusability of COVID-related data*
  Luiz Olavo Bonino da Silva Santos (GO FAIR International Supporting and Coordination Office) (20’)
- *Detailed Methodology and Application Guidelines for “The Global Covid-19 Index (GCI)”*
  Woody Ang Woo Teck (PEMANDU Associates, Malaysia) (20’)
- *Leveraging open data towards containing COVID-19*
  Obwaya Mogire (South Eastern Kenya University) (20’)

- **Discussion**

Friday 25 September 2020

Forum of Early Career Data Scientists in the Asia & Oceania Area (01:00-03:30 UTC)

Conveners: *Akira Kadokura, WDS ECR Network members,
Juanle Wang, Yoshimasa Tanaka, Takashi Watanabe

01:00–03:30 **Forum on Early Career Data Scientists in the A & O Area**

(Chairs: Jesse Xiao, Akira Kadokura)

- *Introduction for WDS ECR Network (Invited)*
  Lianchong Zhang, Jesse Xiao (The University of Hong Kong, WDS ECR Network) (15')
- *Information platform promoting global collaboration in microbial community (Invited)*
  Linhuan Wu (Institute of Microbiology, Chinese Academy of Sciences) (15')
- *Disaster Rapid Damage Mapping with Remote Sensing: benchmarking datasets and methods (Invited)*
  Junshi Xia (Institute of Physical and Chemical Research (RIKEN)) (15')
- *Introduction of FUXI Platform -Global Integrated Observation Data Management and Service System (Invited)*
  Yuanyuan Wang (Zhejiang University) (15')

- *Activities toward Open Data and Research Data Management at NIES/CGER (Invited)*
  Yoko Fukuda (National Institute for Environmental Studies) (15')
- *Standardization of biological sample information database (Invited)*
  Tazro Ohta (Database Center for Life Science, DS, ROIS) (15')
- *Official Micro Data, Causal Inference and Evidence-Based Policy Making (Invited)*
  Junchao Zhang (Center for Social Data Structuring, DS, ROIS) (15')
- *Recent Achievements of Deep Learning on Recognition of Modern Japanese Magazines (Invited)*
  Anh Duc Le (Center for Open Data in the Humanities, DS, ROIS) (15')
- *A data-scientific approach toward understanding of the goldfish genome and morphological diversity (Invited)*
  Tetsuo Kon (Nagahama Institute of Bioscience and Technology) (15')
- *Modelling the Suitability of Parcel Pick-up Lockers Using the Multi-Source Open Data (Invited)*
  Zilai Zheng (Graduate School of Life and Environmental Sciences, University of Tsukuba) (15')

- **Discussion**

Friday 25 September 2020

Promotion of Multi-Disciplinary Data Analysis, Conclusive Adresses (05:00-08:30 UTC)

Conveners: *Takashi Watanabe, Asanobu Kitamoto, Masahito Nose, Ryuho Kataoka

05:00–06:30 **Promotion of Multi-Disciplinary Data Analysis - 1**

(Chair: Takashi Watanabe, Asanobu Kitamoto)

- *Introduction of the Session*
  Takashi Watanabe (International Program Office, World Data System) (5')
- *Developing Open-Source Data Analytics and Tools to Address Interdisciplinary Issues in Political Science, Human Geography, and Environmental Science*
  Joshua Brinks (iSciences, Limited Liability Company, Burlington) (20')

- *Implementation of OA/FAIR Principles in Six Years - A Case Study of Global Change Research Data Publishing & Repository*
  Liu Chuang (Chinese Academy of Sciences) (20')
- *Data analysis and intelligent application in modern agricultural technology extension*
  Ying Deng (National Engineering Research Center for Information Technology in Agriculture, China) (20')
- *A Comparison between Logistic Regression and Decision Tree Methods for Predicting the Pre-term Birth*
  Rakesh Kumar Saroj (SRM University Sikkim-Gangtok, India) (20')

- **Coffee Break**

06:30–08:00 **Promotion of Multi-Disciplinary Data Analysis – 2**
(Chair: Takashi Watanabe, Asanobu Kitamoto)

- *Importance of a Multi-disciplinary Study of Peculiar Environmental and Socio-economic Movements in 18/19 Century*
  Takashi Watanabe (International Program Office, World Data System) (20')
- *A unique website JCDP that aims to disseminate scientific information on historical climate data*
  Takehiko Mikami (Tokyo Metropolitan University) (20')
- *Historical records of red auroras in Japan: Collaboration between literature and science*
  Ryuho Kataoka (National Institute of Polar Research) (20')
- *Research Across Disciplinary Boundaries: Data Challenges and Solutions in the Environmental and Eco-social Sciences*
  Alison Specht (the University of Queensland, Australia) (20')

- **Discussion**

08:00–08:30 **Conclusive Addresses** (Chair: Masaki Kanao; each talk is allocated for 5'-10')

- **Yasuhiro Murayama** (National Institute of Information and Communications Technology (NICT))
- **Myung-seok Choi** (Korea Institute of Science and Technology Information (KISTI))
- **Masaki Kanao** (LOC Chair), summary of symposium & conference proceeding

Extended Abstract only @ conference web-site

Extended Abstracts will be included in the conference "Programme and Abstracts" online booklet.

- *EA-1: Application of machine learning techniques to weather forecasting*
  V. Sakthivel Samy (National Centre for Polar and Ocean Research, India)
- *EA-2: Analyzing the early 19th century's geomagnetic declination in Japanese archipelago from Tadataka Inoh's 67 volumes magnetic survey azimuth ledger Santou-Houi-Ki by Interdisciplinary study*
  Motohiro Tsujimoto (no affiliation)
- *EA-3: Data and metadata sharing among Asian countries*
  Masaki Kanao (Joint Support-Center for Data Science Research, ROIS)

International Symposium
- DSWS-2020 -

“Global Collaboration on Data beyond Disciplines”

ABSTRACTS

Online Conference

23–25 September 2020

Open Access and Data Sharing of Nucleotide Sequence Data

Masanori Arita[1*,2]

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

Summary. Open access, free access, and public domain are different notions. For nucleotide sequence data (NSD), the best practice is the policy of INSDC, or the International Nucleotide Sequence Database Collaboration. INSDC partners, consisting of DDBJ in Japan, EMBL-EBI in Europe, and NCBI in the US, have been the core framework of sharing NSD for over 30 years. Recent virus information, however, are mainly distributed through the GISAID repository supported by the World Health Organization. The historical background of NSD databases can explain the outcome and future prospect of the different publication policies.

Keywords. nucleotide sequence data (NSD), database, repository, INSDC, data policy

1. Open access, free access, and public domain

Open access, free access, and public domain are all different. Although all of them seem to provide access without cost, licensing terms may differ. Open access implies the clear statement of user rights in licensing terms. Free access, often associated with the sentence "all rights reserved", indicates the requirement of an official inquiry when the contents are commercially used or re-distributed. This is why commercial companies do not use 'free' software programs. Lastly, public domain indicates no copy- or other rights; any type of usage is allowed even without credit. This difference is evident for artificial products such as artworks or literature. For nucleotide sequence data (NSD) from nature, however, the best licensing is not always evident.

Empirically, INSDC (International Nucleotide Sequence Database Collaboration) provides a key to viable solutions. The fundamental data-sharing policy of INSDC was published in 2002 by its advisory board: free and unrestricted access to all of the data records, without use restrictions, licensing requirements, or fees on the distribution or use by any party including commercial sectors [1]. INSDC is the core framework for sharing NSD for over 30 years and consists of three partner nodes: the DNA Data Bank of Japan (DDBJ) at the National Institute of Genetics in Mishima, Japan [2]; the European Nucleotide Archive (ENA) at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK [3]; and GenBank at National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health in Bethesda, Maryland, USA [4]. GenBank and Embl-Bank were established in 1982 and followed by DDBJ in 1987. The three institutions keep collecting and providing NSD with unique identifiers called Accession

Numbers (ANs). ANs are mandatory information for sequence studies in the vast majority of life science and medical journals, and the relationship with editorial offices was established to guarantee data accessibility and to assist reproducibility of published results.

2. Traceability of NSD

NSD are often the crucial part of life science studies. At the same time, their country/regional origin may become important for the bioresource access and benefit sharing (ABS). The INSDC system offers metadata information to describe the origin information using the BioSample repository and the /COUNTRY qualifier in the NSD annotation, defined as "locality of isolation of the sequenced organism indicated in terms of political names for nations, oceans or seas, followed by regions and localities". With these metadata with ANs, scientific origin and responsible authors become traceable for NSD. The same is true for patented sequences. In the US, Europe, Japan and Korea, patented sequences are tagged with the ANs and publicized within the INSDC framework. Although tracking every usage is difficult, the current system offers realistic and sustainable solution for data traceability.

3. GISAID and virus sequences

The GISAID repository (https://gisaid.org) by World Health Organization was launched in 2008 on request from virus-affected countries hesitant to publish sequences from the INSDC. GISAID is the default repository of avian influenza and coronavirus with support from World Health Organization. The fundamental difference from the INSDC is its access agreement: GISAID users are forbidden to access, modify, distribute, or even display deposited data in connection with any other database. The strict condition affects availability of data from each nation. The United States and European countries register covid-19 sequences to both INSDC and GISAID. China does to both its national Center (https://bigd.big.ac.cn/) and GISAID. Japan, however, deposit sequences only to GISAID.

This difference is affecting research activities in each country. The number of covid-19 related articles in PubMed database exceeds 50,000, and China is the most prolific country. Among major countries, the number of articles from Japan is relatively small, probably due to the restricted accessibility for Japanese data. Although the number of research paper is not an ideal measure of scientific quality, the size of research body scales well with the article output.

4. Conclusions

Data access and sharing policy affects the size of research body and therefore outputs. Since data traceability is achieved through metadata with unique identifiers even for open data, coronavirus sequences are better publicized without access restriction.

Acknowledgments. The author is grateful to the INSDC members. DDBJ is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, Japan Agency for Medical Research and Development (AMED), and National Bioscience Database Center (NBDC).

References

1. Brunak, S., Danchin, A., Hattori, M., Nakamura, H., Shinozaki, K., et al. Nucleotide sequence database policies. *Science*, 298(5597), 1333, DOI: 10.1126/science.298.5597.1333b, 2002

2. Ogasawara O, Kodama Y, Mashima J, Kosuge T, Fujisawa T. DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Res*, 48(D1), D45-D50, DOI: 10.1093/nar/gkz982, 2020
3. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, et al. The European Nucleotide Archive in 2019. *Nucleic Acids Res,* 48(D1), D70-D76, DOI: 10.1093/nar/gkz1063, 2020
4. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, et al. GenBank. *Nucleic Acids Res*, 48(D1), D84-D86, DOI: 10.1093/nar/gkz956, 2020

Developing Open-Source Data Analytics and Tools to Address Interdisciplinary Issues in Political Science, Human Geography, and Environmental Science

Joshua Brinks[1]*, ***Thomas Parris***[1], ***Robert Chen***[2], ***Eileen Shea***[3], ***Cynthia Crowley***[1], ***Dan Baston***[1]

[1]* *ISciences, LLC, Burlington, VT 05401, United States*
[2] *CIESIN/Columbia University; Manager, NASA Socioeconomic Data and Applications Center (SEDAC), United States*
[3] *CASE Consulting International, North Carolina, United States*
Email: jbrinks@isciences.com

Summary. Environment-security researchers and practitioners must overcome several technical challenges in the course of their efforts. In the absence of an open-community platform, each researcher or institution compiles and curates their own data repositories, analytical tools, and computational environments. This results in significant duplication of effort, limits the replication and validation of research, and diminishes the interoperability of analytic tools. The Data ANalytics and Tools for Ecosecurity (DANTE) Project is a cooperative partnership between the United States Army Corps of Engineers, CIESIN, ISciences, L.L.C., and CASE International that aims to develop a community platform with a suite of open-source tools and content that reduce common barriers to entry for interdisciplinary research in human geography, social and political science, and global environmental change. In addition to providing a curated catalogue of data sets and custom R programming tools tailored to interdisciplinary research, one of the primary goals of the DANTE Project is to demonstrate reproducible research through replication of recent high impact peer reviewed studies.

Keywords. Open-source tools, R, environment-security, gridded data

1. Introduction

Social and political scientists, analysts, and policymakers in the evolving area of environment and security must work closely with Earth scientists to access reliable and targeted data products describing our rapidly changing planet. These include scenarios and predictions of future conditions due to accelerating changes in coupled human-environment systems. Providing this data and information is a tremendous opportunity for both scientific advancement and helping society respond effectively to a range of emerging challenges such as pandemics, extreme climate events, resource conflict, mass migration, and supply chain disruption. We argue that the development of sustained and trusted partnerships between stakeholders from the academic, national security, humanitarian, and private sectors is needed in the environment-security community. Establishing such partnerships also requires contributions from a wide variety of scientific disciplines--including the natural, social, health, engineering, and data sciences--along with new tools to facilitate the integration of science into operational products and services that are reliable and replicable.

Central to this premise, and ultimately success in the environment-security arena, is the recognition of recent trends in scientific inquiry and theory that have witnessed a movement towards replicable science and open data practices. These include open-source code and proper attribution of data sets, software, and packages used for analyses. Although replication is a core component of the scientific method, it is often the most overlooked. In the absence of replication, scientific endeavours may be rewarded by poor methods that result in negative policy implementations and erroneous theoretical assumptions. While close replication presents a significant challenge in psychological research, clinical trials, and field ecology, research efforts focused on standardized global datasets of global climate, migration, political conflict, and agriculture can more easily adhere to open data and open-access science principles to reproduce their research.

2. Conclusions

This presentation will highlight current issues in multidiscipinary partnerships at the intersection of the human geography, environmental stressors, and political science and conflict research sectors. Additionally, we will showcase recent open-science efforts from the DANTE Project. This includes, but is not limited to, spatial analysis tools, open-access R packages, gridded datasets, and materials promoting open-science and replication in the environmental, political, and geographical sciences.

Acknowledgments. This presentation, and the DANTE Project, are funded by United States Army Corp of Engineers STTR program.

The Power of PIDs at the Cambridge Crystallographic Data Centre

Ian Bruno[1*], ***Eric Rogers***[1], ***Suzanna Ward***[1]

[1*] Cambridge Crystallographic Data Centre, *12 Union Road, Cambridge, CB2 1EZ, United Kingdom*
Email: bruno@ccdc.cam.ac.uk

Summary. The Cambridge Crystallographic Data Centre (CCDC) is responsible for the management and curation of the Cambridge Structural Database (CSD) which contains the experimentally-determined 3D crystal structures of over 1 million organic and metal-organic chemical compounds. This workshop presentation outlines the use of persistent identifiers (PIDs) at CCDC. Particular attention is paid to the integration of PIDs into CCDC's data management, preservation and publication workflows and the utility of PIDs for aligning the repository with FAIR data initiatives. It will also outline how PIDs can be used to enable linking between related research objects and data repositories and highlight opportunities for further channelling the power of PIDs.

Keywords. *Persistent identifiers, data repository, Cambridge Structural Database (CSD), chemistry, crystallography*

1. Introduction

Since the establishment of the Cambridge Structural Database (CSD) [1] over 50 years ago, the CCDC has sought to develop and adopt mechanisms for improving the efficiency of how data is stored, processed and shared with its user communities. Throughout this, persistent identifiers (PIDs) have proved to be important tools for enabling the reliable management and preservation of data by the CCDC.

2. Using PIDs to make data FAIR

The value which PIDs can add to research data has been highlighted through the FAIR Guiding Principles for scientific data management and stewardship [2]. As a proponent of FAIR data, the CCDC has adopted PIDs where possible with the aim of rendering the data it holds more findable, accessible, interoperable and reusable.

2.1 Identifiers assigned from deposition to publication

Each structure receives a unique accession ID or *Deposition Number* which is communicated back to the depositor to be cited in their manuscript. Upon publication, additional identifiers are added to datasets, including a *CCDC Data DOI*. By registering the DOI through DataCite, the metadata for the structure is made openly accessible and searchable via DataCite.

Since 2016, users can provide an *ORCID ID* when depositing data which will be shown alongside the data record once published. Using the CCDC online data submission form, depositors can also provide a *DOI link to the raw data* of their structures if this is archived in another repository.

2.2 PIDs for CSD Communications

CSD Communications is a platform that provides researchers with the opportunity to publish datasets independently of a journal article [3]. All *CSD Communications* receive a *CCDC DOI* which is communicated directly to the depositor as soon as the data is made publicly available, allowing for immediate access and citation. In 2019, an *ISSN* was also acquired [4] to help publishers, institutions and researchers track and record citations to *CSD Communications*.

3. Using PIDs to link data to articles

The CCDC has agreements in place with journal publishers who communicate publication information for articles associated with datasets directly to us in a machine-readable format. Using PIDs as a central component for matching records, metadata is automatically updated in the CCDC system so data can be unembargoed and made publicly available alongside an article. Given the amount of X-ray crystal data published each year, these automated systems are vital for accurate addition of metadata to datasets and their timely publication.

The CCDC has been part of the Scholix initiative [5] since its conception. Scholix defines a universal framework for data-article linking that was established through global collaboration between journal publishers, data centres and other service providers. Scholix-based mechanisms are underpinned by the use of PIDs and further enable the automatic identification and publication of links between resources.

4. Future directions for PIDs at the CCDC

Work is currently underway at the CCDC to further harness the opportunities afforded by PIDs.

4.1 Adoption of InChI

The International Chemical Identifier (InChI) [6] provides a way to uniquely and unambiguously identify a particular chemical compound. The CCDC has used InChIs internally to help establish links between the CSD and other chemistry resources and we aim to build on this to enable linking of relevant datasets across chemical and biological domains [7].

4.2 Capturing and linking funding data

Building on the Open and FAIR data movement, international policy makers have put forward proposals for making research more open and transparent. To help those who must monitor compliance with such policies, it is our intention to explore the use of organisational and other identifiers to track the funding and organisations that are associated with published data sets.

5. Conclusions

PIDs enable the reliable identification of datasets, researchers, organisations, and entities such as chemical structures. Adoption of PIDs enables research objects and the links between them to be tracked throughout the research lifecycle and across publication workflows. PIDs help ensure research datasets are discoverable, citable and reusable long after they were first generated.

References

1. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C., The Cambridge Structural Database, *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.*, 72, 171–179, 2016 https://doi.org/10.1107/S2052520616003954
2. Wilkinson M., Dumontier M., Aalbersberg I. *et al.*, The FAIR Guiding Principles for scientific data management and stewardship.

Sci Data, 3, 160018, 2016 https://doi.org/10.1038/sdata.2016.18

3. CCDC, CSD Communications, *The Cambridge Crystallographic Data Centre (CCDC)*. https://www.ccdc.cam.ac.uk/Community/csd-communications/
4. ISSN, CSD Communications, *International Standard Serial Number International Centre*. https://portal.issn.org/resource/ISSN/2631-9888
5. Burton A. *et al.,* The Scholix framework for interoperability in data-literature information exchange. *D-Lib Mag*. 23, 2017. https://doi.org/10.1045/january2017-burton
6. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I., InChI - the worldwide chemical structure identifier standard. *J. Cheminform.* 5, 7, 2013. https://doi.org/10.1186/1758-2946-5
7. CCDC, BioChemGRAPH project will improve data synergy to facilitate drug development, *The Cambridge Crystallographic Data Centre (CCDC),2020* https://www.ccdc.cam.ac.uk/News/List/biochemgraph-project

Adopting Persistent Identifiers in Open Research Infrastructure: ORCID and beyond

Chieh-Chih Estelle Cheng[1*]

[1*] *10411 Motor City Drive, Suite 750, Bethesda, MD 20817, United States*
Email: e.cheng@orcid.org

Summary. One key factor in making science truly open lies in improving data quality and transparency, so that researchers can easily share their work, and organizations can easily acquire the most accurate and up-to-date research information in an open and interoperable way. Such interoperability requires a robust and global-scale infrastructure that enables the open creation of linkages between researchers, research institutions, publishers, and funding organizations. The use of persistent identifiers (PIDs) like ORCID supports the research community in collecting, managing, analyzing and showcasing data transparently and reliably. For Open Science to become a reality, communities and organizations need to collaborate to establish a shared vision and implement policies, technologies, and services.

Keywords. *Persistent identifiers, ORCID, Interoperability, Trusted Connections, Transparency*

1. Introduction

ORCID is a non-profit community driven organization supported by a global community of member organizations, including research institutions, publishers, funders, professional associations, service providers, and other stakeholders in the research ecosystem. ORCID offers an API (Application Programming Interface) that allows systems and applications to connect to the ORCID registry, including collecting data from and connecting data to the ORCID Registry.

Openness is one of the key ORCID values, and ORCID works with the global community and partners to build and sustain a technology infrastructure that supports sharing of research information between systems.

The ORCID Registry supports a wide variety of existing identifiers, which manifests a wide range of connections between ORCID iDs and other PIDs. For instance, ORCID connects with the use of Digital Object Identifiers (DOIs) in publishing and data repositories, and with the use of arXiv identifiers, PubMed and PubMed Central identifiers and most ISBN identifiers, and other organization identifier (OID) like Ringgold ID, GRID ID, LEIs, and Research Organization Registry(ROR). A prime example for how ORCID in alignment with other PIDs to connect researchers, their grants, their contributions, and the organizations is illustrated as follows.

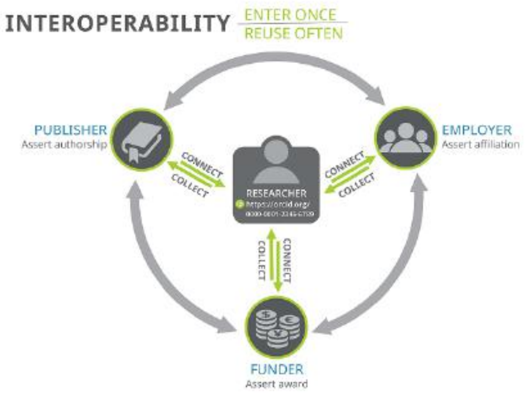


Figure 1. The virtuous circle of interoperability

2. Interoperability and Trusted Connections

Such interoperability and connections among people (researchers and curators), places (research organizations, facilities, funders, repositories, and publishers) and things (such as equipment, facilities, publications, and datasets) are realized through ORCID integrations. Organizations use the ORCID API to link researchers with their affiliations or contributions. All this begins from researchers sharing their ORCID iDs with their affiliated institutes as their personnel record is created, and in turn the institute shares back the researcher's affiliation or education information into the researcher's ORCID record, using the organization ID for the organization (person/ORCID + affiliation/OID). When a researcher goes to a publisher website to submit a manuscript, they can use their ORCID record to share their iD with publishers as well as their institutional affiliation information. And by the time the paper is published, PIDs (person/ORCID + paper/DOI + publisher and research institute/OID) are all interconnected, which can be easily shared with the researcher's home institution or funder [1].

Because the connections are made as a researcher interacts with trusted research information systems, it both imbues verification into the connection and reduces the work needed for researchers and institutes to manage research information and research outputs.

3. ORCID Consortia

As a community-driven organization, ORCID has been in partnership with the international research community, from both a sector perspective collaborating with organizations such as World Data System and DataCite to a national perspective based on regions or countries, where ORCID services and resources can be applied in regional and national contexts. In the Asia-Pacific Region, there are 4 ORCID Consortia: Australia, Japan, New Zealand, and Taiwan [2]. They have joined ORCID's consortia program to take a coordinated national approach to building ORCID into their national research information services or platforms.

3.1 National Apporaches to ORCID in the Asia-Pacific (APAC) Region

Recognizing the importance, many countries have statements on adopting ORCID at a national level, including Australia and New Zealand. The former issued the Joint Statement of Principle: ORCID - Connecting Researchers and Research, with the funders releasing a joint statement on ORCID in which they encouraged all researchers applying for funding to have an ORCID iD. The latter developed the New Zealand ORCID Hub, which allows all New Zealand Consortium members to productively engage with ORCID regardless of technical resources.

4. Conclusions

Each sector in the research community has the power to improve the value and openness of its research information. This will require them to fulfil their roles and responsibilities by adopting PIDs to foster an open research infrastructure.

References

1. Haak Laurel L., Meadows Alice, Brown Josh, Using ORCID, DOI, and Other Open Identifiers in Research Evaluation, *Frontiers in Research Metrics and Analytics.*, 3, 28, 2018 https://doi.org/10.3389/frma.2018.00028
2. ORCID Consortia, https://orcid.org/content/orcid-consortia [accessed on: August 2020]

COVID-19 Data and Knowledge Hub

Liu Chuang[1*]

[1*] *Global Change Research Data Publishing & Repository and GCdataPR -WDS,*
No.11A, Datun Road, Chaoyang District, Beijing , 100101, China
Email: lchuang@igsnrr.ac.cn

Summary. On January 31, 2020, the World Health Organization (WHO) announced the COVID-19 as a Public Health Emergency of International Concern (PHEIC). The Consultant Committees on Communication and Information Technology (CCIT) and Life Sciences and Human Health (CCLH) of China Association for Sciences and Technology (CAST) for the United Nations jointly initiated COVID-19 Knowledge & Data Hub immediately. Professor GONG Ke, President of World Federation of Engineering Organizations (WFEO) and Chairman of CCIT leads the initiative.Member of the leading group includes Prof. LIANG Xiaofeng, Secretary-general of the CCLH, Mr. QIN Jiuyi, Deputy Director of Consultant Committee Office of CAST and Prof. LIU Chuang, member of CCIT. Prof. LIU Chuang was appointed to be the leader of the technical group and Professors. LIANG Xiaofeng and YI Heya, Secretary of the CCLH, as the co-leaders of the information resources group of the COVID-19 KD Hub. The Global Change Science Research Data Publishing & Repository (GCdataPR) (http://www.geodoi.ac.cn) hosts the online system of the COVID-19 Knowledge & Data Hub. The GCdataPR is jointly established in 2014 by the Institute of Geographic Sciences and Natural Resources Research of Chinese Academy of Sciences (IGSNRR/CAS) and the Geographical Society of China (GSC).The GCdataPR is one of Regular Members of the World Data System of the International Science Council (WDS/ISC) and China Data Publishing Center of Global Earth Observation Systems. The China Association for Science and Technology (CAST) is the largest non-governmental organization of scientific and technological professionals in China, serves as a bridge that links the Communist Party of China and the Chinese government to the country's science and technology community. Through its 210 national member societies and local branches all over the country, CAST maintains close ties with millions of Chinese scientists, engineers and other professionals working in the fields of science and technology. CAST attained UN Economic and Social Council (ECOSOC) consultative status in 2004 and took active part in UN activities since then.(http://english.cast.org.cn/).

Keywords.

Data analysis and intelligent application in modern agricultural technology extension

Chunjiang Zhao[1], ***Huarui Wu*** [1], ***Ying Deng*** [1*]

[1*] *National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China*
Email: yingdeng87@163.com

Summary. It is difficult to popularize agricultural technology by offline manpower because of the large population and wide distribution of farmers in China. (Chinese) National Engineering Research Center for Information Technology in Agriculture (NERCITA) has assisted the Ministry of Agriculture and Rural Affairs of the People's Republic of China (MoA) in developing the China Agro-Tech Extension Platform, The CATE, which is dedicated to assisting the implementation of agricultural technology promotion with information technology as the technical support. Through statistical analysis of the data such as agricultural conditions, user logs, Q&A, GPS tracking, Etc., CATE provides the basis for overall decision making for the management department. Throughout monitoring the onsite IoT data, CATE is able find out abnormalities and early warn the farm owners. Through natural language processing, the intelligent question and answer system of agricultural technology is trained by the users' Q&A data, so as to assist experts and agricultural technology promoters to solve farmers' production technology problems more quickly and effectively. According to the user's historical access path, analyse the user's personal interests and areas of expertise in agriculture to carry out personalized recommendation for both agricultural technology personnel and farmers.

Keywords. Agro-tech extension, agricultural condition analysis, NLP, Intelligent Q&A, personalized recommendation

1. Introduction

In order to provide Chinese farmers with higher quality and more immediate agricultural technical guidance services, the MoA has set up agricultural technology extension stations throughout the country to vigorously promote agricultural technology[1]. However, due to the large number and the distribution of a wide range of Chinese farmers, the offline artificial way of agricultural extension is work time-consuming and laborious. With the rapid development and popularization of the Internet, NERCITA has been assisting the MoA to develop the China Agri-Tech Extension Platform, The CATE, which is dedicated to assisting the agricultural technology promotion work with the support of information technology.

2. Core resources

Nearly all the 440,000 front-line agricultural technicians, 2,690 experts and more than 6 million farmers have registered.

CATE has collected a total of 10.35 million agricultural technical questions online so far, of which nearly 90% have been answered, with more than 47 million answers, of which 7.05 million have been answered by other users thumb up, 1.75 million have been adopted by the questioner, and 83.8% of the questions have been approved or responded by the questioner after being answered.

Totally more than 3.9 million valid agricultural condition messages and 14 million valid daily logs were uploaded. And more than 30,000 agricultural condition messages were posted, and 3.5 million kilometres service route GPS data was received from the agro-technicians each day.

3. Data Analysis

Through the analysis of farmers, agricultural experts and agricultural technicians' agricultural condition reporting and posted logs, CATE constructed a nationwide agricultural situation real-time monitoring network, providing data support for the detection, rapid early warning and processing the hot spots of agricultural condition.

The platform analyses the service of agro-technicians based on data such as online Q&A, historical interviews, GPS tracking, etc. So that the administrative department of agricultural technology extension can be aware of the daily service content and service coverage of each agricultural technician, so as to provide decision basis for assessing the performance of each agricultural technician and planning the future work content of agricultural technology extension system.

The platform is connected to onsite IoT devices of thousands of farms receiving the environmental data and real-time images so that CATE is able to monitor and early warn for the meteorology, moisture content, diseases and insect pests, and provides remote control interface of Internet of Things equipment to realize integrated remote management.

4. Make It Smarter

Previously, when a farmer came up with a question he may submit a new post rather than check up from the existing posts due to the low accuracy of the key words search algorithm. It imposed a great burden on the work of agricultural technician and also generated a lot of redundant data. In order to resolve the problem, CATE finds out three ways:

- A semi-supervised garbage text classification NLP model was built to efficiently and quickly remove data content irrelevant to agricultural technology. Then the selected valid Q&A text was iteratively input into the agricultural technology knowledge base.
- An intelligent question answering system of agricultural technology[2] was constructed. According to history dialogue, the model completes and does anaphora resolution to those queries with incomplete structure. And then the recall related knowledge from knowledge base using sematic searching model according to the query words list. Recalled knowledge text list will then be reordered by their relevance using deep learning model. Finally the optimized answer will be generated according to fixed threshold parameters. The intelligent Q&A system replies users' questions immediately and accurately, so that repeated questions and the work of agricultural technicians are reduced.
- The Personalized Recommendation system analyses users' recent browse history, abstracts related keywords and computes the sematic representation. It computes interestingness by users' visit frequency and keep-in-page-seconds along with the forgetting curve correction. Integrated with text similarity weights, the model outputs the ordered recommendation result list to users, and it

auto tunes itself refer to users' satisfaction feedback. Therefore, advanced technical contents are recommended and user experience is improved.

5. Conclusions

Through the analysis of agricultural situation logs and other data, a nationwide agricultural condition monitoring network has been built, and the analysis of agricultural technicians' working data assists administrators making management decisions. Through artificial intelligence methods, the junk filter, intelligent agro-technical Q&A system and personalized recommendation system are constructed, and makes up the real-time, accuracy and advance of the artificial service of CATE platform, so as to further improve the efficiency of agro-technical promotion service.

References

1. Research Group on Reform of Agricultural Technology Extension System in China. Current Situation, Problems and Countermeasures of Agricultural Technology Extension in China [J].Management World, 000(005):50-57,75, 2004, (In Chinese)
2. ZHANG Mingyue,WU Huarui,ZHU Huaji. Analysis of Extraction of Semantic Feature in Agricultural Question and Answer Based on Convolutional Model[J].Transactions of the Chinese Society for Agricultural Machinery, 49(12):203-210, 2018, (In Chinese)

DataCite Commons –
A new service to explore the PID Graph

Martin Fenner[1*]

[1*] *DataCite, Am Welfengarten 1B, Hannover, 30167, Germany*
Email: martin.fenner@datacite.org

Summary. In August 2020, DataCite launched DataCite Commons, an easy to use new discovery service for scholarly works, people and organizations, and their connections. This abstract provides background information about the new service.

Keywords. Persistent identifier, PID Graph, Connections, FREYA

1. Introduction

Persistent identifiers (PIDs) and standardized metadata describing the content identified by these PIDs are centrally important for the discovery, reuse and understanding of scholarly research. The identification and description of research data has been the focus of the non-profit membership organisation and DOI Registration Agency DataCite since its founding in 2009.

Metadata not only describe scholarly resources themselves, but they are also essential for unequi-vocally describing the connections between resources. Examples for these connections include citations, authorship and funding. Together these resources (nodes) and their connections (edges) form a graph. As PIDs are essential for building and exploring this graph, we call it the PID Graph(Fenner & Aryani, 2019) . One focus of the European Commission (EC)-funded FREYA project (*FREYA*, n.d.) that started in December 2017 and wraps up in November 2020 was to identify the most important use cases, agree on the technical architecture, and build production infrastructure to explore this PID Graph. This work builds on earlier work that includes the Research Data Alliance (RDA) Scholarly Link Exchange (Scholix) initiative (Aryani et al., 2017), the Make Data Count initiative (Fenner et al., 2018), work by the EC-funded THOR project (Fenner et al., 2016), and more.

2. User Stories

The driver for all technical work on the PID Graph are user stories, important problems in scholarly communication that can't be adequately addressed with existing infrastructure. One of the first activities of the FREYA project was therefore the collection, discussion and prioritization of these user stories, cumulating in an in-person workshop in August 2018. A total of 45 user stories have been initially documented in GitHub issues, and then made available in the PID Forum for community feedback (*The PID Forum*, n.d.). An example user story that has been documented in detail elsewhere is

As a researcher, I want to get a list of all my research outputs (publications, datasets, software, etc.) supported by grant funding, and how often they are cited, to demonstrate the impact of my work.

3. GraphQL API

The next step in PID Graph development was the exploration of existing PID infrastructure, and how this infrastructure needs to be adopted to address PID Graph user stories. By February 2019 it became clear that extending and connecting the existing PID provider REST APIs was not enough to address the PID Graph user stories in a scalable way. We therefore looked at other technology options, including SPARQL, and eventually picked GraphQL, a query language that is a good fit for the existing user stories, has become a widely adopted open source technology, and can work with existing PID provider backend services such as relational databases and Lucene-based search indexes (Solr, Elasticsearch). We launched a pre-release GraphQL API in May 2019, and a production version in May 2020 (Fenner, 2020a).

4. DataCite Commons

GraphQL greatly simplifies the backend services architecture powering the PID Graph but is an API that few users will directly interact with. We started writing Jupyter notebooks to interact with the GraphQL API, including 10 notebooks that directly address 10 PID Graph user stories, for example (Petryszak et al., 2020).

The next step was then to build a web service powered by the GraphQL API, and on August 27, 2020 the DataCite Commons service was launched for initial public feedback (Fenner, 2020b). The official launch will be in October 2020, as the EC-funded FREYA project is wrapping up.

DataCite Commons was built with React, the currently most popular Javascript framework that integrates seemlessly with GraphQL backend services. The new service provides access to 28 million scholarly resources described by PIDs and associated metadata (**Tab. 1**).

Table 1. Number of PIDs in DataCite Commons as of 6 Sep 2020.

	Identifier	**Count**
Works	DataCite	19,798,202
	Crossref	8,771,341
People	ORCID	9,622,911
Organizations	ROR	98,332
Total		38,290,786

DataCite Commons includes all PIDs from DataCite, ORCID and ROR. For Crossref it is a subset (7.51%) of the total number of DOIs (116,810,696 as of 6 Sep 2020). Many more Crossref DOIs will be imported into the service in the coming months, with the number of Crossref DOIs expected to exceed the number of DataCite DOIs before the end of 2020.

5. Conclusions

DataCite Commons wraps up the FREYA work on launching a production service to explore the PID Graph, but it also is the starting point for a totally new discovery platform for open science.

Acknowledgments. The work on DataCite Commons and the DataCite GraphQL API is supported by funding from the European Union's Horizon 2020 research and innovation program-me under grant agreement No 777523.

References

1. Aryani, A., Burton, A., Manghi, P., Bruzzo, S. L., Stocker, M., Schindler, U., Diepenbroek, M., Fenner, M., & Koers, H. (2017). *Scholix* Framework*: Building a Bridge Between Research Data and Publications*. 304151 Bytes. https://doi.org/10.4225/03/58CA546A18C50
2. Fenner, M. (2020a). *Powering the PID Graph: Announcing the DataCite GraphQL API*. https://doi.org/10.5438/YFCK-MV39
3. Fenner, M. (2020b). *DataCite Commons—Exploiting the Power of PIDs and the PID Graph*. https://doi.org/10.5438/F4DF-4817
4. Fenner, M., & Aryani, A. (2019). *Introducing the PID Graph*. https://doi.org/10.5438/JWVF-8A66
5. Fenner, M., Demeranville, T., Kotarski, R., Dasler, R., *M0cEntyre*, J., de Mello, G., Vision, T., Dappert, A., & Farquhar, A. (2016). *Thor: Conceptual Model Of Persistent Identifier Linking*. Zenodo. https://doi.org/10.5281/ZENODO.48705
6. Fenner, M., Lowenberg, D., Jones, M., Needham, P., Vieglais, D., Abrams, S., Cruse, P., & Chodacki, J. (2018). *Code of practice for research data usage metrics release 1* (e26505v1). PeerJ Inc. https://doi.org/10.7287/peerj.preprints.26505v1
7. *FREYA*. (n.d.). Retrieved 6 September 2020, from https://www.project-freya.eu/en
8. Petryszak, R., Fenner, M., Lambert, S., Llinares, M. B., & Madden, F. (2020). *FREYA PID Graph—Grant Outputs* (1.1.1) [Computer software]. DataCite. https://doi.org/10.14454/QAYM-KT26
9. *The PID Forum*. (n.d.). The PID Forum. Retrieved 6 September 2020, from https://www.pidforum.org/

Activities towards Open Data and Research Data Management at NIES/CGER

Yoko Fukuda[1*], ***Tomoko Shirai***[1]

[1*] *National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki, 305-8506, Japan*
Email: fukuda.yoko@nies.go.jp

Summary. The Center for Global Environmental Research (CGER), National Institute for Environmental Studies (NIES), has been managing and operating the Global Environmental Database (GED) since 2014 as a platform for publication and search of data, focusing on global environmental research. As a data management platform, the NIES/CGER started designing and developing the research data management system (RDMS) in 2018. In this presentation, we introduce activities toward open data and the RDMS at NIES/CGER.

Keywords. Open data, Open science, Research data management

1. Introduction

The Center for Global Environmental Research (CGER), National Institute for Environmental Studies (NIES) conducts research on global environmental issues with particular emphasis on climate change as the core organization of the NIES. The CGER keeps intensive monitoring networks of the atmosphere, ocean, forest ecosystem in the Asia, Oceania, and Pacific regions with international cooperation. The CGER also supports to develop databases for a wide range of fields such as water environment and biosphere in addition to the above fields.

2. Global Environmental Database

To disseminate these data, the CGER constructed Global Environmental Database (GED) in May 2014 as a data publication platform of the CGER. The GED mainly provides research data, real-time data, and analysis support tools. The CGER has been supporting DOI minting to research data in response to the request from researchers since 2016.

3. Research Data Management System for CGER

The CGER started designing of the research data management system (CGER RDMS) in April 2018. The CGER RDMS mainly supports the following works: creating metadata, Licensing, DOI minting, and management of data and supplementary material. The CGER RMDS also enables collaboration of data management with a team regardless of the open/close status of data. Research data and metadata registered in the CGER RDMS can be promptly released from the GED. Research data also can be visualized in Quick Plot which has been developed for both the RDMS and the GED.

In order to collect opinions about what NIES staffs want for the RDMS, test trials of the CGER RDMS were conducted in August 2020. During the last half of FY 2020, the RDMS will be improved based on a user's perspective, and the operation of the CGER RDMS will be partially started in or after FY 2021.

4. Information service of Asia-Pacific Monitoring sites

To promote the effective utilization of observational data on greenhouse gases and air pollutants in the Asia-Pacific region, a website "Asia-Pacific Monitoring Sites" was released in February 2018. This website provides basic information on the observation which has been conducted mainly by NIES such as the location of the observation site and the observation parameters. At the moment, there is no system to download data files directory from this site. To improve the website so that it is easier to use data, further consideration will be needed.

5. Challenges

The author joined NIES after getting a phD in the field of earth science. The experience in observations and data analysis facilitates supporting data activities, especially metadata creation and the RDMS designing from both perspectives of a data user and a researcher. Meanwhile, there are many situations where the expert knowledge and IT skills are lacking. Nevertheless, my works are definitely very rewarding since many NIES researchers show a strong interest in research data management. Remaining concerns are lack of clear career paths and role models. The network of persons who are working on the open science is important.

Japanese DOI RA, Japan Link Center (JaLC) and its collaboration with ORCID

Masashi Hara[1*]

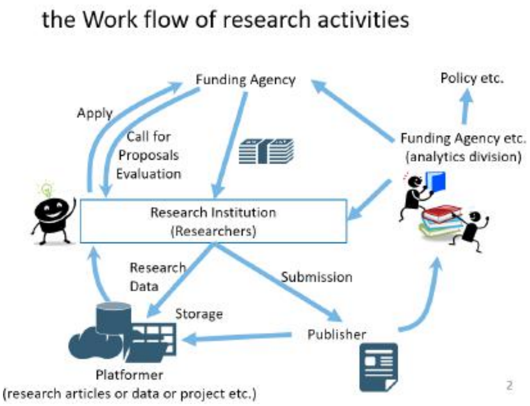
[1*] *Japan Science and Technology Agency, Tokyo, 102-8666, Japan*
Email: mhara@jst.go.jp

Summary. Digital Object Identifier (DOI) is becoming more common recently and is used for not only journal articles but also for research data, research report, e-learning and so on. Japan Link Center (JaLC) started in 2012 with the authorization by the DOI foundation and is a Registration Agency in Japan. JaLC takes a role to register DOIs to academic contents which are produced in Japan and already has registered more than 6 million DOIs. In 2020 JaLC has provided a function in which the authors could link their research outputs to ORCID my-page. JaLC will continue to offer new functions to provide a more efficient academic ecosystem in Japan.

Keywords. DOI, ORCID, Japan Link Center (JaLC), Registration Agency (RA), Persistent Identifier (PID)

1. Introduction

A researcher does research, submits new papers to publishers, applies new research grants and new positions, and reports his/her outputs to FA and/or belonging organizations. In order to do so, it is important for stakeholders like evaluators, reviewers, policy makers to access same unique information. So, the concept of Persistent Identifier (PID) is developed and for example, DOI for articles and ORCID for researchers become common nowadays.



This presentation will include what DOI is, what a registration agency is in Japan, the role of Japan Link Center (JaLC) and its newly function which is linked to ORCID.

2. DOI and Japan Link Center

DOI (Digital Object Identifier) is one of the Persistent Identifiers (PIDs) and registered to approx. 220 million contents as of March 2020. The number of DOIs exceeded 50 million in 2011, 100 million in 2015, and 200 million in 2019, doubled in about 4 years. The number of DOI resolution is 500 million to 600 million every month. These figures show that DOI system is a -fundamental infrastructure.

The International DOI foundation (DOI foundation) was founded in 1997 to develop and manage DOI systems. The DOI foundation and Corporation for National Research Initiatives (CNRI) established DOI system, using the handle system CNRI had developed. After that, DOI became an international standard ISO2632:2012 in 2012.

The DOI operation is governed by the DOI foundation and the DOI Registration Agencies

(RAs). The DOI foundation makes policies and manages RAs. At present, there are 11 RAs, they, register DOIs, assign Prefix and keep the DOI registration. Statistics show that110 million DOIs out of the total 220 million are from the RA Crossref (52%), with 33 million from ISTIC (15%), 21 million from Datacite (10%) andJaLC covers 6 million DOIs (3%).

The JaLC was authorized as a RA in March 2012 by the DOI foundation. The JaLC is governed by the Japan Science and Technology Agency (JST), the National Institute for Materials Science (NIMS), the National Institute of Informatics (NII) and the National Diet Library (NDL) and operated by JST. The purpose of JaLC is to collect the academic contents created in Japan, to promote the usage of DOI in Japan and to bringbetter access to the Japanese research outputs for oversea users JaLC covers DOI registration for the journal articles, books, research reports, research data, e-learning and so on.

3. New Function with ORCID

JaLC released Auto-Update and Search & Link function integrated with ORCID. Nowadays it is becoming more common that researchers manage their research outputs in ORCID and using ORCID in publishing workflow during or during grant proposal submission. These new functions make it possible for researchers to connect their contents which are in JaLC to researchers' ORCID my-page. It enables correct information with ease so we expect this to reduce researchers' burden on data management.

If a researcher sets the Auto-Update function as "on" in advance, as long asthe researcher's authenticated ORCID iD is a part of DOI registration metadata of his/her research outputs, these outputs will be added to the researcher's ORCID record automatically. To do so, the ORCID ID should be included in the JaLC metadata.

In the case of Search & Link function, a researcher searches her/his research outputs in the "JaLC contents search page" and then these outputs can be connected to the researcher's ORCID record one by one. This function does not need the ORCID ID to be included in the JaLC metadata.

JaLC released these functions this spring and with a few months passing by,the login from ORCID web site has counted 10,000 times and the Search & Link function is used about 2,000 times per month. However, the Auto-Update function usage is unfortunately still low.

4. Conclusions

JaLC is willing to bring more services utilizing DOIs to reduce researcher's burden and to bring a better research environment. Next year we are planning to start to provide abstracts which are in the metadata. Furthermore, JaLC is trying to collaborate with more international partners.

References

1. Hara, M., Sato, R., Mimura, N., DOI and JaLC activities. *Journal of Information Science and Technology Association*, vol. 70, 432, 2020
2. Japan Link Center, https://doi.org/10.11502/JaLC_policy [accessed on: August 2020]

Open Science and its Policy towards Open Science Paradigm

Kazuhiro Hayashi[1*]

[1*] *National Institute of Science and Technology Policy, 3-2-2 Kasumigaseki, Chiyoda, Tokyo, 100-0013, Japan*
Email: khayashi@nistep.go.jp

Summary. Open Science is a movement to transform Science, Society, and "Science and Society" with opening up knowledge drastically by the advancement of ICT. Starting from e-journals and Open Access, scholarly communication has been transformed and currently preprint and preprint server is being focused on the COVID-19. Research Data has also been a potential media for scholarly communication, still it needs incentives for researchers. Especially, it needs some arrangement of regulation which allows data providers and users to handle data with confidence. With the acceleration by COVID-19, we are recognizing more and more of the gap between current system and what we are aiming, which is a signal to leap towards Open Science paradigm. From policy perspectives, the essence of Open Science is re-organization of the system and regulations for Science and Society, which is fundamentally for fostering a new culture but with more concrete strategy to mitigate the gap.

Keywords. Open Science, Research Data, regulation, social change, re-organizing

1. Introduction

Open Science has been a hot topic among Science and Technology Policy. In addition to OECD, and G7, UNESCO[1] is now working on for a recommendation. Open Science does not have a sole definition but, most generally, Open Science is a movement to transform Science, Society, and "Science and Society" with opening up knowledge drastically by the advancement of ICT. [2] Under this explanation, we can include most of initiatives related to Open Science.

2. Current evolution and its acceleration by COVID-19

Starting from e-journals and Open Access, scholarly communication has been transformed towards digital native paradigm. E-journals with subscription model and OA model consists an inevitable infrastructure for researchers. Nevertheless, it is still based on the system of peer-review journals. In 2020, preprint and preprint server is being focused on the COVID-19. Preprint provides immediate scientific information, but it casts the basic problem of quality control, and ironically, it also make us sceptical to peer-review system itself with some incidents represented by some papers' retraction. On the other hand, we found that accumulating preprints regarding COVID-19 provides a new and rapid landscape of research without "papers and citations."[3] COVID-19 is revealing something that has been existed but had not been seen.

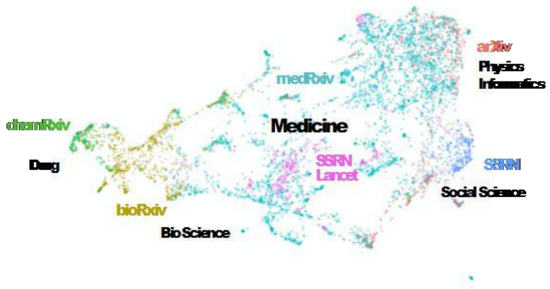


Figure 1. Preprint landscape of COVID-19 with its natural language processing

Research Data has also been a potential media for scholarly communication, which is a central pillar of Open Science Policy of the most countries that are challenging.

3. What policy has done and what it reveals

As of 2020, many countries have Open Access Policy with publicly-funded researches. Funding agency promotes/requires/mandates Open Access to their research articles. However, progress of Open Access provides other issues represented by APC (Article Processing Charge) with embracement of OA by commercial publisher as a business. It seems that whole culture based on peer-review journals have not changed so much.

Regarding Open Research Data, some countries are challenging to share their research data with FAIR Data Principles or the philosophy of "as Open as possible, as Close as necessary." For example, in Japan, the 5th Science and Technology Basic Plan (2016-2020) has promoted Open Science in a general manner followed by the Integrated Innovation Strategy which has specific measures and clear goals focusing on the research data utilization.[4] It supports Data Policy, DMP, Data Infrastructure, Data Repository and Data publication together with pilot programs and monitoring.(Figure 2)

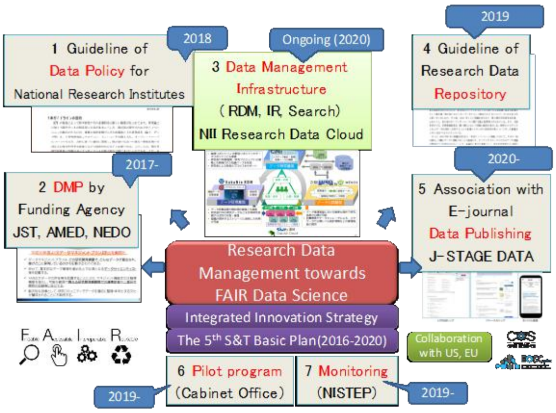

2018
1 Guideline of Data Policy for National Research Institutes
Ongoing (2020)
3 Data Management Infrastructure (RDM, IR, Search) NII Research Data Cloud
2019
4 Guideline of Research Data Repository
2017-
2 DMP by Funding Agency JST, AMED, NEDO
2020-
5 Association with E-journal Data Publishing J-STAGE DATA
Research Data Management towards FAIR Data Science
Integrated Innovation Strategy
The 5th S&T Basic Plan(2016-2020)
Collaboration with US, EU
6 Pilot program (Cabinet Office)
2019-
7 Monitoring (NISTEP)
2019-

Figure 2. Open Science Policy in Japan 2020

Despite of Research Data's potential and its proactive policy development, Open Science Policy and its implementation found that it needed much incentives for researchers and other actors. Especially, it needs some arrangement of regulation which allows data providers and users to handle data with confidence. Science Council of Japan released a recommendation for Open Science and its first pillar of recommendation is "The need for rulemaking in an era where data plays a central role."[5]

With our experiences towards research data sharing and also with its unexpected acceleration by COVID-19, we are recognizing more and more of the gap between our current system and what we are aiming. We've knew it from outset, but in the end, incremental reform is not going to work, which is a signal to leap towards Open Science paradigm.

4. Conclusions

From policy perspectives, the essence of Open Science is re-organization of the system and regulations for Science and Society, which is fundamentally for fostering a new culture but with more concrete strategy with practices to mitigate the gap we have observed or to solve problems for the leap.

References

1. UNESCO Open Science https://en.unesco.org/science-sustainable-future/open-science [accessed on: Sep 2020]
2. Kazuhiro Hayashi. Progress of Open Science and Transforming Citizen Science to Co-creative Research. Trends in Science (Gakujutu no Doko), Vol. 23(11), PP. 11_12-11_29: https://doi.org/10.5363/tits.23.11_12 (Japanese), 2018
3. Hitoshi Koshiba, Kazuhiro Hayashi, Yuko Ito. A Trial of early detection system for research trends through the preprints data — Research status around COVID-19 / SARS-CoV-2. NISTEP Discussion Paper. No 186: http://doi.org/10.15108/dp186
4. Kazuhiro Hayashi. Perspective of Open Science in Japan driven by Integrated Innovation Strategy. International Workshop on Data Science-Present & Future of Open Data & Open Science–:https://ds.rois.ac.jp/wp-content/uploads/2018/10/Abstract_booklet_dsws_2018.pdf [accessed on: Sep 2020]
5. Toward deepening and promoting Open Science: http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-24-t291-1.pdf (Japanese) [accessed on: Sep 2020]

Wastewater-based epidemiology for COVID-19: Perspectives for environmental surveillance of SARS-CoV-2 in wastewater

Ryo Honda[1*], ***Akihiko Hata***[2]

[1*] *Faculty of Geosciences and Civil Engineering, Kanzawa University,*
Kakuma-machi, Kanazawa 920-1192, Japan
[2] *Department of Environmental and Civil Engineering, Toyama Prefectural University,*
5180 Kurokawa, Imizu, Toyama 939-0398, Japan
Email: rhonda@se.kanazawa-u.ac.jp

Summary. Wastewater-based epidemiology (WBE) by monitoring of SARS-CoV-2 RNA in wastewater is expected to enable early warning and prediction of overall status of COVID-19 outbreaks. Detection of SARS-CoV-2 RNA in wastewater and its correlation with the number of clinically reported COVID-19 cases has been reported in several countries. However, the number of clinically reported cases are not static but affected by situations of clinical surveillance because, for example, it does not include patients who do not visit clinics due to asymptomatic and mildly symptomatic infections. Therefore, the advantage of WBE for COVID-19 could be capability of tracking the trend of outbreaks and convergences independent of such clinical situations.

Keywords. SARS-CoV-2 RNA; fecal shedding; wastewater monitoring

1. Introduction

Wastewater-based epidemiology (WBE) is surveillance of epidemiological information of people in a sewer catchment by monitoring of wastewater. WBE is expected to be an effective tool for early warning and prediction of overall status of COVID-19 outbreaks. The potential advantage of WBE is capability of surveying the entire catchment by testing a single wastewater sample, while clinical surveillance needs a large number of samples collected from individuals. SARS-CoV-2 RNA is shed in feces of COVID-19 patients and therefore can be present in wastewater. Detection of SARS-CoV-2 RNA in wastewater has been reported in many countries [1]. However, there are still needs of data accumulation and development of a prediction model to interpret the RNA concentration into epidemic situation. This talk aims to provide the up-to-date situation of research progress on WBE for COVID-19 and discuss its future perspectives.

2. SARS-CoV-2 RNA in wastewater

Detection of SARS-CoV-2 RNA in wastewater were reported in Australia, India, Italy, Japan, Spain, Netherlands, USA, etc [1,2]. Typically, 10-100 mL of untreated wastewater sample is required for detection. The viruses in a wastewater sample was first concentrated and its RNA is extracted. The viral RNA is usually concentrated up to approximately 100 times during this process. The extracted RNA is detected by RT-qPCR assay, which mostly same as clinical tests. There are several options in concentration methods and primer-probe sets for PCR assay. Although currently selection of these methods is likely to depend on preferences of each research group, the efficiency and sensitivity of these methods are being evaluated. The detected concentrations of viral RNA in wastewater reportedly correlates with the number of COVID-19 cases reported in clinical surveillance [3]. Some preprint studies also reported the viral RNA in wastewater was detected 1-week earlier than increase of clinical reported cases [4]. Therefore, it is expected to be applicable as an early warning tool or surveillance of catchment-scale outbreaks.

3. Principle of WBE and uncertainty

Since viral RNA shedding in feces ranged 10^4-10^6 copies/g, viral RNA could be detected in wastewater when no. of COVID-19 infections in a sewer catchment reaches 5-10 cases per 100,000 [5]. Meanwhile, viral RNA concentrations detected in wastewater ranged 10^4-10^6 copies/L. It is often detected at a very low concentration of 10^4 copies/L level, which is slightly above the typical detection limit. In such a low concentration, the detection of SARS-CoV-2 RNA depends not only on the lower limit of detectable concentration but also probability of successful transfer of viral RNA into PCR assay preparation. The number of the clinically reported COVID-19 cases is the most applicable indicator to verify the prediction of COVID-19 outbreak situation by WBE. However, the number of clinically reported cases are not static but affected by situations of clinical surveillance (**Figure 1**). For example, asymptomatic and mildly symptomatic patients may not visit clinics and not be identified by clinical surveillance. The number of the reported cases monotonically increase independent of the epidemic situation when more people are tested by clinical PCR assays. Therefore, WBE dataset obtained as RNA concentration in wastewater is unlikely to simply correlate with the clinically reported epidemic situations. Instead, its advantage is to be able to track the trend of outbreaks and convergences independent of such clinical situations. Therefore, it could be useful to predict the outbreak/convergence from people's mobility data, and to evaluate the effect of lockdown and various countermeasures.

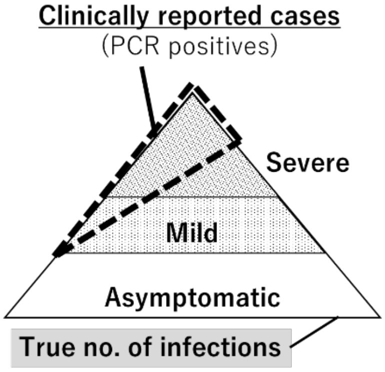


Figure 1 A part of true number of infections are apparent as clinically reported cases.

4. Conclusions and future perspectives

WBE for COVID-19 by environmental surveillance of SARS-CoV-2 in wastewater could be effective tools for early warning and prediction of COVID-19 outbreaks and convergence. However, its limitations and uncertainty are still under investigation. Data accumulation and statistical analysis of WBE data with clinical epidemiological data, people's mobility data, and possible factors which affect spread and convergence of COVID-19 outbreaks are expected as further studies.

References

1. Michael-Kordatou, I., Karaolia, P. & Fatta-Kassinos, D. Sewage analysis as a tool for the COVID-19 pandemic response and management: the urgent need for optimised protocols for SARS-CoV-2 detection and quantification. *J. Environ. Chem. Eng.* **8**, 104306, 2020

2. Hata, A., Honda, R., Hara-Yamamura & Meuchi, Y. Detection of SARS-CoV-2 in wastewater in Japan by multiple molecular assays-implication for wastewater-based epidemiology (WBE). *medRxiv* 2020.06.09.20126417, 2020, doi:10.1101/2020.06.09.20126417
3. Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in The Netherlands. *Environ. Sci. Technol. Lett.* **7**, 511–516, 2020
4. Peccia, J. *et al.* SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics. *medRxiv*, 2020, doi:10.1101/2020.05.19.20105999
5. Hata, A. & Honda, R. Potential Sensitivity of Wastewater Monitoring for SARS-CoV-2: Comparison with Norovirus Cases. *Environ. Sci. Technol.* **54**, 6451–6452, 2020

The Roles of PIDs in Scientific Data Management in China

Xiaoyan Hu[1*], **_Ziming Zou_**[1], **_Lianglin Hu_**[2], **_Jia Liu_**[2], **_Xin Chen_**[2], **_Qi Xu_**[1]

[1*] *National Space Science Center, Chinese Academy of Sciences, NO.1 Nanertiao, Zhongguancun, Haidian district, Beijing, 100190, China*
[2] *Computer Network Information Center, Chinese Academy of Sciences, NO.4 Nansijie, Zhongguancun, Haidian District, Beijing, 100190, China*
Email: huxiaoyan@nssc.ac.cn

Summary. Persistent identifiers (PIDs) play an important role in scientific data management, especially in data publication and citation to support the findability and accessibility of scientific data. Data organizations in China are increasingly aware of PIDs as required for implementation of the FAIR principles, and several good practices have been developed. It is worth noting that China has also established a persistent identifier, namely the China Science & Technology Resource identifier (CSTR). This paper presents a typical case of National Space Science Data Center (NSSDC) embedding four types of PIDs into its workflows, exemplifying the application of PIDs in scientific data management.

Keywords. FAIR data, data ecosystem, data citation, metadata, persistent identifiers

1. Introduction

Persistent identifiers could benefit research in many ways, not just for clear tracking of authors and published results [1], but also for more automated, higher quality scientific data management. The FAIR principles are the most widely accepted theoretical model for data curation and management. Properly managed PIDs with their associated metadata are the primary access point to enable Findability in the FAIR data ecosystem, and to provide a basis for all the other FAIR attributes [2].

2. Global Persistent Identifiers

The most commonly used PIDs in Chinese data repositories and research communities are some general identifiers, e.g. Digital Object Identifier (DOI), Open Researcher and Contributor Identifier (ORCID), and some domain-oriented identifiers.

2.1 DOI

DOI is a well-known digital identifier for objects of any type [3]. More familiar to researchers, DOIs are widely used in academic publication and data publication. Some global platforms such as DataCite are based on DOI for dataset registration and discovery.

2.2 ORCID

ORCID is a persistent digital identifier that distinguishes one researcher from other researchers and a record that supports automatic links among all his professional activities [4]. ORCIDs are currently used to identify specific individuals of data producers, depositors and data users, enabling data provenance and tracking of data usage.

2.3 Domain-oriented identifiers

There are also some domain-oriented PIDs, such as IVOID, SPASE ResourceID, ATCC, etc. Metadata of domain-oriented identifiers prefer greater granularity or richness. These identifiers and associated metadata not only support data findability and accessibility, but also address domain-specific demands for data understanding and data interoperability.

3. CSTR

CSTR is proposed by the Ministry of Science and Technology (MOST) of China for the identification, cataloguing, registration, publication, maintenance and management of all kinds of scientific and technological resources in China [5].

A CSTR name in turn consist of a uniform prefix, a registration authority code, a resource type code and a suffix. The registration authority code indicates the registrant who issues the CSTR. The resource type code identifies 23 types of S&T resources, including scientific equipment, major S&T infrastructure, specimen, scientific data, paper, etc. The suffix is assigned by the registration authority, using a combination of letters, numbers and delimiters of indeterminate length.

An example of a CSTR would be:

CSTR:14800.11.2012-100101-13215-00001-V1

4. NSSDC Practice

Four types of PIDs are embedded in the workflows of NSSDC, which are DOI, CSTR, Handle and ORCID. To support these PIDs, NSSDC self-formulated dataset core metadata covers all elements of DOI, CSTR and Handle metadata, thus all the information required for registration of these PIDs can be derived from this core metadata.

In the phase of deposition, submitting adequate metadata information and standardized introduction document is necessary, and filling in the ORCIDs of data authors is strongly recommended. Then NSSDC will assign a unique internal identifier for the qualified dataset, which will be used throughout the full subsequent process of data management and sharing. When the dataset is released, this internal identifier will be used as the suffix to generate the full PID name, which will be registered via the API interface provided by the service agencies, e.g. CNKI DOI and DataPid. Based on these PIDs NSSDC have published data catalogues on some integrated data platform. Meanwhile, these widely accepted PIDs support normative citation of NSSDC dataset.

5. Conclusions

Data practitioners in China have realized the importance of PIDs in achieving data FAIR. Some best practices have yielded good results, providing a basis for building a healthy research ecosystem.

Acknowledgments. This work was supported in part by the 13th Five-year Informatization Plan of Chinese Academy of Sciences，Grant No. XXH13505 and No. XXH13514.

References

1. Meadows A., DOIs and other persistent identifiers have much more to offer science[J]. *Nature*, 558(7710), 372-372, 2018
2. Juty, N., Wimalaratne, S. M., Soiland-Reyes, S., Kunze, J., Goble, C. A., & Clark, T, Unique, Persistent, Resolvable: Identifiers as the foundation of FAIR. *Data Intelligence*, 2:1-2, 30-39, 2020
3. International DOI Foundation, https://www.doi.org/doi_handbook/1_Introduction.html [accessed on: October, 2015]
4. ORCID, https://support.orcid.org/hc/en-us/articles/360006973993-What-is-ORCID- [accessed on: August, 2020]
5. China National Technical Committee of Standardization for Science & Technology Infrastructure, *Science and technology resource identification*: GB/T32843—2016[S]. China Standards Press, Beijing, 2016 (in Chinese)

Time-series analysis of directional sequence changes in SARS-CoV-2 genomes and an unsupervised explainable AI for studying the corona virus genomes

Toshimichi Ikemura [1,2*], ***Kennosuke Wada***[1], ***Yoshiko Wada***[1], ***Yuki Iwasaki***[1], ***Yasushi Hiromi***[2], ***Takashi Abe***[3]

[1*] *Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga-ken 526-0829, Japan*
[2] *National Institute of Genetics, Mishima, Shizuoka-ken 411-8540, Japan*
[3] *Faculty of Engineering, Niigata University, Niigata-ken 950-2181, Japan*
Email: t_ikemura@nagahama-i-bio.ac.jp

Summary. This study conducted time-series analysis of mono- and dinucleotide compositions for SARS-CoV-2 genomes and found clear time-series changes in these compositions on a monthly basis. We then developed a sequence alignment-free method that extensively searches for advantageous mutations and ranks them in an increase level of their population frequency. Unsupervised machine learning is highly desirable for datamining of big data, and we conducted BLSOM (batch-learning SOM) of the virus genomes, using the oligonucleotide composition in each genome and found the obtained clustering (self-organization) of the genomes is primarily related to known clades. Since BLSOM is an explainable AI, it can tell us what features of the oligonucleotide composition are responsible for the clade separation.

Keywords. COVID-19, big data, zoonotic virus, oligonucleotide, advantageous mutation

1. Introduction

Many host factors are involved in viral growth, and human cells may not present an ideal growth condition for zoonotic viruses (e.g., SARS-CoV-2) that have invaded from nonhuman hosts. Therefore, efficient growth after the invasion will require changes in virus genomes. To study the viral adaptation, we analysed time-series changes in mono- and oligonucleotide compositions for zoonotic RNA viruses, including SARS-CoV-2, and found time-series directional changes that are detectable even on a monthly basis [1-5].

Unsupervised machine learning can discover new knowledge from big data without prior knowledge or particular models and is highly desirable for datamining of big data. We previously established a batch-learning self-organizing map (BLSOM) for oligonucleotide compositions, which can reveal various new characteristics of genome sequences [6]. The present BLSOM study analysed over 40,000 genomes of SARS-CoV-2, which were isolated from December 2019 to June 2020.

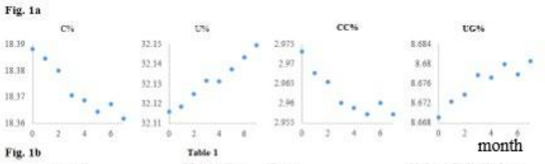
2. Results

2.1 Time-series changes in mono- and dinucleotide composition

In the time-series analysis in Fig. 1a, we tabulated virus strains for each collection month and calculated the mono- and dinucleotide composition (%). The monthly results clearly show monotonic time-series increase/decrease in a part of these nucleotides; for details, see [5].

2.2 Rapidly increasing 20-mers and search for advantageous mutations

For the highly mutable RNA viruses, the time-series analysis of long oligonucleotides such as 20-mers is important for designing PCR primers and therapeutic oligonucleotides that can be used for a sufficiently long period [4]. Since the polyA-tails were removed, all 20-mers and almost all 15-mers are present only once on each genome, and thus their occurrence level (%) in each month's population corresponds to the occurrence level (%) of the strains with the oligomer sequence in the viral population. When we happened to search for 15 and 20-mers that were absent in all strains isolated in December 2019, then emerged and increased in occurrence, we unexpectedly found a group of rapidly increasing 15 and 20-mers. When analysing their time-series occurrence frequency, a very similar pattern of monotonic increase was observed for sixty 15-mers (Fig. 1b), as well as eighty 20-mers. Figure 1c presents mutations responsible for the rapid increasing oligomers; for details, see [5].



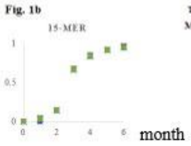
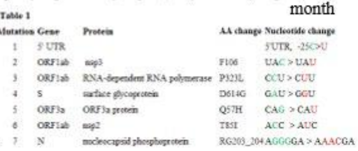


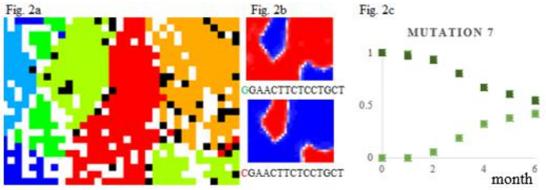
Table 1



Mutation	Gene	Protein	AA change	Nucleotide change
1	5' UTR			5'UTR, -25C>U
2	ORF1ab	nsp3	F106	UAC > UAU
3	ORF1ab	RNA-dependent RNA polymerase	P323L	CCU > CUU
4	S	surface glycoprotein	D614G	GAU > GGU
5	ORF3a	ORF3a protein	Q57H	CAG > CAU
6	ORF1ab	nsp2	T85I	ACC > AUC
7	N	nucleocapsid phosphoprotein	RG203_204	AGGGGA > AAACGA

2.3 Oligonucleotides BLSOM

We next conducted BLSOMs with compositions of short (2~7-mer) and long (e.g., 15-mer) oligomers in the over forty thousands of SARS-CoV-2 genomes; importantly, only the oligomer composition is given in the learning process. To see if the clustering (self-organization) by the BLSOM is related to known clads, grid points containing genomes of a single clade are coloured to indicate each clade, and grid points containing genomes of multiple clades are displayed in black. BLSOMs for 4 ~ 7-mers and 15-mers show a good separation (self-organisation) by clade. Fig. 2a shows the BLSOM of 15-mers, which are rapidly increasing or decreasing in their monthly population. BLSOM can provide information about oligomers responsible for the clustering (self-organisation) according to the clade; the contribution level of each 15-mer in each node can be visualized by a heatmap diagram: high (red), moderate (white) and low (blue). There were many cases in which occurrence levels are reversed for a pair of 15-mers in which only one base has changed. Figure 2b shows the case where the respective mutation corresponds to the No.7 mutation in Table 1 and is responsible for clustering of GR genomes (bright green in Fig. 2a). The increase/decrease pattern of the respective 15-mers is presented in Fig. 2c.



3. Conclusions

The time-series analysis of oligonucleotide composition and an unsupervised-type AI (BLSOM) 0f the composition, which are not dependent on sequence alignment, are shown to be powerful methods for obtaining novel knowledge of viral genome sequences and molecular evolutionary processes.

Acknowledgments. This work was supported by a Grant-in-Aid for Scientific Research (18K07151) from JSPS and by COVID-19 Counterplan Research Project (supervised by Prof. Tatsumi Hirata, NIG) from the Research

Organization of Information and Systems (ROIS).

References

1. Iwasaki Y., et all., Prediction of directional changes of influenza A virus genome sequences with emphasis on pandemic H1N1/09 as a model case. DNA Res. 18, 125-136, 2011
2. Iwasaki Y., et all., Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains. BMC Infect. Dis. 13, 386, 2013
3. Wada Y., et all., Directional and reoccurring sequence change in zoonotic RNA virus genomes visualized by time-series word count. Sci. Rep. 6, 36197, 2016
4. Wada K., et all., Time-series oligonucleotide count to assign antiviral siRNAs with long utility fit in the big data era. Gene Ther. 24, 668–673, 2017
5. Wada K., et all., Time-series analyses of directional sequence changes in SARS-CoV-2 genomes and an efficient search method for candidates for advantageous mutations for growth in human cells. Gene:X ; doi.org/10.1016/j.gene.2020.100038, 2020
6. Abe, T., et all., Informatics for unveiling hidden genome signatures. Genome Res. 13, 693-702, 2003

Collection, Archive and Sharing Space Weather Information in NICT

M. Ishii[1]*, Y. Kubo[1], S. Sakaguchi[1], D. Shiota[1], M. Den[1], M. Nishioka[1], H. Jin[1], H. Ishibashi[1], K. Marubashi[1], K. Fukunaga[1]

[1]* *National Institute of Information and Communications Technology (NICT),*
4-2-1 Nukui-kita, Koganei, Tokyo, 184-8795, Japan
Email: mishii@nict.go.jp

Summary. We, NICT, has a long history to observe the sun and ionosphere, and the amount of observational and analysis data has become huge. Some parts of the database are obtained during high solar activity period

and include precious information for discussing social impact with extreme events, which is difficult to get in low solar activity periods, such as cycle 23 and 24. Over the past decade, we have been digitizing ionogram data recorded in microfilms which is easy to lose in the long days. Recently, we add the action of data rescue for optical and radio solar activity data observed at Hiraiso observatory. In addition, we began to digitize the data named "Solar Activity Chart" in which the real time data of solar disk, geomagnetic field, radio black out and cosmic ray with every cycle of 27 days, and a part of data made possible to show in the website freely. Another remarkable topic is to provide space weather information to ICAO as a part of global centers since

Nov. 7, 2019. In the present status, three global centers use their own dataset independently, but now we discuss to unify the dataset and evaluate prediction models for providing with the same quality. We will present the detail of present status and issues to be solved.

Keywords. Space Weather, legend data retrieve, real time data sharing, ICAO

1. Introduction

Space weather is electro-magnetic phenomena near the Earth occurred by solar activities which affects on radio utility, Global network of satellite system, electric power grid, etc. NICT has a long history for space weather monitoring and forecasting [1]. We have been observing ionosphere not only in Japan but also in Asian region since 1930s, publishing radio alert just after the world war II for stable radio use and started operational space weather forecast service in 1988 as a member of International Space Environment Services (ISES)[2].

Since the beginning of 21st century, International Civil Aviation Organization (ICAO) has been discussing the use of space weather information in civil aviation. In addition to the preparation of regulations, we worked for the selection of organization who provide space weather information to ICAO. As a result of multiple audits, three global canters, US PECASUS (Consortium of Finland (leader), UK, Austria, Belgium, Cyprus, Germany, Italy, Netherland and Poland) and ACFJ (Consortium of Australia, Canada, France and Japan) were selected in 2018 and the operational service started in 2019.

2. Digitization of Legend data

We started ionospheric observation in 1924 temporality and 1934 operationally. Because the ionospheric information was important in military mission, many ionospheric observatories (14 sites in operational, 15 sites in plan) were operated in Asian region during World War II. After the WWII, we started operational ionospheric measurement in Wakkanai, Fukaura, Shibata, Kokubunji and Yamagawa since 1946-47. In present, we operate four stations, Wakkanai/Sarobetsu, Kokubunji, Yamagawa and Ogimi.

We had operated solar radio and optical measurement for monitoring the solar activities in Hiraiso observatory. The solar radio telescope has been launched in 1962 for measuring 200MHz solar radiation. We added antennas for expanding measuring bandwidth from 100MHz to 9.5GHz. With closing Hiraiso Observatory in 2016, The function of solar radio measurement moved to Yamagawa observatory. The optical measurement of H alpha started in 1986 in Hiraiso observatory. The system improved in 1994 and continued the observation until closing.

There are various kinds of data for radio alert/space weather forecast, e.g., solar surface, geomagnetic field, radio black out, cosmic ray, and upstream events effect on the downstream ones. We have been preparing a diagram named "Solar Activity Chart" for visualizing the relation among these parameters since 1951, which is a kind of relational database and include important information as sun-solar wind-magnetosphere-ionosphere interactions.

These data have been recorded on micro-films and papers and can be exposed dismissing in long years. We started to save these legend data with digitizing for succeeding to the next generation. In the present status, ionosphere and solar data are already digitized and open on our web. Now we try to digitize the Solar Activity Chart. [3-5]

3. International Cooperation for Space Weather monitoring in Asia-Oceania Region

We started an international project named SEALION (South East Asia Low-latitude Ionospheric Network) in 2003 for monitoring and forecasting equatorial ionospheric disturbances, especially plasma bubbles (EPB) which affect on the use of satellite positioning seriously. In present, we have a cooperative observation in Chiang Mai, Phuket and Chumphon in Thailand, Bac Lieu in Vietnam, Kototabang in Indonesia, and Cebu in Philippine. The observed data are sharing in real time and shown in web site.[6] We built a new VHF radar in Chumphon, Thailand for real time monitoring EPB as a part of SEALION in Jan. 2020.

4. Data Sharing in ICAO Space Weather Centers

As described in the introduction section, the space weather information service already started in Nov. 2019 with three global centers. However, each center has it's own observation database independently from others, and the prediction model is also same situation. This situation could be a cause of discrepancies in the space weather forecast information, and it is quite necessary to harmonize the dataset and validate the models using same dataset. Now NICT prepares the data server for gathering real time ionospheric data observed by ionosondes, GPS and satellite occultation. This system will be available since the early 2021 and hope to contribute to improve the precision of global space weather information for aviation before the next high solar activity period start.

5. Conclusions

NICT has a huge amount of data since the early 20th century not only in domestic but also Asia Oceania region. We started to digitize the data and already Opened them via websites. As a new attempt, we try to share real time data in ICAO global centers for improving space weather forecast services for aviation.

Acknowledgments. We appreciate all legend leaders who started the observation for space weather monitoring sometimes under the difficult situations. A part of this work is supported by the spectrum user fee, the Ministry of Internal affairs and Communications.

References

1. https://swc.nict.go.jp/en/
2. http://www.spaceweather.org/
3. http://wdc.nict.go.jp/IONO/wdc/index.html
4. http://solarobs.nict.go.jp/
5. http://wdc.nict.go.jp/hsv/
6. https://aer-nc-web.nict.go.jp/sealion/

Data and metadata sharing among Asian 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 and publish the data involving Japanese research activities as one of a National Antarctic Data Center (NADC). At the International Polar Year (IPY2007-2008), a significant number of multi-disciplinary data have been compiled. These collected data/metadata have a tight collaboration with the Global Change Master Directory (GCMD), the Polar Information Commons (PIC), as well as several data centers belonging to the World Data System (WDS). In terms of data activities in polar communities of the Scientific Committee on Antarctic Research (SCAR) and the International Arctic Science Committee (IASC), tighter linkages of data/metadata sharing within the Asian Forum for Polar Sciences (AFoPS) countries has been discussed and should be further promoted by the involved Asian countries, in particular China, India, South Korea, Malaysia and Japan.

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

1. Introduction

Diverse data accumulated by several 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 one of the National Antarctic Data Center (NADC). A tight collaboration has been established between PDC/PEDSC and the Global Change Master Directory (GCMD) of NASA, the Polar Information Commons (PIC), and newly established World Data System (WDS) after the IPY. In this Abstract, a history of data management in polar region by Japan is summarized, focusing on the last one decade.

The PDC/PEDSC has been performed a significant function of the 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 influence on the global environmental research. In order to construct an effective framework for long-term strategy of the polar data management, the data should be made available promptly and new Internet technologies such a repository network service, cloud system must be employed. In addition to the activities in polar science communities of SCAR and the International Arctic Science Committee (IASC), tighter linkages have to be established with other cross-cutting data science initiatives under ISC, such as CODATA, and WDS.

2. Asian collaboration perspective

In terms of a collaboration in activities involving polar sciences, the Asian Forum for Polar Sciences (AFoPS) has been 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 the 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 polar regions and the need to work closely with other national operators, together with 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 the 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 symposium 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 the polar sciences.

Regarding the data related activities in polar communities of SCAR and the International Arctic Science Committee (IASC), addition, close linkages in the data sharing among AFoPS countries should be promoted by involved members of the Asian countries. Establishing a new portal server inside the GCMD, for example, the first step for improving data sharing and inter-operability. The metadata from involved Asian countries have been compiled into the Antarctic Master Directly (AMD) within the GCMD, the sharing the metadata from AFoPS nations could be rather smoothly conducted prior to the actual data compilation.

Detail discussion and consultation have been carried out in last few years at the SCAR general meetings and relating Polar Data Forums in Canada (2015), Finland (2018), Norway (2021 as planned), as well as the WDS initiative conferences (Kyoto, 2017). Continuous discussion might be necessary among Asian communities so as to construct the whole data and metadata sharing system. This online conference of DSWS-2020 could be one of the mile-stone to achieve the common goals.

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

Historical records of red auroras in Japan: Collaboration between literature and science

Ryuho Kataoka[1*]

[1*] *National Institute of Polar Research, 10-3 Midori-cho, Tachikawa, 190-8518 Tokyo, Japan*
Email: kataoka.ryuho@nipr.ac.jp

Summary. During the largest magnetic storms, red auroras can be seen by naked eyes even in Japan regardless of the low magnetic latitude. Such a large magnetic storms are rare (e.g., one in 100 years), and it is hard to investigate the basic properties in detail by modern observations alone. Collaboration project between Japanese literature and polar science, called aurora4d project, was therefore conducted to identify such historical records of red auroras. Several successful examples are briefly reviewed in this article.

Keywords. space weather, magnetic storm, aurora, pre-modern Japanese text

1. Introduction

The oldest astronomical record of Japan is the "red sign in the heaven" in AD 620, which can be red aurora occurred during the greatest magnetic storm [1]. Such a characters of red sign, or *Sekki*, can be found in many other old texts in Japan, although careful analyses are always needed in terms of both historical aspects and scientific aspects to identify the real auroral events. We report our latest efforts to make the meaningful analysis to the historical accounts.

2. Meigetsuki event

One of the most popular examples of *Sekki* is the event on 1204 Feb 19-21, recorded in Meigetsuki [2]. The observation site was Kyoto, and the magnetic latitude was 36 deg when the geomagnetic north pole largely declined to Japan.

The most notable characteristic of this auroral event is the prolonged activity, i.e. successive occurrence over 3 nights. More than 10 similar events, prolonged *Sekki*, can also be found in the 300-year history book of China, called *Soushi*, and the occurrence pattern is found to be associated with the 11-year solar cycles [2].

Modern space observations revealed that the largest solar flares hardly stop by a single eruption. The prolonged property of *Sekki*, or successive occurrence of largest magnetic storms, is therefore the natural consequence of the largest eruptive solar flares. In other words, we can learn that such a nature of eruptive solar activity was hold even in 1000 year ago.

3. Seikai event

Graphical evidence in literature sometimes changes the game. A fan-shaped red sky was painted in the book called *Seikai*, on 1770 September 17 in Kyoto. Although the magnetic latitude was only 24 deg, it was also one of the most famous red auroral event in Japan [3].

From the geometrical analysis of inclined geomagnetic field lines over Kyoto, it is found that the field-aligned emission pattern of auroras makes the shape of fan. The reconstructed geometry of the auroral oval then gives an estimate of the possible amplitude of the magnetic storm.

The estimated amplitude is approximately 10^3 nT in the Dst index, which is comparable to or even larger than the Carrington event, the historically largest magnetic storm occurred in September 1859.

4. Taro-Jiro event

Connection of the old data to the modern data is a challenging work. At the end of the aurora4d project, we noticed that in Japan, the first photograph of aurora was taken on 1958 February 11. From the modern analysis of the successive photograph of 1 min cadence, we revealed the westward drift of the fan-shaped aurora [4].

Fortunately, hand-written diaries and paintings also existed in this event [5], and it gives the firm connection among the words, paintings, and photographic modern data. The most surprising present for the author was that the high-school student who wrote ad painted the red sky 60 years ago, Kazama-san, recently visited our institute, bringing another aurora painting. He recently read a newspaper article of the press-released products of our aurora4d project, and noticed that his 60-year-old painting can contribute to the new field of science. This may be an interesting example of citizen science.

5. Conclusions

A variety of examples of historical auroral study are reviewed, in which the close collaboration between literature and science was essentially important and meaningful.

Acknowledgments. The author would like to thank all the members of aurora4d project, supporters, and Kazama-san.

References

1. Kataoka, R., et al., Pheasant Tail: Consideration of the shape of the red sign in the Nihon-Shoki, *SOKENDAI Review of Cultural and Social Studies*, 16, 17-28, 2020
2. Kataoka, R., et al., Historical space weather monitoring of prolonged aurora activities in Japan and in China, *Space Weather*, 15(2), 2017
3. Kataoka, R., and K. Iwahashi, Inclined zenith aurora over Kyoto on 17 September 1770: Graphical evidence of extreme magnetic storm, *Space Weather*, 15, 1314-1320, 2017
4. Kataoka, R., et al., Fan-shaped aurora as seen from Japan during a great magnetic storm on 11 February 1958, 9, A16, 2019
5. Kataoka, R., and S. Kazama, A watercolor painting of northern lights seen above Japan on 11 February 1958, *J. Space Weather Space Clim.*, 9, A28, 2019

Sharing literature annotation regarding COVID-19 through PubAnnotation

Jin-Dong Kim[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: jdkim@dbcls.rois.ac.jp

Summary. To accelerate collaboration for fighting Covid-19, collections of scientific literature related to Covid-19 have been made available, by publishers and several initiatives, for research purposes. Consequently, many research groups which have expertise in natural language processing or text mining have produced annotations to the literature collections, to enable computer-aided access to the content. Toward integration of the datasets, the Covid19-PubAnnotation project was launched to provide an open platform for voluntary contribution and integration, based on the PubAnnotation system. So far, contributions from 6 groups have been integrated, which are publicly available in various ways e.g., download, Web API, or SPARQL search.

Keywords. literature, annotation, textmining, search

1. Background

Covid-19 is giving unprecedented challenges to mankind on almost the whole world. While people are struggling to find ways to overcome it, many scientists are also taking their efforts to extend knowledge about the virus, for the development of treatments, medicines, and so on. Particularly, the community of science is highly motivated to share useful resources, e.g., data, and tools. In the same line, many publishers are opening the content of scientific literature. There are also public initiatives, which are compiling and releasing collections of Covid-19-related literature. To name a few, the *LitCovid* dataset by *NCBI* [1], and the *CORD-19* dataset by the *Allen Institute of AI* [2] are widely recognized as important resources for sharing uptodate scientific knowledge, and for mining potential further discoveries.

Since the release of the datasets, many groups which had expertise in natural language processing (NLP) or text mining (TM) worked on pre-processing the literature [3, 4], e.g., for semantic indexing, to enable ways for computer-aided access to the content of literature. Such pre-processing is often called *annotation*, and the results are called *annotations datasets*.

Those annotation datasets are useful for timely access to specific pieces of information which otherwise might have been hidden in the large amount of literature. However, when they are only accessible from their individual project websites, separately from each other, their utility is limited, missing the chance for an integrative use. It is expected that collecting the annotation datasets and providing an integrated interface to them would make them

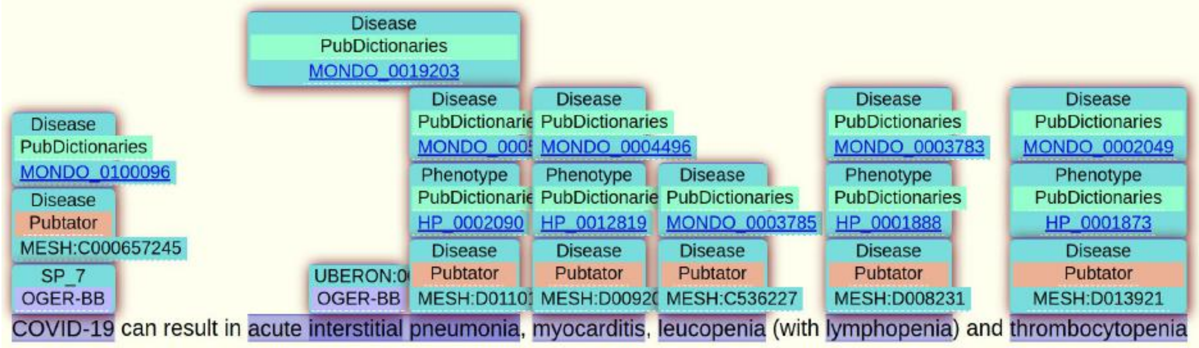


Figure 1. Example of integrated annotations contributed by PubTator, OGER-BB, and PubDictionaries.

even further useful, opening a chance for synergic use of them.

2. Covid19-PubAnnotation

The *Covid19-PubAnnotation* project was launched to address the need for integration of annotation datasets made to the Covid-19-related literature collections. *PubAnnotation* is an open repository of annotations of biomedical literature whose goal is to collect and integrate annotations contributed by the global NLP community [5]. One of the core functions which the system features is the automatic alignment function, thanks to which annotation datasets contributed by various groups can be aligned together.

The PubAnnotation system was set up to collect annotation datasets to the two literature collections, and relevant research groups were invited to contribute. For the call, 6 groups responded and contributed their annotation datasets. Figure 1 is an example of integrated annotation datasets, which includes annotations contributed by three initiatives: PubTator [3], OGER-BB [4], and PubDictionaries [5]. It was found that the contributed annotation datasets are highly complementary to each other, showing the usefulness of the setup. The results of the collected and integrated annotation data sets are publicly available for search, visualization, and fine-grained access.

It is an ongoing project. Based on the analysis on the 6 integrated datasets, the platform is improved particularly for sustainability, and the second call for contribution will be soon issued to solicit more datasets.

Acknowledgments. This work has been financially supported by National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

References

1. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research, *Nature*, 2020;579(7798):193, 2020
2. Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: The Covid-19 Open Research Dataset, *ArXiv*, 2020;arXiv:2004.10706v2, 2020
3. Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu, PubTator: a web-based text mining tool for assisting biocuration, Nucleic Acids Res. 2013 Jul; 41(Web Server issue): W518–W522. 2013
4. Marco Basaldella, Lenz Furrer, Carlo Tasso, and Fabio Rinaldi, Entity recognition in the biomedical domain using a hybrid approach, Journal of Biomedical Semantics, 8 (51), 2017
5. Jin-Dong Kim, Yue Wang, Toyofumi Fujiwara, Shujiro Okuda, Tiffany J Callahan, and K Bretonnel Cohen, Open Agile text mining for bioinformatics: the PubAnnotation ecosystem, Bioinformatics, 35 (21), 4372–4380, 2019

Do we all need to become data scientists?

Jens Klump[1*]

[1*] *CSIRO Mineral Resources, 26 Dick Perry Avenue, Kensington, WA, 6151, Australia*
Email: jens.klump@csiro.au

Summary. The digital transformation changes the way we do research. A previously unimaginable increase in the availability of data and compute resources enables us to put the data first and search for emerging patterns in highly complex datasets, thus accelerating research and innovation. Our response to this transformation needs to balance domain knowledge with digital skills in order to leverage the power of the new research tools. To benefit from this trend, we need to increase the digital literacy of our entire workforce in research. In exchange, we gain new insights through data-driven research.

Keywords. A list of up to 5 words describing the main concepts in the abstract, separated by commas

1. Introduction

Over the last decade, the way we conduct research has changed. We have conducted millennia of empirical research and centuries of theory development. Computing added simulation as a new way of conducting research, and ten years ago, data-driven research was postulated as the "fourth paradigm" of research [1]. In recent years we have seen the "fourth paradigm" evolving into data science, powered by a previously unimaginable increase in the availability of data and compute resources. Are we prepared to make the best use of the new resources data science adds to our toolbox? How will data science change the way we do research?

2. From the "Fourth Paradigm" to Data Science

While exploratory data analysis is not a new concept, data-driven research goes a step further in its use of digital tools. Dramatic increases in computing power and data volumes have made machine learning and artificial intelligence available as common tools for research and engineering. The shift brought forward by the "Fourth Paradigm" of data-driven research is the ability to put the data first and search for emerging patterns in highly complex datasets rather than stating the hypothesis first and then prove or disprove it. In this way, data-driven analysis supplements hypothesis-driven analysis.

Data-driven research is now part of the way we do research, and it comes with its own set of tools. As in any discipline, researchers need to be expert users of their research tools, even if they are not the ones building the next generation. This distinction between the expert user and the developer of new tools might be useful in describing the relationship between the disciplinary and digital domains.

3. Digital Transformation

The tools of data-driven research rely heavily on new digital tools that go beyond the spreadsheet and sometimes beyond the power of the desktop computer. While this shift in scale might sound intimidating, many sophisticated data analysis tools are now available as web-based services.

The barrier to broader adoption of these new tools is both technical and social. The pace of technological change has become so fast that we can no longer wait for it to be absorbed by generational change.

Research organisations like CSIRO in Australia or the Helmholtz Association in Germany have realised that for solving research questions, disciplinary and digital expertise have to go hand in hand. At the same time they also recognised that it takes a concerted effort to upskill their entire workforce to increase their digital literacy.

4. Organisational Response

In its Digital Megatrends report [2], CSIRO recognised that the scale and velocity of the digital transformation of industry, society and science will continue to accelerate over the coming decade. This trend is likely to impact three areas: (1) Data-driven scientific discovery, (2) opportunity for industrial innovation, and (3) key enablers for innovation.

To raise the level of digital literacy in their organisations, CSIRO and the Helmholtz Association have both initiated training programmes called "Digital Academies". CSIRO goes even further than the Helmholtz Association in aiming these digital training programmes not only at early career researchers but also offering digital literacy training for senior researchers and executive managers.

The digital literacy programme in CSIRO is part of a broader initiative, Digital+Domain, that includes expanding and innovating the digital infrastructure, and a research programme. An important part of the research programme is the Machine Learning and Artificial Intelligence Future Science Platform (MLAI FSP). The MLAI FSP consists of about twenty-five postdoctoral researchers matched by an equal number of researchers from CSIRO's business units, thus pairing digital specialists with domain specialists.

5. Conclusions

The digital transformation changes the way we do research. In our response to this transformation, we need to balance domain knowledge with digital skills in order to leverage the power of the new research tools. In exchange, we gain new insights through data-driven research.

Acknowledgments. The author would like to thank his colleagues in CSIRO, ARDC, Helmholtz Digital Earth, and in the Research Data Alliance for the fruitful discussions that informed this abstract. Special thanks go to the Digital Transformation Teams in CSIRO.

References

1. Hey, T., Tansley, S., & Tolle, K. (Eds.). The Fourth Paradigm: Data-Intensive Scientific Discovery (1.1). Redmond, WA: Microsoft Research. 2009
2. Hajkowicz SA, Dawson D. Digital Megatrends: A perspective on the coming decade of digital disruption, CSIRO Data61, Brisbane. 2019

The IGSN Global Sample Number - A PID for Physical Objects

Jens Klump[1], Kerstin Lehnert[2], Sarah Ramdeen[2], Lesley Wyborn[3]*

[1] CSIRO Mineral Resources, 26 Dick Perry Avenue, Kensington, WA, 6151, Australia*
[2] Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, 10964, USA
[3] Australian National University, 143 Ward Road, Acton, ACT, 2601, Australia
Email: jens.klump@csiro.au

Summary. Physical samples are at the heart of many research disciplines, yet it is challenging to unambiguously identify and trace samples outside of their immediate institutional context. The IGSN Global Sample Number transferred the concept of persistent identifiers championed through the use of Digital Object Identifiers by DataCite and applied it to physical samples. The application of IGSN allows persistent identification of physical samples across institutional boundaries. Through linking IGSN with DOI of data and literature, physical samples are linked to scholarly communication and become part of the record of science.

Keywords. Persistent Identifier, Metadata Catalogues, Collection Management, Knowledge Graph

1. Introduction

Physical samples have been a critical component of research since the beginning of scientific investigations, especially in the natural and environmental sciences, material sciences, agriculture, physical anthropology, archaeology, and biomedicine. Millions of samples have been acquired from all over the globe and beyond.

Traditionally, research has seen these samples described, locally processed and then typically stored within individual institutions. Over the last centuries, most universities and government agencies valued their collections and locally curated and catalogued them. In recent decades, as costs rose, administrators argued that preservation of samples had become too expensive and that samples could be recollected when required. Samples were no longer valued and became buried in desk drawers, sheds, or basements, where they remained unknown and inaccessible, supporting neither the transparency of current research nor future science. In many cases, sample sites can no longer be accessed (e.g. mine sites) or were too expensive to recollect (e.g. lunar samples, remote locations, ocean drilling). As time progressed, single institutions no longer had full suites of analytical instruments: samples were shipped around the globe for analysis and reuse by other research communities. However, unless the samples were uniquely identified, confusion reigned over exactly which sample was analysed, by whom and when, and it was often not known which institution the samples came from nor who had originally collected them.

2. Sharing Information About Samples

A solution lies in the IGSN Global Sample Number (IGSN) [1], which provides a globally unique identifier for physical samples. It allows a researcher to establish links between samples (or the digital representation of them), data acquired on these samples, and any publications that result from these data. It enables any researcher, institution or funder that supported the collection of the sample to be acknowledged and credited and can demonstrate value for the contribution that a particular sample has made to scientific research. More importantly, it can ensure that credit is given not only for the collection of the samples but also for the efforts and expertise invested by institutions into sharing them and preserving them for reuse in future science.

Originally developed for the solid Earth Sciences with funding from the US National Science Foundation, the IGSN has evolved into an international PID system that is now adopted by a growing number and range of stakeholders worldwide, and by other disciplines that need to uniquely identify physical samples. More than seven million samples have been registered so far.

3. Linking Samples to the Record of Science

Persistent Identifiers (PID) are a foundational element of research data infrastructure and scholarly communication. PIDs are essential to reference digital and non-digital objects, making these citable and more interoperable, and facilitating unambiguous identification of resources in machine-to-machine communication between information systems. If broadly adopted, they lead to convergence of diverse data systems in this highly fragmented ecosystem. PIDs such as DOI and ORCID have been successfully implemented globally and across disciplines to identify, track and relate digital objects and persons, respectively.

IGSN follows the example of these successful PID systems. The use of IGSN as globally unique, persistent, and resolvable identifiers for samples ensures unambiguous identification of samples and actionable links from publications to online metadata profiles (landing pages) and to other data generated by other studies of the same sample, thus linking physical samples to scholarly communication and the record of science [2].

4. Outlook

The recent expansion of the IGSN beyond geosciences confirms the power of its concept and implementation but imposes substantial pressures on the existing capacity and capabilities of the IGSN architecture and its governing organization.

IGSN is not the first PID service provider to make the transition from project to product and there are lessons to be learnt from other PID services. To this end, the project invited experts in the field of research data infrastructures to develop an organisational and technological strategy and roadmap towards long-term sustainability of the IGSN, a project funded by the Alfred P. Sloan Foundation.

5. Conclusions

Using IGSN as PID links physical samples to data and literature. It shares a common technological base with other PID systems like DOI. This opens new ways for embedding physical samples in the record of science.

Acknowledgments. Building the IGSN Global Sample Number has been supported over the years by the member organisations of the IGSN Implementation Organization (IGSN e.V.)

References

1. IGSN e.V. https://www.igsn.org
Lehnert, K. A., Vinayagamoorthy, S., Djapic, B., & Klump, J. The Digital Sample: Metadata, Unique Identification, and Links

to Data and Publications. *EOS, Transactions, AGU*, 87(52, Fall Meet. Suppl.), Abstract IN53C-07. 2006

General-purpose research-data management service for international research collaboration

Yusuke Komiyama[1*], ***Mao Tsunekawa***[1]

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

Summary. There is an information technology infrastructure system for a research-data management (RDM) service called the National Institute of Informatics' Research Data Cloud (NII RDC) in Japan. The NII RDC has three stages of research-data platforms—discovery, publication, and management—and they work harmoniously together. The web service that manages the closed data under research among the NII RDCs is called GakuNin RDM (GRDM). The GRDM is based on the Open Science Framework by the Center for Opens Science in the US; we modified to fit the GRDM to Japan's current open science policy. In particular, we added a function of research-data trail preservation from the perspective of research integrity and a storage customization function for academic institutions into the GRDM. As of 2019, the GRDM is in the demonstration phase and is in the progress of implementation in collaboration with domestic and international academic institutions. In this article, we introduce the current development progress of the GRDM in Asia and Oceania.

Keywords. research-data management, data sharing, data platform, open science, research collaboration

1. Introduction

To date, research-data management (RDM) services have been required by Japanese academic institutions to avoid scientific misconduct in the context of corporate governance [1]. Recently, the demand for research-data infrastructure is increasing to the research promoting open science policymaking in Japan. For example, the guidelines for the large-scale research & development project "Moonshot" have described the need for advanced research management and research-data infrastructure systems [2].

At present, the national research and education network of Japan—the National Institute of Informatics (NII)—is trying to establish RDM services over the high-speed science network—SINET [3] —for all academic institutions across the country. There is an information technology infrastructure system for an RDM service called NII Research Data Cloud (NII RDC) in Japan. The NII RDC has three stages of research-data platforms—discovery, publication, and management—and they work harmoniously together. The web service that manages the closed data under research among the NII RDC is called GakuNin RDM (GRDM) [4].

2. System Design

2.1 Enhancing RDM service from Open Science Framework

Firstly, We developed a GRDM based on the OSF [5] by Center for Opens Science in the US and modified it to fit with Japan's current open science policy. In particular, we added a function of research-data trail preservation from the perspective of research integrity and a storage customization function for academic institutions into the GRDM. In addition, the GRDM can build a data-sharing system by merely configuring the identity provider of overseas research institutions with eduGAIN [6] —a global authentication and authorization infrastructure (AAI) federation service— because the GRDM has been registered to it as a service provider.

2.2. Linking between RDM service and data management plan (DMP)

Secondly, we designed features that would enable RDM service to be compliant with the DMP. DMPs are an essential part of the research-data life cycle. The RDM service needs to have the functionality to prepare an appropriate research environment compliant with the data management plan. Therefore, we are customizing a data management planning tool that is called ReDBox 2.0 [7] and linking to GRDM as a plugin. ReDBox 2.0 is an open-source software provided by the Queensland Cyber Infrastructure Foundation (QCIF) of Australia. The software has functions that not only support to create DMPs but also orchestrate repositories and RDM service to be compliant with the DMPs. In the initial stage of development, we achieved to provide researchers functions for deploying research-data projects compliant with DMPs into GRDM and extracting research data from GRDM for defining research data sets.

3. Development

3.1. Domestic deployment

We are running a demonstration of the GRDM service for domestic use to academic institutions, starting in April 2019. As of 2020, 19 academic institutions in Japan were participating in the GRDM trial service. Several interdisciplinary projects used the demonstrations service of the GRDM. For example, there is a case of the GRDM usage in neurology by a collaboration of medical doctors and mathematicians. There are also applications in the life sciences where the RDM services are combined with business process systems for reviewing invalid images in a paper before publication in an organization.

3.2. International deployment

Focusing on the Asian region, we are providing the source code of the GRDM as open-source software to the Malaysian Research & Education Network (MYREN). We held an international workshop in Tokyo in February 2020 with MYREN members. They modified the GRDM source code for the installation to enable the RDM service to run on an on-premises virtual machine. They are currently running the same system as the GRDM for 12 institutions in Malaysia.

4. Conclusion

We are building the NII RDC. It is a nationwide RDM service platform on a high-speed science network in Japan, which supports each phase of the research-data life cycle. In particular, the GRDM is a platform that focuses on handling sensitive research data. We have adopted the OSF as the basis for the GRDM and were enhancing it to fit with Japan's open science policy. Besides, the GRDM is being used by research institutions in Japan and Malaysia. The NII RDC is a global standard RDM service for advanced research management that works with a DMP tool that called RedBox 2.0 and supports AAI federation services. Therefore, we believe the NII RDC is useful for international research collaboration.

References

1. Funamori, M., Hayashi, M., Komiyama, Y., Tsuchiya, M. & Yamaji, K. Requirements Analysis of System for Research Data Management to Prevent Scientific Misconduct. in *7th IIAI International Conference on Advanced Applied Informatics (IIAI AAI 2018)*, 382–389, 2018
2. CabinetOffice. Moonshot Reseach and Development Program - Science, Technology and Innovation- Cabinet Office Home Page. https://www8.cao.go.jp/cstp/english/moonshot/top.html, 2020
3. Kurimoto, T. *et al*. SINET5: A low-latency and high-bandwidth backbone network for SDN/NFV Era. in *2017 IEEE International Conference on Communications (ICC),* 1–7, 2017
4. Komiyama, Y. & Yamaji, K. Nationwide Research Data Management Service of Japan in the Open Science Era. in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI 2017)* 129–133, 2017
5. Sullivan, I., DeHaven, A. & Mellor, D. Open and Reproducible Research on Open Science Framework. *Curr. Protoc. Essent. Lab. Tech.* **18**, 2019
6. Michael, S. & Ziegler Jule, A. An Identity Provider as a Service platform for the eduGAIN research and education community. in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM),* 739–740, 2019
7. Nairnsey, J. for Research Data Management Plans (RDMPs), Connecting to cloud services, and Archiving Using ReDBox 2. in *eResearch Australasia 2019*, 1–19, 2019

A data-scientific approach toward understanding of the goldfish genome and morphological diversity

***Tetsuo Kon*[1]*, *Yoshihiro Omori*[1]**

[1]* Laboratory of Functional Genomics, Graduate School of Bioscience, Nagahama Institute of
Bioscience and Technology, Nagahama, Shiga, 526-0829, Japan
Email: t_kon@nagahama-i-bio.ac.jp

Summary. Goldfish strains show highly diversified phenotypes in morphology and coloration. However, the genetic basis underlying these phenotypes is still waiting for elucidation. We recently reported a *de novo* assembly of the goldfish genome. The goldfish genome has 50 chromosomes, which are further divided into two asymmetrically evolved subgenomes consisting of 25 chromosomes each. Furthermore, we performed whole-genome sequencing of 27 goldfish strains and wild goldfish. We found more than 60 million genetic variations and established a population structure of major goldfish strains. Genome-wide association studies and analysis of strain-specific variants showed genetic loci associated with several goldfish phenotypes, including the dorsal fin loss, long-tail, telescope-eye, albinism, and heart-shaped tail. Our results suggest that accumulated mutations in the asymmetrically evolved subgenomes led to the generation of diverse phenotypes in goldfish strains. Taken together, we conclude that data-scientific approaches are effective for understanding the genetic basis of the phenotypic diversity among goldfish strains.

Keywords. Goldfish strains, Diverse phenotypes, Genome assembly, Whole-genome duplication, Genome-wide association study

1. Introduction

The goldfish (*Carassius auratus*) is a domesticated cyprinid teleost, which is closely related to the crucian carp. There are at least 180 variants and 70 genetically established goldfish strains [1]. They show highly diverse phenotypes in body shape, colouration, scales, fin, eye and hood morphology. These phenotypes contain biologically interesting phenotypes that have not been observed in zebrafish or medaka mutants. In addition to showing diverse phenotypes, goldfish are interesting from evolutionary perspectives because their genome experienced a recent whole genome duplication (WGD) event [2]. However, whether and how WGD contributes to genetic and phenotypic diversity in goldfish strains have been largely unknown. To solve this problem, we adopted a data-scientific approach. First, we established a whole genome assembly of the goldfish genome [3]. On top of that, we performed whole-genome sequencing of 27 domesticated goldfish strains [4]. These studies provide us clues why goldfish strains show diversified phenotypes in morphology and coloration.

2. *De novo* assembly of the goldfish genome

By using long read DNA sequencing technology, we generated a high-quality draft sequence and gene annotations of the goldfish genome [3]. The total length of this genome assembly was 1.82 giga base pairs. It contains the 50 chromosome-scale scaffolds which are further divided into two subgenomes, the L- and S- subgenome. We found that the L-subgenome is preserved to stay more similar to the ancestral state, whereas the S-subgenome experiences more gene losses, and changes in levels of gene expression [4]. This phenomenon is called as asymmetric subgenome evolution.

3. Whole-genome sequencing of 27 goldfish strains

To identify genes associated with diverse phenotypes in goldfish strains, we performed whole-genome sequencing of 48 individuals from 27 strains bred in Japan and a wild goldfish lineage. We compared the genome sequencing data of each individual and our goldfish reference genome sequence, and identified more than 60 million variant sites [4]. Genome-wide association studies and analysis of strain-specific variants showed genetic loci associated with several goldfish phenotypes, including the dorsal fin loss, long-tail, telescope-eye, albinism, and heart-shaped tail. These mutations were more related to the S-subgenome than the L-subgenome. This suggests that the asymmetric subgenome evolution contributes to genetic and phenotypic diversity in the goldfish strains.

4. Conclusions

By collecting and analysing the huge genomic data, we suggest the genetic basis of phenotypic diversity in the goldfish strains.

References

1. Omori Y., Kon T. Goldfish: an old and new model system to study vertebrate development, evolution and human disease. *J Biochem*, 165, 209-218, 2019
2. Braasch I. Genome Evolution: Domestication of the Allopolyploid Goldfish. *Curr Biol*, 30, R812-R815, 2020
3. Chen Z.*, Omori Y.*, Koren S., Shirokiya T., Kuroda T., Miyamoto A., Wada H., Fujiyama A., Toyoda A., Zhang S., Wolfsberg TG., Kawakami K., Phillippy AM., NISC Comparative Sequencing Program; Mullikin JC., Burgess SM. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Sci Adv*, 5, eaav0547, 2019 (*equally contributed)
4. Kon T., Omori Y., Fukuta K., Wada H., Watanabe M., Chen Z., Iwasaki M., Mishina T., Matsuzaki SS., Yoshihara D., Arakawa J., Kawakami K., Toyoda A., Burgess SM., Noguchi H., Furukawa T. The genetic basis of morphological diversity in domesticated goldfish. *Curr Biol*, 30, 2260-2274, 2020

Recent Achievements of Deep Learning on Recognition of Modern Japanese Magazines

Anh Duc Le[1*]

[1*] *The Center for Open Data in the Humanities, DS, ROIS*
Email: anh@ism.ac.jp

Summary. Inspired by the recent successes of deep learning on computer vision, natural language processing, we present our recognition of modern Japanese Magazines by using deep learning. The recognition system has two main stages: text line detection and text line recognition. We employ Character Region Awareness Network for text line detection and Attention-based Encoder-Decoder for text line recognition. The system can convert an image of modern Japanese to a text file. It can be the main component for many applications such as text retrieval from image documents, document storage. We also open the source code of the recognition system for research communities.

Keywords. Modern Japanese Magazines, text line detection, text line recognition, deep learning

1. Introduction

Since historical documents are an invaluable resource for historians in exploring social aspects, lifestyles, even weather in the previous era, many countries have been preserved their historical documents. The popular method is to construct digital libraries to preserve historical documents and made them available to the public. This helps researchers from domestic and aboard to access historical documents. The traditional method is scanning books to images. Then, experts read them and provide transcriptions. This approach is labor-intensive, time-consuming, and requiring many experts. So, it is not feasible to provide transcription for a large number of historical documents in libraries. Document analysis and recognition can speed up the digitalization process.

The goal of this research is to develop a recognition system for the full page of modern Japanese magazines. The project will inherit the recent deep learning techniques such as Character Region Awareness Network (CRAFT) for text line detection and Attention-based Encoder-Decoder (AED) for text line recognition.

2. Overview of Recognition System

The system has two main modules: text line extraction and text line recognition. The detailed architectures are shown in Figures 1 and 2.

2.1 Text line detection by CRAFT

CRAFT [1] inherits the UNet model to detect text characters and affinity between characters. We combine the bounding boxes of characters and affinity between them to create the bounding boxes of text lines.

2.2 Text line recognition by AED

The AED model [2] is used for the sequence to sequence. We employ the AED model to convert an image of text line to text. It has two main parts: DenseNet for extracting features from an image of a text line, and an LSTM integrated with an attention model for generating the output text.

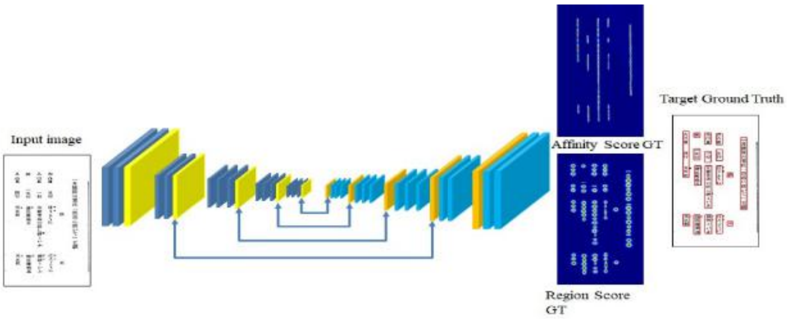


Figure 1. The architecture of text line detection.

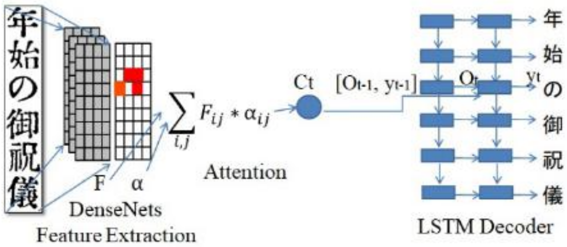


Figure 2.The architecture of text line recognition.

3. Experimental results

We employ a dataset of modern magazines in Japan from 1870 to 1945. It has 922 pages, and we randomly select 80% of pages for training, 10% of pages for validation, and the rest for testing. The number of categories is 5,398, which contains many character categories that do not use in the current Japanese character system.

Table 1 shows the Precision, Recall, and F1 for text line detection. Our retrain model achieves 90.8% of F1 score.

Table1. The result of CRAFT for text line detection.

Method	Precision	Recall	F1
CRAFT	89.6	92.0	90.8

Table 1. The results of AED for text line recognition.

Method	Character Error Rate (%)
AED	22.52

Table 2 shows the CER of our AED model for text line recognition. We achieved 22.52% of CER on the testing sets. Since the training dataset is imbalanced between character classes, we have to improve the accuracy of AED in the tfuture. Moreover, we have released our source code for the research communities [3].

4. Conclusions

In this paper, we have presented the recent achievements in the recognition of modern Japanese magazines. Our recognition system is able to recognize a full page of modern Japanese magazines. We hope our source code is useful for the communities.

References

1. Baek Youngmin et al., Character Region Awareness for Text Detection, CVPR, pp. 9365--9374, 2019
2. Anh Duc Le et al., Recognition of Japanese historical text lines by an attention-based encoder-decoder and text line generation. In Proceedings of the 5th International

Workshop on Historical Document Imaging and Processing, pp. 37–41, 2019

3. Kindai OCR: https://github.com/ducanh841988/Kindai-OCR

The RDA Community Response to COVID-19

Mark Leggott[1*]

[1*] *Research Data Canada, 45 O'Connor St., Suite 1150, Ottawa, ON K1P 1A4, Canada*
Email: mark.leggott@rdc-drc.ca

Summary. In March 2020 The European Commission approached the Research Data Alliance (RDA) to develop guidelines for data sharing for research into the COVID-19 pandemic. By the end of March RDAlaunched a process using its open platform for community engagement, and within a few weeks over 600 community members had volunteered their time to develop the guidelines. The final version of the "COVID-19 Recommendations and Guidelines on Data Sharing" (the Guidelines) was released on June 30, and provides advice in 4 thematic areas (Clinical, Omics, Epidemiology, Social Sciences) and 4 cross-cutting areas (Community, Indigenous Data Guidelines, Research Software, Legal and Ethical). Since their release, the Guidelines have been endorsed by a wide range of stakeholder groups, and will help inform the research into the COVID-19 pandemic. In addition to the Guidelines, a series of supporting navigational aids (infographic, mindmap, decision wizard) and journal articles are being developed.

Keywords. COVID-19, data sharing, research data management, best practices

Data Stewardship: Indian Perspective

Devika P. Madalli[1*]

[1*] *National Documentation Research and Training Center, Indian Statistical Institute,*
8th Mile Mysore Road, Bangalore 560 059, India
Email: devika@drtc.isibang.ac.in

Summary. The talk intends to give an overview of Data initiatives in India. India is diverse and host to a large number of research institutes and universities. There are a number of national nodal centers to foster work in particular domains such as Agriculture, Astrophysics, Medical sciences, cultural studies among many others. It is evident that the data landscape is wide and varied.

Data sharing at national level is fostered by the National Data Sharing and Accessibility Policy (NDSAP) by the Government of India. The national level mandate is binding on all government departments, institutes and projects funded by the central government. Basically the policy ensures that data generated out of public funds will be openly available. One of the significant implications of the policy and mandate is that it resulted in the National Data sharing portal http://data.gov.in which is the national platform for sharing data in various domains. The portal hosts over 400,000 resources in various domains such as agriculture, health, transport, education etc; all openly accessible.

Other than the national portal there are domain based communities that are data intensive. The centre initiated task forces on data management in domains such as climate, health and biosciences. As a response to the present pandemic situation data.gov.in has focused data on COVID19 and presents data (granular), infographics and visualizations.

The Government of India has initiated the Smart Cities Mission. The mission aims to aide urban planners to plan and implement all facilities and citizen services based on needs assessment of communities. The mission is data intensive. Smart cities planning is aided by arrays of datasets such as governance data, demographic data, transport data, employment data, water, sanitation among others.

There are however, issues in data management such as provision of metadata, ensuring quality of data, data curation methods and multilingual issues that need to be addressed. The talk intends to discuss some the issues and possible solutions by adopting best practices for the same.

Keywords.

A unique website JCDP that aims to disseminate scientific information on historical climate data

***Takehiko Mikami*[1*], *Masumi Zaiki*[2], *Junpei Hirano*[3]**

[1*] *Tokyo Metropolitan University, 1-1 Minami-Osawa, Hachioji City, Tokyo, 192-0397, Japan*
[2] *Seikei University, 3-3-1 Kichijoji-Kitamachi, Musashino City, Tokyo, 180-8633, Japan*
[3] *Teikyo University, 359 Otsuka, Hachioji City, Tokyo, 192-0395, Japan*
Email: mikami@tmu.ac.jp

Summary. JCDP (Japan-Asia Climate Data Program) is a unique website that aims to disseminate scientific information on historical climate data. Historical climate data include various kinds of meteorological documentary records, such as old instrumental temperature and pressure observations, since the 19th century before the JMA official data were available. In Japan, daily weather records in feudal domains during the Edo period (the 1600s-1860s) provide detailed climate information from all over the country. Our website JCDP will contribute to the development of historical climatology not only in Japan and Asia but also in Europe and America, where meteorological observations started earlier than those in Asian countries.

Keywords. climate data, documentary record, JCDP

1. Introduction

In Japan, official meteorological observations started at Hakodate Local Meteorological Observatory in 1872, founded by the present Japan Meteorological Agency (JMA). However, our research group discovered dozens of documents which include sub-daily meteorological data observed throughout locations in Japan, such as in Nagasaki, Kobe, Osaka, Yokohama, Tokyo, Mito, Hakodate and so on during the 19th century. Also, in East Asia and Southeast Asia, a vast amount of meteorological records, which cover the 19th and 20th centuries up to World War II, have been kept safe in libraries and museums, or in some cases left to decay.

Apart from meteorological data, there are several kinds of historical documentary sources which enable reconstructions of climate variations before the 19th century in and around Japan. Historical documents, including diaries of individuals, logs of clan offices, government documents, and reports from temples and shrines, have been preserved in local libraries and museums. These often contain daily weather descriptions such as "cold", "fine", "rainy" and "windy", and mention peculiar climate-related natural phenomena such as "lake freezing" and "flower blooming". Information about severe climate-related extreme events such as floods, droughts, and heavy snowfalls are also contained in historical documents.

JCDP aims to disseminate scientific information on historical climate data.

2. Contents of JCDP

JCDP has six drop-down navigation menus; Climate Information (Historical Documents, Old Instrumental Data, and Lighthouse

Observations), Data (Instrumental Meteorological Data, Reconstructed Climate Indices, Historical Weather Database, and Daily Diary Weather Records), Publications, Column, Links, and People (Fig.1). You can download various types of climate data in the text file (EXCEL) format.

The JCDP site has been constantly accessed since it was opened in February 2018, and 42% of the access is from overseas. The number of access from overseas is higher from the United States (26.6%), France (4.3%), United Kingdom (2.4%), etc., and access has been made from 72 countries in total so far.

3. Conclusions

In Japan, we have a tremendous amount of historical documents that include various kinds of climate information, such as weather conditions, natural disasters, and so on. However, almost all such documents were written in Japanese and Chinese characters, which prevents foreign climate scientists from understanding and examining climate changes in Japan during the historical period.

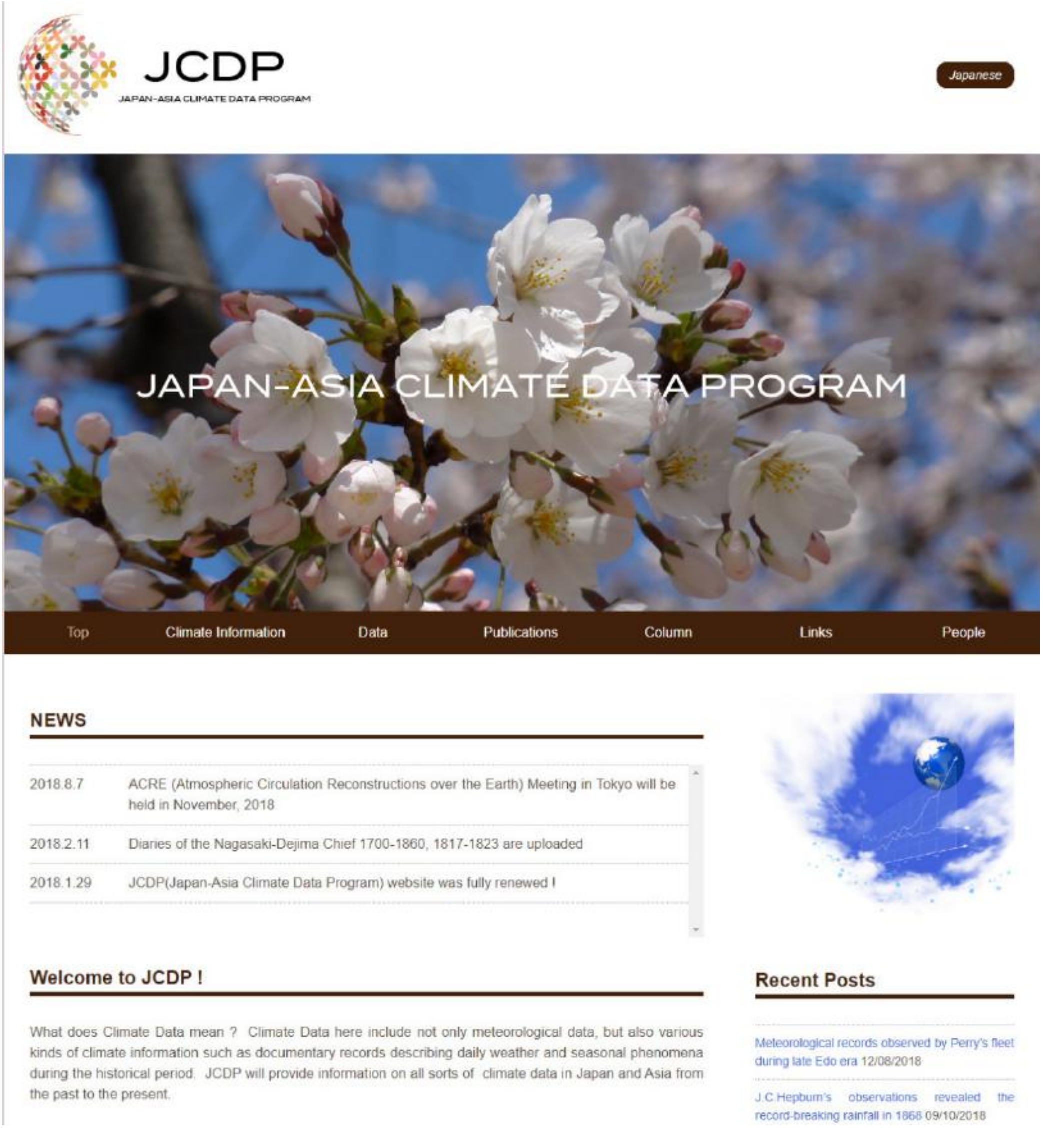
In contrast, collaborative research efforts in Europe have made remarkable progress in historical climatology with the aid of common language English (e.g., Brazdil et al.,2005). Therefore, we will continue disseminating scientific information on Japanese historical climate data in English.

Notes

URL of JCDP website:
https://jcdp.jp/

References

1. Brazdil,R., Pfister,C., WannerH., Von Storch,H. and Luterbacher,J., Historical Climatology in Europe – The state of the Art. Climatic Change, 70, 363-430, 2005



NEWS

2018.8.7	ACRE (Atmospheric Circulation Reconstructions over the Earth) Meeting in Tokyo will be held in November, 2018
2018.2.11	Diaries of the Nagasaki-Dejima Chief 1700-1860, 1817-1823 are uploaded
2018.1.29	JCDP(Japan-Asia Climate Data Program) website was fully renewed !

Welcome to JCDP !

What does Climate Data mean ? Climate Data here include not only meteorological data, but also various kinds of climate information such as documentary records describing daily weather and seasonal phenomena during the historical period. JCDP will provide information on all sorts of climate data in Japan and Asia from the past to the present.

Recent Posts

Meteorological records observed by Perry's fleet during late Edo era 12/08/2018

J.C.Hepburn's observations revealed the record-breaking rainfall in 1868 09/10/2018

Figure 1. A top page of the JCDP website

Spatial Analysis of COVID19 using Compositionally-warped Gaussian Process

Daisuke Murakami[1*]

[1*] *Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo, 190-8562, Japan*
Email: dmuraka@ism.ac.jp

Summary. This study first develops a regression approach to model number of COVID19 infectious by prefectures, which are noisy and non-Gaussian distribution. A compositionally-warping function is used therein. After that, factors determining ease of infectious are analysed using the developed approach. The result demonstrates that commuting, concentration of young people in metropolitan area, people concentration in commercial area, and other factors are likely to increase infectious.

Keywords. COVID19, Spatial analysis, Additive model, Compositionally-warped Gaussian process

1. Introduction

More and more spatial data are becoming available owing to the development of senor and other technologies. Together with that, statistical methods for spatial data have been developed. Yet, these methods are not necessarily adaptive to recent data that are noisy and have non-normal distribution.

Daily number of COVID19 infected people, which we will analyse, is one of such data. As shown in Figure 1, the data have severely skewed (i.e., non-normal) distribution because of the rareness of infected people. Statistical method modelling such data is required to appropriately analyse factors behind COVID19 spread. Given that, we first develop a regression model for such non-normal data. Then, it is applied to an analysis of COVID19.

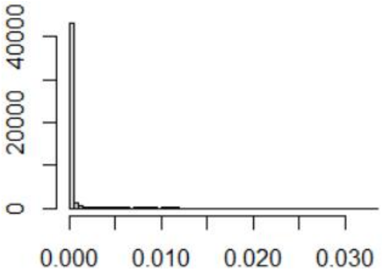


Figure 1: Histogram of # of infectious per area

2. Model for number of infectious

This study analyses the number of infectious per area $y_{i,t}$ on t-th day in i-th prefecture using the following regression model:

$$\varphi(y_{i,t}) = \sum_{k=1}^{K} x_{i,t,k} b_{i,k}(p_i, x_{i,k}) + \varepsilon_{i,t} \quad \varepsilon_{i,t} \sim N(0, \sigma^2)$$

where $x_{i,t,k}$ is the k-th covariate. The coefficient $b_{i,k}(p_i, x_{i,k})$ is assumed to vary depending on prefecture p_i and covariate value $x_{i,k}$. Given the assumption, map pattern such as high risk nearby Tokyo, and covariate-dependent pattern such as high risk of aging people (if age is used as a covariate) are estimated.

A difficulty is the non-normality of $y_{i,t}$. While logarithmic has typically been used to transform non-normal variable $y_{i,t}$ to normal one $\varphi(y_{i,t})$, we use a compositionally-warping approach, concatenating transformation functions to achieve better transformation and model accuracy.

3. Analysis

We analyse the daily number of infectious by prefecture between March 1 and June 3 (source: https://gis.jag-japan.com/covid19jp/). The covariates are week, age, age × prefecture, mean people density in residential area at 7PM (Pop), and day-night ratio of population density at 7PM (DayPop).

Figure 1 summarizes estimated effects from week, days of the week, and age. The major findings are as follows: (a) infectious smoothly varying over week; (b) smaller number of infectious on Monday and Sunday; (c) high risk of people aged between 20 and 59.

Based on (c), commuting might increase infection risk. Figure 2 plots estimated influence from age by prefecture. Interestingly, (A) prefectures nearby Tokyo and Osaka have different properties from (B) the other prefectures. 20s and 30s tend to have higher risk in (A) whereas 40s and 50s has the highest risk in (B). Concentration of young people might spread infection in (A). On the other hand, high risk in (B) might mean commuting is the major source of COVID19 spread.

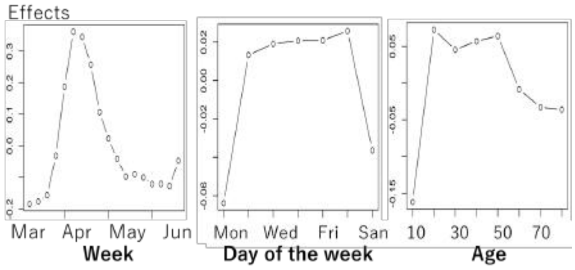


Figure 1: Effects from covariates

Regarding Pop and DayPop, both are estimated to increase number of infectious. The result suggests that, while the number of infectious increases as population density increases, this tendency gets severe in commercial area with high daytime population.

Figure 3 plots residual factors that could not be explained by the covariates. This figure demonstrates that number of infectious is extraordinarily high nearby Tokyo even subtracting the effect of population density.

In summary, commuting, concentration of young people in Tokyo and Osaka, people concentration in commercial area, and location nearby Tokyo, are factors increasing infectious.

Acknowledgement

People density data is provided by NTT DoCoMo Inc.

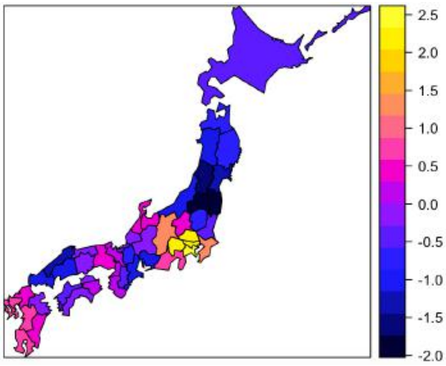
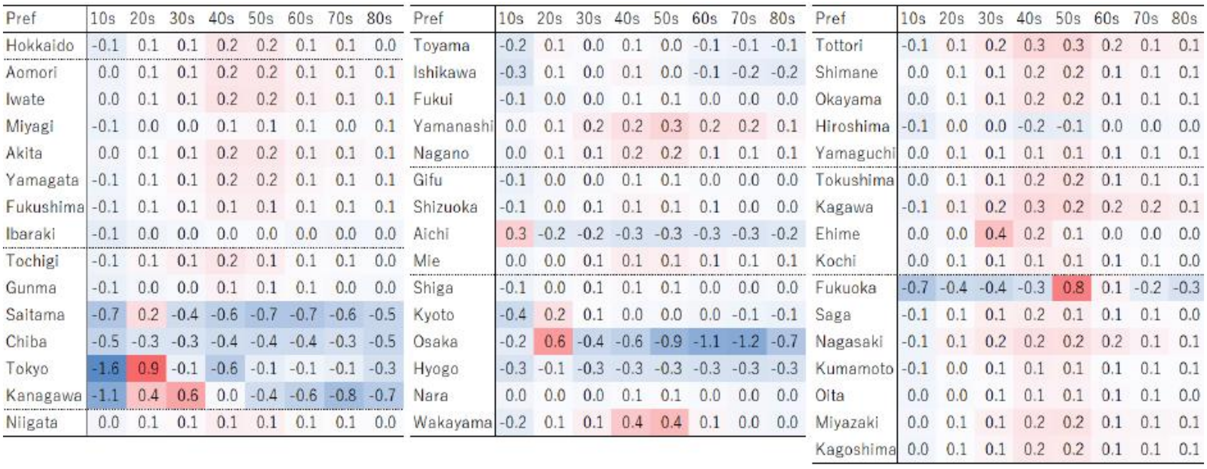


Figure 3: Residual spatial process (ease of infectious that are not explained by the covariates).



Pref	10s	20s	30s	40s	50s	60s	70s	80s
Hokkaido	-0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.0
Aomori	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Iwate	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Miyagi	-0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.1
Akita	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Yamagata	-0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Fukushima	-0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Ibaraki	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tochigi	-0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0
Gunma	-0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0
Saitama	-0.7	0.2	-0.4	-0.6	-0.7	-0.7	-0.6	-0.5
Chiba	-0.5	-0.3	-0.3	-0.4	-0.4	-0.4	-0.3	-0.5
Tokyo	-1.6	0.9	-0.1	-0.6	-0.1	-0.1	-0.1	-0.3
Kanagawa	-1.1	0.4	0.6	0.0	-0.4	-0.6	-0.8	-0.7
Niigata	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.0
Toyama	-0.2	0.1	0.0	0.1	0.0	-0.1	-0.1	-0.1
Ishikawa	-0.3	0.1	0.0	0.1	0.0	-0.1	-0.2	-0.2
Fukui	-0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
Yamanashi	0.0	0.1	0.2	0.2	0.3	0.2	0.2	0.1
Nagano	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Gifu	-0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0
Shizuoka	-0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.0
Aichi	0.3	-0.2	-0.2	-0.3	-0.3	-0.3	-0.3	-0.2
Mie	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1
Shiga	-0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.0
Kyoto	-0.4	0.2	0.1	0.0	0.0	0.0	-0.1	-0.1
Osaka	-0.2	0.6	-0.4	-0.6	-0.9	-1.1	-1.2	-0.7
Hyogo	-0.3	-0.1	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3
Nara	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0
Wakayama	-0.2	0.1	0.1	0.4	0.4	0.1	0.0	0.0
Tottori	-0.1	0.1	0.2	0.3	0.3	0.2	0.1	0.1
Shimane	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Okayama	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Hiroshima	-0.1	0.0	0.0	-0.2	-0.1	0.0	0.0	0.0
Yamaguchi	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Tokushima	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Kagawa	-0.1	0.1	0.2	0.3	0.2	0.2	0.2	0.1
Ehime	0.0	0.0	0.4	0.2	0.1	0.0	0.0	0.0
Kochi	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.0
Fukuoka	-0.7	-0.4	-0.4	-0.3	0.8	0.1	-0.2	-0.3
Saga	-0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.0
Nagasaki	-0.1	0.1	0.2	0.2	0.2	0.2	0.1	0.1
Kumamoto	-0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1
Oita	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.0
Miyazaki	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1
Kagoshima	0.0	0.1	0.1	0.2	0.2	0.1	0.1	0.1

Figure 2: Effects from age by prefecture

Monitoring Sustainable Development Goals Amidst COVID-19 Through Big Data, Deep Learning and Interdisciplinarity

Kassim Said Mwitondi[1*]

[1*]*Sheffield Hallam University, College of Business, Technology and Engineering; Sheffield S1 2NU. United Kingdom*
Email: k.mwitondi@shu.ac.uk

Summary. As the coronavirus disease 2019 (COVID–19) ravaged across the globe, in the first half of 2020, the world was once again reminded of the huge gaps in our knowledge, despite our current scientific and technological capacities. The pandemic has had a severe impact on our ways of life, and despite its devastating impact, it has presented us with an opportunity for paying greater attention to the challenges we face. It is in that context that we associate the fight against COVID-19 with monitoring Sustainable Development Goals (SDG). Considering each SDG as a source of Big Data, we present a generic framework for combining Big Data, machine learning and interdisciplinarity to address global challenges. The work delivers descriptive and prescriptive findings, using data visualisation and animation techniques, on the one hand, and predictive results, based on convolutional neural networks, on the other. The former is based on structured data on cases and deaths from COVID–19 obtained from the European Centre for Disease Prevention and Control (ECDC) and data on the impact of the pandemic on various aspects of life, obtained from the UK Office of National Statistics. Predictive findings are based on unstructured data–a large COVID–19 X–Ray data, 3181 image files, obtained from Github and Kaggle. The results from both sets are presented in the form that resonates with cross disciplinary discussions, opening novel paths for interdisciplinary research in tackling global challenges.

Keywords. Big Data, Convolutional Neural Networks, Data Science, Data Visualisation, Interdisciplinarity, Predictive Modelling, Sustainable Development Goals

1 Introduction

The United Nations Sustainable Development Goals (SDGs) SDG [2] were signed up by 193 member states in 2015, outlining measurable targets and indicators to be attained by the year 2030, hence its commonly used pseudonym, agenda 2030. They were drawn to address the global challenges–poverty, inequality, climate change, environmental degradation, peace and justice, with a target set for the year 2030.

The complex interactions of the SDGs, the magnitude and dynamics of inherent data attributes and the deep and wide socio–economic and cultural variations across the globe provide both a challenge and an opportunity to the SDG project [8, 9]. In the light of the impact of COVID–19 on SDGs and the computing power, each SDG is a source of Big Data [8-12]. This work is motivated by the foregoing aspects, and it seeks to answer the question: How can interdisciplinarity, Big Data and deep learning combine to deliver sustainable solutions in the wake of the COVID–19 pandemic? We set the following objectives.

1 To provide a descriptive mapping of skills and interdisciplinary knowledge for addressing global challenges.

2 To illustrate the impact of COVID–19 based on structured and unstructured data.

3 To demonstrate the efficacy of combining data, modelling techniques and skills in an interdisciplinary context.

4 To present analytical results through data visualisation, animation and modelling.

2 Methods

The general framework for this work is illustrated in Figure 1, in which the intersections highlights the setup for developing robust solutions for global challenges like COVID–19.

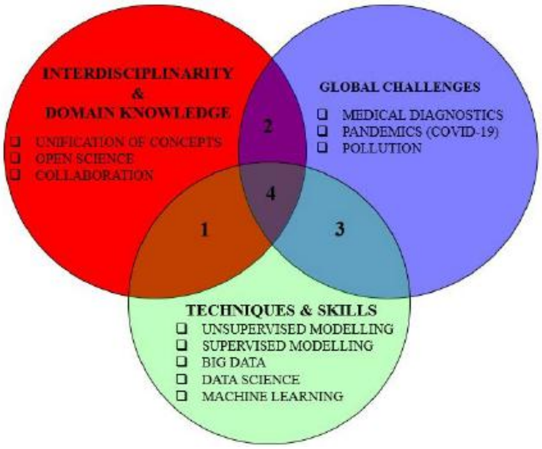


Figure 1: Iintersection of challenges, data & skills

2.1 Data Sources

The work is based on structured and unstructured data. The former came from the European Centre for Disease Prevention and Control (ECDC) [13] and the UK Office of National Statistics [14]. Unstructured data is a combination of 1840 COVID–19 X–Ray image files, downloaded from a Github[15] and 1341 normal data collected from Kaggle [16].

2.2 Data Visualisation

Data visualisation and animation techniques are used on structured data to reflect the impact of COVID–19 in a descriptive and prescriptive way. Results provide clear insights to stakeholders in both the battle against COVID–19 and in monitoring SDGs. We mainly focus on the key parameters, before and during the pandemic. Visualisation and animation codes are adaptable to handle new cases.,

2.3 Convolutional Neural Networks

We apply Convolutional Neural Networks (CNN)-a machine learning technique, for classifying the X–Ray data, as in Figure 2.

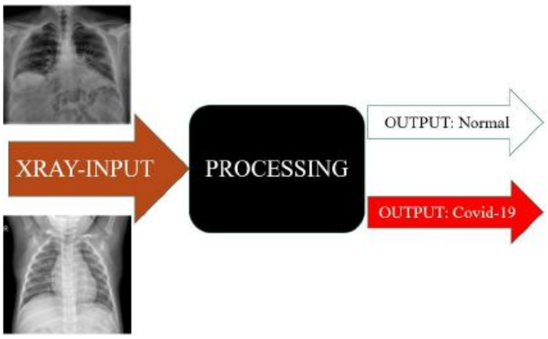


Figure 2: A CNN model for classifying imagery data

The mechanics of CNN are well-understood [17, 18, 19]. Its architecture is illustrated in Figure 3.

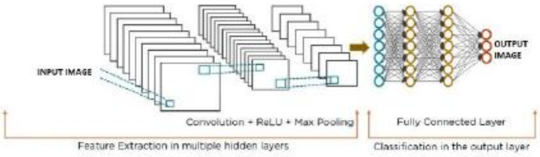
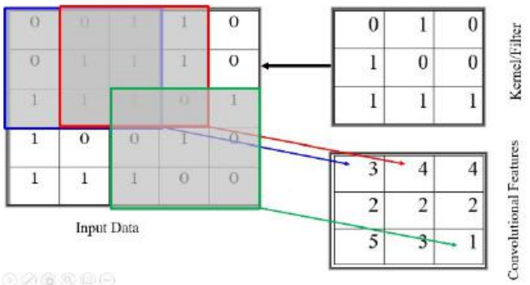

INPUT IMAGE
OUTPUT IMAGE
Convolution + ReLU + Max Pooling
Fully Connected Layer
Feature Extraction in multiple hidden layers
Classification in the output layer

Figure 3: A typical CNN architecture

The CNN output, $Y_{i,j,k}$ denotes the neuron output in the i^{th} row and the j^{th} column of feature map k of the l^{th} convolutional layer. We highlight the technical challenges in obtaining the convolutional values, demonstrating them via sliding the kernel over the input data, as shown in Figure 4.



0	0	1	1	0
0	1	1	1	0
1	1	1	0	1
1	0	0	1	0
1	1	1	0	0

Input Data

0	1	0
1	0	0
1	1	1

Kernel/Filter

3	4	4
2	2	2
5	3	1

Convolutional Features

Figure 4: Convolutional values obtained via kernel sliding

The vertical and horizontal strides in Figure 4 as they impinge on the model's capability of feature capturing. The pooling layer in Figure 3 reduces the dimensionality of the rectified feature map, using different filters to identify different parts of the image. The Fully Connected layer then receives this as input, for classifying the image. The Rectified Linear Unit (ReLU) applies the activation function such that the output is zero for all negative inputs $(x < 0)$ and x otherwise. The sigmoid and the hyperbolic tangent are also commonly used.

3 Analyses

For structured data, analyses are presented as graphical images captured from animated patterns using the Gapminder library in R. For unstructured data, a CNN model in Python is implemented, with adaptable parameters tested via a Sample-Measure-Assess algorithm.

3.1 Data Visualisation

Multiple illustrations are given, based on UK data on production and employment, as well as on COVID-19 and non-COVID-19 related deaths. Figure 5 shows COVID-19 related cases and deaths for May 2020, in Brazil, France, Italy, Japan, South Africa, US and the UK.

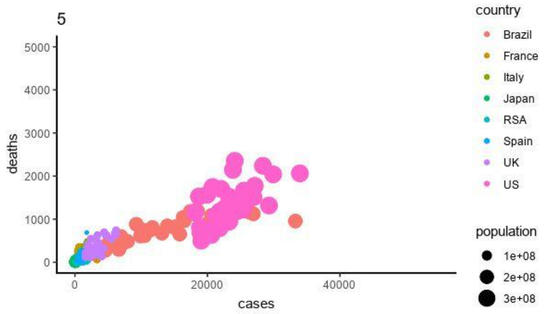


Figure 5: COVID-19 related cases and deaths in May

3.2 Convolutional Neural Networks

The two images in Figure 6 represent two COVID–19 X–Ray data described in Section 2.1. The image on the left-hand side is from the COVID–19 positive sample of 1840 cases, while the one to the right is from the 1341 non–COVID–19 samples. Under pandemic conditions, doctors and radiologists are under pressure to distinguish such images.

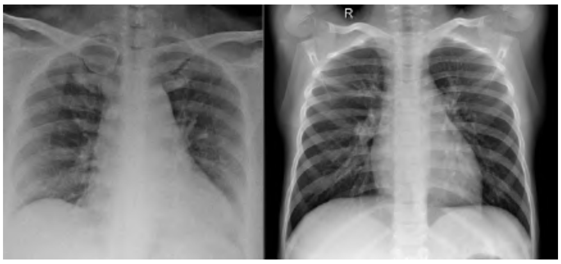


Figure 6: Covid-19 positive (L) and normal (R)

While the application of CNN in modelling imagery data is not new, challenges have remained relating to accuracy and generalisation. We demonstrate these challenges and opportunities relating to CNN modelling in an interdisciplinary context.

4 Concluding Remarks

This paper highlighted the potentials of combining underlying domain knowledge, on the one hand, and data science–technical skills and soft skills, on the other. It underlined the role of interdisciplinarity in addressing global challenges, as described in the SDGs. While CNN models can detect patterns that might go unnoticed to the human eye, for all their power and complexity, they do not provide thorough interpretations of the imagery data. Further, they may perform poorly on new data.

The ECDC acknowledges that the data might not be very accurate the calculations by the ECDC Epidemic Intelligence are affected by variations in national testing strategies, laboratory capacities effectiveness of surveillance systems. Thus, any comparisons should be made with care, possibly in combination with other factors like "...testing policies, number of tests performed, test positivity, excess mortality and rates of hospital and Intensive Care Unit (ICU) admissions." In particular, such comparisons must be done by teams of data scientists, epidemiologists, and other medical and social experts.

Acknowledgments. The author would like to thank the Joint Support-Centre for Data Science Research (ROIS-DS) for inviting him to all its events in the last few years. The partnership has significantly enhanced his involvement in interdisciplinary research.

References

1. Rothan, H. A.; Byrareddy, S. N. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Journal of Autoimmunity, 109, 102433, 2020
2. SDG, Sustainable Development Goals. 2015; https://www.un.org/sustainabledevelopment/sustainable-development-goals/
3. Zambrano-Monserrate, M. A.; Ruano, M. A.; Sanchez-Alcalde, L. Indirect effects of COVID-19 on the environment. Science of The Total Environment, 728, 138813, 2020
4. Bartik, A. W.; Bertrand, M.; Cullen, Z.; Glaeser, E. L.; Luca, M.; Stanton, C. The impact of COVID-19 on small business outcomes and expectations, 117, 17656–17666, 2020
5. Pan, S. L.; Zhang, S. From fighting COVID-19 pandemic to tackling sustainable development goals: An opportunity for responsible information systems research. International Journal of Information Management,102196, 2020
6. Wang, C. J.; Ng, C. Y.; Brook, R. H. Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing, 323, 1341–1342, 2020
7. IUCN, In the spirit of nature. 2018; https://www.iucn.org/news/europe/20181/spirit-nature-everything-connected
8. Mwitondi, K.; Munyakazi, I.; Gatsheni, B. Amenability of the United Nations Sustainable Development Goals to Big Data Modelling. International Workshop on Data Science-Present and Future of Open Data and Open Science, 12-15 Nov 2018, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall, Mishima, Shizuoka, Japan, 2018
9. Mwitondi, K.; Munyakazi, I.; Gatsheni, B. An Interdisciplinary Data-Driven Framework for Development Science. DIRISA National Research Data Workshop, CSIR ICC, 19-21 June 2018, Pretoria, RSA, 2018
10. Kharrazi, A. Challenges and Opportunities of Urban Big-data for Sustainable Development. Asia-Pacific Tech Monitor, 34, 17–21, 2017
11. Kruse, C. S.; Goswamy, R.; Raval, Y.; Marawi, S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Medical Informatics, 4, e38, 2016
12. Yan, M.; Haiping, W.; Lizhe, W.; Bormin, H.; Ranjan, R.; Zomaya, A.; Wei, J. Remote sensing big data computing: Challenges and opportunities. Future Generation Computer Systems, 51, 47–60, 2015
13. ECDC, COVID-19 Data. 2020; https://www.ecdc.europa.eu/en/publications-data
14. ONS, Office for National Statistics. 2020; https://www.ons.gov.uk/
15. Cohen, J. P.; Morrison, P.; Dao, L. COVID-19 image data collection. arXiv 2003.11597 2020
16. Kaggle, Chest X-Ray Images (Pneumonia). 2020; https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia
17. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36, 193–202, 1980
18. LeCun, Y.; Jackel, L. D.; Boser, B.; Denker, J. S.; Graf, H. P.; Guyon, I.; Henderson, D.; Howard, R. E.; Hubbard, W. Handwritten Digit Recognition: Applications of Neural Net Chips and Automatic Learning. IEEE Communication, 41–46, invited paper, 1989
19. LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation, 1, 541–551, 1989
20. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. In Advances in Neural Information Processing Systems 25; Pereira, F., Burges, C. J. C.,

Bottou, L., Weinberger, K. Q., Eds.; Curran Associates, Inc., 1097–1105, 2012

21. Impact of the digital divide in the age of COVID-19. Journal of the American Medical Informatics Association, 27, 1147-1148, 2020

Standardization of biological sample information database

Tazro Ohta[1*], Shuya Ikeda[1], Takatomo Fujisawa[2], Shuichi Kawashima[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*
[2] *DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411-8540, JAPAN*
Email: t.ohta@dbcls.rois.ac.jp

Summary. Genomic data sharing is an essential concept for biological and medical sciences. Researchers submit the biological sample information as metadata of their genome sequence data to the public repository called BioSample, where researchers can freely utilize the information. However, the sample information is often described without using controlled vocabulary, which causes problems in searching public data. To provide a better form of sample information, we designed a data model in Resource Description Format (RDF) to improve interoperability and added ontology terms to the existing entries to standardize the description.

Keywords. Data sharing, Genome sequencing, Biological sample information, Resource Description Framework, Ontology

1. Introduction

Sharing the genomic data among the researchers have advanced biological and medical sciences [1]. Since the adoption of the Bermuda principles, most journals require publication of genomic data as a condition for paper publication. Even some research fundings require researchers to release data before the paper publication. The genomic data shared by the researchers all over the world have been collected by three organizations, DNA Data Bank Japan (DDBJ) in Japan, European Bioinformatics Institute (EBI) in Europe, and National Center for Biotechnology Information (NCBI) in the United States. Anyone can use the collected genomic data for any purpose free of charge. Thus, many secondary databases have been built based on shared data [2]. The secondary databases would be utilized to generate new data, which is then shared again. This ecosystem of genomic data sharing has been a key to the success of the field.

However, still there is a big challenge in sharing genomic data. In the past 10 years, the rapid advance of the DNA molecule measurement equipment so-called DNA sequencer resulted in the enormous amount of data produced per project [3]. Genomic data, which used to be collected for single species, are now available at a finer granularity, such as cell line, cell culture, or numerous single cells. The low-cost DNA sequencing methods accomplished by the latest technologies resulted in the massive growth of the amount of the genomic data per project. Researchers now get hundreds of samples to sequence, which occurs sample information management tasks, a new bottleneck for data sharing.

The public DNA database accepts various kinds of data. The target species are from human to microorganisms. The sequencing methods capture DNA, RNA, or those molecules enriched by a special experimental protocol. Therefore it is not feasible to control all the vocabularies used to describe the used samples. Sample information is written in natural language, often in the form of key-value pairs. This lack of a fully standardized way to describe the sample information causes a problem when a user searches public data based on sample information [4]. Many sample information descriptions include synonyms and typographical errors. For example, a protein name Oct4 has many different forms; for instance, a hyphenated style "Oct-4", its synonym "POU5F1", or a Microsoft Excel induced error "October 4th". It is not likely that all the researchers look for hundreds of thousands of samples to find those differences between sample information to obtain a set of samples by an attribute.

To solve the problem of the sample information description, we developed an improved sample information database with a standardized format and controlled vocabularies. In this research, our target is the BioSample database, the largest biological sample information database that principally collects samples of genomic sequencing experiments.

2. Result

First, we designed a data model based on the Resource Description Framework (RDF), a W3C standard to enable interchange between different resources available on the web [5]. We employed RDF because the BioSample database is a hub to connect various biological databases and knowledge bases. Many biological databases publish their data in RDF format for its interoperability (cite: KERO, ChIP-Atlas). Modeling BioSample data in RDF can facilitate the merging of public genome data and the secondary data generated by third-party database developers.

Second, we added ontology terms to the existing BioSample entries based on the description by the original data submitters. This enables us to identify different words that indicate the same concept. We developed a pipeline to process the entries by using a tool used to develop MetaSRA, a database that organizes the public information of RNA sequencing of human samples [6]. The current implementation of the pipeline can process only human samples and process over 400 million samples in nearly 10 hours.

3. Conclusion

Providing sample information in RDF format and assigning ontology terms can solve the two big challenges in utilizing public biological information. This allows users to get the sample information without preprocessing data which costs crucial. Our approach also helps use cases of modern machine learning applications that require input data with standardized metadata.

References

1. Lakiotaki K, Vorniotakis N, Tsagris M et al. BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. Database 2018;2018
2. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet, 14, 89–99, 2012
3. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics, 107,1–8, 2016
4. Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data;6, 2019
5. Kawashima S, Katayama T, Hatanaka H et al. NBDC RDF portal: a comprehensive

repository for semantic data in life sciences. Database 2018

6. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Wren J (ed.). Bioinformatics, 33, 2914–23, 2017

Susceptibility to COVID-19 infection: Insights from mathematical modelling

Ryosuke Omori[1*], **_Ryota Matsuyama_**[2], **_Yukihiko Nakata_**[3]

[1*] *Research Center for Zoonosis Control, Hokkaido University, Kita-20 Nishi-10 Kita-ku, Sapporo, Hokkaido, 001-0020, Japan*
[2] *Hiroshima University, 1-2-3 kasumi, Minami-Ku, Hiroshima, Hiroshima, 734-8553, Japan*
[3] *Aoyama Gakuin Univeristy, 5-10-1 Fuchinobe, Sagamihara, Kanagawa, 252-5258, Japan*
Email: omori@czc.hokudai.ac.jp

Summary. Since the emergence of COVID-19, its pandemic has been observed all over the world. In spite of the large heterogeneities in the number of confirmed cases and deaths by COVID-19 between countries, the trend in the age distribution of mortality is quite similar between countries; most deaths are elderly persons. To understand the mechanism of a common trend in age distribution of mortality, the assumptions regarding the cause for age-dependency of mortality by COVID-19 were evaluated by employing a mathematical model describing transmission process and natural history of COVID-19. Our results suggest that the susceptibility to COVID-19 is not differed largely by age, but the age-distribution of mortality by COVID-19 can be explained simply by the age-dependent mortality or the rate becoming severe cases among all cases.

Keywords. COVID-19, Mathematical modelling, Epidemiology

1. Introduction

Since the emergence of novel coronavirus (COVID-19), the pandemic of COVID-19 has been observed around the world; 382 cases per million population in Italy, 507 cases in Spain, and 13 cases in Japan have been confirmed until 29th May 2020. Also, 29525 deaths in Italy, 18818 deaths in Spain, and 400 deaths in Japan have been confirmed until 12th May 2020. Both of confirmed cases and deaths are varied largely among countries. Estimates of R_0, the expected number of secondary cases from a single primary case, are also varied among countries, it suggests that transmissibility of COVID-19 are varied largely among countries.

On the other hand, a common trend in the age-distribution of mortality by COVID-19 among countries has been observed; most deaths are elderly persons. Between Italy, Spain, and Japan, where the number of deaths and confirmed cases are varied largely, the quite similar pattern in the age distribution of mortality by COVID-19 have been observed. This suggests a weak association between the transmissibility of COVID-19 (measured by R_0) and the age distribution of mortality by COVID-19.

The causes for the age-dependency of mortality by infectious diseases can be classified into two groups, one is the heterogeneity in susceptibilities (the rate of infected cases among exposed cases to COVID-19) and another is the heterogeneity in the rates becoming severe cases among infected individuals. We evaluated that which factor shapes a common trend in the age-distribution of mortality by COVID-19 among countries by fitting a mathematical model describing a transmission process and natural history of COVID-19 with the observed data of the age-distribution of mortality by COVID-19.

2. Model

Since the existence of latent period, recovery or death after infection of COVID-19 are known, a mathematical model, SEIRD model (an extended model from SIR model) was employed to model a transmission process and natural history of COVID-19;

$$S'_n = -\beta\sigma_n S_n(\Sigma_m k_{n,m} I_m), \tag{1}$$

$$E'_n = \beta\sigma_n S_n(\Sigma_m k_{n,m} I_m) - \varepsilon E_n, \tag{2}$$

$$I'_n = \varepsilon E_n - (\gamma+\delta_n)I_\mathrm{n}, \tag{3}$$

$$R'_n = \gamma I_n, \tag{4}$$

$$D'_n = \delta_n I_n, \tag{5}$$

S_n, E_n, I_n R_n and D_n represent the proportions of susceptible, exposed, infected, recovered and death among age group *n*. β, ε, γ and δ_n are transmission coefficient, the rate becoming infectious, recovery rate, and mortality rate among age group *n*. Among the causes for the association between R_0 and the age distribution of mortality, a most important cause is the heterogeneity of contact rate between different age groups described as $k_\mathrm{n,m}$ in the model. It can be considered that $k_\mathrm{n,m}$ can be varied among countries and this heterogeneity can affect to the result, $k_\mathrm{n,m}$ was parameterized using the estimates per country in the previous study [1]. As for the modelling of heterogeneity of susceptibility by age, we modelled the susceptibility among age group n as

$$\sigma_\mathrm{n} = cn^{\varphi}. \tag{6}$$

If $\varphi = 0$, the susceptibility is independent with age. From the estimated values of φ explaining the age-distribution of mortality by COVID-19, we evaluated whether the age-distribution of mortality by COVID-19 comes from the age-dependency of susceptibility or not.

3. Results

First, we evaluated whether the age-distribution of mortality by COVID-19 can be explained by only the age-dependency of susceptibility or not, by estimating φ with assuming no age-dependency in mortality (δ_ns are constant between age groups). If it is assumed that δ_ns are constant between age groups, the age-distribution of mortality $D_n(\infty)/\Sigma_m D_m(\infty)$ is equal to $R_n(\infty)/\Sigma_m R_m(\infty)$ regardless of the value of δ_n. Since the number of asymptomatic and mild cases are difficult to estimate, δ_ns are also difficult to estimate. If this assumption is suitable, the estimate of δ_n is not necessary to argue the association between age-dependency of susceptibility (can be measured by φ) and age-distribution of mortality ($D_n(\infty)/\Sigma_m D_m(\infty)$). From numerical analyses of equation (1)-(5), we observed that the age-distribution of mortality is dependent with φ. However, the age-distribution of mortality is also dependent with R_0, contradicting to the observed low sensitivity of age-distribution of mortality to R_0.

On the other hand, if we assume the age-dependent mortality δ_n but no age-dependency in the rate becoming symptomatic cases among all cases, the dependency of age-distribution of mortality to R_0 is quite low. Here we estimated δ_n from the observed data of the mortality rate among confirmed cases.

Also, estimated values φ to explain the age-distribution of mortality in Italy, Spain and Japan with the each of two assumptions as above are quite different between three countries. Although the age-dependency in susceptibility can be differed by country, the difference is unrealistically large. This suggests that these two assumptions i) no age-dependency in mortality ii) no age-dependency in the rate becoming symptomatic cases among all cases is unrealistic to explain the age-distribution of mortality.

4. Conclusions

The age dependency of susceptibility to COVID-19 is difficult to explain the age-distribution of mortality with low sensitivity to R_0 values. This suggests that the susceptibility to COVID-19 is not differed largely by age. Also, the age-distribution of mortality by COVID-19 can be explained simply by the age-dependent mortality or the rate becoming severe cases among all cases. Further detail can be found in [2].

References

1. Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comput Biol. 2017;13(9):e1005697. doi:10.1371/journal.pcbi.1005697
2. Omori R., Matsuyama R., Nakata Y., Does susceptibility to novel coronavirus (COVID-19) infection differ by age? : Insights from mathematical modelling, medRxiv, doi:10.1101/2020.06.08.20126003

Open Science for Asia and the Pacific

Mazlan Othman[1*]***, Sufyan Aslam***[2]

[1*] *Director, ISC ROAP, 902-4 Jalan Tun Ismail, 50480, Kuala Lumpur, Malaysia*
[2] *Science Officer, ISC ROAP, 902-4 Jalan Tun Ismail, 50480, Kuala Lumpur, Malaysia*
Email: mazlan.othman@council.science

Summary. Open Science is a global movement for doing science in the future. This movement is being championed by UNESCO, which is undertaking a survey for the development of an international standard-setting instrument on Open Science. The APEC platform provides the political foundation for collaboration within Asia Pacific.

Keywords. Open Science, International Science Council, UNESCO, APEC

I. Introduction

Open Science represents an approach to the scientific process which is based on cooperative work and new ways of disseminating knowledge by using digital technologies and new collaborative tools. The idea captures a systemic change to the way science and research have been carried out for the last fifty years: complementing the standard practices of publishing research results in scientific publications by sharing and using all available knowledge at an earlier stage in the research process.

The recent response of the scientific community to the COVID-19 pandemic has vividly demonstrated the importance of Open Science and how it can accelerate the creation of scientific solutions to a global challenge. The genetic sequence of the SARS-CoV-2 virus was posted in an open access repository and made freely available for all researchers. Several companies made the designs for protective face shields open source, allowing these shields to be freely 3D printed in cities and societies where they are needed the most.

II. International Science Council's Vision for Open Science

The International Science Council (ISC), in its Action Plan 2019-2021 (1), highlighted the importance of addressing Open Science, particularly in the Global South. Open Science in the Global South aims to allow scientists and science systems in the Global South to collaborate and position themselves at the cutting edge of data-intensive open science. In collaboration with CODATA, the Council has been working with its Regional Offices to start the discussions on regional Open Science Platforms that will convene and create regional interests, ideas, people, institutions and resources needed to advance data-intensive, solutions-oriented research in the Global South (2).

A pilot study for a Pan-African Open Science Platform (AOSP) was launched in December 2016 with the support of the South African Department of Science and Innovation and in collaboration with the Academy of Science of South Africa and the South African National Research Foundation (3). In April 2020, the National Research Foundation agreed to host the African Open Science Platform Project Office for the next 3 to 5 years.

III. UNESCO's Recommendation on Open Science

In spite of the encouraging open science actions in response to COVID-19, and the growing number of national and regional initiatives, there is currently no international framework nor common policy guidance for open science globally. During the 40th session of UNESCO's General Conference, 193 Member States tasked the Organization with the development of an international standard-setting instrument on Open Science in the form of a UNESCO Recommendation on Open Science (4). The Recommendation is expected to define shared values and principles for Open Science and identify concrete measures on Open Access and Open Data, with proposals to bring citizens closer to science and commitments facilitating the production and dissemination of scientific knowledge around the world.

IV. Open Science for Asia and the Pacific

In looking for a political platform to advocate for Open Science, ISC Regional Office for Asia and the Pacific (ROAP), together with the Academy of Sciences Malaysia (ASM) and the Ministry of Science, Technology and Innovation (MOSTI) Malaysia, decided to work through the Asia-Pacific Economic Cooperation (APEC) Policy Partnership on Science, Technology and Innovation (PPSTI) Working Group, which aims to strengthen the STI ecosystem and connectivity network in an open, accessible, and effective way to accelerate knowledge sharing and technological collaboration (5).

The initiative will strengthen the capacity of practitioners in APEC Economies in applying: (i) best practices in APEC Economies and international policies for platform governance; and (ii) a joint statement to promote engagement and knowledge sharing among APEC Economies through Open Science. This will enable Economies to participate more fully in the global scientific enterprise since an Open Science approach would assist in bridging the gap of knowledge divide between APEC Economies.

The key objectives to the initiative include:

- Building a community of purpose that works together to identify platform governance measures and technical know-how;
- Sharing of best practices in APEC Economies and international policies for platform governance;
- Developing an APEC PPSTI Statement on Open Science to promote engagement and knowledge sharing among APEC Economies through Open Science.

V. Way Forward

Open Science is a necessary transformation to the scientific practice in adapting to the changes, challenges and opportunities of the 21st century digital era to advance knowledge and to improve our world.

With UNESCO currently developing its Recommendation, it is imperative for all stakeholders to engage with and contribute towards the consultations in defining shared values and principles for Open Science. The ISC will continue to champion collaboration and strengthen partnerships with the multi-stakeholder members involved in the UNESCO Global Open Science Partnership.

APEC PPSTI has endorsed the Open Science Statement and ISC ROAP will continue to push the Open Science agenda through PPSTI. This will include future capacity building initiatives amongst APEC Economies and beyond.

References

1. ISC Action Plan 2019 - 2021. *International Science Council.* [Online] https://council.science/actionplan/

2. ISC Action Plan, Open Science. *International Science Council.* [Online]

https://council.science/actionplan/open-science/

3. African Open Science Platform: Strategy and Vision. *CODATA*. [Online] https://codata.org/initiatives/strategic-programme/african-open-science/

4. Open Science. *UNESCO.* [Online] https://en.unesco.org/science-sustainable-future/open-science/recommendation

5. Policy Partnership on Science, Technolpogy and Innovation (PPSTI). *Asia-Pacific Economic Cooperation (APEC).* [Online] https://www.apec.org/Groups/SOM-Steering-Committee-on-Economic-and-Technical-Cooperation/Working-Groups/Policy-Partnership-on-Science-Technology-and-Innovation

Past, present and future of global data collaboration

Mark A. Parsons[1*]

[1*] CODATA Data Science Journal
https://orcid.org/0000-0002-7723-0950
parsons.mark@gmail.com

Summary. International data collaboration has been going on for centuries, but it really exploded with the dawn of the digital age and especially in the last decade. Although we continue to grapple with many of the same issues, increased mediation has allowed us to deal with ever increasing complexity through additional layers of abstraction. Global data collaboration is more critical than ever.

Keywords. Global data collaboration, standards, mediation, socio-technical relationships, FAIR Digital Objects

With global data collaboration, you are never done. Consider biological classification, a data standardization exercise we have been pursuing for centuries. Linnaeus published his *Systema Naturae* in 1735. Darwin published *On the Origin of Species* in 1859 changing conceptions of what a species is. Today, as genomics again transforms our conception of species, we continue to work on the Encyclopedia of Life, which was first released in 2007. At the same time, we are increasingly recognizing how global standards can obscure valuable local knowledge [1]. Biological classification, it turns out, is really difficult, and it is but one example of the challenges of data collaboration.

This general desire to standardize contributed to the establishment of many international scientific organizations in the late 1800s and early 1900s. These organizations were the precursors of many contemporary organizations that are grappling with data issues today. It was not until the post-war era, however, that organizations emerged specifically devoted to collaborative issues around data, notably the World Data Centers System in 1957 and the Committee on Data (CODATA) in 1966. Both under the auspices of what is now the International Science Council. Then, of course, with the birth of digital data and especially the internet, things started changing very rapidly, and global collaboration became increasingly imperative. The late 1990s saw the growth of the open data and open access movements, and initial realization that data stewardship was essential to science and society.

With the dawn of the new millennium, we began to see changing expectations in how science was conducted. In 2004, the International Council of Science conducted a comprehensive review of its efforts related to scientific data and information [2], and the United Nations launched the World Summit on the Information Society in 2003. In 2007, Jim Gray put forth his vision an entirely new “fourth paradigm” of scientific exploration [3]. Gray called for new tools and new infrastructure to enable this data-driven science, and the global community has responded.

In the last decade, there has been an explosion of collaborative data activity. In 2009, the World Data System (WDS) and CODATA were reformed and reinvigorated to meet the new needs and expectations of data-driven science [4]. In 2011, the FORCE11 manifesto was published for "Improving future fesearch communication and e-scholarship" [5]. The Research Data Alliance (RDA) launched in 2013 to enable the necessary data infrastructure and did much to create and unite myriad data collaboration communities [6]. In 2016, a large influential community published the FAIR (Findable, Accessible, Interoperable, Reusable) principles [7], which have created a common framework for discussion on the detailed specifics of machine interoperability and a focus on implementation through the GOFAIR initiative.

Today, the four major international data organizations (CODATA, GOFAIR, RDA, and WDS) have outlined their "joint commitment to work together to optimize the global research data ecosystem and to identify the opportunities and needs that will trigger federated infrastructures to service the new reality of data-driven science" in a joint "Data Together" statement[1]. This has resulted in the Virus Outbreak Data Network (VODAN) and the RDA COVID-19 Recommendations and Guidelines on Data Sharing [8]. These are important and timely accomplishments, but if the COVID-19 pandemic has shown us anything, it is that we have a long way to go before we have robust and open data sharing and collaboration.

Just like the 19th century promise of standardization was never fully realized, the openness of the 20th century remains a challenge. Now as we focus on FAIRness to enable machines to automate the exploration and exploitation of massive data, we continue to struggle with many of the same challenges while some of the old challenges now seem trivial as they have disappeared under layers of abstraction.

We are dealing with much greater complexity of both data and the insights they reveal, but we have dealt with this through ever increasing layers of complex mediation [9]. The nascent concept of precisely specified and identified "FAIR digital objects" is the latest example of a conceptualization that allows a new layer of cyber-mediation, and it shows great promise[10]. Nonetheless, socio-technical relationships among people, among machines, and between people and machines remain central issues. The imperatives for global data collaboration are greater than ever.

References

1. Tsing, A., *Friction: An ethnography of global connection*. Princeton University Press, 2005
2. ICSU. *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information.* ICSU, Paris, 2004
3. Hey, T., Tansley, C. and Tolle, K., *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research, Redmond, 2009
4. ICSU, *Decisions, 29th ICSU general assembly, Maputo, Mozambique, 21-24 October 2008.* ICSU, Paris 2009
5. Bourne, P. et al., *Force11 white paper: Improving the future of research communication and e-scholarship*. 2011
6. Lannom, L., Special issue on the research data alliance. *D-Lib* 20, 1-2, 2014
7. Wilkinson, M. D. et al., The FAIR guiding principles for scientific data management

[1] https://www.go-fair.org/wp-content/uploads/2020/03/Data-Together_March-2020.pdf

and stewardship. *Scientific Data,* 3,160018–160018, 2016

8. RDA COVID-19 WG, *RDA COVID-19 recommendations and guidelines on data sharing (version 1.0)*, 2020
9. Borgman, C. et al., *Fostering learning in the networked world: The cyberlearning opportunity and challenge. Report of the NSF Task Force on Cyberlearning.* NSF, Washington, 2008
10. Wittenburg, P, Strawn, G, Mons, B., Boninho, L., and Schultes, E., *Digital objects as drivers towards convergence in data infrastructures*. 2019

Sharing data for a coordinated response during COVID-19 pandemic

Priyanka Pillai[1*],

[1*] *The University of Melbourne, Parkville, Melbourne, Australia, Victoria 3010 Australia*
Email: priyanka.pillai@unimelb.edu.au

Summary. This keynote speech will describe the composition of the infectious disease data ecosystem and highlight some challenges from the past outbreaks associated with building the data ecosystem for a response. This speech will also describe how making data consistent and shareable has strengthened preparedness and response activities in present-day scenario of the ongoing COVID-19 pandemic.

Keywords. COVID-19, Infectious Diseases, Data Management, Data Sharing, Health Informatics

Globally, the health and medical research strategic plans emphasise enhanced data collection, efficient reporting systems and building advanced infrastructure to respond to infectious diseases emergencies in a timely manner. How does data support preparedness towards infectious disease emergencies? What information is needed to identify the start of an outbreak? How does data inform the potential severity and spread of an outbreak? What is the role of a data specialist?

The infectious diseases data ecosystem is comprised of information from a wide range of sources like general practices, jurisdictional surveillance systems, clinical research, emergency departments, diagnostic laboratories, epidemiology studies and genomics. The carefully distilled knowledge from this diverse data ecosystem enables better preparedness for and response towards an outbreak. A data specialist is a bridge or a central point of information for clinicians, public health practitioners, virologists, microbiologists, epidemiologists, public health researchers and policymakers.

Past infectious disease outbreaks have demonstrated several challenges associated with rapid aggregation, integration and sharing of data to inform a response during an outbreak. It is essential to improve data collection, facilitate data sharing and support data usage for decision-making in the infectious diseases community.

The conclusion is that the challenges in sharing and aggregating data can be addressed by building trust among data custodians, promoting collaboration and implementing data stewardship practices. The infrastructure solutions to leverage big data in infectious diseases should be agile, comply with ethical requirements and legislation, facilitate equitable data access and expedite cross-border data sharing globally. There are many lessons learnt from past pandemics that have strengthened the preparedness and response capabilities in the present-day scenario of the ongoing COVID-19 pandemic.

Making historical data Re-useable:
a case study of the challenges and success of Bangladesh Bureau of Statistic's historical data conversion project

Chandra Shekhar Roy [1*]

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

Summary. Bangladesh starts journey focusing on ICT with a view to realize the vision of 'Digital Bangladesh 2021'. Thus, at the long run census and survey data can be used exhaustively in the planning process to transform into digital Bangladesh by 2021. In Bangladesh, several types of official data are released under the Statistics Act. BBS played a significant role in the field of historical Statistical data preservation. After the independence of Bangladesh in 1971, there was a rich repository of statistical microdata in IBM 360 to ES/9000 model mainframe tapes. Almost 8600 nine-track ½ inch spool tapes were used to preserve those data. Recently BBS has converted all those data from EBCDIC format to ASCII format. Near about 165 data set recovered which has been declared as digital asset. Since independence, 2,391 of BBS surveys and census publications have been digitized and converted to e-book system. Alternative back-up for data reservation 200 kilometers away from NSO.

Keywords. Historical data, Data preservation, Statistical microdata, Statistics Act.

1. Introduction

1. a. importance of historical data for analysis

Bangladesh Bureau of Statistics (BBS) has vowed to convert Census/Surveys micro data for the next generation technology format processed in IBM proprietary technology. 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 Sustainable Development Goals (SDG).

1.b. specific data and their value (who will use & why they should use it)

The overarching objective is to strengthen the prevailing national statistical archiving system, academic and scholarly debates can take place taking cognizance of historical data. By revisiting time series data, it is hoped that well-informed and meticulous policies can be designed and formulated in the future. Several types of legacy micro-data are released across a variety of formats and based on data confidentiality and user needs.

2. The process of data conversion

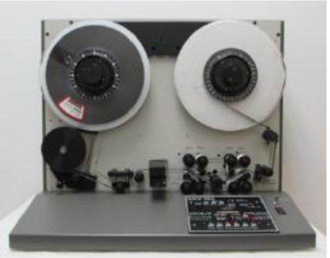


Figure 1: 9-track tape cleaner

2.a. Institutional challenges

- Total 8600 of 9-track spool data tapes are needed to be read and convert.
- Technology outdated
- More than 20 years old backup magnetic-tape
- Historical data from 1972 to year 2019 is needed for developing time series data

2.b. technical challenges

- In-depth understanding of the functionality of the COBOL code structure.
- Source data quality, if poor, needs to be cleansed in order to be successfully converted.
- Metadata preparation and data cataloguing.
- Microdata documentation.

2.c. the success of the project

- IBM mainframe backward compatible technology exists;
- EBCDIC to ASCII conversion software is available in the USA;
- All the backup data was stored in structured way (from year 1972 to 1999);
- All the statistical published reports (2391 e-book & .pdf) have been digitized (from year 1972 to 2018);
- A total of 165 Microdata sets including 50 censuses and 115 surveys have been created/transformed with the help of existing COBOL programming and BBS publication. 159 set of legacy-data documentation has been prepared for BBS data portal. It is fully human intervention with tedious job. Some Cobol programmers were associated with this activities.

3. History of data publication of BBS

Since the post-independence period, BBS has been compiling various census and survey based statistics and conducting surveys and publishing various reports / publications. All these published reports could not be saved properly due to lack of proper technology. Reports published by BBS are available only in official Library. Due to insufficient copies of these publications, they cannot be made available to the ordinary people and these valuable publications cannot be provided as per the demand of local and foreign students-teachers, researchers, stakeholders and various institutions.

In the interest of the overall development of the country, there is a need for historical information on determining the course of socio-economic conditions and policy making. There is no alternative to digitization of all the reports published by BBS to meet the demand for all these valuable information needs. Moreover, digital version of information / publication is now the time demand in building a digital Bangladesh.

Since independence, 2,391 of BBS surveys and census publications have been digitized and a web based software titled 'Digitization of BBS Publications' has been developed and data searching optimization have been incorporated.

Consideration of launching the digitization of BBS Publication started in April, 2015. all those digital version reports have been converted to e-book system and integrated into web based software.

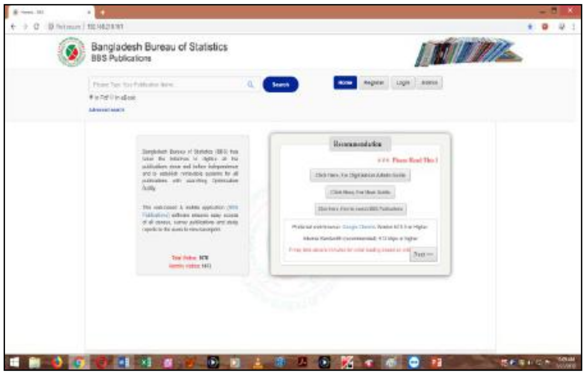
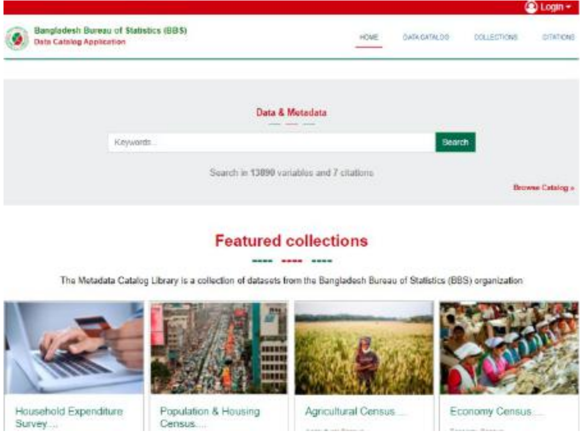


figure: 2 e-book website in BBS domain.

4. Statistical description/metadata of data that were restored

Census & Survey Metabase= (Metadata + Database) for big data analysis Two interoperable database and application have been developed by the country local vendor. A web portal called "data.bbs.gov.bd" has been created and following developing tools has been used.

Linux; PHP, Codigniter framework, Metabase; Apache; Oracle, MySQL; HTML reporting.



Search in 13890 variables and 7 citations

The Metadata Catalog Library is a collection of datasets from the Bangladesh Bureau of Statistics (BBS) organization

Figure 3 data & metadata portal (data.bbs.gov.bd)

5. Digital preservation (DP) in earthquake free area in the country

Jessore is one of the remote districts of Bangladesh which is internationally declared as earthquake free zone

Three tier data backup called-

- Online storage, both in NAS and adequate storage server has been installed
- Online digital archiving, 50GB capacity optical blue ray disk has been used for the purpose.
- Offline LTO-5 tape backup, 1.5 TB capacities is being used for off-the- shelf data backup. Low power consumption has been ensured by calculating upcoming data volume for five years, while the small footprint can help keep capex low. Initially 80 terabyte storage capacity has been installed with one tape backup facility. The following initiatives have been taken for DPC.

1)On-grid Solar power backup system; 2) In-house power generator; 3) Intelligent human Access Control System; 4) Fire suppression system for DPC; 5) C^3- S.P.E.A.R Data preservation unit. (next generation data Centre solutions); 6) Power and environment control EMS unit.; 7) NSO data center to DPC (P2P) link.

Fig: 4

6. Future work

Data community in Bangladesh is realizing it as 'Digital Assets' of the country. This project can be shared in the international community both data rescue, Big Data and dissemination groups.

- DDI can be initiated in BBS where 47 years of data are available.
- SDMX can be implemented easily by using all those data along with current data.
- Data preservation is a vital issue in a developing country. Special attention may need to support in this data-security field.

7. Conclusions

The methods used for data conversion for official statistics in Bangladesh are largely based on those developed in USA. In order to promote the secondary use of official statistics in Bangladesh, further research into data sharing methods for official statistical data should be pursued.

References

1. Bangladesh bureau of statistics, http://www.bbs.gov.bd
2. BBS Data portal, http://data.bbs.gov.bd
3. Digitization of BBS Publication, http://203.112.218.73:8082
4. Online data Dissemination portal, http://redatam.bbs.gov.bd

Application of machine learning techniques to weather forecasting

V. Sakthivel Samy[1]*, **Veena T.**[2]

[1]* *National Centre for Polar and Ocean Research, Goa, 403804, INDIA*
[2] *National Institute of Technology, Goa, 403401, INDIA*
Email: vssamy@ncpor.res.in

Summary. The Polar climate system is paid more attention to atmospheric scientific community as it is regarded to be sensitive for anthropogenic induced climate changes, ozone depletion and melting of ice shelf due to global warming. The temperature difference between equator-tropical region and polar region is the major driver of the general circulation of whole atmosphere. The meteorological and polar data sets collected from Polar Regions such as Antarctica, Arctic, Southern Ocean and Himalaya are being archived at National Polar Data Center (NPDC) for ease access and retrieval of the same.

Keywords. Machine Learning, Data Analytics, Antarctica, Long Short-Term Memory

1. Introduction

At present, weather forecasting relies on numerical weather prediction (NWP) models and the following factors: (i) Past weather data (ii) Present atmospheric conditions (iii) Pattern of climatic conditions based on history. But here atmospheric conditions are chaotic, which is non-linear in nature.

In nutshell chaotic nature of atmosphere, complexity involved in numerical weather predictions, assumptions of initial conditions which are prone to error, issues involved in parameterizations make the forecast less accurate. It has been also established by India Meteorological Department (IMD) that no model except the Artificial Neural Network (ANN) predicts the long- range forecasting so accurately districts-wise. ANN's also work effectively for a non-liner input condition [2, 3]. So, the objective of this work is to analyse the significance of machine learning approaches over other traditional methods in weather forecasting.

2. Objective

The proposed application of machine learning methodologies to solve problems and find alternative solutions to weather forecasting tasks. Weather forecasting is currently based on a combination of explicitly-solved physical numerical and empirical models that represent the current state and evolution of the atmosphere. Using the data produced by these models as benchmark, we aim to identify machine learning methodologies that can improve the accuracy of the forecasts and automate activities in the forecasting process.

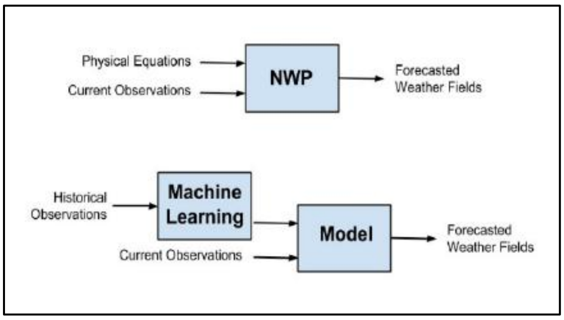


Fig.1: The traditional NWP and machine learning approaches to weather forecasting

ITEMS	DETAILS
Prediction Target	Temperature forecasting for 210 / 240 days
Input Variable	Temperature data of meteorological station. Hourly data is resampled to Daily Data
Training Parameter	Learning rate = 0.001; Optimizer = adam; RMS prop Number of Units: 180,200 ;Number of Epochs: 700, 100

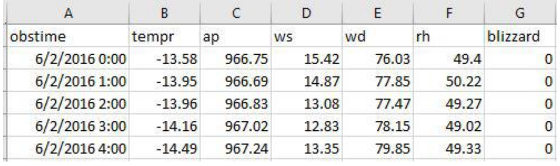
Table 2: LSTM model details

3. Analyzing the time series data

Observations of air temperature, air pressure, wind speed, wind direction, relative humidity, and other several synoptic variables are carried out at Maitri Station (70°45′52″S, 11°44′03″E), Antarctica, and at Bharati Station (69°24′41″S, 76°11′72″E), Antarctica. Using these data, collected over the years 1985-2018(varying for different stations), analysis is done on the data count and visualization is performed through analysis with their Trends, Behavioural, Seasonal and Wavelet Spectrum Analysis. For this study we have used the observations carried out at Bharati station, collected by India Meteorological Department (IMD), New Delhi. These records are available from National Polar Data Center (http:/npdc.ncaor.gov.in). The records used in this study cover the period Feb 2015 to Feb 2020.

Identifying and eliminating anomalies in the data: The statistical data of hourly TEMP (Temperature), WS (Wind Speed), WD (Wind Direction) and AP (Air Pressure) time series used in this study.

Table 1: Sample data format



A	B	C	D	E	F	G
obstime	tempr	ap	ws	wd	rh	blizzard
6/2/2016 0:00	-13.58	966.75	15.42	76.03	49.4	0
6/2/2016 1:00	-13.95	966.69	14.87	77.85	50.22	0
6/2/2016 2:00	-13.96	966.83	13.08	77.47	49.27	0
6/2/2016 3:00	-14.16	967.02	12.83	78.15	49.02	0
6/2/2016 4:00	-14.49	967.24	13.35	79.85	49.33	0

4. Proof of Concept using Long Short-Term Memory (LSTM) Model

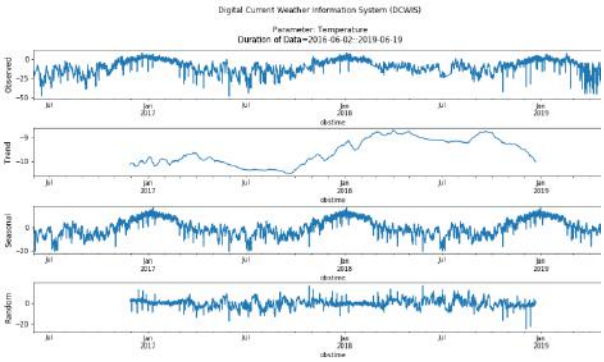
Carried out proof of concept using Long Short-Term Memory model and the predicted results of the LSTM model is depending on the actual collected data at the Bharati Station, Antarctica. The LSTM models use the input data to update numerous values in the internal cell states. However, the LSTM models learn these physical principles during the training and calibration processes from the input data and observed data, and are optimized to forecast the discharges as accurately as possible. The setup details are summarized in Table 2 as given above:-

As per the model setup details are summarized in the above Table 2, the several training options and parameters of the model have changed to get best results. The adaptable parameters of the neural network were also updated depending on a given loss function of the iteration step.

This input data layer must be reformatted into three-dimensional (3D) vectors to match the architecture of the LSTM model. The input vector (3D) comprises of (samples, time steps, and features) with the shape num_samples, num_timesteps, num_features, respectively. The num_samples are the data rows or the total of time steps collected. The num_timesteps are the past observation for the feature, i.e., a lag variable.

The data / input vector is (1131, 240, 1) where 1131 is the total samples, 240 is the time steps and 1 is the feature. Therefore, the sequence is created with 240 data elements in one batch. Hence based on the past observation of 240 timestamps the future values of 240 timestamps is predicted.

Training cases:
Case 1: Training Year: 2016 – 2018 Prediction Year: 2019



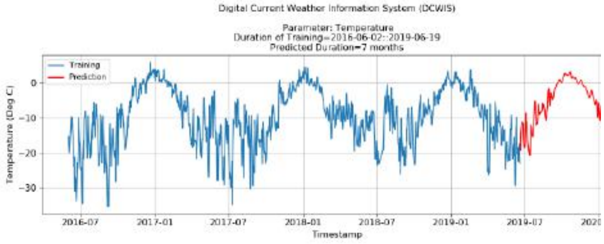
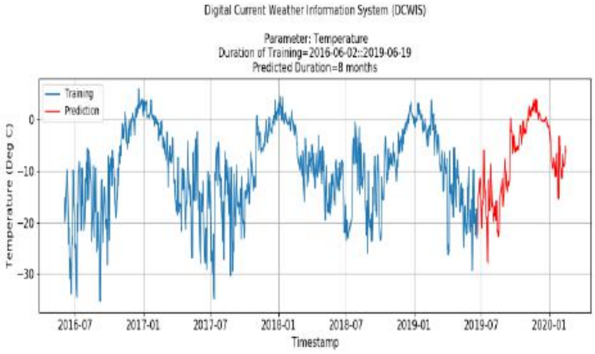
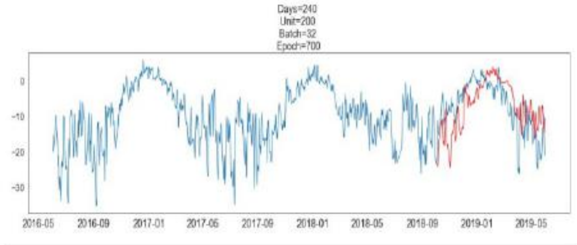


Table 3: Summary of the results and parameters of the day model forecast

Forecast for	Case	Batch Size	No. of Units	No. of Epochs	RMSE
210 days	Case 1	32	200	700	-
240 days	Case 1 Case 2	32 32	200 200	700 500	10.82 12.72





References

1. Shrivastava, G., Karmakar, S., Kowar, M., & Guhatakurta, P., Application of artificial neural networks in weather forecasting: A comprehensive literature review. International Journal of Computer Applications, 51(18), 17–29, 2012
2. Somasundar, Ministry of earth sciences, Data.gov.in, Jan 2017
3. Crosby, D. S., & Ferraro, R., Estimating the probability of rain in an SSM/I FOV using logistic regression. Journal of Applied Meteorology, 34, 2476–2480, 1995

VODAN-in-a-box: a FAIR toolkit for improving reusability of COVID-related data

Luiz Olavo Bonino da Silva Santos[1*,2,3], ***Barend Mons***[1,2]

[1*] GO FAIR International Support and Coordination Office, Leiden, the Netherlands
[2] Leiden University Medical Center, Leiden, the Netherlands
[3] University of Twente, Enschede, the Netherlands
Email: luiz.bonino@go-fair.org

Summary. The COVID-19 pandemic highlighted the need for improvements in data management and reuse to allow for quicker and better understanding of the situation, which leads to better informed decision making and ultimately a better response to the challenges imposed by the outbreak. With the experience of using the FAIR principles as guidelines to tackle these issues, the GO FAIR community launched the Virus Outbreak Data Network initiative with the aim of applying the FAIR principles and related technologies to increase the availability of FAIR data and services focusing on the needs of virus outbreak in general and the SARS CoV-2 pandemic in particular. The first result of VODAN is the VODAN-in-a-box, a toolkit containing a data entry tool compliant with the World Health Organization's COVID-19 Rapid Case Report Form, a semantic data model for this CRF and a FAIR Data Point to expose the metadata of the dataset containing the case reports.

Keywords. FAIR, data, interoperability

1. Introduction

Since the publication of the FAIR principles [1] in March 2016 they have been used as guidelines for improving data stewardship. From the initial period of the 2020 pandemic of SARS-CoV 2, researchers, decision makers, health care providers and to a certain extend the world population struggled with the difficulties in finding, accessing, interoperating and, ultimately, reuse data related to the outbreak. These data were, and still are, needed for many different purposes including early detection of the outbreak, understanding of the biological implications of the infection and discovery of preventive and curative measures.

The aforementioned difficulties are, not coincidently, the very reason for the emergence of the FAIR principles. With this realisation, the GO FAIR community, with is inherent bottom-up approach, launched the Virus Outbreak Data Network (VODAN) [2,3] initiative to reuse the existing FAIR-related methods and technologies to help improving the availability of FAIR data and services relevant to the reaction against virus outbreaks.

The first deliverable of VODAN is the VODAN-in-a-box, which is a toolbox consisting of:

- A data-entry application named CRF Wizard to allow health care providers to enter the information about their COVID-19 cases. The CRF Wizard uses the case report form (CRF) template defined by the World Health Organization (WHO) for the COVID-19 pandemic.

- A semantic data model for the WHO COVID-19 CRF. This semantic data model provides machine-actionable semantic information about the questions and answers contained in the WHO CRF. The CRF Wizard uses this semantic data model to transform the data entered in its web form into properly annotated RDF, which is stored in a triple-store.
- A VODAN FAIR Data Point (FDP). The FDP is a tool to expose FAIR metadata about datasets. The VODAN FDP is the component of the toolkit that exposes the metadata of the available COVID-19-related datasets. The CRF Wizard is configured to interact with the related VODAN FDP so that, when a new case is reported, the metadata of the dataset is automatically updated.

The VODAN-in-a-box is available under the MIT license and can be deployed using the available Docker images. The documentation of the toolkit can be found at https://docs.vodan.fairdatapoint.org/en/lastest.

2. Adoption of the toolkit

As of September 11 2020, the VODAN-in-a-Box has been already deployed in 8 locations in 6 different countries. More locations are in the process of validating the solution for use. Health care personnel are being instructed to enter the information of their case reports.

One characteristic of the VODAN-in-a-box that has been mentioned as decisive for its adoption is the deployment freedom. Health-related data are highly sensitive. Even though the WHO COVID-19 CRF does not contain many identifiable data, the fact that the whole toolkit can be deployed on-premise gives confidence to the hosting entity that their data is kept under control, so that compliance with GDPR (or other privacy-related regulations) can be guaranteed.

3. Future work

The work around VODAN-in-a-box will continue to not only support other entities to deploy and use the toolkit but also to expand the package with semantic models of other types of COVID-19-related datasets, richer metadata schemas and improvements in the findability of the indexed metadata.

The adoption of the toolkit already demonstrated improvements in some aspects of data stewardship such as metadata management, (meta)data findability and interoperability, among others. However, to further improve on interoperability and reusability we need to adapt and develop end-user tools that can make use of this emerging FAIR infrastructure.

4. Conclusions

The experience with the development, deployment and usage of the VODAN-in-a-box demonstrated that the realisation of the FAIR principles in terms of a toolkit supporting the improvements of data stewardship is feasible.

Work will continue on improvement of the existing tools as well as on expansion of the toolkit, metadata schemas, semantic data model and end-user applications.

Acknowledgments. We would like to acknowledge the contribution of ZonMW and the Philips Foundation for funding the initial

development and deployment support of the VODAN-in-a-box.

References

1. Wilkinson, M., *et al*, The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9, 2016
2. The Virus Outbreak Data Network (VODAN): https://www.go-fair.org/implementation-networks/overview/vodan
3. Mons, B., The VODAN IN: support of a FAIR-based infrastructure for COVID-19, *European Journal of Human Genetics*, 28, 724-727, doi: 10.1038/s41431-020-0635-7, 2020

“A Comparison between Logistic Regression and Decision Tree Methods for Predicting the Preterm Birth"

Rakesh Kumar Saroj[1]*, Madhu Anand[2]

[1]*_Department of Mathematics and Statistics, SRM University Gangtok, Sikkim/Gangtok -737 102, India_
[2]_Department of Chemistry, Dr. B.R. Ambedkar University, Agra, India_
Email: rakeshkumarsaroj.b@srmus.edu.in

Summary. Pre-term birth is an increasingly prevalent complex condition with multiple risk factors including environmental pollutants. This is also leading cause of neonatal death of developing countries. Machine learning techniques play a key role in predicting the factors in the pre-term birth.
The objective of this study is to compare the performance of logistic regression and decision tree classification methods and to find the significant determinants that cause pre-term birth.
For this study we have used a case–control study of 50 cases of full-term births and 40 cases of pre-term births. We have been tested the logistic regression and decision tree classifier methods in this dataset and to evaluate the accuracy of the logistic regression and decision tree methods through various indices.
The logistic regression is determined to be suitable classifier model for this dataset as compare to decision tree methods. The variables like alpha HCH, total HCH and MDA are the most influential factors with respect to association with preterm birth. The final result revealed that logistic regression classifier is accurate model to predict the pre-term birth with better accuracy.

Keywords. Machine learning, Pre-term birth, Classifier, Decision tree, Logistic regression

1. Introduction

Preterm birth is the birth of a baby at less than 37 weeks' gestational age, as opposed to the usual about 40 weeks. Preterm birth (PTB) has found through various reason including low socio-economic status, smoking, race and consumption of alcohol [1]. Previous research have suggested the associations between organochorines and increased risk of abortion, small for gestational age babies, minor malformations, , cryptochildism and hypospadias in the infants have been reported [2-3] There are various machine learning prediction and classification models like regression, logistic regression, principal component analysis, decision tree and maximum likelihood method have been used to improve maternal and child health. The machine learning scientists have worked on interpretable prediction techniques in several places [4-6]. The previous research article recommended that if early risk factors are identified through classification methods in acute kidney injury then can be

Overall aim of this study to construct the logistic and classification models to predict the high-risk groups for PTB based on several factors and covariates in order to reduce the risk of PTB and compare the predicting performance of LR and DT classification methods.

2. Material and Methods

For this study we have used a case–control study of 50 cases of full-term births and 40 cases of pre-term births at Dr. Bhim Rao Ambedkar University (located at Agra, Uttar Pradesh, India).

The dependent variable in this study is preterm birth and it is classified between full term birth and pre-term birth. The risk factors considered in this study include variables age, BMI, number of children, lactation duration, addiction, residence, pesticide exposure, drinking water sources; dietary habits baby gender and organ-chlorine pesticides in the placenta of the females.

3. Statistical Analysis

In this research paper, we used the decision tree model in whole dataset and train datasets because decision tree creates a binary tree and it is very useful in classification problems. The next step is measure to rank of the variables from the data sets through information gain technique. Finally, we have used the decision tree and logistic regression both classifier model in the training dataset (70% of cases) and evaluate the model using the test sample (30% of cases).

4. Results

In this study the variable rank shows in the Figure 1 using information gain measure. Higher values of information gain indicate those variables are important for preterm birth. The alpha HCH, total HCH, MDA, p_p_DDT, beta HCH and p_p_DDE are the highly ranked variables. In other words, it declares that these variables play important role in preterm birth. The higher rank variables like alpha HCH, total HCH and MDA are the most influential factors with respect to association with preterm birth. We developed the logistic and decision tree classification model for prediction with the help of 70% of training dataset including all variables. The logistics and decision tree models are tested on 30% testing dataset and model evaluation results are given in Table 1.The result shows that logistics classifiers with the better accuracy of predictions compared to decision tree. The accuracy of logistic regression for classifying preterm birth is 0.96, significantly different from the decision tree method. The comparison of the performance of the both classifier models include all variables reveals that logistic regression performs the better in terms of metrics (precision = 0.92, F1-score = 0.96 and AUROC = 0.97), while decision tree performs the poor (precision = 0.75, F1-score = 0.86 and AUROC = 0.87).The Figure 2 shows the receiver operating characteristic (ROC) curve for all variables and it is found that logistic regression model is better than decision tree model.

After that we have used the top six ranked variables and apply the both model. The reason behind repeat the same experiment with top six ranked variables is to emphasize how efficient is the use of information gain measures in the data. The results obtained using the top six variables are given in Table 2. The result found that top six variables do not affect the accuracy performance of decision tree method and again the accuracy of logistic regression for classifying preterm birth is 0.78, significantly different from the decision tree method. The performance gain is shown by logistic regression (precision = 0.65, F1-score = 0.79 and AUROC = 0.81) and decision tree performs (precision = 0.56, F1-score = 0.69 and AUROC = 0.71). In figure 3, we present ROC curves for both classifiers models with the help of top six ranked variables. It shows that that logistic regression model is again better than decision tree model.

5. Discussion

While the final objective of specialist classification/prediction machine learning technique development is to predict preterm birth risk, the definition of preterm birth and data needed to analyze preterm birth risk are less amenable to study presently. Therefore, the purpose of this study is to determine the feasibility of using classification/prediction machine learning to generate expert system (knowledge-base) rules for prediction of preterm birth. In this study we use the logistic regression and decision tree method on preterm birth data and it is found that LR is the most accurate in terms of predicting preterm birth. With this method, scientists, researchers and practitioners are able to predict and detect the preterm birth that is at a higher of data sets.

6. Conclusion

This study is limited due to small amount of data because the machine learning techniques give better result in big datasets. The result finds that preterm birth is significantly associated with only γ-HCH and MDA. Finally the results revealed that the logistic regression is better accuracy in classifying preterm birth compared to the decision tree method.

References

1. Metzger, M. J., Halperin, A. C., Manhart, L. E., & Hawes, S. E., Association of maternal smoking during pregnancy with infant hospitalization and mortality due to infectious diseases. The Pediatric infectious disease journal, 32(1), e1,2013
2. Birnbaum, S. C., Kien, N., Martucci, R. W., Gelzleichter, T. R., Witschi, H., Hendrickx, A. G., & Last, J. A. Nicotine-or epinephrine-induced uteroplacental vasoconstriction and fetal growth in the rat,Toxicology, 94(1-3), 69-80,1994
3. Hosie, S., Loff, S., Witt, K., Niessen, K., & Waag, K. L., Is there a correlation between organochlorine compounds and undescended testes.European Journal of Pediatric Surgery, 10(05), 304-309,2000
4. Jacob Bien, Robert Tibshirani, et al. Prototype selection for interpretable classification. The Annals of Applied Statistics, 5(4):2403–2424, 2011
5. Emilio Carrizosa, Amaya Nogales-G´omez, and Dolores Romero Morales. Strongly agree orstrongly disagree?: Rating features in Support Vector Machines. Information Sciences,329:256–273, 2016
6. Amin Emad, Kush R Varshney, and Dmitry M Malioutov. A semiquantitative group testing approach for learning interpretable clinical prediction rules, In Proc. Signal Process. Adapt. Sparse Struct. Repr. Workshop, Cambridge, UK, 2015

.

Table 1. Evaluation of classification models using all factors.

Performance metrics	Model	
	LR	DT
Accuracy	0.96	0.85
Sensitivity (Recall)	1.00	1.00
Precision (Positive predictive value)	0.92	0.75
F1-score	0.96	0.86
AUROC	0.97	0.87

Table 2. Evaluation of classification models using the important factors.

Performance metrics	Model	
	LR	DT
Accuracy	0.78	0.67

Sensitivity (Recall)	1.0	0.91
Precision (Positive predictive value)	0.65	0.56
F1-score	0.79	0.69
AUROC	0.81	0.71

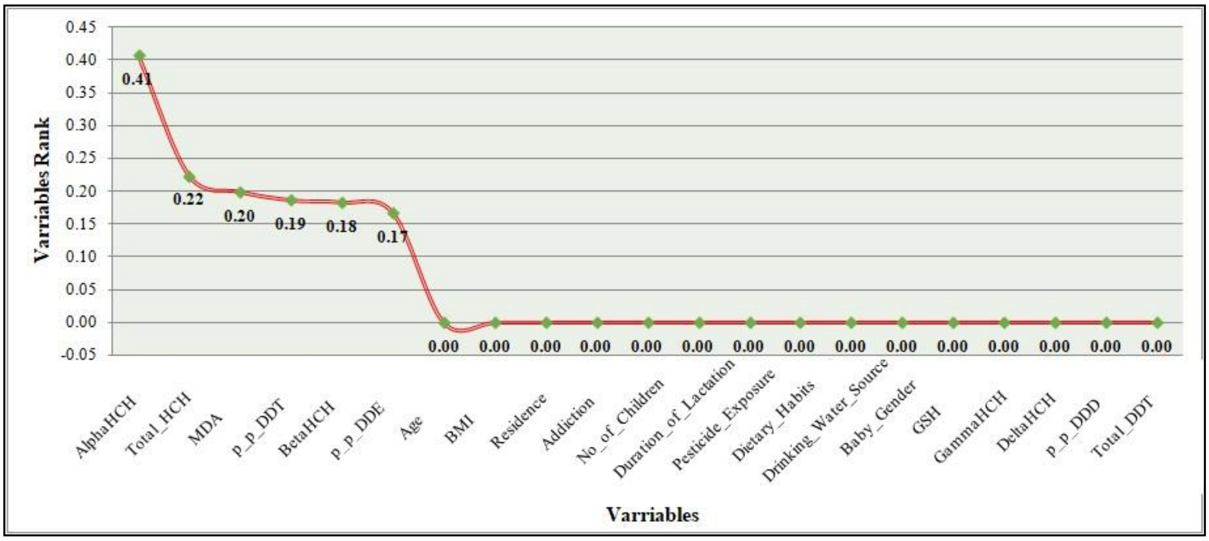


Figure 1. Variables ranking according to information gain

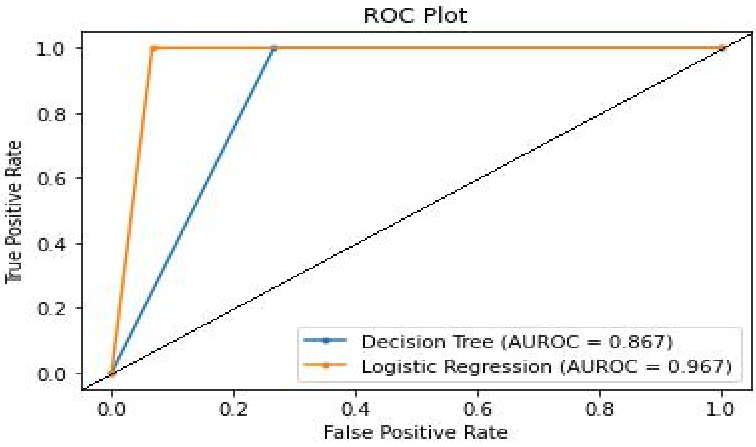


Figure 2 . Area under the curve analysis of the logistic regression and decision tree methods using all factors

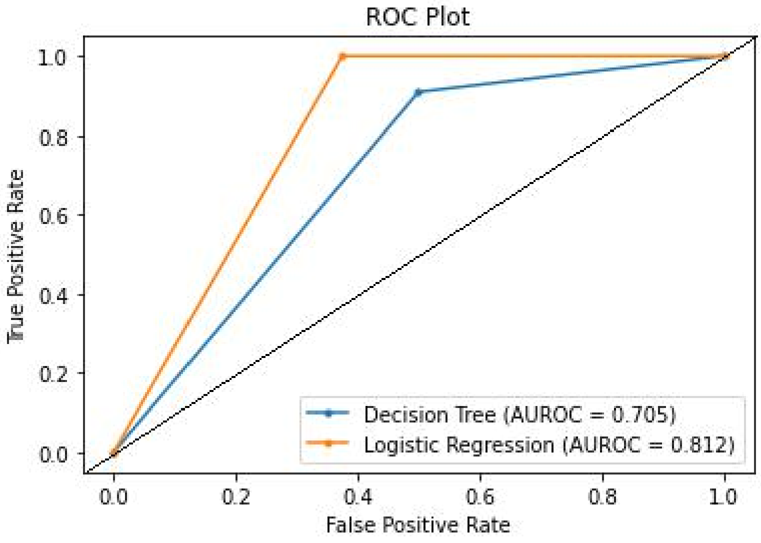


Figure 3. Area under the curve analysis of the logistic regression and decision tree methods using important factors

Machine Learning based Peptide Therapeutics for Covid-19

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. The work aims to target the main protease enzyme (M^{pro}) of SARS-CoV-2, a variant of coronavirus causing disease Covid-19. As we all are aware that the current pandemic situation has been caused by SAR-CoV-2 and due to the presence of large number of mutants, development of an effective drug or vaccine is not yet possible. Thus, we studied these diverse mutants of SARS-CoV-2 virus to identify conserved sequential motifs in main protease enzyme (M^{pro}). M^{pro} is a key enzyme responsible for maturation of many proteins or transcription factors from just two polypeptides, hence acting as a very promising target. We applied a novel technique of combinatorial machine learning algorithm with an optimization algorithm to design novel set of peptides (library of 97 peptides) which on one hand are conserved in coronavirus species and on the other hand are non-toxic to humans. The computational studies have confirmed the good binding affinity of two peptides with M^{pro} protein. Thus, blocking the functionality of M^{pro} may result in regulating the replication of viral RNA. Also, by studying the evolvability of virus of all the mutants available till date, machine learning strategy ensures the efficiency of the peptides against viral variants that might arise in future. In brief, the machine learning and optimization approach used by us might provide an efficient way to combat against SARS-CoV-2 virus that warrants further experimental validation.

***Keywords.*.**

Korea Research Data Management Platform and PID

Sa-kwang Song+1*

1+*Korea Institute of Science and Technology Information, Research Data Sharing Center, 245 Daehak-ro, Seogu, Daejeon, 34141, Republic of Korea*
1* *University of Science and Technology, Department of Data and HPC Science, 217 Gajeong-ro, Yuseong-gu, Daejeon, 34113, Republic of Korea*
Email: esmallj@kisti.re.kr

Summary. Recent rapid increasing interest on research data in Korea makes the government to prepare for constructing research data platform designated for Korean R&D environment as well as the regulation related to sharing and utilization of research data. Korea Institute of Science and Technology Information(KISTI) has been developing Korea research data platform, DataON, since 2018 and opened its first version to the researchers in Korea in Jan 2020. it employed three persistent identifiers for datasets, researchers, and projects, which are based on DOI, Researcher Number and Project Number of NTIS(National Institute of Science and Technology Information, Korea). For the global users, the DataON allows ORCID to be added in the metadata as well.

Keywords. Research Data Management Platform, PID, DOI, Data Management Plan, ORCID

1. Introduction

Considering the recent research trends such as Open Science, Big Data, and Artificial Intelligence, various movements are underway in the sharing and utilization of research data in various countries, including EU's OpenAIRE, Australia's ARDC, Japan's NII, etc. In response to the change to the global fourth-generation research paradigm, the Korean government recognized the importance of data and has implemented various policies. Especially, domestic interest in the open science paradigm for easy access and utilization of publicly funded research results has been rapidly increasing. As a result, led by the Ministry of Science and ICT in Korea, a strategy to promote sharing and utilization of national research data has been promoted in 2017, and the National Science and Technology Research Association in Korea has also been preparing a plan to promote the conversion of research data into big data since 2018. In order to meet the these kinds of demands of the times, Korean government has been promoting legalization related to research data from 2018 and establishing DataON, a national research data platform, and officially opened the first version in January 2020.

2. DataON: Research Data Management System

DataON has been designed to gather and utilize research data managed by Institutional Data Registries(IDRs) or Domain specific Research Data Platforms(DRDPs) in various fields such as bio, materials, chemistry, geology, etc. since 2017. It is an aggregated research data management environment to preserve, share, and reuse research data that is especially essential for multidisciplinary research. It supports convenient search of research data which is distributed both domestically in IDRs & DRDPs and overseas aggregators such EU's OpenAIRE, Australia's ARDC, Japan's NII, etc. Moreover, it provides secure sharing and collaborative analysis environment for supporting individual and But all information created during research, including raw, intermediate, and final data could be part of its convergent research.

2.1 Government Strategy

DataON has started based on the government strategy entitled "Research data sharing and utilization strategy " by the ministry of Science and ICT in 2017. Its goal is to promote sharing and utilization through knowledge capitalization of national research data. The strategy comprises four key parts including changing legal system, fostering human resource, constructing infrastructure, activating research communities related to research data.

2.2 Research Data

Definition of Research Data in Korean regulation is the essential and objective factual data for reproduction of research results, obtained through various experiments, observations, surveys, and analyses of national R&D projects (Regulations on management of national R&D projects, Article 25 Section 28). But all information created during research, including raw, intermediate, and final data could be part of it.

2.3 Data Management Plan(DMP)

DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed. The main elements of DMP in the Government Policy are

- Summary(Data Type, Producing Method, etc.)
- Storing & Preserving Plan (Storage, Management, Backup, Method)
- Sharing Plan(Scope, Method, Time, DOI)
- Management & Sharing Director (Name, Contact)

The ministry of Science and ICT has three project management organizations which have planned to apply the DMP process to 300 or more projects by the end of 2020.

3. PIDs on DataON

Actually, DataON is the first step of its roadmap. However, it employed three persistent identifiers for datasets, researchers, and projects though it needs further elaborate design.

As a PID for datasets, DataOn adopted DOI identification because KISTI is the first DOI RA(registration agency) in Korea. Whereas Researcher Number of National Institute of Science and Technology Information (NTIS) is adopted as a PID for researcher identification in Korea. The Researcher Number is a unique number that is essentially given to participants in the Korean national R&D project created and managed by the NTIS, commissioned by the Korean government to KISTI. For the global users, the DataON allows ORCID to be added in the metadata as well. At last, a unique identifier for national R&D projects, Project Number, has been adopted for persistent identifier for a project.

4. Conclusions

KISTI has been developed Korean research data platform, DataON, since 2018, in which three persistent identifiers for datasets, researchers, and projects has been applied. Each of them are based on DOI, Researcher Number and Project Number imported from NTIS respectively.

Research Across Disciplinary Boundaries: Data Challenges and Solutions in the Environmental and Eco-social Sciences

Alison Specht[1]*, Jeaneth Machicao[2], Shelley Stall[3], Danton Vellenich[2], Pedro Corrêa[2] and the PARSEC consortium[4]

[1]* *School of Earth and Environmental Sciences and TERN, the University of Queensland, 4072, Australia;*
[2] *Department of Computer and Digital Systems, Av. Prof. Luciano Gualberto, 158, University of Sao Paulo, S-P, 05508-010, Brazil;*
[3] *American Geophysical Union, 2000 Florida Ave., N.W., Washington DC, 20009, USA;*
[4] *www.parsecproject.org.*
Email: a.specht@uq.edu.au

Summary. The synthesis of data across disciplinary domains is required to achieve desperately needed understanding about modern, complex, global challenges. Complex questions in today's world such as the challenges of climate change, carry with them a complex range of data, all of which need to be harmonised to permit analysis. In the PARSEC project [1], a team of synthesis and data scientists has been assembled from across the globe. We come from varying disciplines, including data science, the management of protected areas, ecology, remote sensing, AI, sociology, economics, and modelling. We discuss some of our solutions to enable successful collaboration as we: (1) blend data of different scales, time-frames and types; (2) work together across large geographic boundaries; and (3) ensure the data management is of international standard from the beginning of the project to publication.

Keywords. inter-disciplinary, PARSEC, DDOMP, data sharing, workflow

1 Introduction

The PARSEC project (Building New Tools for Data Sharing and Re-use through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas [1]) has a unique cross-cutting objective: to conduct interdisciplinary, transnational synthesis science while developing and testing novel approaches to the management and preservation of environmental and socioeconomic data. PARSEC is funded through the Science-driven e-Infrastructure Innovation (SEI) Collaborative Research Action of the Belmont Forum [2] by four countries, Japan, the USA, France and Brazil, with associates from Australia, the UK and elsewhere. The project provides a unique insight into the challenges of: (1) blending data of different scales, time frames and types; (2) working together across large geographic boundaries; and (3) ensuring data management is to international standards from the start to publication.

2 Challenges and solutions

Our scientific objective is to detect changes in the socio-economic status of communities in the environs of Protected Areas (PA's) in response to the creation of the PA. We are using remotely sensed spatial data, trained using Artificial Intelligence techniques to detect changes in detectable spatial characteristics that have socio-economic meaning.

2.1 Blending data of different scales, time frames and types

To analyse the impact of PAs on nearby regions we are using a data science framework (Figure 1) whose central machinery is a deep neural network model to map inputs to outputs. We are accessing data from several sources such as geostatistical national census data, which is at different spatial and temporal scale and resolution from the remotely sensed data, e.g. from decades for the socio-economic data, to days for the remote-sensed data. The first challenge, therefore, is to disaggregate these heterogeneous data to establish a common granularity which can be used across countries. We need the most fine-grained data available, but there will be various cycles to refine the main goal, and we may need to adjust the granularity to reveal the relationships.

2.2 Working across geographic boundaries.

The project relies on co-ordinated practice across the project teams, who conduct different components of the analysis. Development of algorithms is primarily occurring in France, but the workflows and large data sets need to be shared with the other teams safely, robustly, and in a timely manner. We intend to use Amazon Web Services (AWS) infrastructure, for its substantial working space, AWS Organization to create team space, and AWS Lake Formation for satellite image storage. Amazon EC2 will be used for machine learning experiments, and a GitHub account [4] has been created to handle versioning.

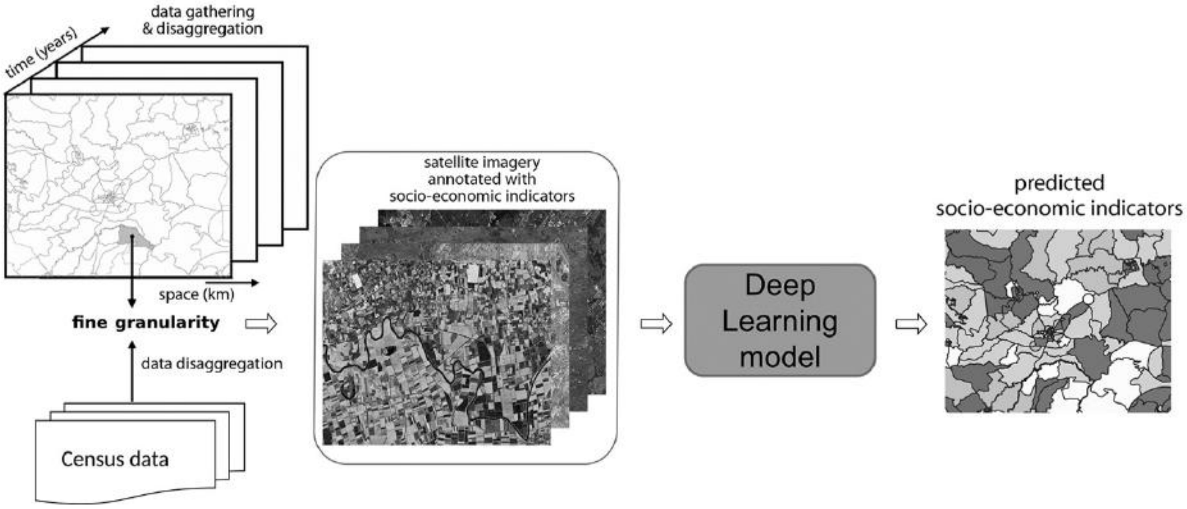

time (years)
data gathering & disaggregation
space (km)
fine granularity
data disaggregation
Census data
satellite imagery annotated with socio-economic indicators
Deep Learning model
predicted socio-economic indicators

Figure 1: Proposed workflow to handle acquisition from different scales and types. A common data granularity, as fine as possible, will be used to annotate the satellite image data, and then inputted into the deep learning model for training and testing.

An important aspect of co-working in a transnational consortium is good communication. As well as email, we have a system of regular Zoom meetings, multiple Slack channels, a dedicated Google Drive and a Zotero community. Time zone differences dictate the use of multiple communication modes (verbal, written, synchronous and asynchronous).

2.3 Ensuring data management is of international standard.

The PARSEC project, as a grantee of the Belmont Forum, is required to implement the new Data and Digital Output Management Plan (DDOMP) format [3, 5]. One of the objectives of PARSEC is to create guidance for researchers that transforms the DDOMP into a living document with guides and templates on how to make decisions concerning creation and management of scientific data. We build upon the existing information provided by the Belmont Forum and created checklists for PIs to give their researchers to follow during the project, suggestions on how to communicate and use resources for better tracking of data and digital objects during the project. As a case study, the PARSEC project has implemented these tools for the use of our researchers. This makes yearly reporting requirements much easier as well as communication across the team. We look forward to sharing these tools with the Global Collaboration on Data community.

3 Conclusions

This project, which exclusively uses existing data, has another two and a half years of life. We expect we shall need to adjust many plans along the way. By following open work practices where we can (we anticipate limitations on some of the socio-economic data), we intend to share what we learn along the way.

Acknowledgements. This research is a product of the PARSEC group funded by the Belmont Forum as part of its Collaborative Research Action (CRA) on Science-Driven e-Infrastructures Innovation (SEI), with partners, the ANR (France), FAPESP (Brazil) and JST (Japan), NSF (US), and the synthesis centre CESAB of the FRB (France).

References

1. Building New Tools for Data Sharing and Re-use through a Transnational Investigation of the Socioeconomic Impacts of Protected Areas <www.parsecproject.org>
2. Science-driven e-Infrastructure Innovation (SEI) Collaborative Research Action of the Belmont Forum <www.belmontforum.org>
3. Bishop, B., Gundermann, H., Davis, R., Lee, T., Howard, R., Samors, R., Murphy, F., Ungvari, J. Data curation profiling to assess data management training needs and practices to inform a toolkit. Data Science Journal, 19(1), 4-11, 2020. doi: 10.5334/dsj-2020-002
4. PARSEC github: https://github.com/PARSECworld
5. Data And Digital objects Management Plan: https://www.belmontforum.org/resources/data-and-digital-outputs-management-plan-ddomp/

Detailed Methodology and Application Guidelines for "The Global Covid-19 Index (GCI)"

Woody Ang Woo Teck[1*]

[1*] *PEMANDU Associates, Level 21, Sunway Putra Tower, Jalan Putra, 50350 Kuala Lumpur, Malaysia*
Email: woody.ang@pemandu.org

Summary. The Global Covid-19 Index (GCI) is a fully data-driven and objective approach to assess the true severity and recovery progress of any given country. It is meant to be instructive to Governments in answering the key questions of what measures have been applied globally, and from there, what appears to be working and which ones do not. The GCI is derived from our very own proprietary algorithms that incorporate all the metrics that truly matter from well recognised and validated open source data. Furthermore, unlike other indexes in the world that are published at a slow frequency, the GCI can be updated LIVE on a daily basis.

Keywords. COVID-19; Index; Severity; Recovery; WHO; Pathfinder

1. Introduction

The Global COVID-19 Index (GCI) was developed by PEMANDU Associates in partnership with the Ministry of Science, Technology and Innovation of the Government of Malaysia. The intent of the GCI was to identify countries that were making positive progress in their battle against COVID-19, and to be able to identify common best practices that were implemented by these countries.

2. GCI – Severity and Recovery Indices

The GCI is intentionally designed to be comprised of 70% dynamic components of which data is updated daily whilst the remaining 30% is comprised of less dynamic parameters that may only be updated once or twice yearly.

2.1 GCI Severity Index

- The GCI Severity Index is designed to exhibit a 'scarring' characteristics so that countries that have been affected badly from a health perspective by COVID-19 can be compared with countries that have been similarly affected to compare their progress in tackling the pandemic. This also enables more relevant comparisons as some countries may have been more successful without having to recover from a significant number of infections and may not be directly comparable with those that were late to react and may be looking for more relevant best practices.

The following key dynamic parameters contribute 70% of the GCI Severity Index:

Confirmed Cases per Population: This parameter is intended to ensure cases are seen in relative to the size of each country. This data is updated daily currently based on Johns Hopkins University's Center for Systems Science and Engineering (JHU CSSE) Open Data sources[1]. However, we will be migrating to utilise the World Health Organisation's (WHO) database instead now that the WHO COVID-19 dashboard is also live. Population data is retrieved from World Bank[2] based on the 2019 revision.

Proportionate Death Rate due to COVID-19: This parameter takes into consideration the crude death rates of each country and factors its population size, and then compares this to the death rates due to COVID-19 since the first case was reported in the country. This gives a true reflection of how death rates are being affected by COVID-19 in these countries. On a daily basis, the number of deaths is cumulatively tracked whilst the denominator will increase by a single day.

- We apply an outlier checks and min-max normalisation methods to apply a relative score to each country in the GCI sample for the above parameters.
- To eliminate outliers, we determine each component's interquartile range (IQR), after which we flag any samples which fall outside either limits. These outliers will be given a score of 0 for those below the lower limit and 100 for those above the upper limit.
- All other non-outlier samples will then be analysed and normalised using the min-max statistical method which will give each country a relative score for each dynamic component, and this enables us to identify countries performing better or worse than their peer.
- The remaining 30% of the Severity Index is comprised of semi-dynamic components which are aimed at enabling a relative comparison on other comprehensive considerations that expose the risk and recovery ability of each individual country. We will cover this in further detail in Section 3.1.

In summary, the GCI Severity Index components are as follows:

$$\begin{aligned}&GCI\ Severity\ Index\\&=\frac{Confirmed\ Cases}{Population}\\&+Proportionate\ Death\ Rate\ due\ to\ COVID19\\&+UHC\ Billion\ +HEP\ Billion+HPOP\ Billion\\&+INFORM\ (Socioeconomic\ Vulnerability\\&+Governance+Communication\\&+Phyiscal\ Connectivity)\end{aligned}$$

2.2 GCI Recovery Index

- The Recovery Index is a measure of how well the country is containing and recovering from the epidemic based on the health perspective and takes into consideration the recoveries, testing efforts, and remaining active cases (the latter of which represents the number of people still infectious) in the country.
- The following key dynamic parameters contribute 70% of the GCI Recovery Index:

Active Cases per Population): This parameter is calculated by removing recoveries and deaths from the total confirmed cases. It also visually enables the representation of the actual flattening and reduction of the COVID-19 epidemic curve. This data is updated daily from the JHU CSSE Open Data repository.

Recoveries per Confirmed Case. This parameter factors a country's success in treating patients that have been diagnosed COVID-19 positive. It is also updated daily based on data from JHU CSSE.

Tests Conducted per Confirmed Case: This reflects how much effort a country has invested in testing. This metric normalises for both mass testing and controlled testing as the lesser that is found for the number of tests conducted would usually indicate that a better recovery to countries which are finding a high number of confirmed positive cases. Test data currently is obtained from Our World In Data[3], an Open Data platform project of the Global Change Data Lab based in the United Kingdom.

Country Measures on Testing and Contact Tracing: We utilise the data from Oxford COVID-19 Government Response Tracker (OxCGRT)[6] by the Blavatnik School of Government of Oxford University. The OxCGRT tracks the ongoing Testing and Contact Tracing Policies implemented by Governments of each country. We have applied a scoring method to the information that the OxCGRT tracks as follows:

Government policies on who has access to PCR testing:

- 0: No testing policy
- 1: Only those who both (a) have symptoms AND (b) meet specific criteria (eg key workers, admitted to hospital, came into contact with a known case, returned from overseas)
- 2: Testing of anyone showing Covid-19 symptoms
- 3: Open public testing (eg "drive through" testing available to asymptomatic people)

Government policies on contract tracing after a positive diagnosis:

- 0: No contact tracing
- 1: Limited contact tracing; not done for all cases
- 2: Comprehensive contact tracing; done for all identified cases

- As of 20th August 2020, at least 29 WHO member states score a maximum score for this component:

  Australia Austria Bahrain
  Canada China Cuba
  Cyprus Denmark Djibouti
  Dominica Gabon Germany
  Ghana Greece Iceland
  Kazakhstan Kenya Luxembourg
  Malaysia Mauritius Qatar
  Russia Rwanda San Marino
  Saudi Arabia South Africa South Korea
  Switzerland United Arab Emirates

- For all the components above, the same outlier checks and min-max statistical methods for the GCI Severity Index as described in Section 2.1 were applied to obtain a relative score for each country in the GCI sample for the above Recovery Index Parameters.
- The remaining 30% of the Recovery Index is comprised of semi-dynamic components which reflect the inherent abilities of the country's healthcare system to respond to the pandemic. We will cover in further detail in Section 3.1.

  In summary, the GCI Recovery Index components are as follows:

$$\begin{aligned}&\text{GCI Recovery Index}\\&=\frac{\text{Active Cases}}{\text{Population}}+\frac{\text{Recoveries}}{\text{Confirmed Cases}}\\&+\frac{\text{Tests Conducted}}{\text{Confirmed Cases}}\\&+\text{Country Measures on Testing \& Contact Tracing}\\&+\text{UHC Billion }+\text{HEP Billion}+\text{HPOP Billion}\\&+\text{INFORM (Socioeconomic Vulnerability}\\&+\text{Governance}+\text{Communication}\\&+\text{Phyiscal Connectivity)}\end{aligned}$$

3. Upcoming Improvements in the GCI Methodology

We have taken on board many recommendations based on the discussions with the World Health Organisation team of data experts. This section will elaborate all new components to the GCI Methodology that have been built as part of the new version that will be made available soon.

3.1 Change in Semi-Dynamic Indicators

- We recognise that the originally applied subindicators from the Global Health Secruity Index (GHSI) funded by the Bill and Melinda Gates Foundation may have some disputed results. We have taken on WHO's recommendation to explore other Indices that would achieve the same objectives of assessing a country's preparedness and ability to respond to a pandemic. For the new prototype, we have now applied the following sub-indicators as part of the new GCI methodology.
- ***WHO's Triple Billions Concept***: Based on the Thirteenth General Programme of Work (GPW13) Methods for Impact Measurement[4] published by the WHO, 3 new outcome indicators have been proposed as part of the Triple Billions Concept; the Universal Health Coverage (UHC) Billion, Health Emergencies Protection (HEP) Billion and Healthy Populations (HPOP) Billion. We have applied individual weightages to the GCI from these individual Billion scores:
- ***Universal Health Coverage (UHC) Billion***: The UHC Billion aims to ensure that an additional 1 billion people receive the quality health services they need without financial hardship. It uses both Average Service Coverage and Financial Hardship as its key components. This data will be updated on an annual basis from all WHO Member States.
- ***Health Emergencies Protection (HEP) Billion***: The HEP Billion aims for 1 billion more people to be better protected from health emergencies. Its key indicators rely on IHR State Party Self-assessment Annual Reporting (SPAR), vaccine coverage of at-risk groups for epidemic- or pandemic-prone diseases; and timely detection and response to potential health emergencies. This data will be updated on an annual basis from all WHO Member States.
- ***Healthier Populations (HPOP) Billion***: The HPOP Billion goal is for 1 billion more people to enjoy better health and well-being. The HPOP Billion marks the first time that WHO has created a single measure of change in the domain of the behavioural, environmental and socially determined healthiness of global populations. It utilises 16 subindicators which are linked to Sustainable Development Goals (SDGs) and WHO's resolution WHA66.10. This data will also be updated on an annual basis from all WHO Member States.
- ***European Commission's (EU) Disaster Risk Management Knowledege Centre (DRMKC) INFORM Risk Index***[5]: The INFORM model is the DRMKC's efforts to establish an index based on the risk concepts of: hazards & exposure, vulnerability and lack of coping capacity dimensions. Each country is scored based on these 3 risk areas, of which a comprehensive list of components are utilised. For the purpose of the GCI, we have zoomed in specifically on the aspects of Socio-Economic Vulnerability, Governance, Communication Infrastructure and Physical Connectivity Infrastructure. These scores

are currently updated on a half-yearly basis by the EU DRMKC.

- ***INFORM Socio-Economic Vulnerability Score***: The INFORM Soci-Economic Vulnerability Score ranks countries on their level of development, inequality and dependency on foreign aid. A higher score indicates a higher vulnerability that the member state is exposed to, and will mean they are at a higher risk should a pandemic affect them. We apply this directly to the GCI Severity Index but inverse it for the GCI Recovery Index.
- ***INFORM Governance Score***: The Governance score is derived from two lenses; the Government Effectiveness Index and Corruption Perception Index. This is a measure of existing country's Governance mechanisms to effectively respond to a national-level hazard such as a health epidemic. A higher score indicates that a country's existing governance will likely be more effective in handling COVID-19. We apply this directly to the GCI Recovery Index but inverse it for the GCI Severity Index.
- ***INFORM Communication Infrastructure Score***: The Communication and Infrastructure score considers Electricity, Internet, Mobile Cellular Access and Literacy Rates. This is a measure of how quickly and widespread public communication efforts can reach a country's population. A higher score indicates better infrastructure. We apply this directly to the GCI Recovery Index but inverse it for the GCI Severity Index.
- ***INFORM Physical Connectivity Score***: The Physical Connectivity score considers Road Density, Water and Sanitation Access. This is a measure of how easy it will be for a country's population to connect to needed utilities and also for help to reach them. A higher score indicates better infrastructure. We apply this directly to the GCI Recovery Index but inverse it for the GCI Severity Index.
- We believe utilising a combination of WHO Billions and INFORM risk has further strengthened the semi-dynamic section of the GCI and also leverages on an ongoing commitment by both the WHO and EU to keep these assessments updated at least on an annual basis.

3.2 Change in Monitored Duration of GCI Index Dynamic Calculation

- We recognise that since the initial outbreak in many countries, some are now seeing waves beyond the original COVID19 outbreak within their respective borders. Currently, the GCI records dynamic components since the start of the pandemic. In order to better reflect the ever-changing circumstances in all member states, we have implemented a 90-day version of our GCI Methodology. This means that all the dynamic components will be tracked and updated for only the past 90 days, and better reflects the current situation, especially if new waves are commencing in countries that may have been successful before.

4. Final Observations and Conclusions

We believe the new version will further strengthen the original objective of the Global COVID-19 Index; comprehensively combining data from highly reputable data sources, which now includes the WHO Billions Concept, and the EU Inform Risk Ratings.

We remain committed together with the Malaysian Government to keep the GCI live and updated for the remainder of 2020 and 2021 and hope to be able to use it to enable the development of more insight in addition to the full Global Pathfinder Report which is now publicly available fpr download on the GCI website and we hope more member states will be able to benefit from the learnings in this report:
(https://covid19.pemandu.org/download/GPF%20Full%20Report.pdf)

We are also committed to continue working closely with the World Health Organisation so that the GCI continues to be a tool that can be helpful to member states and will be more willing to extend our assistance in its ongoing efforts to battle this global pandemic.

Acknowledgments. On behalf of PEMANDU Associates and the Government of Malaysia, we would like to accord our special thanks to the GCI Advisory Council Members for the feedback on the original GCI model, and subsequently World Health Organisation team comprising Dr Nirmal Kandel, Dr Bochen, Dr Amit Prasad, and Dr Brett Archer for the multiple constructive consultations and inputs in our efforts to bring about a new version of the GCI for in-principle endorsement. We would not have been able to develop new version at this pace without their valued inputs.

References

1. John Hopkins Center of Science and Center for Systems Science and Engineering
2. World Bank Population Data
3. Our World in Data – COVID19 Testing Data
4. World Health Organisation (WHO) Thirteenth General Programme of Work (GPW13) Methods for Impact Measurement (Billions Concept)
5. European Commission's Disaster Risk Management Knowledege Centre (DRMKC) INFORM Risk Index
6. Oxford COVID-19 Government Response Tracker (OxCGRT) by the Blavatnik School of Government of Oxford University

Analyzing the early 19 th century's geomagnetic declination in Japanese archipelago from Tadataka Inoh's 67 volumes magnetic survey azimuth ledger Santou-Houi-Ki by interdisciplinary study

Motohiro Tsujimoto[1*]

[1*] *Sakai Shi Osaka Japan 592-8347*
Email: motori-7map@wine.ocn.ne.jp

Summary. The "Santou-Houi-Ki" is recorded by Japanese cartographic surveyor Tadataka Inoh in 1800 to 1816 consist of 67 volumes survey ledger to produce the first survey map of Japan, called "Coastal map of Japan" or Inoh Map (1:3600.1:21600.1:432000). In the Santou-Houi-Ki estimate 200,000 magnetic compass survey azimuth data by accuracy of 0°05'unit were written ,with the name of survey reference point and Target points. The surveyed region extends from North eastern Hokkaido to Yakushima island in western Japan. We execute interdisciplinary and simultaneous analysis of geomagnetic declination, real azimuth, precise position of the survey execution point and target points (latitude and longitude less than 0.1 second). And add the historical importance of each precise position of survey reference point. Calculate backward the precise position of survey reference point, where the value of magnetic declination, subtracting the magnetic survey azimuth from the true azimuth to any target points is similar or approximate. When I compare the analysis of magnetic declination from Santou-Houi-Ki to NOAA's Historical declination viewer, NOAA's progress of declination west in Kyushu and in western Honshu is more than 3 to 10 years late than the analysis from Santou-Houi-Ki and 20 years late in Tsushima Island in particular. I must continue to analysis in Kyushu and Tsushima Island and supply analysis data from Santou-Houi-Ki to NOAA's historical declination Viewer.

By conducting verification by analysis that fuses the disciplines of humanities and sciences that have never been learned with each other, we can obtain epoch-making highly accurate data that has never existed before, and create a chain that can be reused in various fields of humanities and sciences to be born.

Keywords. Santou-Houi-Ki, Tadatakaka Inoh, magnetic compass survey, magnetic declination, NOAA's Historical Declination Viewer

Scientifc Data Standard System Construction and its Application for Data Archiving in China

Juanle Wang[1*], ***and Yujie Wang***[1]

[1*] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, A11, Datun Road, Chaoyang District, Beijing, 100101, China*
Email: wangjl@igsnrr.ac.cn

Summary. Scientific data standards are the basis for the long-term acquisition, processing, preservation, and sustainable access and utilization of scientific data. The degree of standardization of scientific data is also a significant index to measure the accumulation and effective utilization of scientific data resources in the world. This paper analyzes the current status of the scientific data standards system and its development, closely follows the standardization requirements of scientific data management, analyzes the full life circle process of scientific data, and establishes a reference model of scientific data standard system. This system is established based on the three levels of basic standard, general standard and specialized standard. On this basis, some suggestions are put forward, such as perfecting the reference model of scientific data standard system, covering the whole life cycle chain, implementing the system step by step according to the plan, coordinating the multilevel standards, and publicizing and implementing the standard application.

Keywords. standards system, scientific data, scientific data centers, data management and sharing, reference model

1. Introduction

Scientific data standard and its standardization run through the whole life cycle of scientific data. In order to strengthen and enhance the standardization level of scientific data, it is necessary to carry out top-level design and system construction for the standard system of scientific data. The lack of overall coordination of the standard system will make the standards among different scientific data centers or repositories be incompatible, forming new data barriers. This study will explore a scientific data standard system that includes multidisciplinary content, covers multi domain platforms, and runs through the life cycle of scientific data.

2. Method.

The standardization requirements of scientific data management activities are analyzed from the aspects of organization of content, life cycle management, security guarantee, application service and assessment and evaluation of scientific data. The reference model of scientific data standard system is constructed under the guidance of hierarchical structure. The standard is divided into three categories: basic standard, general standard and specialized standard. According to the methods of function centralization and life cycle, the specific contents of the three kinds of standards are designed. Based on the basic business process of scientific data management standardization, the concept of life cycle after appropriate simplification is integrated into the analysis, and the content and function classification of scientific data standard specification are established.

3. Results.

In Summary, the scientific data standard system consists of eight subsystems: definition and guide standard, scientific data description standard, scientific data collection and processing standard, scientific data archiving standard, scientific data preservation and maintenance standard, scientific data sharing and service standard, scientific data assessment and evaluation standard, and scientific data security standard. Based on this model, data archiving standards are drew up and practiced for scientific research program data archiving in China.

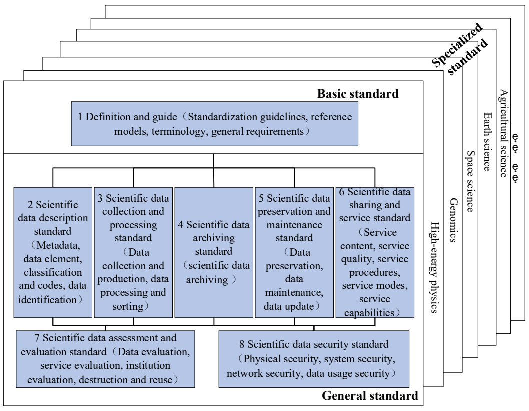

Specialized standard
Basic standard
1 Definition and guide（Standardization guidelines, reference models, terminology, general requirements）
2 Scientific data description standard（Metadata, data element, classification and codes, data identification）
3 Scientific data collection and processing standard（Data collection and production, data processing and sorting）
4 Scientific data archiving standard（scientific data archiving ）
5 Scientific data preservation and maintenance standard（Data preservation, data maintenance, data update）
6 Scientific data sharing and service standard（Service content, service quality, service procedures, service modes, service capabilities）
7 Scientific data assessment and evaluation standard（Data evaluation, service evaluation, institution evaluation, destruction and reuse）
8 Scientific data security standard（Physical security, system security, network security, data usage security）
General standard
High-energy physics
Genomics
Space science
Earth science
Agricultural science
……

Figure 1. Reference model of scientific data standard system

4. Conclusions.

This study fully considers the business requirements of scientific data management standardization and the common characteristics of scientific data in different disciplines and fields, and coordinates with the whole standard system of science and technology platform. According to the three levels of basic standard, general standard and specialized standard, the reference model of scientific data standard is established. Eight sub-systems including definition and guide standard, scientific data description standard, scientific data collection and processing standard, scientific data archiving standard, scientific data preservation and maintenance standard, scientific data sharing and service standard, scientific data assessment and evaluation standard, and scientific data security standard are constructed. This standard system is expected to provide standardization reference for National Science Data Centers and related scientific data management.

Acknowledgments. A special acknowledgement should be expressed to Special Foundation of National Science & Technology Infrastructure Center "Survey and analysis of global typical databases in key fields" （ No. 2020WT22 ） and the 13th Five-year Informatization Plan of Chinese Academy of Sciences (No.XXH13505-07).

Introduction of FUXI Platform——
Global Integrated Observation Data Management and Service System

***Yuanyuan Wang*[1,2]*, *Zhenhong Du*[1]**

[1]* *School of Earth Sciences, Zhejiang University, 38 Zheda Road, Hangzhou 310027, China;*
[2] *Ocean Academy, Zhejiang University, 1 Zheda Road, Zhoushan 316021, China*
Email: wangyuanyuanxy@zju.edu.cn

Summary. Global Integrated Observation Data Management and Sharing Service Platform—FUXI is a new generation of intelligent platform for remote sensing observation data, which integrates super-large-scale remote sensing comprehensive observation data results, massive artificial intelligence models and algorithms, and online high-performance computing and analysis services in one. FUXI has the capabilities which were independently innovated of efficient management of EB-level observation results, artificial intelligence-driven knowledge modelling, and constructing online template chains for application scenarios supported by microservices. It will support platforms for international shared services, such as the Cooperation on the Analysis of carbon Satellite data (CASA) and Global Ecosystems and Environment Observation Analysis Research Cooperation.

Keywords. Remote Sensing, Data Management, Data Service

WDS-led Activities on Data in the Asia-Oceania Area

***Takashi Watanabe*[1*]*,Juanle Wang*[2]*,Toshihiko Iyemori*[3]*, Yasuhiro Murayama*[4]**

[1*] *WDS International Programme Office, Koganei, Tokyo, 184-8795, Japan*
[2] *Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science*
1A, Datun Road, Chaoyang District, Beijing, 100101, China
[3] *Kyoto University, Kyoto, 606-8501, Country*
[4] *National Institute of Communications Technology, Koganei, Tokyo, 184-8795, Japan*
Email: takashi.watanabe@worlddatasystem.org

Summary. As a part of the activities of the World Data System (WDS) of the International Science Council (ISC), two WDS Asia–-Oceania Conferences were held in Japan in 2017 and in China in 2019 respectively, These Conferences were co-organized by the collaboration of the WDS National Committee of Science Council of Japan, the WDS community of China, and the WDS International Programme Office. The principal objective of the conferences was to build a WDS-oriented network of research-data repositories in Asia–Oceania comprising of existing WDS Members and other data-focused organizations. Collaboration with the committees of ISC's Committee on Data (CODATA) and the ISC Regional Office for Asia–-Pacific (ROAP) will also be important to progress such activities, as well as including the involvement of government data repositories, particularly in Southeast Asian countries.

Keywords. Data Network, WDS, Asia, Oceania

1. Background

To identify current problem on long-term preservation and open provision of research data in the Asia–Oceania area, a survey of current data-oriented activities, mainly in the field of Environmental Sciences in Asia-Oceania was performed by searching for data portals on the Internet and by visiting several institutional and governmental repositories. As of result of this survey, the following problems shown below were discovered.

(1) The majority of well-established data repositories, especially in the Southeast Asian countries, are governmental in nature; for example, they manage satellite remote sensing topological images or environmental data. The values of Open Science and Open Data are not high on the agenda of such repositories at this stage because they are funded at the national level by mandated programmes.

(2) Only very limited number of data repositories were found to be operated by researchers working in universities or other research institutions.

(3) Various observational data are being obtained in Southeast Asia as part of international research programmes. However, the majority of these data are hosted and shared by data repositories in more advanced countries.

2 Past and future WDS Asia–-Oceania Conferences

To ascertain the current status and the problems of Research Data Management (RDM) in Asia–-Oceania, two WDS Asia-Oceania Conferences focusing on the region were organized through a collaboration among the WDS National Committee of the Science Council of Japan, the WDS Community of China, and the WDS International Programme office in Tokyo, Japan.

2-1. WDS Asia-–Oceania Conference 2017, Kyoto, Japan[1]

The first WDS Asia–-Oceania Conference was held on 27–29 September 2017 at Kyoto University, Japan, hosted by the WDC for Geomagnetism, Kyoto. More than 110 data-related people from 16 countries attended the 2017 Conference (Figure 1), The Conference brought together data practitioners, data repositories managers, and data-oriented researchers to identify data issues and discuss a future plan to help strength data activities, and build a platform to connect this al activity to the international data community. The resulting network was considered valuable also in helping to support other international research programmes under ISC, namely, Future Earth.

World Data System Asia-Oceania Conference 2017
(Kyoto University, September 27-29)

Figure 1. Participants of the WDS Asia–-Oceania Conference 2017.

2-2. WDS Asia-–Oceania Conference 2019, Beijing, China[2]

The second WDS Asia-Oceania Conference was held on 6–7 May 2019 at the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Science, in Beijing, China. On this occasion, we had more than 120 participants from 14 countries were in attendance (Figure 2), and about 50 papers were presented. The main objective of the 2019 Conference was to reinforce the collaborations in Asia–-Oceania that began in the first Conference, and at the same time progress activities of strategic importance to the WDS mission-oriented activities. The realization by the 2019 Conference was that building the envisaged network to involve more people from more countries would be a gradual process; especially, when it comes from Southeast Asia. With the 20th Meeting of the WDS Scientific Committee held in conjunction with the Conference, this was a notable opportunity to impart information to the Conference participants on the projects of WDS, including the procedure to become certified as trustworthy by obtaining the CoreTrustSeal[3] for Data Repositories.

World Data System Asia-Oceania Conference 2019
May. 7-8, 2019
World Data System Asia-Oceania Conference 2019

Figure 2. Participants of the WDS Asia–-Oceania Conference 2019

2-3. Future Activities

It will be important to continue our activity having meetings regularly, at least biannually. Collaborations with ROAP and CODATA National committees in the Asia-Oceania area will be important to involve governmental repositories in this are to our activity.

References

[1]http://wdc2.kugi.kyotou.ac.jp/wds2017/program20170917.html

[2]http://www.wds-china.org/meeting201905.htm

[3]https://www.coretrustseal.org/

Importance of a Multi-disciplinary Study of Peculiar Environmental and Socio-economic Movements in 18/19 Century

Takashi Watanabe[1*]

[1*]*World Data System International Programme Office, Koganei, Tokyo, 184-8795, Japan*
Email: takashi.watanabe@worlddatasystem.org

Summary. The 18/19 Century was an interesting period in terms of environmental and socio-economic aspects. The period is known to have been a relatively cold period in the last phase of the Little Ice Age starting in 16 Century. The level of solar activity was generally low in the period of the Mounder Minimum (1500-1700) and the Dalton Minimum (1790-1830). Global influences of extensive volcanic activities were also important. It is found that this interval will be important to promote a multi-disciplinary data analysis to study environmental influences on socio-economic activities. To do this, collaborations between data providers and users (scientists) will be important to promote multidisciplinary usages of data.

Keywords. Data Analysis, Space Weather, Environmental Science, Social Science

1. Introduction

The importance of multidisciplinary usage of data has been stressed in various disciplines but there still exists a lot of obstacles to do. To identify problems in multi-disciplinary usage of data, a trial is performed on environmental and economic data during the 18th-19th Centuries because this interval will be interesting to many scientists in various research fields including the space weather, the atmospheric sciences, the history, the economics, the agriculture, etc. Socio-economic movements in this period were taken place under the peculiar environmental situation. This period was located in the last half of the so-called Little Ice Age (LIA) which was a period of generally lower global temperatures expanding from 16th to the mid of 19th centuries. In the LIA, the Maunder Minimum (MM) of solar activity (Eddy 1976) expanding from 1645 to 1715 was a remarkable period of globally cold climate. Following the MM, a relatively short minimum of the solar activity, called Dalton Minimum (DM), was taken place in 1790-1830. Since considerable depression of socio-economic activities, e.g., the Industrial Revolutions in England, the French Revolution (1789-1799) and the Napoleonic Wars (1799-1815) were taken place in Europe under the relatively cold environment and high volcanic activities. A collaborative study of this interval by scientists in a wide range of disciplines will be important to know the socio-economic movements in the advent of the industrial world under the "basic" natural environment of the Earth, just before the anthropological environmental changes have been clearly recognized. Such collaborative works will be useful to create a strong motivation to improve the accessibility and usability of research data to meet the FAIR (Findable, Accessible, Interoperative, and Reusable) Guiding Principle research data for scientific data management and stewardship.

2. Why the 18th/19th century?

Time changes of the sunspot number and the annual wheat price in England in 18th/19th Centuries are shown in Figure 1. A depression of amplitude of the 11-year sunspot cycle in the interval of 1790-1830 is called the Dalton Minimum (DM). It is difficult to find a clear correlation between the sunspot 11-year Schwabe Cycle and the British wheat price in this figure but the price showed a decadal enhancement during the DM. The rise of wheat price in question is suggested to have been caused by poor harvesting known to have happened in England during the period of the Napoleonic Wars (1799-1815), accidentally happened in the DM, and several spiky enhancements of the price were mainly caused by shocks of the war, including blockages of British and French seaports. A general increase in wheat price in 1815-1846 was partly caused by the influence of the British Corn Laws.We will perform our study to find environmental influence on the price history because similar increases in prices of crops can be found also in other European countries (e.g., France, Germany, and Austria), North America, and in Asian countries including India and China. This means that the rise of the crop price is suggested to have been taken place in a global manner by some common environmental changes.

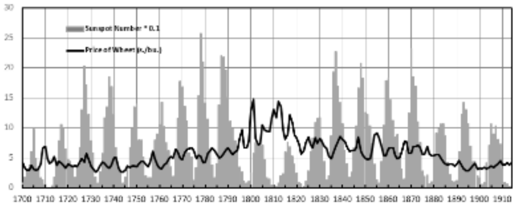


Figure 1. Time series of yearly sunspot numbers (x 0.1) and the wheat price in England (shillings per bushel) in the interval of 1700-1910.

3. Data Analysis

A provisional correlation analysis is performed for the parameters shown in Figure 2. In this analysis, no time lag is applied although this procedure has been employed in several correlation analyses e.g. the correlation between solar activity and precipitation in Europe. The notable tendencies found by correlation analyses are: (1) The precipitation data for the summertime (JJA) show a significant positive correlation (the correlation coefficient is +0.6) with the long-term (>11 years) variations of sunspot numbers. The temperature shows no significant correlation with the sunspot number in the period of present analysis (1770-1860), although a week correlation (+0.2) is seen in the wider interval (1700-1900), (2) The wheat price has a significant negative correlation (-0.7) with the wheat yield. This suggest that the most important factor to control the wheat price in this period was the wheat yield, (3) The wheat yield in the DM era was largely controlled by the precipitation (+0.4). No significant correlation with the averaged JJA temperature but an enduring cold summer in the interval of 1809-1817 is suggested to partly contribute to the reduction of yield, (4) The JJA precipitation in England-Wales showed a negative correlation (-0.5) with the positive (westerly) PLWI (full names of datasets are given in Figure 2) throughout the period of the DM, (5) The JJA temperature did not show a significant correlation with the sunspot number but very weak correlation with PLWI is seen, (6) Major volcanic activities in the interval of analysis (e.g. Tambola in1815) produced clear short-term depressions of temperature. Weak correlations seen between TEM and VOI were produced by these volcanic events.

On the data issue, efforts to convert data on papers to machine-readable formats and to standardize reconstructed data are highly desirable. The collaborative multi-disciplinary study of the 18th/19th century will be pertinent to stimulate the efforts and useful to predict consequences of the current anthropological environmental change.

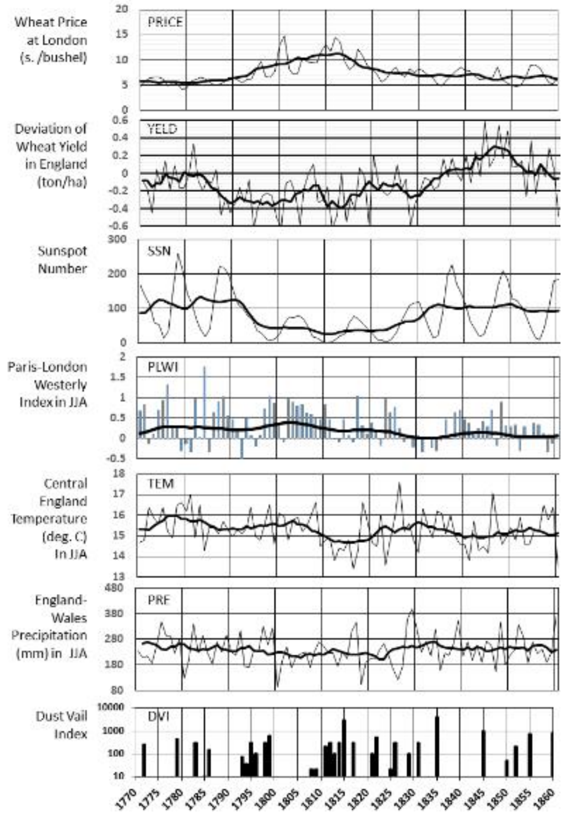


Figure 2. Time series of economic and environmental data in the 18th-19th Centuries: (a) wheat price at London (PRICE), deviation of the wheat yield (YELD) in England from a linear regression line in the interval, , yearly sunspot number (SSN), yearly mean temperature (deg. C) in the central England (TEM), Paris-London Westerly Index (PLWI) as a measure of the North-Atlantic Oscillation, and the Dust Vail Index (DVI) which is a measure of the effect of major volcanic eruptions in the world.

COVID-19 Data Sharing Initiative by WDCM and NMDC

Linhuan Wu[1*], **_Juncai Ma_**[1]

[1*] *Institute of Microbiology, Chinese Academy of Sciences, NO.1-3 Beichen West Road,*
Chaoyang District, Beijing 100101, China
Email: wulh@im.ac.cn

Summary: As the COVID-19 virus are widely spreading on the world, the World Data centre for Microorganisms (WDCM) and NMDC (National Microbiology Data Center), as national scientific data center in China which is responsible for data collections, data sharing and data service in the field of microbiology, are making efforts on provide public available data sharing fighting for COVID-19 including resources and efforts for SARS-CoV-2 data sharing and promoting global cooperation.

NMDC established the Novel Coronavirus National Science And Technology Resource Service System releasing the first electron microscope picture of the virus and strain information at Jan 24, 2020. And online data analysis tools were developed in the system including Phylogenetic analysis and display and COVID-19 multidimensional analysis platform. A total of 237077 users from 176 countries and regions worldwide made 10.27 million visits for the system, 834 thousand of which were by 28372 overseas users.

Keywords. COVID-19, genome sequences, data sharing

Information platform promoting global collaboration in microbial community

Linhuan Wu[1*]

[1*] *Microbial Resource and Big Data Center, Institute of Microbiology,*
Chinese Academy of Sciences, Beijing 100101, China
Email:wulh@im.ac.cn

Summary: World Data Centre for Microorganisms (WDCM) is the most important physical resource data platform of microorganisms worldwide, which was hosted by Institute of Microbiology, Chinese Academy of Sciences including several databases - Culture Collections Information Worldwide, Global Catalogue of Microorganisms, Analyzer of Bio-resource Citations and Reference Strain Catalogue, GCM2.0 Type Strain Genome Database and gcMeta Microbiome Research Platform. GCM international cooperation program was established for providing effective data support for all aspects of physical resources of microorganisms. GCM2.0 Global Microbial Type Strain Genome and Microbiome Sequencing Project is expected to complete genome sequencing of more than 10,000 microbial type strains, and set up a globally authoritative reference database and data analysis platform of microbiome. As the COVID-19 virus are widely spreading on the world, WDCM are making some efforts on fighting for COVID-19. As some culture collections are temporarily shut down, WDCM has host several international online conference and workshop to support culture collections to work and fight the epidemic.

Keywords. A list of up to 5 words describing the main concepts in the abstract, separated by commas

National Cross-Domain Activities on Research Data in Australia and Connections with International Activities

Lesley Wyborn[1*], Danny Kingsley [2], Adrian Burton[3], Simon Cox[4], Steven McEachern[5], Ginny Barbour[6]

[1*] _Australian National University, 56 Mills Road, Acton, ACT, 2601, Australia_
[2]_Australian National University, 42A Linneaus Way, Acton ACT 2601_
[3]_Australian Research Data Commons, 9 Liversidge Street, Acton, ACT, 2601, Australia_
[4]_CSIRO Land and Water, Address, Private Bag 10, Clayton South, Victoria, 3169, Australia_
[5]_Australian Data Archive, 146 Ellery Crescent, Acton, ACT, 2601, Australia_
[6]_AOASG, UNSW Library, UNSW Sydney, NSW 2052, Australia_
Email:lesley.wyborn@anu.edu.au

Summary. In the past decade, many data collections of value to Australian research have been made accessible as a result of significant, targeted investments in digital data infrastructures, data standards and protocols, as well as new funding of physical infrastructures to store and process data. In Australia the FAIR principles are becoming accepted, as is 'Intelligent Openness'. But to date, investments in data tend to be siloed with respect to discipline and sector, and it is now timely to focus on enabling cross-domain data infrastructures. To realise the full potential of data-intensive research, any new data initiatives now must be done within the context and predicted capabilities of next generation computational infrastructures, which by 2030 will be at exascale.

Keywords. FAIR data, metadata, transdisciplinary research, 2030 computing, research ecosystem

1. Introduction

Over the last decade, Australia has had many successful data initiatives to ensure the persistence of publicly funded research data. The Findable, Accessible, Interoperable and Reusable (FAIR) principles are becoming accepted and are now a requirement of some funding schemes. 'Intelligent Openness' is recognised, acknowledging that there may be valid reasons for withholding data or limiting its reuse. Metadata catalogues, persistent identifiers (PIDs, e.g., DOI, Handle, ORCiD, IGSN, Grant-ID, etc) are progressively being implemented and there are agreed minimal metadata standards for research and government data. Data of value to the research community can be collected by publicly funded academic projects as well as by initiatives of government research agencies: it is also becoming easier to gain access to relevant data generated by industry and citizen science.

But there are noticeable gaps. Initiatives to date have mostly focussed on domains with larger volume collections, particularly those in terabytes and petabytes, which need specific storage infrastructures (e.g., climate, geophysics, astronomy). Heterogenous, small scale (<1GB) datasets in the long tail communities are not as well managed. Data is more commonly organised within institutional/domain silos, and sector silos are present mainly due to varying capabilities/ capacities and differing stakeholder/community needs. Many online data delivery mechanisms are biased towards only making highly processed/ downscaled data products accessible: rawer versions of datasets can be hard to access.

In 2020 as physical infrastructures grow in capacity (including both instruments to collect as well as computational power to process and store data) traditional scientific paradigms are being transformed in varying ways, to produce or use increasingly large or complex streams of data that can only be explored and analysed with modern data analytic methods. This is allowing research to ask new and bigger questions and create greater impact, but at the same time, these developments are posing significant challenges that need to be addressed to fully realise the undoubted potential of data-intensive research. As the science paradigm changes, we need to ensure that data storage and management systems are developed in the context of 2030 science infrastructures and societal drivers (e.g., the UN Sustainable Development Goals).

This paper will outline some national cross-domain activities in research data led by the Australian Academy of Science (AAS) and the Australian Research Data Commons (ARDC): it will also highlight some key Australian connections to international data infrastructure initiatives.

2. Australian Academy of Science (AAS)

The AAS provides independent, authoritative scientific advice and promotes international scientific engagement. The AAS is currently developing a paper on 'Advancing Data Intensive Research in Australia' that is exploring issues, challenges and opportunities afforded by 'Big Data' in Australian research. The AAS also provides strong linkages to the International science agenda and supports the National Committee for Data in Science (NCDiS) [1], which is also the National Committee for CODATA. The NCDIS is one of 22 National Committees (NCs) [2], many of which have connections to the relevant International Unions of the International Science Council (ISC).

3. Australia Research Data Commons

The ARDC is an initiative supported by the Australian Government through the National Collaborative Research Infrastructure Strategy. ARDC seeks to enable the Australian research community access to nationally significant, leading edge data intensive eInfrastructure, platforms, skills and collections of high-quality data. It works with both the research sector and with government and industry partners to build a coherent national and collaborative research data commons. In 2020 it launched the National Data Assets [3] program to leverage existing investments to ensure ongoing persistence, sustainability and stewardship of data collections of value to Australian research.

4. Australian Participation in International Data Initiatives

In addition to the AAS connections with CODATA and ISC, many Australians are connected to international data initiatives. For example, there is active engagement of Australians in Working/ Interest Groups of the Research Data Alliance, in the development of PID systems (e.g., DataCite, IGSN, RAiD) and in data standards bodies (e.g., W3C and OGC). The key challenge for any participant in these groups is in being able to bring international perspectives back into national initiatives.

5. Conclusions

Australia is moving from a more siloed, discipline- and/or sector-based approach towards a national research ecosystem for data generation, management, curation and stewardship that also recognises the importance of linking data to any other research outputs. Wherever possible, national activities are integrated with international initiatives. Any new data infrastructures will need to be planned/developed in the context of 2030 science infrastructures to ensure that future generations of research can contribute to the societal challenges of the next decade.

Acknowledgments. All authors of this paper are members of the NCDiS and each acknowledges the support of both the Academy of Science and of our home institutions in undertaking this work.

References

1. AAS National Committee for Data in Science, https://www.science.org.au/supporting-science/national-committees-science/national-committee-data-science [accessed September 2020].
2. AAS National Committees, https://www.science.org.au/supporting-science/national-committees-science [accessed September 2020]
3. ARDC National Data Assets, https://ardc.edu.au/collaborations/strategic-activities/national-data-assets/ [accessed on: September 2020].

Disaster Rapid Damage Mapping with Remote Sensing: benchmarking datasets and methods

Junshi Xia[1*]

[1*]*RIKEN, Chuo City, Tokyo, 103-0027, Japan*

Email: junshi.xia@riken.jp

Summary. Disasters such as earthquakes, hurricanes, and flooding are responsible for large-scale infrastructure damages and loss of human lives. Immediately after disaster strikes, one of the most critical and challenging tasks is accurately assessing the extent and severity of the disaster. Due to the development of deep convolutional neural networks (CNN), it is urgent to develop or construct the benchmarking datasets for damage mapping. This work first brief introduces the xView2 datasets and then presents a damage mapping framework using remote sensing imagery.

Keywords. Building damage mapping, xView2, Remote Sensing, CNN

1. Introduction

When disaster strikes, accurate situational information and fast, effective response are critical to saving lives. Remote Sensing is one of the most effective tools to provide rapid damage mapping for natural disasters. The main issue is how to develop a robust method to obtain building damage mapping [1].

Recently, deep convolutional neural networks (CNN) architectures have shown significant advantages for damage mapping using remote sensing imaging [2]. In previous studies, the researchers often analyze the damage mapping of one disaster type, such as tsunami, earthquake. In this context, deep learning techniques demand a vast amount of training data for proper generalization ability. This essential factor hampers the applicability of CNN-based frameworks for emergency response. Due to the rapid development of machine learning and the increasing number of natural disasters, a large number of remote sensing datasets, including all kinds of disasters, are urgently needed.

In this work, we will briefly introduce the xView2 datasets and which kinds of methods can be used for building damage mapping using xView2 datasets.

2. xView2 datasets

xView2 dataset is designed for the building damage mapping and covers 19 disaster events, such as tsunamis, earthquakes, floods, hurricanes, tornados, wildfires, and volcanic eruptions [3]. All the datasets are derived from the Worldview image of Maxar. The datasets contain the pairs of pre- and post-events with a size of 1024*1024 pixels. Besides, the building footprint of the pre-event and the damage levels (i.e., no-damage, minor-damage, major-damage, and destroyed) are provided. The objective of xView2 datasets is to provide not only the buildings but also the building damage levels. Figure 1 has shown examples of xView2 datasets (including pre-event, post-event, building footprints, and damage grading).

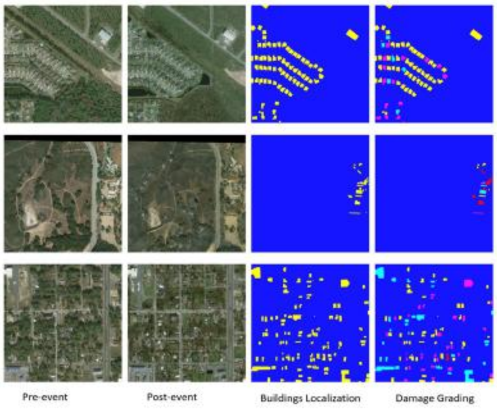


Figure 1 Examples of xView2 datasets.

3. Building damage mapping

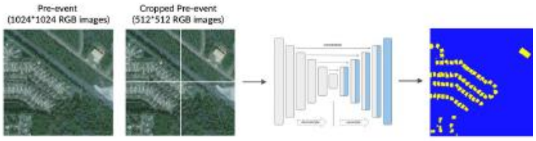


Figure 2 Unet for building footprint extraction

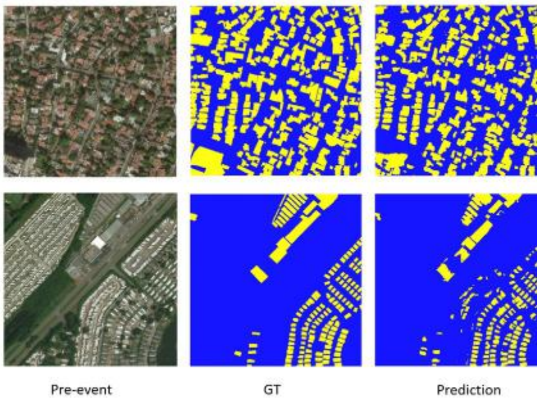


Figure 3 Results of building footprint extraction

First, we use the widely used Unet with attention [4-5] for the building footprint extraction (as shown in Figure 2). In this case, we selected the three powerful encoders: Resnext50_32x4d, Densenet, Efficientnet B7, and ensemble their outputs. The F1 score is 0.8613 (as shown in Figure 3), which is better than the individual ones of 0.8562.

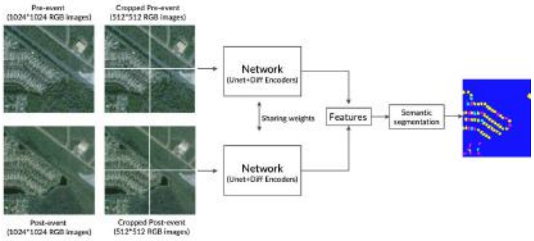


Figure 4 Siamese Unet for building damage mapping

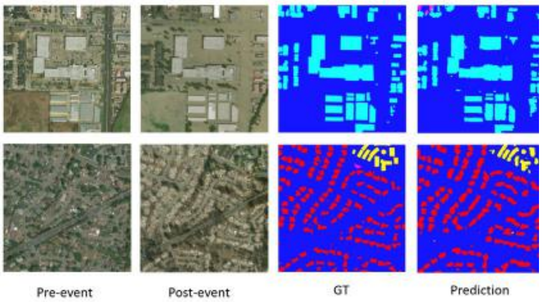


Figure 5 Results of building damage mapping

For the building damage mapping, we proposed to use Siamese Unet. The two-stream Unet shared the same weights. We used the prvious Unet for building segmentation to fine-tuning the Siamese Unet.The two-stream features produced by the U-Net of pre-event and post-event are stacked together, and then fed to the final layer for the final semantic segmentation. Due to the unbalanced samples of four damage levels, we added the class-specific weights for the loss functions. Finally, we got the F1 score of 0.7594 (see the results in Figure 5).

4. Conclusions

In this paper, we present the ensemble of Siam-UNetself-attention for building segmentation and damage mapping. Experinmental results have shown the proposed model accomplishes both damage classification and building segmentation more accurately than other approaches with the xView2 dataset

Acknowledgments. The authors thank the Defense Innovation Unit to provide xView2 dataset.

References

1. M. Pesaresi, A. Gerhardinger, and F. Haag, Rapid damage assessment of built-up structures using vhr satellite data in tsunami-affected areas, International Journal of Remote Sensing, vol. 28, no. 13-14, pp. 3013–3036, 2007
2. Y. Bai, E. Mas, and S. Koshimura, Towards operational satellite-based damage-mapping using u-net convolutional

network: A case study of 2011 tohoku earthquake-tsunami, Remote Sensing, vol. 10, no. 10, 2018

3. R. Gupta et al. Creating xbd: A dataset for assessing building damage from satellite imagery. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 2019
4. O. Oktay et al, Attention u-net: Learning where to look for the pancreas, CoRR, vol. abs/1804.03999, 2018
5. S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, Aggregated residual transformations for deep neural networks, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 5987– 5995, 2017

Official Micro Data, Causal Inference and Evidence-Based Policy Making

Junchao Zhang[1*]

[1*] *Center for Social Data Structuring, DS, ROIS*
Email: zhang@ism.ac.jp

Summary. Detailed personal information (e.g., exact date of birth, community location, etc.) is very important in identifying causal effects of policies and programs. However, these information are generally not collected or excluded in public use micro data due to privacy reasons . Without these information, we cannot correctly define the treatment variable indicating whether one was exposed to a particular policy or program. Using high-quality official micro data, this study discusses the problems in causal inference and offers useful evidence for policy making.

Keywords. micro data, causal inference, EBPM

1. Introduction

In the last decade, evidence-based policy making (EBPM) has been of great interest to policy makers and academics. Before EBPM become widespread, the policy makers actually do not know whether they are doing the right things. To effectively and sufficiently allocate governmental resources, how to quantitatively evaluate the causal effects of policies and programs is a fundamental challenge.

In hard sciences, a randomized controlled trial (RCT) is considered as the gold standard. Using RCT, researchers randomly assign test subjects into treatment and control groups, and can effectively determine whether the treatment group fares better or worse when exposed to the intervention. Researchers can evaluate the average treatment effect (ATE) even with a simple OLS estimand if the trial is properly randomized.

However, we practically meet the following issues in policy studies. First, RCT may not be a feasible option because of law and ethics. The estimates in observational studies probably reflect correlation rather than causality. As an alternative method, policy researchers exploit natural experiments(e.g. policy changes, weather events, natural disasters, etc.) that occasionally assign people into different groups for causal inference.

Second, detailed personal information(exact date of birth, community location, etc.) are very important in identifying causal effects. Without these information, we cannot define the treatment variable indicating whether one was exposed to a specific policy intervention. These information are generally not collected or excluded in public use micro data due to privacy reasons.

In this study, we use high-quality official micro data, with huge sample size and detailed personal information, to perform a causal study. The huge sample size allows us to explore people's heterogenous response even when our identification strategies rely on rare events. Heterogenous effects may help policy makers to target potential treatment population and improve policy design.

2. Model

Suppose we are interested in the effect of number of children on maternal labor market outcome. A benchmark linear model is specified as follows:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where y_i is a binary outcome variable indicating married women's labor force participation. $\beta_1$is the coefficient of interest, capturing the effect of number of children x_i.

However, $Cov(x_i, u_i) = 0$ condition is probably not satisfied as number of children are not randomly assigned to women. Unobserved factors(u_i) such as preferences regarding childbearing or career ambition could be correlated with women's decision making. The OLS estimand will be no longer unbiased and consistent. β_1 reflects simple correlation rather than causality.

Instead, we use a natural experiment that induces exogenous variation(z_i) in the number of children to estimate the causality. z_i indicates whether the child is a twin.

$$\hat{\beta}_{IV} = (\boldsymbol{z}'\boldsymbol{x})^{-1}\boldsymbol{z}'\boldsymbol{y}$$

where $\boldsymbol{z}, \boldsymbol{x}, \boldsymbol{y}$ are $N \times 1$ vectors. Mothers gave birth to twins and non-twins naturally have different numbers of children. This identification strategy rely on very huge sample size as twin birth rates are only about 9 to 20 per 1000 deliveries across different countries. To explore the heterogeneity, we also estimate the IV estimand by different sub-samples(birth parity and time passed since last child birth).

3. Results

We use Japanese census to estimate the IV estimates. IV estimates are reported in Table 1, OLS estimates are also reported for comparison. OLS and IV estimates are quite different across birth parity and over time passed since last child birth.

According to data restriction, previous studies can only estimate on unconditioned sample. Huge sample size of official micro data allow us to perform IV estimation using different sub-samples to detect people's different decision-makings. According to Table 1, negative effects of number of children on women's labor supply are only significant for first parity.

4. Conclusions

This study estimates the causality between children and maternal labor supply. We find that the effect of fertility varies substantially with the time elapsed since the last childbirth. In contrast to conventional wisdom, the effect of fertility on maternal labor suppy is not monotonically decreasing in the number of children. These results imply that policy-makers should design accurate policies to target those who will benefit more from policies or programs.

Acknowledgments. This work is supported by JSPS KAKENHI Grant Number 19K13710.

References

1. Angrist, J. D., & Evans, W. N., Children and their parents' labor supply: Evidence from exogenous variation in family size. American Economic Review, 88(3), 450–477, 1998
2. Li, H., Zhang, J., & Zhu, Y., The quantity-quality trade-off of children in a developing country: Identification using Chinese twins. Demography, 45(1), 223–243, 2008
3. Rosenzweig, M. R., & Wolpin, K. I., Testing the quantity-quality fertility model: The use of twins as a natural experiment. Econometrica, 48(1), 227–240, 1980

Table 1. Effects of number of children on maternal labor supply by birth parity and time since last child birth.

	Unconditioned		**No more than 3 years**		**No more than 1 year**		**No more than 3 months**	
	OLS	**IV**	**OLS**	**IV**	**OLS**	**IV**	**OLS**	**IV**
Birth parity: 1	-0.003 ***	0.000	-0.005 ***	-0.047 ***	-0.027 ***	-0.031 ***	-0.056 ***	0.004
Birth parity: 2	-0.027 ***	-0.002	0.012 ***	0.004	0.015 ***	0.009	-0.008 ***	-0.001
Birth parity: 3	-0.037 ***	0.050 **	0.002	0.066 **	0.012 **	0.026	0.010	0.072

Notes: All specifications control for age, age squared, education attainment, husband's education attainment, husband's labor force participation, co-residence with elder parents, and prefecture dummies. In all panels, upper bounds on the number of children are not imposed. *** p<0.01, **p<0.05, * p<0.1. Robust standard errors are not reported because of space constraint.

Introduction for WDS ECR Network

Lianchong Zhang [1*], **Jesse Xiao** [2]

[1*] *Aerospace Information Research Institute of the Chinese Academy of Sciences*
[2] *The University of Hong Kong*
Email: zhanglc@radi.ac.cn

Summary. The World Data System (WDS) is a body of the International Science Council (ISC) that helps to coordinate and support research data centres and data services worldwide. WDS activities span all disciplines, and are designed to ensure that research data are preserved and openly disseminated to safeguard the integrity of science. WDS is also concerned with the availability to scientists and policymakers of the critical information necessary to manage Earth's resources wisely. Recognizing the important role of Early Career Researchers in developing and promoting best practice in data management, data analysis and data sharing, WDS establishes a Network of Early Career Researchers and Scientists (ECRs), to help foster better communication among ECRs, and to design activities targeting their interests and concerns. WDS-ECR Network members will have an opportunity to participate in regular webinars and yearly training to develop their skills and share their knowledge related to the key areas of research data management. The Network will facilitate connections between the members and more experienced global leaders within WDS and beyond, as well as regularly profile its members in its quarterly newsletters and other publications.

Keywords.

Modelling the Suitability of Parcel Pick-up Lockers Using the Multi-Source Open Data

Zilai Zheng[1*], ***Takehiro Morimoto***[2]

[1*] *Graduate School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai,
Tsukuba, Ibaraki, 305-0006, Japan*
[2] *Faculty of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki,
305-0006, Japan*
Email: zhengzilai23@gmail.com

Summary. Parcel pickup locker (PPL) is an effective way of solving the last-mile problem and becomes a hot topic in logistics geography. In this study, a GIS-based multivariate logistic regression(MLR) model was developed to modelling the suitability of PPLs using the multi-source data, such as the point-of-interest(POI) data from a popular daily navigation application named Gaode and the road data from crowdsourcing project OpenStreetMap(OSM). A spatial database of 27 potential predictor variables was constructed under Geographical Information System (GIS) environment. Eight variables were chosen by the forward-stepwise selection method of the model, and the most crucial variable was the distance to the nearest residential quarter. This study proved that conducting the GIS-based MLR model using multi-source open data to model the suitability of the facility is efficient and with high accuracy. Furthermore, it gives the data support for decision-makers to decide the new location of PPLs.

Keywords. parcel pick-up locker(PPL), multivariate logistic regression model(MLR), multi-source, point-of-interest(POI), Geographical Information System (GIS)

1. Introduction

The location planning of PPL is the most complicated issue in the optimization topic of PPL. From the literature review, the micro-scale studies on PPLs focused only on the facility or location selection in a small area and have not pointed out how to find suitable places for PPLs [1]. The macro-scale studies focused mainly on the relationships between the number of facilities and influencing factors. Few studies have conducted location analyses from the pixel scale to find the relationship with the factors.

This study attempted to conduct a GIS-based machine learning(ML) model using multi-source data to model the suitability of PPLs in Guangzhou city, where the amount of parcel ranks the first in China from 2014 to 2019. Suitability problems can be considered as a binary classification problem, and logistic regression is regarded as an efficient and straightforward way in ML to solve this problem.

2. Materials and methods

A comprehensive range of detailed, high-resolution GIS data was collected from different open sources, including topographic data (30 m), population density raster data (100 m), land price, road networks, and point of interest(POI) data. Comparing the POI data extracted from Gaode Map [2] by the Python code and that downloaded from OSM [3], the data from

Gaode Map is more detailed and with higher quality. After the data cleaning, the POI data layers of the PPL, bus stop, metro exit, parking lot, commercial office building, commercial-residential building, residential quarter, villa, dormitory and community centre contain 679, 6778, 808, 9882, 5658, 825, 7619, 280, 2031 and 353 features, respectively. The analysis of Euclidean distance in Esri's ArcMap was used to calculate the distance from each cell in the raster data to the closest POI, different types of roads, and water area. The kernel density function was employed to calculate the amount of the feature in one cell. A spatial database of 27 potential predictor variables was constructed at the unified resolution of 100 m. The methodology is shown in Figure 1.

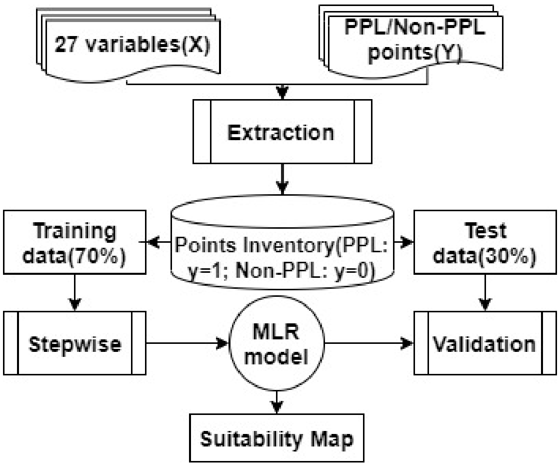


Figure 1. The methodology of the study

3. Results

3.1 Selection the significant variables

The forward-stepwise method was applied in the MLR model to choose some significant (p-value<0.005) predictor variables for the final model and to make the model optimization with the least-squares criteria. Eight variables were selected by the MLR model. According to the Wald-value of each variable in the final model, the most significant factor was the distance to the nearest residential quarter, followed by the land price, the distance to the nearest bus stop, and the density of commercial buildings.

3.2 The performance of the model

The performance of the MLR model was usually evaluated by their discrimination and calibration [4]. As shown in table 1, the classification accuracy of the training data in the model was very high. The model's bias was minimal, and the AUC value greater than 0.9 was considered outstanding [5]. The model validation was conducted with the test data, and the performance of the model was slightly better than the training data. It can be seen that the suitability model with these coefficients is good-fit the real data.

Table 1. The evaluation of the suitability model

Data	Discrimination				Calibration
	Precision	Recall	Accuracy	F-Measure	AUC
Training	90.6%	86.5%	88.2%	88.5%	0.95
Test	92.0%	89.8%	90.8%	90.9%	0.96

The results show that 8.7% of the total area in Guangzhou city is a suitable area for PPLs, contains 91% of the amount of PPLs.

4. Conclusions

This study proved that conducting the GIS-based MLR model using multi-source open data to model the suitability of the facility is efficient and with high accuracy. It gives a data reference for decision-makers to decide the new location of PPLs. The POI data of the Gaode Map, a daily navigation map, is more detailed and higher quality than the crowdsourced OSM.

References

1. Zheng, Z., Morimoto, T., Murayama, Y., Optimal Location Analysis of Delivery Parcel-Pickup Points Using AHP and Network Huff Model: A Case Study of

Shiweitang Sub-District in Guangzhou City, China. ISPRS International Journal of Geo-Information, 9(4), 193,2020

2. Gaode Map, https://ditu.amap.com/ [accessed on: August 2020]
3. OSM Ontology, http://wiki.openstreetmap.org/wiki/OSMonto [accessed on: August 2020]
4. Pearce, J., Ferrier, S, Evaluating the predictive performance of habitat models developed using logistic regression. Ecological modelling, 133(3), 225-245,2000
5. Hosmer, DW., Lemeshow, S., Applied Logistic Regression, John Wiley and Sons, New York, 160 –164, 2000

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)

SCIENCE COUNCIL OF JAPAN
日本学術会議

NICT
National Institute of
Information and
Communications
Technology

WORLD
DATA SYSTEM

ORCID
Connecting Research
and Researchers