Report of the Europe Research Data Alliance Meeting on Data Provenance Approaches

The RDA Europe meeting on Data Provenance Approaches took place during the 15th and 16th January 2018 ath the Barcelona Supercomputing Center, aiming at presenting some of the ongoing activities on data provenance in public and private European institutions or projects, discuss about the main challenges in the field, and identify best practices or possible ways for collaboration.

Organization committee

Nicolau Manubens - Barcelona Supercomputing Center
Pierre-Antoine Bretonnière - Barcelona Supercomputing Center
Alasdair Hunter - Barcelona Supercomputing Center
Kim Serradell - Barcelona Supercomputing Center
Amalia Hafner - Barcelona Supercomputing Center

Participants

Pierre-Antoine Bretonnière - Barcelona Supercomputing Center
Paolo Missier - Newcastle University
Joaquín Bedia - PREDICTIA Intelligent Data Solutions
Sanaz Moghim - Sharif University of Technology
Alessandro Spinuso - Royal Netherlands Meteorological Institute
Javier Vegas - Barcelona Supercomputing Center
Barbara Magagna - Environment Agency Austria
Ge Peng - North Carolina State University, CICS-NC/NCEI (remotely)
Stian Soiland-Reyes - University of Manchester (remotely)
Francisco Doblas-Reyes - Barcelona Supercomputing Center, ICREA
Nicolau Manubens - Barcelona Supercomputing Center
Kim Serradell - Barcelona Supercomputing Center
Alasdair Hunter - Barcelona Supercomputing Center
Antonio Cofiño - Universidad de Cantabria

Number of participants from the private sector: 1 (7%)
Number of participants from the public/investigation sector: 13 (93%)

Minutes

Pierre-Antoine Bretonnière opened the event with an overview of the **RDA** structure and activities.

Paolo Missier followed with an explanation on the activities conducted by the **RDA Provenance Patterns Working Group**[1] **(WG)**. In tight collaboration with W3C and its PROV model for encoding provenance information, the WG has built the **Provenance Patterns Data Base (PPDB)**[2], a database with recommended patterns and anti-patterns (those recommended not to follow), including examples and indications for interested users to easily choose the most beneficial pattern for their case. **The participants were encouraged to contribute to the database** with use cases and feedback. Paolo also presented the activities and solutions for Earth sciences implemented by **DataOne**[3], including an extension of PROV called ProvOne, and a graphical web interface to explore provenance of generated products. The **provenance templates**, helpful to represent complex provenance patterns with repeating patterns (developed at King's College London) and the **ReComp project** for re-running a certain part of a workflow when its inputs have been altered (developed at Newcastle University) were also presented.

Discussion*: it was discussed how to foster future collaboration with the PPDB. A solution proposed was the inclusion of contribution of use cases to the PPDB as part of next proposals.*

Joaquín Bedia followed with an explanation of **METACLIP**, a provenance model and tool developed to provide reproducible climate analysis products with attached provenance information, and which can be dragged/dropped into a web application that displays the provenance information in a human-friendly way. The solution builds upon the PROV model, which is translated into the **Resource Description Framework** notation (another W3C standard), and extended with other RDF models (a.k.a. vocabularies) developed either by the community or specifically for the project. R software implementing the METACLIP model has been developed.

Discussion: *METACLIP will be further evolved to align as much as possible with the PROV data model.*
Discussion: *PROV-O + RDF is one of the most mature and commonly used approach, although other tools exist to work with the PROV-N notation (the python module ProvToolbox).*
Discussion: *the solution explained in this talk followed one of the anti-patterns presented in the previous talk. According to the recommendations of the RDA Provenance WG, provenance information should not be attached directly to generated products, but rather in a database which allows for search and discovery.*

[1] https://www.rd-alliance.org/groups/provenance-patterns-wg
[2] http://patterns.promsns.org/
[3] https://www.dataone.org/

Discussion: *METACLIP will check compliance (or not) with patterns in the PPDB and provide feedback.*
Discussion: *the speaker pointed out the current challenge with trying to keep provenance information size as reduced as possible.*

Alessandro Spinuso followed with an explanation of the DARE project and **S-ProvFlow**, a provenance model and management system which results from extending PROV and ProvOne and focuses on data-intensive platforms (computing environments with complex architectures where large amounts of complex data are processed potentially in parallel, and where tracking the behavior of complex streaming operators in the context of the application domain is important to analyse the results). Python software is available to define workflows following the S-ProvFlow model and to execute them on computing platforms of different kinds. The use of workflows together with provenance mechanisms allows to follow the so-called **FAIR principles**[4] (Findable, Accessible, Interoperable, Reusable). S-ProvFlow offers a web application to query and visualize the workflow's lineage interactively. It enables fine-grain analysis and comprehensive views. Configurable wheel plots allow users to explore re-use of resources in collaborative working environments.

Discussion: *there is no concrete plan yet for implementing a provenance solution in ESMValTool. Alessandro is acting as an advisor to the ESMValTool development team.*

Barbara Magagna explained the progress done in **ENVRIplus**[5], a European project with several environmental research infrastructures (RI) involved, aiming at providing solutions to the research community. She introduced the **ENVRI-Reference Model (RM)**, a conceptual framework used to describe problems and solutions to problems from different viewpoints (e.g. scientific, computational, …) using shared vocabularies (ontologies) between research infrastructures. The information is finally ingested and served by a database called the Knowledge Base. The ontologies are available online in the **Ontology Web Language** format, another W3C standard. The option of extending the ENVRI-RM with provenance-specific capabilities will be explored, and efforts will be done to make the RM compatible with the PROV-O standard. ENVRIplus is in place to identify provenance requirements and use cases and to foster interaction on provenance matters between projects involved and external projects (e.g. EUDAT, RDA).

Discussion: *ENVRIplus is not planning to implement a governance solution, the RIs involved will implement separately their own provenance solution with the guidance (assessment of best standards / tools) offered by ENVRIplus.*
Discussion: *ENVRIplus will collaborate with the RDA PPDB by checking compliance (or not) with the available patterns, and will provide feedback, collected requirements and use cases. ENVRIplus will be represented at the next 11th RDA Plenary Meeting in Berlin.*

---

4 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
5 http://www.envriplus.eu/

Discussion: *action will be taken in February/March to properly adapt the ENVRI RM to the PROV model.*

Ge Peng introduced the **ESIP Information Quality Cluster[6] (IQC)**, an association with representatives from a large number of entities and projects at the USA and Europe levels. **The participants were encouraged to join the IQC**. Membership is open and participation is voluntary. The IQC aims at becoming an authoritative and internationally recognized resource for data and information quality, promoting standards and best practices. IQC activities include monthly teleconferences (with invited/contributed talks discussing about use cases, issues and guidelines for large data stakeholders and administrators). The IQC also organizes and participates in events on information quality, and issues publications on the work done. Some of the most relevant efforts consist in prompting **maturity matrices** that allow to quantify the grade of maturity of a product in terms of scientific, product, stewardship and service quality.

Discussion: *the ESIP IQC can be a testbed for work done at various institutions in the USA. ESIP IQC does not have authority to dictate implementation at individual organizations but provides a place to obtain community feedback and develop guidelines on consistent implementation of community best practices, particularly those on ensuring and improving Earth science data quality.*

Stian Soiland-Reyes followed with an explanation on the variety of workflow managers and languages, and introduced the **Common Workflow Language[7]** as a generic workflow definition language. Version Control Systems such as Git on GitHub were proposed as solution to uniquely and concisely identify versions of workflow definitions, and the use of container technologies (such as Docker[8] and Singularity) as a solution to uniquely identify tools used in a workflow as well as specific computational environment configuration. He explained how to cleanly represent complex workflows with **PROV-N[9]** (the W3C standard notation for provenance graphs). Finally he presented the **Research Object[10]** project as a solution to the large variety of standards for representing and serializing provenance information. The so-called Research Object manifest allows to declare which inputs and standards were used in the generation of a research outcome (document, image, …) as well as general provenance information. He also introduced **BagIt[11]**, a mechanism that can complement Research Object and is useful to safely transfer and archive the generated products.

Francisco Doblas-Reyes explained the need for provenance and information quality control in the context of the Climate Change Service[12] being developed as part of the **Copernicus** project. Reproducibility, traceability and user guidance are key to **build trust** on the products

[6] http://wiki.esipfed.org/index.php/Information_Quality
[7] http://www.commonwl.org/
[8] http://docker.com/
[9] https://www.w3.org/TR/prov-n/
[10] http://www.researchobject.org/
[11] https://en.wikipedia.org/wiki/BagIt
[12] https://climate.copernicus.eu/

provided in a service, and this is why quality assurance is becoming more relevant in this kind of projects.

Sanaz Moghim and Javier Vegas explained the activities carried out at Sharif University of Technology and Barcelona Supercomputing Center, respectively, stressed out the need for data provenance solutions and pointed out some of the **provenance-related issues** experienced: the need for a comprehensive solution that is able to track provenance in a heterogeneous environment, which in turn does not exclude non-provenance-friendly tools; the large size of the provenance information when it is stored at too high resolution; and the need for provenance information to be always accessible and attached somehow in the products.