Using data for system-level science: A provenance perspective

Barbara Magagna[1] Malcolm Atkinson[2] Markus Stocker[3,4]

[1]Environment Agency Austria, Spittelauer Lände 5, 1090 Vienna, Austria, barbara.magagna@umweltbundesamt.at

[2]University of Edinburgh, School of Informatics, 10 Crichton Street EH8 9AB Edinburgh, Scotland, malcolm.atkinson@ed.ac.uk

[3]TIB Leibniz Information Centre for Science and Technology and Leibniz University of Hannover, 30167 Hannover, Germany, markus.stocker@tib.eu

[4]PANGAEA Data Publisher for Earth & Environmental Science, MARUM Center for Marine Environmental Science, 28359 Bremen, Germany, mstocker@marum.de

Introduction

The quantities researchers report in scientific literature, say summary statistics such as 8:00 for the mean duration of a studied phenomenon, are generally the result of complex workflows. While not always obvious from reading the reported materials and methods, such values may be derived from numbers generated by an instrument of an observatory; acquired, curated, and published by a research infrastructure; processed using one or more computational models; and interpreted by a postgraduate student supervised by a postdoc who may ultimately derive the reported summary statistics. In using environmental data for system-level science we have thus much provenance as a side product. Unfortunately, such provenance is seldom recorded systematically. Building on a use case in aerosol science, specifically the study of new particle formation events, we discuss one approach for how infrastructure can support the specification and execution of complex workflows "as a service" to research communities.

Workflow

Primary (observational) data for particle size distribution at given spatio-temporal locations are published by research infrastructures and can be obtained programmatically, for instance using the SmartSMEAR API (https://avaa.tdata.fi/web/smart/smear/api) of the Station for Measuring Ecosystem Atmosphere Relations (SMEAR) research infrastructure [1]. Using a computational environment of their choice, researchers visualize primary data (Figure 1) to determine the occurrence of a new particle formation event at the given spatio-temporal locations. The result of primary data interpretation is secondary data describing the event, in particular when and where it occurs, its classification, duration, growth rate and other attributes. Finally, secondary data are used to compute, e.g., summary statistics, such as mean duration of events (Figure 2). These are tertiary data that may be reported in the scientific literature.

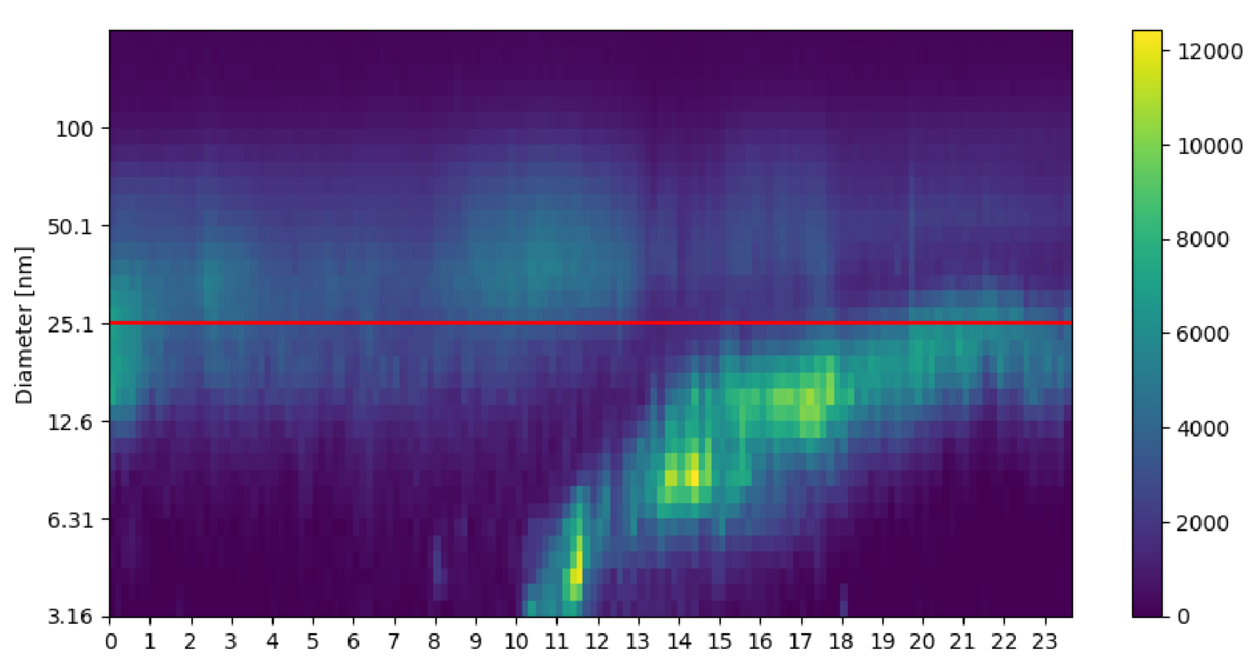

Figure 1: Visualization of primary data.

Provenance

Primary, secondary and tertiary data are entities. In our workflow, mean duration of events (tertiary data) are entities derived from a set of event descriptions (secondary data) which themselves are derived from particle size distribution data (primary data) (Figure 3). Various agents and activities are involved, in particular human (researchers) and computational agents and the 'data visualization' and 'averaging data transformation' activities. Relationships between such entities, agents and activities can be acquired, curated and potentially published and processed by infrastructure.

```
In [1]: from pynpf.processing.statistics import duration
        from pynpf.factory import events, record

        # Compute the average duration of events, possibly on a specific day and/or place
        d = duration(events(), fun='avg',
                     prov={'agent': 'https://orcid.org/0000-0001-5492-3212'})

        print(d.value())

8:00:00
```

Record the computed average duration, for instance if it ought to be published in a paper as a result.

This records the computed average duration as average value with scalar value specification, that is a numeric duration with unit type hour, whereby the average value is about the dataset of events for which the average duration was computed. This also records the provenance of the average value as it was derived from the dataset of events, including involved agent and activity of averaging data transformation.

As a result, the computed average duration is an identified resource and could potentially be referred to in published literature.

```
In [ ]: record(d)
```

Figure 2: Recording of tertiary data.

Problems

In the data use phase of the research data lifecycle, researchers currently tend to download data as they are published by research infrastructures onto a local computational environment. This raises issues:

- Infrastructural discontinuity
- Systematic recording of provenance
- Heterogeneity of secondary data
- Systematic acquisition of secondary and tertiary data

```
In [2]: query("""
            select ?entity2 ?entity1 ?activity where {
              ?entity2 prov:wasDerivedFrom ?entity1 .
              ?entity2 prov:wasGeneratedBy [ rdfs:label ?activity ] .
              ?entity2 prov:wasAttributedTo <https://orcid.org/0000-0001-5492-3212> .
            } order by desc(?activity) limit 3
        """)

        query("""
            select ?p ?o where {
              smear:eb1ad ?p ?o .
            }
        """)

     entity2       entity1                        activity
  0  smear:eb1ad   file:2013-04-04-hyytiaelae.csv data visualization
  1  smear:5db1b   smear:dc3cd                    averaging data transformation


     p                         o
  0  prov:wasAttributedTo      https://orcid.org/0000-0001-5492-3212
  1  smear:hasClassification   smear:ClassIa
  2  prov:wasGeneratedBy       http://purl.obolibrary.org/obo/OBI_0200111
  3  rdf:type                  linkedevents:Event
  4  rdf:type                  prov:Entity
  5  linkedevents:atTime       smear:92be5
  6  prov:wasDerivedFrom       file:2013-04-04-hyytiaelae.csv
  7  linkedevents:inSpace      smear:7f885
  8  linkedevents:atPlace      geonames:656888
```

Figure 3: Provenance between primary, secondary, and tertiary data.

Implementation

For the presented use case in aerosol science, we propose a Jupyter [2] based workflow implementation operated "as a service" to the research community on the European Grid Infrastructure (EGI). Operated "as a service," the federated infrastructure involving both research infrastructures and e-Infrastructure is connected. It avoids (primary) data being downloaded and is "aware" of the workflows executed. It can thus systematically record provenance. Furthermore, it harmonizes the representation of secondary and tertiary data, specifically descriptions about new particle formation events and computed quantities such as mean duration of events. Finally, secondary and tertiary data are systematically acquired by research infrastructure, guaranteeing the curation and, possibly, the publication of such data, thus enabling their further processing—and the closure of the research data lifecycle. We adopt semantic web technologies and represent secondary and tertiary data in RDF. Following a concept of the Ontology for Biomedical Investigations (http://purl.obolibrary.org/obo/OBI_0000679), tertiary data are data items produced as the output of an averaging data transformation (the activity) representing the average value of the input data (the entity, here a set of event descriptions). Provenance of entities, involved agents and activities is represented using the PROV Ontology [3].

Discussion and Conclusion

We are attempting to actively involve the research community. First, the community should agree on how to represent secondary data describing new particle formation events. A first step towards harmonized representation was taken by introducing a relevant concept in the Environment Ontology (http://purl.obolibrary.org/obo/ENVO_01001085). Second, the research community should ultimately adopt the proposed service and perform their data driven science workflows on research infrastructure, rather than on local computational environments. These are arguably major steps for this research community, steps that require addressing further issues including the systematic publication of secondary data and the collaborative development and use of software but also the maturity of the approach.

References

[1] Hari P. et al. (2013) Station for Measuring Ecosystem-Atmosphere Relations: SMEAR. In: Hari P., Heliövaara K., Kulmala L. (eds) Physical and Physiological Forest Ecology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-5603-8_9

[2] Pérez, F. Granger, B. E. (2007) IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering, 9(3):21-29. https://doi.org/10.1109/MCSE.2007.53

[3] Lebo, T. et al. (2013). PROV-O: The PROV Ontology. W3C Recommendation.

Acknowledgements

This work has received support by the ENVRIplus project, which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182.