Subject: Research Data Alliance (RDA) 7th Plenary Meeting (Japan Science and Technology Agency -JSTA- Tokyo, Japan, 1-3 March 2016)

From: Pierre-Antoine Bretonnière

To: BSC-ES

Copy: Fabrizio Gagliardi, Sergi Girona, Francesca Arcara (BSC), Emmanouil Chaniotakis (HIT)

Introduction

- Objective: These RDA plenary meetings are held every 6 months around the world and are the opportunity to discuss all the general RDA topics as well as the advancements of the work of the different Interest Groups (IG) and Working Groups (WG).

  For us in Earth Sciences, we went to this meeting to present, promote and get feedback about the creation of an IG on weather, climate and air quality that we want to launch within RDA. For this purpose, a Birds of a Feather (BoF) on the subject was scheduled the last day where I presented the state of the work. The specific and extended summary of the BoF can be found below (day 3, 11.00-12.30).

- Funding: BSC RDA

- Attendants: around 400 people from all around the world and many different scientific communities: agriculture, linguistics, publishing business and a large representation of the Earth Sciences.

  From BSC: Fabrizio Gagliardi as RDA OAB member, Sergi Girona as organizer of the upcoming 9th Plenary that will be held at the BSC.

- Agenda: https://rd-alliance.org/plenary-meetings/seventh-plenary/programme.html

  The meeting was a mix between plenary sessions and separate working meetings.

Day 1:

<u>9.30-11.00: Opening Plenary—Welcome from the organizers</u>

Mark Paersons, RDA Secretary General

We are at a period where there is a global need for more policy guidance on data. RDA, even if not a “policy organization” can bring the infrastructure for these policies, playing both local and global, at geographical and disciplinary levels.

Yuko Harayama (Executive Member, Council for Science Technology and Innovation Cabinet Office)

The way to address to science and the way we team up (scientists, citizens, engineers) has drastically changed over the last years thanks to computing power increase. In this meeting, we should propose new ways to do science for scientists and policy makers. Japan wants to use these discussions to make decisions at policy level.

Yashuo Kishimoto (Deputy Director General Science Technology Policy Bureau, Ministry of Education, Culture, Sport, Science and Technology)

This meeting, in addition to provide an opportunity for stakeholders to meet should increase awareness for the need of building a system based on open science.

Saturo Ohtake (Principal Fellow, JSTA) – organizer

The investment in science is fundamental to achieve discoveries that will serve the whole society but data is what will "feed" the science. In this context, open science and open data must be encouraged.

Michael Hager (Head of Commissioner Gunther Oettinger's Cabinet, EC) – video message

This talk was a general overview of the importance of data ("We must take actions for a better future and the future is data") and a reminder that the European Commission (EC) wants to work with RDA on interoperability of data to foster the collaboration with industries. He also reminds the collaboration with Geant on HPC and data sharing.

Panel discussion on general RDA engagement and questions from the audience

Presentation of different data use cases: European Open Science Cloud (EOSC)[1] is one of the commissioner's priority actions for 2016 under open science. It will federate existing and emerging data infrastructures and provide 1.7M European researchers free and open services for data storage, management, and analysis.

The other talks from the panel included the presentation of the Australian National Data Service (ANDS) infrastructure (from Ross Wilkinson) and a general discussion on Big Data (can mean "big interdisciplinarity"), open science (need for open science articles). It was emphasized that data analytics were needed as much as data itself.

11.30-13.00 Breakout Group (BoG) 1 – Data Fabric and National Data Services (Peter Wittenburg)

The other BoG were: "IG Agriculture Data: Results of the IGAD Pre-Meeting RDA P7: adopting RDA outputs", "IG Chemistry Research Data: Mapping the Chemistry Data Landscape", "Joint meeting of IG Data Fabric, IG National Data Services: National Data Services - responsibilities and activities in testing RDA Outputs", "IG Education and Training on handling of research data: Focus on Research -Data handling related competences and skills", "IG PID", "IG Vocabulary Services: Community Use Cases for Vocabulary Services".

This IG is called Data Fabric IG and is the union of former IGs on data fundamental terminology, data registry types, metadata standards directory, PID info types and practical policies. The data fabric includes all what concerns the creation of collections of new datasets, from "scratch" or from existing collections through processing, publishing, registering or metadata editing. All the data fabric cycle has to be improved because it costs

[1] - https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

too much time (a recent survey showed that up to 75% of the time of a researcher can be occupied by data management work), principally by integrating this cycle into a general infrastructure. Different levels of infrastructures can be identified: data generators, project infrastructures (NoMad[2], DOBES), domain infrastructures (CLARIN[3], ELIXIR[4]), and e-infrastructures (EUDAT,[5] OpenAire[6],...).

RDA Data Fabric should try and unify or at least get them together (for example as TCP/IP for Internet).

There is also a need to test the different pieces of software that would eventually come out of this IG as "RDA recommendations"[7] by defining a "testbed". The national data services (NDS) have a major role to play here.

14.30-16.00 BoG2—Joint meeting of IG Metadata, WG Metadata Standards Catalog, IG Data in Context, WG Metadata Standards Catalog: Domain Discussion on Initial Standard Canonical Set of Metadata Elements.

The other BoG were: "BoF Initial Breakout for the Data Typing Working Group", "Health Data IG: Health Data Challenges", "IG Agriculture Data: Joint Session between the RDA Interest Group on Agricultural Data (IGAD) & The Global Open Data in Agriculture and Nutrition (GODAN)", "IG Brokering: Interoperability challenges - Extending collaboration in the Broker Framework", "IG Repository Platforms for Research Data", "Joint meeting of IG Education and Training on handling of research data, WG RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World: International Coordination of Research Data Education and Training Requirements", "Initial Standard Canonical Set of Metadata Elements", "WG QoS-DataLC Definitions: Building an initial set of QoS metrics", "Organisational Assembly (OA) Meeting"

These IG/WG are trying to come up with a draft list of general metadata elements, trying to understand the metadata used and desired by each domain. They want to raise a profile of metadata about data but also about everything involved in the life of the data: softwares, physical users, computing ressources, etc...

These IG/WG stand between infrastructure groups (data fabric, PIDs, ...) and domain groups.

A first draft of a metadata catalogue has been published and contains, for example, the following fields: organization, person, projects, name/title, keywords, spatial coordinates, temporal coordinates, classification term, indicator,...

This is still a work in progress so they invite every domain to provide use cases and list of metadata that could or should be evaluated for this list.

2 - https://www.hashicorp.com/blog/nomad.html

3 - http://clarin.eu/

4 - https://www.elixir-europe.org/

5 - http://www.eudat.eu/

6 - https://www.openaire.eu/

7 - https://rd-alliance.org/groups/working-group-process-task-force/wiki/criteria-rda-recommendations.html

16.30-18h BoG3 - BoF on Data Search

The other BoG were: "IG Archives and Records Professionals for Research Data", "IG Geospatial", "IG Global Water Information", "IG Long tail of research data: Managing diverse data sets: challenges and incentives", "IG Data Fabric: Machine readable Repository Registries for large Federations", "Joint meeting of IG ELIXIR Bridging Force, WG BioSharing Registry: Reference data in the life sciences".

This session aims at identifying if there is a need for an IG on this topic and, if so, define what should be its objectives.

Several data search engines were exposed:

- research data switchboard (Australian National Data Center).
- biocaddie[8] (biomedical and health care data discovery index ecosystem) - based on Json and ElasticSearch.
- BCube crawler[9] (for geosciences).
- PANGAEA (about earth sciences data) [10]
- earthchem.org search for chemistry

After a very technical debate, the discussion ended up with the idea that the mission of an IG on Data Search could be to create a list of typical search engines used by the different RDA communities and build a test suite to compare them.

Day 2

9.00-10.30: Plenary Session – Chaired by Herman Stehouwer, RDA Secretariat and Max Planck Computing and Data Facility (MPCDF) RDA Recommendations & Outputs Session

This session consisted in a series of brief presentations of the different "RDA recommendations" from the diverse IG and WG, often in collaboration with the working groups from the World data system organization (WDS)[11].

09.00-09.10: Introduction to RDA Recommendations and Outputs, Herman Stehouwer, RDA & Mustapha Mokrane, ICSU-WDS

09.10-09.20: RDA/WDS Publishing Data Bibliometrics Recommendations, Kerstin Lehnert, Columbia University

09.20-09.30: RDA/WDS Publishing Data Services Recommendations, Adrian Burton, ANDS & Hylke Koers, Elsevier

09.30-09.40: RDA/WDS Publishing Data Workflows Recommendations, Amy Nurnberger, Columbia University

[8] - https://biocaddie.org/

[9] - http://earthcube.org/forum/bcube/access-geoscience-data-services-discovered-bcube-crawler

[10] - www.pangaea.de

[11] - https://www.icsu-wds.org/

09.40-09.50 OECD Adoption of the Cost Recovery Models, Carthage Smith, OECD

10.00-10.10: RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World Interim Recommendations, Simon Hodson, CODATA

10.10-10.20: RDA Outputs as ICT Technical Specifications, Hilary Hanahoe, RDA Europe & Trust-IT Services Ltd.

<u>11.00-12h30 BoG 4 – IG Big Data: Considerations in applying Big Data technologies</u>

The other BoG were: "IG Data Foundations and Terminology: expanding vocabulary coverage and services", "IG Digital Practices in History and Ethnography: Updates and Working Group Initiative", "IG Research data needs of the Photon and Neutron Science community: Open data sharing - User Facility Opportunities and Challenges", "Joint meeting of IG Data Rescue, IG Data Fabric, IG Preservation e-Infrastructure, IG Domain Repositories, IG Libraries for Research Data: Rescuing, Re-Using and Sharing Data At Risk", "WG Data Citation - Making Dynamic Data citable: Adoption of the Recommendations", "WG RDA/NISO Privacy Implications of Research Data Sets".

This session consisted in a general presentation of the technology of the array database, followed by some practical considerations about how to increase the exchanges within this WG.

The array database technologies are databases supporting queries on massive n-D (generally up to 5 or 6) arrays, highly scalable and tested on datasets of 130+ TB. It could be particularly interesting for us in Earth Sciences as it is used in various Big Data projects from our community (Copernicus, Earth server2), allows direct visualization and see the dataset no longer as a series of files but as a "data cube". There is also an EUDAT WG on the subject. Contact has been made with the developers to see how/if this could be a good option for our local research on Big Data.

<u>14.00-15.00 Plenary Session – Keynote presentation: Masaru Kitsuregawa, Director General, National Institute of Informatics (NII), Power of Data: from scientific discoveries to Societal Benefits</u>

The NII is a research institute also dedicated to education in informatics and, in addition, provides an internet network (SINET) for a large part of Japan. The presentation was a series of practical use cases of data in everyday life:

- How to reduce the number of car accidents in Tokyo.
- Data driven flood mitigation (based on last years extreme events in Japan).
- How to improve the nurse work analysing their most time consuming activities through data analysis.

<u>15.00-16.30 BoG 5 – BoF on Research Data Repository Interoperability: Preparation of the Research Data Repository Interoperability WG</u>

The other BoG were: "BoF on IG Management and Curation of Physical Samples: Developing Physical Samples Management and Curation Best Practices", "IG National Data Services", "IG RDA/CODATA Legal Interoperability: Testing the 'Principles' and 'Implementation Guidelines' for the Legal Interoperability of Research Data", "IG Vocabulary Services: Access Methods Review of Existing Vocabulary Services", "Joint meeting of IG RDA/CODATA Materials Data, Infrastructure & Interoperability, IG Chemistry Research Data, IG Research data needs of the Photon and Neutron Science community: A Discussion on general and domain-

specific metadata approaches for Chemistry, Materials, and Photon & Neutron Data", "WG Brokering Governance: Broker (and software) sustainability - models and options".

Started in the P6 in Paris, this WG has for goal to establish standards for interoperability between different research data repositories. Some use cases could be the deposit of digital objects in a data repository, their retrieval, pulling objects from different repositories and using them all at the same time.

The work of this WG is at the stage of the definition of a precise workflow of things that should be addressed on a list (to be defined) of different repositories. There has been a big debate during this session to know if this WG had to concentrate on a concrete list of repositories and technologies or aim for bigger and more general without pre-defined examples in mind and just try to create a sort of "super-tool" capable of acting above the existing technologies.

Day 3

9.00-10.30: BoG 6 – IG Data Fabric: From testing RDA output to widely agreed recommendations—Beth Plale

The other BoG were: "BoF on Metadata Standards for attribution of physical and digital collections stewardship", "BoF on Text and Data Mining: Text and Data Mining: Defining the Challenges and Actions", "IG Libraries for Research Data: Applying Global Information-sharing and Collaboration in Libraries to Local Practice", "IG RDA/CODATA Materials Data, Infrastructure & Interoperability: Current & Future Efforts", "Joint meeting of IG RDA/WDS Publishing Data Cost Recovery for Data Centres, IG Domain Repositories: Business Models for Data Repositories, an OECD Global Science Forum Project", "Joint meeting of IG Vocabulary Services, IG Data Foundations and Terminology: Exploring Use Cases for Data Foundation and Terminology Vocabulary Services within the RDA", "WG RDA/WDS Publishing Data Workflows: Incorporating Publishing Data Workflows into the Research Cycle", "WG Data Security and Trust: Start building the trust".

The objective of this IG is to determine how to make the community outside of RDA adopt the recommendations from RDA in general and the Data Fabric in particular. To do this, the main strategies and points discussed during this session are the following ones:

- run an inductive examination of fabrication composition, the starting point preferably being a RDA recommendation with a reference software, to go to RDA recommendations that are purely human consumption and actions.
- the success of the data fabric will likely run on possibly distributed e-infra (EUDAT, NDS), serve scholarly domain as domain infrastructure, support multiple projects within that domain and result in cross-domain research.
- identify incentives for e-infrastructures providers to provision core compositions in experimental mode.
- assess will of group to take on challenge and determine next steps.

11.00-12.30: BoG 7 – BoF on Creation of a RDA IG on weather, climate and air quality

The other BoG were: "BoF e-Infrastructure for Global Change Research", "BoF on IG-VRE (Virtual Research Environment): Kick-Off Meeting to establish IG", "IG Data Fabric: Data Fabric and Common Components - state *and perspectives”, “IG RDA/WDS Certification of Digital Repositories: Reaping the fruit and sowing the seeds”, “WG RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World: CODATA-RDA schools in Research Data Science”, “WG RDA/WDS Publishing Data Services: We proudly present an open, universal literature/data interlinking service”.*

The slides from the presentation are available on the wiki[12]

Around 15 people attended the BoF, from different branches of the Earth Sciences (climate, air quality, climate services, ...) and from several institutes around the world.

After a quick round table to know where people were coming from and what their interest in this session was, we introduced the motivation for such an IG and presented the conclusions of the workshop held at the BSC the 11th of February[13]. From the summary of the discussion of the Barcelona workshop, started a debate on the points discussed. The main conclusions of the discussions are the following ones:

- Even if storage is cheap, it should be considered when talking about storage, not only about the data storage at the central repositories but also about the one that every users use when they download the data.
- The conclusion that data should be brought to the compute has been emphasized, the actual need is no longer data centres but data analysis centres. Then some dedicated storage solutions could be designed specifically for our usage. Then came the issue about “who is paying for that when you need data analytics”. A point was made that, in a same way as there are international calls for HPC computing hours, there could be some dedicated calls to get “data analysis hours” on HPCs.
- The point that was mentioned during the BSC workshop about data reproduction and the reflection about what to keep and what simulations to redo was also discussed.
- About metadata and at which level the standards and requisites should be defined, it was mentioned that several sets of metadata could be used and defined: one general and cross community (following the work of the metadata IG) and one a community level.

After this first discussion, we had the presentations of 2 use cases of weather, climate and air quality data: one from Emmanouil Chaniotakis from the National Technical University of Athens about “Using weather, climate and air quality data in transportation and disaster management” and the other one from Varsha Khodiyar from Scientific Data[14] on publication of scientific (in particular climate) datasets.

Then, we had a final discussion on if there was a real interest from the community on such an IG and the next steps to follow.

Except from one person that thought that this IG could fit in the Data Arrays IG, the idea seemed to be that such an IG would be useful to discuss data issues particular to our communities and gather people from these sectors.

The next steps to follow will be:

[12] - https://earth.bsc.es/wiki/lib/exe/fetch.php?media=library:external:20160303_bretonniere_rda_bof.pdf

https://earth.bsc.es/wiki/lib/exe/fetch.php?media=library:external:20160303_bretonniere_rda_bcn_workshop.pdf

[13] - https://www.bsc.es/about-bsc/press/bsc-in-the-media/bsc-organises-rda-earth-sciences-interest-group-workshop

[14] - http://www.nature.com/sdata/

- Refine the objectives from the IG following the BSC workshop and this BoF
- Circulate a summary of this BoF to the participants and through different mailing lists
- Once we reach a consensus with the users on the precise objectives of the IG, fill in the official papers from RDA to create officially the IG.

12.30-13.00: Closing of the meeting

- Thanks from Hilary Hanahoe and presentation of the 2 next plenaries (Denver[15] and BSC[16]).

Results

- In addition to the specific points mentioned above in this mission report and the results from the IG, here is a list of points/potential collaborations that came out of this meeting:
- A potential collaboration with Emmanouil Chaniotakis from the HIT, who gave a presentation during our BoF might be considered. They seem to have activities related to the ones of the Earth system services team here at BSC. They are looking for partners to participate in future European calls. Further information will be provided to Albert Soret.
- The WG on Array Database (former Big Data) might be worth following as the technological solutions they are investigating seem to be well designed for Earth Sciences
- The publication of some of our work (SPECS for example) on data in journals such as Scientific Data should be considered. It can be a great opportunity to give the project another kind of visibility.

[15] - https://rd-alliance.org/plenary-meetings/rda-eighth-plenary-meeting.html

[16] - https://rd-alliance.org/plenary-meetings/rda-ninth-plenary-meeting.html