TC_17 Connecting the particle formation research community to research infrastructure

1. Background

1.1 Short description

Particle formation is an atmospheric process whereby at specific spatial locations aerosol particles form and grow in diameter size over the course of a few hours. Particle formation is studied for its role in climate change and human respiratory health.

To study these processes, particle formation needs to be detected for where and when it occurs. Having detected particle formation, the processes are characterized for their qualities, e.g. duration, growth rate and other attributes. The detection and characterization of atmospheric particle formation relies on the measurement of particle size distribution, typically using an instrument called Differential Mobility Particle Sizer (DMPS).

In the context of particle formation research, particle size distribution as measured by a DMPS is *observational data* – in other words primary, uninterpreted data. For each day and location, observational data are processed and interpreted to detect and characterize particle formation. Observational data processing and interpretation are carried out by one or more human experts (typically postgraduate students). This constitutes an *in silico* (i.e., performed on computer) and *human-in-the-loop* scientific workflow. In the context of particle formation research, the output of such workflow is *information* describing (i.e., about) individual particle formation processes.

Information is truthful, meaningful, well-formed data (Floridi, 2011) – in other words secondary, interpreted data. Information is commonly also referred to as “data + meaning” and is also known as “data product.” Meaning is created in workflow execution, in which human experts also ensure that the resulting meaningful well-formed data are truthful. Information describing individual particle formation is further processed into summary statistics, e.g. the average duration. Such summary statistics are ultimately reported in scientific literature.

The use case aims to, primarily, (1) harmonize the information describing particle formation; (2) represent information, specifically the meaning of data, using an appropriate computer language; and (3) acquire and curate information in infrastructure.

1.2 Contact

Background	Contact Person	Organization	Contact email
ICT	*Markus Stocker*	*TIB, PANGAEA*	*markus.stocker@gmail.com* *markus.stocker@tib.eu* *mstocker@marum.de*
RI-Domain	*Jaana Bäck*	*University of Helsinki*	*jaana.back@helsinki.fi*
RI-Domain *RI-ICT*	*Markus Fiebig*	*NILU*	*markus.fiebig@nilu.no*
e-Infrastructure	*Yin Chen*	*EGI*	*yin.chen@egi.eu*
e-Infrastructure	*Yann Le Franc*	*EUDAT*	*ylefranc@gmail.com*
ICT	*Leonardo Candela*	*CNR*	*leonardo.candela@isti.cnr.it*
ICT	*Robert Huber*	*UniHB, PANGAEA*	*rhuber@uni-bremen.de*
ICT	*Paul Martin*	*UvA*	*p.w.martin@uva.nl*
ICT	*Barbara Magagna*	*EAA*	*barbara.magagna@umweltbundesamt.at*

1.3 Use case type

Test Case

1.4 Scientific domain and communities

Scientific domain

Atmosphere

Community

Data Use, Data Acquisition (primarily)

Data Curation (secondarily)

Data Publication (possibly)

Behavior

Relevant Data Use Community Behaviors

- Co-create: Aerosol scientists, possibly in collaboration with ICT specialists, design and plan the analysis of particle size distribution observational data. RI specialists design and plan the collection, preservation, and possibly publishing of information describing particle formation.
- Collaborate: Aerosol scientists participate in interpretation of particle size distribution observational data. RI specialists participate in the collection, preservation, and possibly publishing of information describing particle formation.
- Contribute: Aerosol scientists directly interpret particle size distribution observational data held by research infrastructures (i.e., SMEAR, ACTRIS), according to a predefined protocol. RI specialists directly collect, preserve, and possibly publish information describing particle formation.
- User Working Space Management: ICT and e-Infrastructure specialists support work spaces that allow the interpretation of particle size distribution observational data and the acquisition of information describing particle formation in infrastructure.

Relevant Data Acquisition Community Behaviors

- Data Collection: The information describing particle formation is created by a human expert (acting as “sensor”) and is a result of interpreting particle size distribution observational data. The Data Collection behavior is performed by the RI Data Collector.

Relevant Data Curation Community Behaviors

- Data Quality Checking: The RI Data Curator detects and corrects (or removes) corrupt, inconsistent or inaccurate information describing particle formation.
- Data Preservation: The RI Data Curator deposits over long-term the information describing particle formation.
- Build Local Conceptual Model: The Semantic Curator, an ICT specialist, builds an ontology design pattern for information describing particle formation.
- Data Annotation: The Semantic Curator supports linking information describing particle formation with the ontology design pattern (local conceptual model). This behavior is performed by software agents of the scientific workflow.

Relevant Data Publication Community Behaviors

- Data Publication: Performed by the RI, this behavior provides information describing particle formation by following specified publication and sharing policies.
- Semantic Harmonisation: Since the information describing particle formation is semantically harmonized, this behavior is performed by software agents of the scientific workflow, before Data Collection. Assumed is community agreement for the ontology design pattern.

Roles

Relevant Data Use Community Roles

- Scientist: An active role, a researcher who executes the scientific workflow, thereby interpreting particle size distribution observational data to detect and characterize particle formation, thereby generating information describing particle formation.
- Engineer: An active role, an ICT specialist who develops the required RI components to acquire, curate, and possibly publish information describing particle formation. Furthermore, a person who implements and deploys the scientific workflow as a service.
- Data Use Subsystem: A passive role implementing the scientific workflow to fetch particle size distribution observational data from RIs (i.e., SMEAR, ACTRIS), interpret observational data to detect and characterize particle formation, and represent and collect information describing particle formation in infrastructure.

Relevant Data Acquisition Community Roles

- Data Collector: A passive role, namely a software agent of the scientific workflow that submits information describing particle formation to infrastructure.
- Data Acquisition Subsystem: A passive role of the RI providing functionalities for automated information acquisition.

Relevant Data Curation Community Roles

- Data Curator: An active role, a person of the RI who verifies the quality of (acquired) information describing particle formation.
- Semantic Curator: An active role, an ICT specialist who designs and maintains the ontology design pattern for information describing particle formation. The annotation of data with this pattern is done automatically in information representation by a software agent of the scientific workflow.
- Data Curation Subsystem: A passive role of the RI that stores, manages, and ensures access to information describing particle formation.

Relevant Data Publication Community Roles

- Data Originator: A passive role, an RI component that provides information describing particle formation to be made available for (public) access.
- Data Repository: A passive role, an RI component that is the facility for the deposition of published information describing particle formation.
- Data Publisher: An active role, a person of the RI in charge of supervising the information publishing processes.
- Data Publishing Subsystem: A passive role of the RI that enables the discovery and retrieval of information describing particle formation.

2. Detailed description

Section 1.1 provides a summary of the primary aims of this use case. We begin this section by providing a more detailed description of the aims. Where applicable, we discuss how these aims align with FAIR Principles (Wilkinson, 2016). Aims marked optional will be addressed if time permits.

In detail, the use case aims at the following:

- 1.As a community effort, harmonize the information describing particle formation. Specifically, harmonize the used vocabulary. This aim addresses the following FAIR principles: Data and metadata use vocabularies that follow FAIR principles (I2); Data and metadata are richly described with a plurality of accurate and relevant attributes (R1); Data and metadata meet domain-relevant community standards (R1.3).
- 2.Link information describing particle formation with other relevant information, specifically external vocabularies (e.g., for time and space) as well as related descriptions (e.g., locations as provided by a gazetteer such as GeoNames[2]). This aim addresses the following FAIR principle: Data and metadata include qualified references to other data or metadata (I3).

- 3.Represent information describing particle formation using a computer language for information representation. Specifically, represent meaning (in addition to data). This aim addresses the following FAIR principle: Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation (I1). The Web Ontology Language[3] is considered as the language of choice in this use case.
- 4.Implement and publish an ontology design pattern that reflects the harmonized description, information linking, and information representation proposed in aims 1-3. It is proposed for this pattern to be part of the Environment Ontology[4].
- 5.Adopting the solutions proposed in aims 1-4, implement the scientific workflow in Jupyter[5] (Perez and Granger, 2007) and deploy the implementation on e-Infrastructure. Specifically, expose the scientific workflow as a service used by the particle formation research community, thereby connecting the research community to infrastructure.
- 6.Systematically acquire and curate information describing particle formation in infrastructure. Specifically, research infrastructures, e-Infrastructures, data centers such as PANGAEA[6], or similar. If not achievable, provide a concept for systematic acquisition and curation on institutional systems (possibly including individual workstations).
- 7.(Optional) Support computing summary statistics (or other processing) on curated information describing particle formation. Implemented in Jupyter.
- 8.(Optional) Represent, acquire and curate summary statistics in infrastructure. The Statistics Ontology[7] is the specialized language of choice in this use case for representing summary statistics.
- 9.(Optional) Represent, acquire and curate provenance relating (summary statistics) to information describing particle formation and to particle size distribution observational data, as well as the involved agents (e.g., researchers) and activities (e.g., data interpretation). This aim addresses the following FAIR principle: Data and metadata are associated with detailed provenance (R1.2). The PROV Ontology[8] is the specialized language of choice in this use case for representing provenance.

Objective and Impact

There exist multiple, institutionally and geographically distributed, research groups that perform the scientific task of interpreting particle size distribution observational data to detect and characterize the occurrence of particle formation at determinate spatiotemporal locations. Two groups well-known to the authors of this use case are the Atmospheric Aerosol Physics[9] research group at the University of Eastern Finland and the Aerosol Cloud Climate Interactions[10] research group at the University of Helsinki.

The first objective is to harmonize how these groups describe particle formation – i.e., harmonize the information (as data + meaning) created as a result of observational data interpretation via the scientific workflow. The objective addresses aims 1-4. To catalyze this predominantly community driven work, we aim to organize a workshop (Q2 2018, Helsinki). The workshop brings together representatives of the research community, research infrastructures, e-Infrastructures, and ICT specialists. The objective of the workshop is to advance, primarily, Aim 1, and address, secondarily, Aim 5 and Aim 6. The results of the workshop will inform Aim 4, i.e. the development of an ontology design pattern.

A concrete proposal for such a pattern to build on has already been developed. As part of this work, the Environment Ontology has recently introduced a concept for “formation of particles in an atmosphere”[11]. Figure 1 describes the concept as visualized by the EMBL-EBI Ontology Lookup Service.

Figure 1: The concept "formation of particles in an atmosphere" of the Environment Ontology, as visualized by the EMBL-EBI Ontology Lookup Service. The concept is described as an atmospheric particle formation process that occurs in an atmosphere and has some aerosol as output. The concept is specialized in formation of liquid droplets or solid particles.

It is proposed that this use case builds on and extends this concept. This approach ensures that information describing particle formation conforms to the FAIR I1, I2, I3, R1, and R1.3 principles. The approach is thus expected to contribute substantially to improving the interoperability and reusability of information describing particle formation, i.e. data and meaning created by the research community in data interpretation performed using the scientific workflow.

The second objective is to expose the scientific workflow as a service used by the particle formation research community, thereby connecting the research community to infrastructure. Addressing Aim 5, the objective adopts the solutions proposed in aims 1-4 and implements the scientific workflow in Jupyter as well as deploys the implementation on e-Infrastructure.

Figure 2: Prototype Jupyter implementation of the scientific workflow. Following the initialization of day and location, particle size distribution observational data are fetched via the SmartSMEAR API and plotted. Such visualization is used by researchers to detect and characterize particle formation at the specified day and location. Information describing particle formation is recorded. Information is represented following the developed ontology design pattern and is acquired and curated by infrastructure (specifically, acquired via a SPARQL endpoint and curated in an RDF database, both deployed on EGI).

Substantial work has already been conducted toward this aim. In fact, a prototype Jupyter implementation has been implemented and deployed on EGI. Figure 2 provides an overview of the graphical interface that exposes the scientific workflow to the research community.

The workflow provides specialized functions to fetch and plot particle size distribution observational data as well as to acquire information describing particle formation in infrastructure. The details of fetching observational data and converting them into a (Python) native data structure as well as the details of representing and acquiring information are taken care of by the infrastructure, also by means of a Python library with specialized functions (e.g., fetchdata). As such, the research community can focus on the primary task at hand, namely the interpretation of observational data.

The prototype Jupyter implementation can be extended in order to address aims 7-9. Curated information describing particle formation can be processed to compute summary statistics. Such computation can be performed by extending the Jupyter Notebook, and made easy by extending the Python library with specialized functions.

Addressing Aim 6, the third objective is for infrastructures (RIs, e-Infrastructure, data centers, or similar) to systematically acquire and curate information describing particle formation. In this regard it is interesting to determine the kind of infrastructure best equipped to acquire, curate, and possibly publish information describing particle formation. Currently, such information is curated as data (with little or no formal meaning) on the computer hard drives of researchers. Since such information generated by the research community is extremely valuable, e.g. for integrated particle formation analysis, as well as essential for the reproducibility of summary statistics published in literature, it is evident that the systematic acquisition and long term curation of such information is important. A further interesting aspect is arguably the mode of information acquisition. In the proposed Jupyter implementation, acquisition occurs via a specialized function each time information describing particle formation is recorded. Technically, recording can occur on any infrastructure, including the researcher’s workstation. The research community should ultimately decide with which acquisition (and publication) mode as well as which infrastructure it wants to operate.

Realizing these objectives will have a couple of interesting impacts. First, harmonized information describing particle formation will result in (more) interoperable and reusable (FAIR) data that can be integrated for further processing, e.g. spatiotemporal visualization of particle formation or their statistical analysis. This is expected to hold across distributed research groups and is a result of adopting a community-agreed ontology design pattern, the harmonized representation of information, and the exposure of the research groups within the community to a common scientific workflow for particle size distribution observational data interpretation.

The second impact is the possible systematic acquisition and curation of explicit and formal (i.e., machine actionable) meaning of data (in addition to the data themselves). Rather than merely acquiring data products in form of, e.g., visualizations such as maps or plots (with implicit information content not available to machines) this use case aims to set an example for how infrastructures can systematically acquire and curate truthful, meaningful, well-formed data (i.e., information) whereby meaning is explicit and formal. Furthermore, we expect that harmonized information generated by distributed research groups will be easier to acquire for infrastructure, and thus curate and possibly publish. As such, the use case contributes to advancing infrastructures from the current data systems to information and knowledge-based systems (Stocker, 2017) that manage information about natural worlds and their phenomena of interest (in addition to information about people, organizations, instrumentation, publications, etc.).

Challenges

A key challenge is to bring together representatives of the research community studying particle formation and come to an agreement for how to harmonize the information describing particle formation. It is unclear whether such agreement is desired and achievable. At this stage it is also unclear whether the required people can be motivated to attend the planned workshop.

A further challenge is to motivate the use of the scientific workflow implementation across research groups in the community. The greatest benefit of the proposed approach will result from research groups adopting the service, rather than individually implementing their own.

A third difficulty is the lack of clarity for whether it is possible for infrastructure to systematically acquire, curate and potentially publish the information describing particle formation as envisioned in this use case.

Detailed scenarios

The basic scenario is for research groups, specifically individual researchers, of the atmospheric aerosol particle formation research community to be served with a service that implements a scientific workflow for particle size distribution observational data interpretation and the systematic acquisition, curation and possible publishing of information describing particle formation, resulting from observational data interpretation.

The service should enable researchers, members of distributed research groups, to execute a scientific workflow that fetches and visualizes observational data from one or more (selected) research infrastructures. The workflow should support the detection and characterization of (i.e., extraction of information describing) particle formation. Such description should reflect a community-agreed ontology design pattern.

Furthermore, the scientific workflow should support the further processing of information describing particle formation, e.g., to obtain summary statistics about particle formation at specific spatiotemporal locations.

Advanced scenarios include the possibility for the research community to inspect provenance relating summary statistics published in literature to information describing particle formation and particle size distribution observational data as well as the to the relevant involved agents and activities.

Of interest to advanced scenarios is also the possibility to openly publish information describing particle formation as well as the support for functionality relevant to data publishing, such as persistent identification and citation of information describing particle formation.

Technical status and requirements

The required components are Jupyter, the implementation of the scientific workflow as a Jupyter Notebook, an RDF database with SPARQL endpoint, as well as a Python library with specialized functions. Figure 2 shows a visualization of the prototype implementation of the scientific workflow. The components are containerized using Docker and can easily be deployed on infrastructures such as EGI. Indeed, this has already been tested with the deployment at http://212.189.145.31:8888. Recently, we have adopted JupyterHub[12] in order to support authentication of multiple users and management of individual notebooks.

Most of these components exist, specifically Jupyter, JupyterHub and the RDF database with SPARQL endpoint. These are all open source projects of high technical readiness.

The scientific workflow should be extended with further functionality. Some are already planned while others will result from obtaining research community requirements. Prototype functionality exists for automated machine detection of particle formation. This is supported by a trained machine learning classifier. We plan to extend this to support automated machine characterization of particle formation with functionality designed to extract information about detected particle formation, such as duration and growth rate. This intermediate step of the scientific workflow aims at supporting the research community in observational data interpretation by providing an automated machine extraction of information describing particle formation, which can subsequently be reviewed by researchers. As such, the automated extraction results in semantic *content* while the expert review results in semantic *information* (Floridi, 2011). Semantic information is truthful, in addition to meaningful well-formed data. Truthfulness is determined by human experts.

Scientific workflow functionality depends on the specialized functions provided by the Python library[13]. The maturity of this library and, thus, the scientific workflow implementation is prototype. It works reliably but would benefit from a thorough redesign based on software engineering principles. The library is publicly available on GitHub[14].

The use case involves observational data for particle size distribution. Such data are currently obtained via the SmartSMEAR API[15] of the SMEAR research infrastructure. Envisioned functionality may support the selection of observational data sources, one of possible several research infrastructures. Such selection could be supported as an additional (configuration) step in the scientific workflow. We are also considering linking this functionality with the ENVRIplus Knowledge Base developed by Theme 2. The Knowledge Base is designed to manage and support the querying of research infrastructure descriptions. Assuming descriptions of relevant infrastructures, such as SMEAR and ACTRIS, a demonstrator may succeed in linking the scientific workflow discussed here to the Knowledge Base to support selection of observational data sources as well as automated configuration of the corresponding API call to fetch observational data.

Overall, the use case is arguably already in a fairly advanced stage. While further technical advances are possible, the more critical advancements now rely on collaborative work with the research community, such as achieving agreement on representing information describing particle formation and adoption of the scientific workflow as a service.

Implementation plan and timetable

We envision the following implementation plan. First, we plan to organize the aforementioned workshop during Q1 2018 and hold the workshop during Q2 2018, possibly in April ahead of the next ENVRIweek, which would allow for presenting results on aims 1-3 during ENVRIweek. The successful execution of the workshop is a milestone for this use case.

During Q1 and Q2 2018, we plan to make further improvements to the Python library of specialized functions and thus to the implementation of the scientific workflow. Specifically, we plan to create a deployment based on JupyterHub; update the information representation to reflect the ontology design pattern currently published by the Environment Ontology; improve the functionality for automated extraction of information about particle formation; develop a concept for integrating provenance. Completing these steps are all milestones for this use case.

In the second half of 2018, ahead of the Fall ENVRIweek, we plan to address aims 4 and 5. During the ENVRIweek we plan to have a demonstration of the use case.

The 2018 ENVRIweeks will serve to address Aim 6. Relevant research infrastructures for Aim 6 include SMEAR and ACTRIS as well as EUDAT and data centers (e.g. PANGAEA).

Assuming we have achieved the milestones as planned, the remaining time of the ENVRIplus project during 2019 will be used to complete the aims 7-9. The first half of 2019 should also lead to concrete results in research groups of the community using the proposed service. Furthermore, we will complete the report that results from addressing Aim 6. Depending on the developments with infrastructures, this report may discuss a concrete implementation of information acquisition, curation, and possibly publishing in a sustained infrastructure. The final results and conclusions of the use case will be presented during the Spring ENVRIweek in 2019.

The developed concept for provenance and its implementation may serve as a demonstrator to ENVRIplus Theme 2 WP 8.

By connecting a research community with infrastructure in the data use phase of the research data lifecycle, the use case relates to ENVRIplus Theme 2 WP 7. The results of this use case may serve as demonstrator to the WP insofar as it supports a concrete research community in data analysis, specifically data interpretation, and connects the output of analysis, i.e. information, with the data acquisition phase of the subsequent iteration of the lifecycle.

Finally, linking the scientific workflow with the ENVRIplus Knowledge Base in order to support selection of observational data sources and, possibly, automated retrieval of data required in workflow execution also relates to Theme 2 activities and the implementation may serve as a demonstrator in this context.

Expected output and evaluation of output

The use case expects the following (primary) outputs:

- 1.A community-agreed ontology design pattern for information describing particle formation published as a concept of the Environment Ontology. The pattern can be extended to be adopted in the proposed solution to represent information. The output is deemed a success if the community-agreed ontology design pattern is published by the Environment Ontology.
- 2.Jupyter implementation and deployment on infrastructure of the scientific workflow as a service to be used by the research community. The output is deemed a success if (1) a functioning implementation is deployed as a service on infrastructure and (2) at least two groups of the research community are using the service.
- 3.A report that discusses which infrastructure is best suited to acquire, curate, and possibly publish the information describing particle formation, as well as derived summary statistics (if applicable). This output is deemed a success if the report is published.
- 4.A demonstrator that shows how the proposed solution enables integrated (e.g., statistical) analysis of information describing particle formation, generated by distributed research groups. This output is deemed a success if such a demonstrator is delivered.

References

Floridi, L. (2011). The Philosophy of Information. Oxford University Press.

Perez, F., Granger, B. E. (2007). IPython: A System for Interactive Scientific Computing, in Computing in Science & Engineering, vol. 9, no. 3, pp. 21-29. https://doi.org/10.1109/MCSE.2007.53

Stocker, M. (2017). Advancing the Software Systems of Environmental Knowledge Infrastructures. In Abad Chabbi and Henry W. Loescher (Eds.), Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities, pp. 399–423. Taylor & Francis Group, CRC Press. ISBN: 9781498751315 https://doi.org/10.1201/9781315368252-16

Wilkinson, M. D., Dumontier, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3. https://doi.org/10.1038/sdata.2016.18

[1] Please contact Yin Chen (yin.chen@egi.eu), Wouter Los (W.Los@uva.nl) or Zhiming Zhao (z.zhao@uva.nl) if you need help filling this template.

[2]http://www.geonames.org/

[3]https://www.w3.org/OWL/

[4]http://www.environmentontology.org/

[5]http://jupyter.org/

[6]https://www.pangaea.de/

[7]http://stato-ontology.org/

[8]https://www.w3.org/TR/prov-o/

[9]https://www.uef.fi/web/aerosol

[10]https://www.helsinki.fi/en/researchgroups/aerosol-cloud-climate-interactions

[11]https://www.ebi.ac.uk/ols/ontologies/envo/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FENVO_01001085

[12]https://jupyterhub.readthedocs.io/en/latest/

[13] Strictly speaking these functions are not necessary but they simplify the workflow

[14]https://github.com/markusstocker/pynpf

[15]https://avaa.tdata.fi/web/smart