Science Demonstrator 6: New particle formation event analysis on interoperable infrastructure (Use Case TC_17)

Overview

For the scientific community in aerosol sciences that studies atmospheric new particle formation events (NPFEs), this service aims to prototype how the scientific community can be deeply integrated with interoperable Research Infrastructures and e-Infrastructures (unless specified otherwise henceforth referred to as infrastructures). The result is a knowledge infrastructure[1] i.e., a robust network of scientists, artefacts such as virtual research environments and research data, and institutions such as research infrastructures and e-Infrastructures that acquire, maintain and share scientific knowledge about the natural world.

The service demonstrates how data analysis can be exposed to researchers as a Web based service while interoperable infrastructures orchestrate everything else, specifically: (1) loading primary observational data into computing environments for subsequent analysis by researchers; (2) representation of data derived in analysis using data models that employ domain-relevant community vocabularies and capture machine readable data semantics i.e., information[2] (“meaningful data”); (3) systematic and automated acquisition of derivative information in infrastructures; (4) registration of derivative information in catalogues.

Scientific Objectives

This demonstrator aims to prototype the future of scientific data analysis on interoperable infrastructure. It showcases how well-engineered infrastructures can provide to science communities data analysis as a service while taking care of everything else e.g., data conversions, curating data derived in analysis - as described in this section.

Here, the scientific community focuses on what they are most interested in and most enjoy doing (i.e., address scientific hypotheses with data analysis and interpretation) and the infrastructure guarantees that their data are FAIR[3] i.e., findable and accessible by systematic and automated acquisition and cataloguing; interoperable by using a formal, accessible, shared, and broadly applicable language for knowledge representation and vocabularies that follow FAIR principles; and reusable by rich description of data using domain-relevant community vocabularies and by their release with a clear data usage license.

An important objective is to entirely erase the need for manual data download and upload by researchers. The download of data from research infrastructures is “considered harmful” in most cases[4]. Indeed, the practice of downloading data perpetuates the infrastructural discontinuity between local computing environments (e.g., researchers’ workstations) and engineered infrastructures. Such discontinuity makes it difficult or impossible for engineered infrastructures to monitor workflows and executed activities, retain information about the involved primary and derivative data, as well as to systematically acquire derivative data. We want to demonstrate what a knowledge infrastructure may look like when manual download and upload is not an option.

A second objective is to unravel what occurs in the data use phase of the research data lifecycle. Studied for a concrete use case in aerosol science involving infrastructures and the relevant scientific community, we analyse the details of a scientific data analysis workflow and the roles of both the scientific community and infrastructures as elements of knowledge infrastructures. The demonstrator showcases how primary observational data acquired, curated and published by a research infrastructure are analysed by the scientific community in the data use phase and how such analysis generates derivative data. Traditionally, derivative data are poorly standardized in the community and generally reside on the workstations of researchers. The demonstrator shows how, when data analysis is performed on interoperable infrastructures - thus avoiding the aforementioned infrastructural disconnect - infrastructures can guarantee systematic and automated acquisition of derivative data, thus ensuring a strong link between the data use phase and the (derivative) data acquisition phase of the research data lifecycle. The demonstrator emphasises that there indeed is a cycle for research data from primary data to scientific knowledge communicated in scholarly literature.

A third objective is to connect, i.e., deeply integrate, a scientific community with well-engineered infrastructures. In the future, this social objective needs to be given more emphasis. While there continue to be issues to iron out, the technical elements of knowledge infrastructures are today mature enough to move into full-scale operation. The experiences collected with this use case underscore that the social elements are lagging behind and are at best only slowly starting to be receptive to the novel approaches developed here, at least in the earth and environmental sciences and especially for the long tail of science i.e., for communities that do “small science” with “little data”. This demonstrator is a contribution to integrating the social and technical elements of knowledge infrastructures.

Description

Usage Scenarios

Jaana and Mikko are two fresh graduate students of a Finnish research group that, among other things, study new particle formation events. New particle formation events are atmospheric events whereby aerosol particles form and grow over the course of a day at specific spatial locations. These events are studied to increase our understanding for the formation process and to quantify the formed aerosol.

Prior to Jaana and Mikko, earlier generations of students have developed Python software and published the codes as a Jupyter Notebook on GitHub. These codes have become the *de facto* standard software used by the scientific community. Jaana and Mikko are instructed by their supervisor to use these codes for their analysis and have obtained further instructions on how to execute the analysis on e-Infrastructures by their postdoc colleagues.

Figure 1. Jupyter Notebook for New Particle Formation Event Classification As Seen From the D4Science VRE. The VRE Includes the EGI Jupyterlab Computing Environment.

The two graduate students create an account on D4Science and are given access to the relevant Virtual Research Environment (VRE). The VRE gives them access to JupyterLab, serviced by EGI. The students use the JupyterLab Terminal to clone the required Jupyter Notebook from GitHub into their own working space. Now the students are ready to analyse primary data (i.e., particle size distribution data) in order to detect and describe new particle formation events that may have occurred at specific places and days. Figure 1 displays the Jupyter Notebook as seen by Jaana and Mikko. All they need to do is to select a day and place and interpret the corresponding visualization of primary data. For days and places at which an event occurred, the result of primary data interpretation is a description of the event, recording in particular the beginning and end times as well as the classification of the event, which follows a scheme accepted by the relevant scientific community.

Architecture

Figure 2 provides an overview of the architectural design of the service implementation. Researchers access JupyterLab operated on the EGI e-Infrastructure (provided by WP9) in order to analyse primary data for the purpose of new particle formation event detection and description. JupyterLab is accessible from the corresponding D4Science Virtual Research Environment (VRE). Having cloned the required Jupyter Notebook from GitHub, researchers can start to analyse primary data to detect and describe new particle formation events.

Figure 2. Architectural Design of the Service Implementation.

Workflow and Interfaces

The analysis consists of two main steps. Both are implemented as D4Science Data Miner algorithms and are accessed from within the Jupyter Notebook, programmatically via a WPS (OGC Web Process Service) interface. Given a day and place, as configured by the researcher, the first step fetches and visualizes primary data. The primary data are published by SmartSMEAR, a “data visualization and download tool for the database of continuous atmospheric, flux, soil, tree physiological and water quality measurements at SMEAR research stations of the University of Helsinki.” SmartSMEAR is developed and provided in collaboration with CSC (https://www.csc.fi/home), the Finnish national supercomputing center, who also host the SMEAR data. SmartSMEAR is thus an (software) artifact of the SMEAR (Station for Measuring Ecosystem-Atmosphere Relations) research infrastructure (RI). SmartSMEAR provides an API for data access. The primary data can thus be fetched and loaded into Python data structures in a programmatic manner.

Given the primary data, the Data Miner algorithm creates a visual representation (see Jupyter Notebook in Figure 1). This visualization is used by researchers to decide whether a new particle formation event occurred on the selected day and place as well as to describe the event for its properties e.g., beginning and end times, classification, among others. The visualization is a conventional PNG image which is made accessible by D4Science. The Data Miner algorithm returns to Jupyter Notebook the URL for the location of the image. The notebook then visualizes the image by retrieving it from the given location.

Assuming an event occurred on the selected day and place, the result of interpreting the visualization is a description of the event. Since researchers here study new particle formation events, in this context an event description is information i.e., meaningful data. Such data are thus rich in semantics.

In the second step, an additional Data Miner algorithm records the event description. The researcher merely records the day (e.g., 2013-04-04), place (e.g., Hyytiälä), beginning (e.g., 11:00), end (e.g., 12:30) and classification (e.g., Class Ia). Rather than recording these strings into a row of a table, the algorithm creates an RDF (Resource Description Framework) description of the event. The meaning of the strings is thus recorded as well. The result is a self-describing information object (see Figure 4). The current implementation uses the LODE ontology, which provides a concept Event and relations for time and space. We are currently working with the scientific community to develop a more appropriate concept of new particle formation events. This concept will be part of the Environment Ontology. When this process is completed, we will modify the implementation to reflect the conceptualization developed by the scientific community.

Finally, the RDF description is registered as a resource on the CKAN based D4Science catalogue. This is done automatically by the Data Miner algorithm. The data derived in analysis are thus automatically catalogued with corresponding metadata that support search. Figure 3 shows a number of catalogued resources for Hyytiälä (FI) on various days. Figure 4 shows the metadata of a selected resource, specifically the one for the description of the new particle formation event that occurred at Hyytiälä on April 4, 2013. Here, users are also given the location (URL) from which the event description can be accessed. Finally, Figure 5 shows the RDF description (in Turtle syntax) of the new particle formation event that occurred at Hyytiälä on April 4, 2013, obtained by accessing the URL on D4Science.

The RDF description is machine-readable, uses formal languages for knowledge representation (RDF, RDFS, OWL), and represents the semantics of the data (the day, place, beginning, end and classification) derived in analysis using domain-relevant community vocabularies. Specifically, beginning and end data elements (10:30 and 12:00, respectively) are described as an OWL-Time Interval with beginning and end OWL-Time Instants, relating to the respective timestamps as XSD DateTime. Also notable is that the place is identified as a GeoNames resource (http://sws.geonames.org/656888/) for Hyytiälä, Finland. The data derived in analysis are thus richly annotated. Indeed, the description adopts Linked Data principles to relate to resources defined elsewhere (here, GeoNames.org).

Figure 3. New Particle Formation Event Descriptions Registered as Resources in the CKAN based D4Science Catalogue.

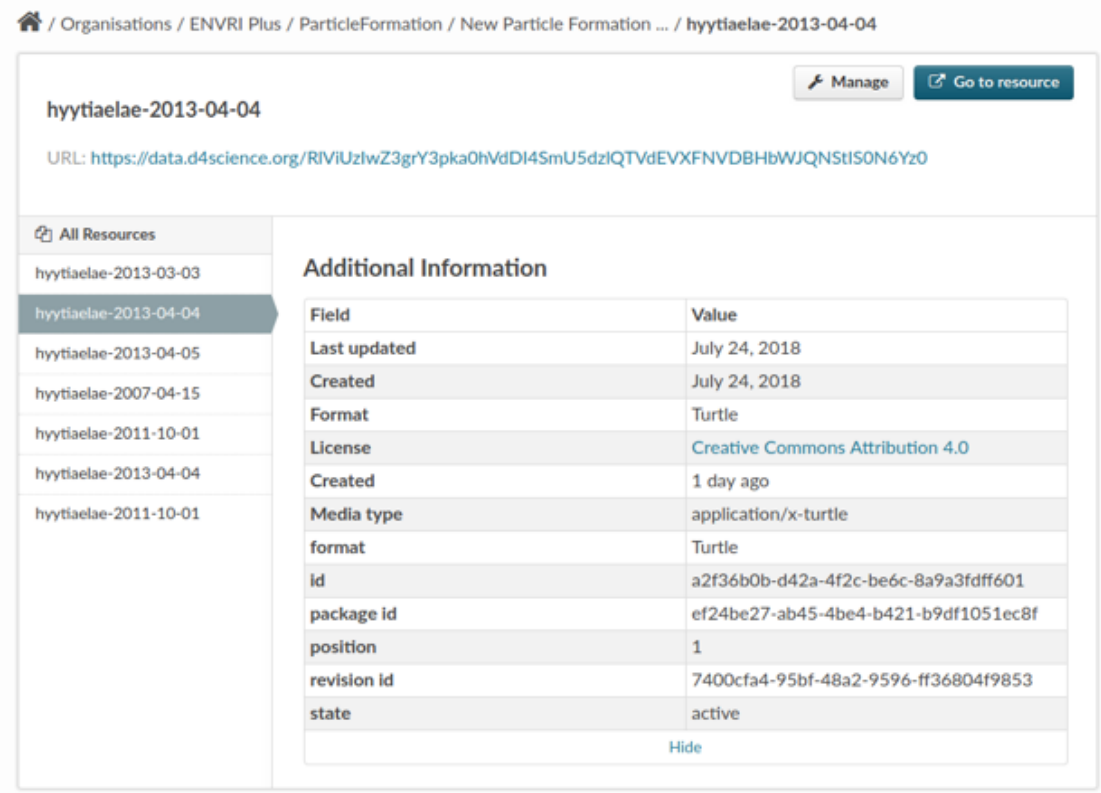


Figure 4. Metadata and Access URL of the Resource Describing the New Particle Formation Event that Occurred at Hyytiala (FI) on April 4, 2013.

Figure 5. The RDF Description of the New Particle Formation Event that Occurred at Hyytiala on April 4, 2013 as Retrieved from the D4Science Catalogue.

Advantages

The idea of transforming data into knowledge is popular among research infrastructures. Among others, the Integrated Carbon Observation System (ICOS) research infrastructure uses the tagline "knowledge through observations". The European Multidisciplinary Seafloor and water column Observatory (EMSO) suggests that the research infrastructure plays "a major role in supporting the European marine sciences and technology [...] to enter a new paradigm of knowledge in the XXI Century". As an example beyond research infrastructures, the European Open Science Cloud (EOSC) is envisioned to be an environment that enables turning ever increasing amounts of data "into knowledge as renewable, sustainable fuel for innovation in turn to meet global challenges".

Beyond the specifics of the developed use case in aerosol science, this demonstrator is a clear contribution to this idea. It demonstrates a possible architecture of an infrastructure that “transforms data into knowledge”. Essential factors of such knowledge infrastructures are (1) the deep integration of science communities with research and e-Infrastructures; and, as an important technical factor, (2) the curation of formal (i.e., machine-readable) data semantics. The deep integration of science communities is essential because, never mind the Age of Artificial Intelligence, in science it is researchers that transform data into knowledge. As this demonstrator underscores, deep integration with infrastructures allows for a range of novel possibilities, in particular enable researchers to focus on data analysis and interpretation while leaving data access and transformation from and to systems, the representation of data and their semantics following community standards, the capture of provenance information, and other infrastructural aspects to infrastructures.

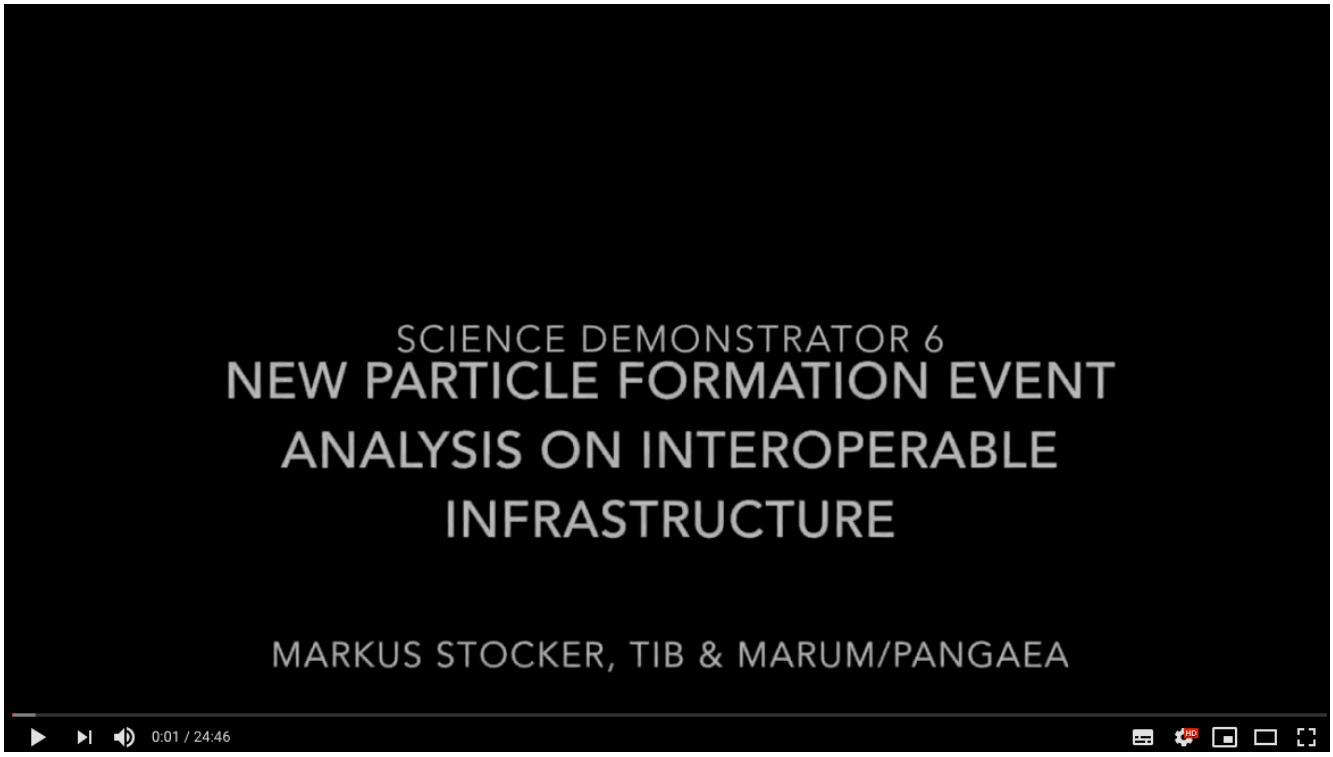
The curation of data semantics is an additional essential, technical, factor. Information is inherently semantic and becomes knowledge through learning (i.e., internalization). When researchers interpret primary data, the resulting derivative *information* is thus rich in meaning (relative to the context of data analysis). As demonstrated here, the data tuple (2013-04-04, Hyytiälä, 11:00, 12:30, Class Ia) has a specific semantic. In the Age of Semantics, it is paramount to move on from data structures that merely capture such a tuple of values to data models that also support the machine readable representation of the meaning of these values. We understand that it can’t be expected from researchers to (manually or, to this date also, programmatically) on broad disciplinary scale translate data tuples such as the one above into a description as shown in Figure 16. Indeed, a key point of this demonstrator is to showcase the possibility of engineering such translation into infrastructures so that the whole is entirely invisible to researchers. Such invisibility reflects well the nature of infrastructure[5].

Analysing 17 ENVRIplus research infrastructures, Hardisty et al.[6] have reported that at the time (2016) only three identify core competencies in the data use phase of the research lifecycle. As per the definition of the ENVRI Reference Model, data use is the phase in which researchers use data, potentially producing new research data. This description best describes the activity supported by this demonstrator, which thus focuses on the data use phase. While only few ENVRIplus research infrastructures identify core competencies in the data use phase, it is surely correct that all research infrastructures serve the data use phase. Indeed, their very existence is to serve this phase, serve researchers' data use. A key kind of use is arguably data analysis and interpretation. Again, beyond the specifics of the developed use case in aerosol science, we argue that this demonstrator showcases one possible interplay between research infrastructures and e-Infrastructures in the data use phase. We argue that the developed architecture is applicable to other (ENVRIplus) research infrastructures and the science communities they serve. Hence, the principles developed by this demonstrator are broadly applicable. Naturally, the specifics (e.g., the vocabularies, the notebook, the cataloguing, etc.) need to be adapted to meet the requirements of other scientific communities. The architecture and implementation principles (e.g., the design and technologies used) are, however, broadly applicable and thus of relevance to other or probably most if not all (ENVRIplus) research infrastructures.

Since FAIR data is on the global agenda of infrastructures, funders and other institutions, we underscore that this demonstrator significantly contributes to implementing this agenda by promoting the notion of "FAIR by Design" - weaving data FAIRness into infrastructures' fabric. The demonstrator builds on the principle not to leave making data FAIR to researchers but to guarantee it by design of well-engineered infrastructures. We argue that the removal of manual download and upload of data from and to systems is a crucial factor to this effect.

Naturally, the demonstrator is first and foremost of primary interest to a specific scientific community, namely the one consisting of the various aerosol research groups that study new particle formation events. To the best of our knowledge, the globally most renown research group in this area is the one led by Prof. Markku Kulmala at University of Helsinki[7]. Prof. Kulmala and some of the postdocs in his group have been involved in the developments of this demonstrator. Most importantly, postdocs have been actively involved in the development of a conceptualization of new particle formation events and a corresponding concept of the Environment Ontology. Naturally, in its current stage the demonstrator is a prototype to showcase to the scientific and infrastructure communities what is possible using state of the art interoperable infrastructures. A transition in practice from how data analysis is currently done to such infrastructures as demonstrated here requires further work as well as further acceptance by the scientific community. While we think to have reached an important milestone with this demonstrator, we cannot claim to know if and when such a transition will occur, for this scientific community or beyond. Clearer is, however, the imperative of the transition toward a practice as delineated by this demonstrator.

Link to the Demonstrator



- Youtube video is at: https://youtu.be/ra9W7b5DbgI
- Instructions: https://github.com/markusstocker/pynpf-d4science/blob/master/README.md
- Virtual Research Environment https://services.d4science.org/group/particleformation/
- Blog post http://markusstocker.com/data-analysis-on-interoperable-infrastructure/

Contributors

- Dr Markus Stocker, TIB and MARUM/PANGAEA, markus.stocker@tib.eu
- Dr Markus Fiebig, NILU, markus.fiebig@nilu.no
- Dr Leonardo Candela, CNR, leonardo.candela@isti.cnr.it

- Giuseppe La Rocca, EGI, giuseppe.larocca@egi.eu
- Dr Enol Fernandez, EGI, enol.fernandez@egi.eu
- Alex Hardisty, CU, hardistyar@cardiff.ac.uk

Reference

[1] Edwards, Paul N. 2010. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. MIT Press.

[2] Floridi, L.: The Philosophy of Information. Oxford University Press (2011)

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

[4] Atkinson, M., Filgueira, R., Spinuso, A., Trani, L.: Download considered harmful (2018), manuscript in preparation.

[5] Star, S.L.: The ethnography of infrastructure. American Behavioral Scientist 43(3), 377–391 (1999). https://doi.org/10.1177/000276499219955326

[6] Hardisty, A. et al. (2016). A definition of the ENVRIplus Reference Model. ENVRIplus Deliverable 5.2. http://www.envriplus.eu/wp-content/uploads/2015/08/D5.2-A-definition-of-the-ENVRIplus-Reference-Model.pdf