DATA SCIENCE JOURNAL

Stocker, M, et al. 2018. Curating Scientific Information in Knowledge Infrastructures. *Data Science Journal,* 17: 21, pp. 1–16, DOI: https://doi.org/10.5334/dsj-2018-021

RESEARCH PAPER

Curating Scientific Information in Knowledge Infrastructures

Markus Stocker[1,2], Pauli Paasonen[3], Markus Fiebig[4], Martha A Zaidan[3] and Alex Hardisty[5]

[1] TIB Leibniz Information Centre for Science and Technology, Welfengarten 1 B, 30167 Hannover, DE

[2] MARUM Center for Marine Environmental Sciences, PANGAEA Data Publisher for Earth & Environmental Science, Leobener Strasse 8, 28359 Bremen, DE

[3] Institute for Atmospheric and Earth System Research/Physics, 00014 University of Helsinki, FI

[4] NILU – Norsk Institutt for Luftforskning, Dept. Atmospheric and Climate Research, Instituttveien 18, 2007 Kjeller, NO

[5] School of Computer Science and Informatics, Cardiff University, Queens Buildings, 5 The Parade, Cardiff CF24 3AA, UK

Corresponding author: Markus Stocker (markus.stocker@tib.eu)

Interpreting observational data is a fundamental task in the sciences, specifically in earth and environmental science where observational data are increasingly acquired, curated, and published systematically by environmental research infrastructures. Typically subject to substantial processing, observational data are used by research communities, their research groups and individual scientists, who interpret such primary data for their meaning in the context of research investigations. The result of interpretation is information—meaningful secondary or derived data—about the observed environment. Research infrastructures and research communities are thus essential to evolving uninterpreted observational data to information. In digital form, the classical bearer of information are the commonly known "(elaborated) data products," for instance maps. In such form, meaning is generally implicit e.g., in map colour coding, and thus largely inaccessible to machines. The systematic acquisition, curation, possible publishing and further processing of information gained in observational data interpretation—as machine readable data and their machine readable meaning—is not common practice among environmental research infrastructures. For a use case in aerosol science, we elucidate these problems and present a Jupyter based prototype infrastructure that exploits a machine learning approach to interpretation and could support a research community in interpreting observational data and, more importantly, in curating and further using resulting information about a studied natural phenomenon.

Keywords: Data Use; Data Interpretation; Linked Data; Semantic Information; Environmental Research Infrastructures; Environmental Knowledge Infrastructures; Informatics; Data Science

1 Introduction

Environmental research infrastructures in atmospheric, marine, solid earth, and biodiversity domains are maturing their support for the acquisition, curation, publishing, processing, and use of data. For many infrastructures (specifically, observation systems) acquired data are primarily collected in observation activities and are thus *observational* data. Examples include the European Integrated Carbon Observation System[1] (Paris et al., 2012); the US National Ecological Observatory Network[2] (Keller et al., 2008); the Argo global array of ocean temperature/salinity profiling floats[3] (Roemmich et al., 2009)—among many others.

[1] https://www.icos-ri.eu/

[2] https://www.neonscience.org/

[3] http://www.argo.ucsd.edu

Research infrastructures that build on sensor network based operational observation systems are approaching full automation in observational data acquisition (Hellström et al., 2016). Acknowledging the need for greater data interoperability, curated observational data are gradually becoming standardized (Pearlman et al., 2016; Vossepoel & Murray, 2016). Standardization is facilitated by harmonizing the vocabularies used to structure data and has been a high priority in recent years.

Contrasting the standardization of curated observational data, scientists have so far had little research infrastructure support to curate information resulting from interpreting observational data. Of specific concern is that meaning acquired in interpretation is generally implicit when information is curated—as numbers in spreadsheets, text files, or the bitmap of a raster graphic. With increasing automation of observational data acquisition and standardization of curated observational data, we argue that the current frontier for environmental research infrastructures must address the challenge of curating machine-readable meaning of data that result from data interpretation. This implies prior acquisition of meaning in machine readable format.

Here, we present a Jupyter (Kluyver et al., 2016) based prototype infrastructure for a use case in aerosol science, namely the study of atmospheric new particle formation (NPF) events using polydisperse aerosol particle size distribution observational data. Such an infrastructure could support a research community, its research groups and individual scientists, in interpreting particle size distribution observational data and, more importantly, in acquiring, curating and further processing resulting information about NPF events—both data about events and the meaning of data, in machine readable format.

The use case shows how observational data for particle size distribution evolve to but are different from information about NPF events; how particle size distribution data and NPF event information are input and output, respectively, to the activity of data interpretation performed, largely manually, by scientists; and how, in conventional representation, meaning acquired in data interpretation is implicit when information about NPF events is acquired and curated.

This work makes four contributions. First, we make first attempts at grounding the "data to information, and knowledge" discussions, increasingly common among research infrastructures, in an existing theoretical framework. We suggest that the community needs to better understand what the terms 'data', 'information' and 'knowledge' mean in the context of research infrastructures. Too common are phrases such as "by data we mean information." There exists a wealth of literature to build on. Second, we demonstrate how artifacts of knowledge infrastructures, in particular software systems, can integrate semantic technologies in data analysis environments to support the representation, acquisition, curation and processing of data *semantics* without requiring researchers (users) to significantly modify their data analysis workflows. Third, as a side effect of the second contribution, we demonstrate how the link between the primary data use and the derivative data acquisition phases of the research data lifecycle can be strengthened by infrastructures that ensure the systematic acquisition of the latter through removing the requirement of manual down and uploading of data from and to systems. Together these contributions serve better reproducibility of science and improved semantic interoperability between systems, as well as offering a step towards a future proposed programming paradigm beyond Turing/von Neumann approaches, in which information assumes a role as a first class object to be manipulated (Schubert & Jeffery, 2015). Our work represents a contribution towards this future scenario.

2 Motivation

Data, information, and, increasingly, knowledge are terms commonly used in earth and environmental sciences, as well as in informatics supporting these sciences. We present a few examples. The Lindstrom et al. (2012) Framework for Ocean Observing highlights the "challenge of delivering ocean information for societal benefit" and suggests that a key framework concept is to promote the "transformation of observational data organized in [Essential Ocean Variables] into information." This is mirrored in the natural history realm where Essential Biodiversity Variables are part of an information supply chain, conceptually positioned between raw data (i.e., primary biodiversity data observations) and synthetic indices (indicators for reporting biodiversity change) (Kissling et al., 2017). The ICOS research infrastructure uses "Knowledge through observations"[4] as its succinct tag line. Writing about Oceans 2.0, Ocean Networks Canada highlights[5] that the system is able to mine "data streams to detect trends, classify content and extract features [...] thereby turning raw data into information and setting the stage to allow the information to be transformed into knowledge."

[4] https://www.icos-ri.eu/our-mission

[5] http://www.oceannetworks.ca/innovation-centre/smart-ocean-systems/ocean-observing-systems/oceans-20

Arguably inspired by the 2017 Motto[6] of the Helmholtz Association of German Research Centres—namely, "From Data to Knowledge"—several recent international conferences hosted by German institutions adopted this idea in their conference theme. Examples include the Göttingen-CODATA RDM Symposium 2018 with theme "The critical role of university RDM infrastructure in transforming data to knowledge"; the RDA 11th Plenary Meeting with theme "From data to knowledge"; and the 10th International Conference on Ecological Informatics with theme "Translating ecological data into knowledge and decisions in a rapidly changing world." The idea of transforming data into knowledge is popular, indeed.

Surely, there is broad agreement that knowledge can be obtained from data. The details on how this occurs; what the entities 'data', 'knowledge' and presumably 'information' are, and how they relate; the agents and activities involved in evolving data into information, and knowledge; or how infrastructures support agents and activities is, however, less obvious and less well understood.

The notion of a logical progression from data to knowledge, via information, has been described as "fairytale" (Zins, 2007). Indeed, information is represented as data in (computer) systems, which could suggest a "circularity" from information to data. If we qualify data as observational, experimental or computational (Borgman, 2007)—for simplicity, primary data—and information as about the unit of analysis e.g., a natural phenomenon under investigation, the logical progression can be more defensible. Information about the unit of analysis is thus obtained from primary data; e.g., data that result from the activity of observation, carried out by, for instance, sensing devices. The logical progression from primary data to information about the unit of analysis seems to be defensible, since derived data resulting from representing information in a computer system are of a kind other than primary data. Note that the unit of analysis is contextualized: information about it may well be primary data in a different context.

Meaning plays a central role in the evolution of primary data into information about units of analysis, and possibly knowledge. According to its standard definition, information is meaningful well-structured data (Floridi, 2011). Interpretation is the activity that transforms data as uninterpreted symbols with "no meaning for the system concerned" (Aamodt & Nygård, 1995) into information i.e., "data with meaning." As suggested by Aamodt & Nygård, central to this is the ability to "determine the contextual meaning of data," which is generally attributed to human agents. Thus, people are essential in the first instance in evolving data to knowledge. Arguably though, this ability can also be exercised by computer agents (Jennings et al., 1998).

This essential role for meaning demands a conceptualization that unifies people and infrastructures, including research infrastructures, e-Infrastructures, and related infrastructures such as university research data management infrastructures. Knowledge infrastructure, described by Edwards (2010) as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds" may be a core concept of a unifying conceptualization. Here, researchers as members of research communities, together with infrastructures form networks that generate through, among other activities data interpretation scientific knowledge about the human and natural worlds. As elements of knowledge infrastructures, research infrastructures are institutions that operate and maintain artefacts such as scientific instruments, data and software.

We argue that the concept of knowledge infrastructure (Edwards et al., 2013; Borgman et al., 2015; Karasti et al., 2016) can help to identify and organize some of the challenges faced by research infrastructures, e-Infrastructures, university research data management infrastructures, digital libraries, etc. as elements of networks that transform data into knowledge. A key challenge faced by such infrastructures is the curation of information i.e., caring for and presenting data and their meaning. Concretely, the meaning of data resulting from human-in-the-loop primary data interpretation activities should be explicit and formal, and thus, machine readable. This implies systematic acquisition, by infrastructures, of the meaning of data resulting from primary data interpretation activities. In other words, the predominantly data based systems of current infrastructures should evolve into information and knowledge based systems that curate data and their meaning (Stocker, 2017).

3 Information

We adopt the framework by Floridi (2010, 2011). Building on a widely adopted General Definition of Information (GDI), Floridi develops a definition of semantic information. GDI defines information in terms of "data + meaning." Floridi proposes a more precise formulation that borrows the term *infon* (Barwise & Perry, 1981; Devlin, 1991), a discrete item of information. The infon σ is an instance of information, understood as

[6] https://www.helmholtz.de/ueber_uns/die_gemeinschaft/mission/motto_2017/

semantic content, if and only if σ consists of n data, $n \geq 1$; the data are well formed; and the well-formed data are meaningful (i.e., of significance to some person, situation or machine).

Thus, information is made of data, and data are structured according to a syntax and must comply with a semantics. Of specific interest here is *factual* semantic content i.e., semantic content about a situation or fact that can be qualified as either true or false. Only semantic content that is true is informative. Thus, Floridi suggests that p qualifies as factual *semantic information* if and only if p is well-formed, meaningful, and *truthful* data. While factual semantic content can be false, factual semantic information needs to be true. Floridi uses the term 'truthful' instead of 'true' because well-formed and meaningful data can constitute constructs other than natural language sentences, for instance formulae, maps, diagrams, or videos.

Floridi discusses a classification of types of data, of which two are of importance here. *Primary data* are the principal data stored, for example in a database. In environmental research infrastructures, primary data are often numerical values resulting from observation activities, which a database may organize along temporal and spatial dimensions as time series, arrays, or data cubes. *Derivative data* are data that "can be extracted from some data whenever the latter are used as indirect sources in search of patterns, clues or inferential evidence about things other than those directly addressed by the data themselves." Here, information about the environment is derivative information, constituted by derivative data, acquired by interpretation of (i.e., addition of meaning to) primary data.

We borrow the notion of *data interpretation* from the unified definitional model of data, information, and knowledge proposed by Aamodt & Nygård (1995). Here, data interpretation is the activity carried out by an interpreter through which data becomes information. Data are uninterpreted symbols with "no meaning for the system concerned." Thus, Aamodt & Nygård have *meaning* as the key feature distinguishing data from information, in common with Floridi. Interpretation occurs "within a real-world context and for a particular purpose." The interpreter thus determines the contextual meaning of data. Aamodt & Nygård emphasise that to interpret data, an interpreter must possess knowledge.

4 Use Case

Our use case, building on earlier related work (Stocker et al., 2013, 2014, 2015; Stocker, 2015, 2017) is in aerosol science, specifically for the study of atmospheric new particle formation (NPF) events using polydisperse aerosol particle size distribution observational data as measured by a differential mobility particle sizer (DMPS) (Aalto et al., 2001). Central to NPF events is the formation of aerosol particles at specific spatio-temporal locations and the growth of particle diameter size over the course of a few hours (Kulmala et al., 2004). NPF events are studied for their relevance in climate science and human health.

Aerosol scientists interpret observational data to detect and describe NPF events, by visualizing and analysing the data for a day and a spatial location (Dal Maso et al., 2005). The observational data are measured at the SMEAR II (Station for Measuring Ecosystem-Atmosphere Relations) in Hyytiälä, Southern Finland. The data are accessible via SmartSMEAR[7] (Junninen et al., 2009), a Web service providing access to (processed) observational data acquired, curated, and published by the SMEAR research infrastructure (Hari & Kulmala, 2005), for locations in Finland over multi-year timespans.

Detected NPF events are typically described by their attributes. In addition to obvious ones, such as days and locations at which events occurred, scientists may also classify events using a classification scheme. Dal Maso et al. (2005) proposed two main classes: Class I for events for which particle growth rate and further new particle formation rate can be determined with a good confidence level and Class II for events for which it is not possible to determine these quantities with high enough confidence level, typically due to inhomogeneities in air masses or contributions of other nearby aerosol sources. It was proposed to further divide Class I into Class Ia and Class Ib. Class Ia for events that are very clear because the concentration of particles produced during the NPF event clearly exceeds the concentration of pre-existing particles. Class Ib for events for which the contribution of pre-existing particles is significant. Alternative classifications have also been proposed (Hamed et al., 2007). Other event attributes extracted by experts include duration and growth rate. Of interest here is that different research groups within the research community adopt different classification schemes. Hence, information about NPF events results in data with heterogeneous syntax and semantics.

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

Listing 1. Research Group A

```
# Matlab datenum
# Event Class Ia
# Event Class Ib
# Event Class II

735328 1 0 0
```

Listing 2. Research Group B

```
# Date
# Class

2013-04-04,1
```

Listing 3. Research Group C

```
# Date
# Matlab datenum
# Class

04/04/2013,735328,1
```

Figure 1: Conventional representation of information about NPF event classification.

As a selected example, we discuss the NPF event that occurred at Hyytiälä, Finland, on April 4, 2013. **Figure 1** highlights how information about the classification of a NPF event is conventionally represented, for three different research groups of the community. For Research Group A (Listing 1), the value `735328` is of type MATLAB `datenum`.[8] It is more intelligible as `2013-04-04`. The remaining values are a bit encoding for whether a NPF event of Class Ia, Class Ib, or Class II, respectively, occurred. In our example, the second value 1 signifies that a NPF event of Class Ia occurred on that day. For each analysed day, the data matrix is extended by an additional row. Research Group B (Listing 2) uses a different classification scheme consisting of the classes 0, 1, 2, 3, and 4. The first value is the date and this is followed by the class label (1 in the example). Research Group C (Listing 3) includes the date in two different formats and uses yet another classification scheme, one that merges Class Ia and Class Ib into a single class 1.

Notably, other information about the NPF event e.g., location and duration is curated separately from such classification data. Hence, there is no integrated record for the data encoding all information available for the NPF event. Rather, the information is scattered across multiple files, data and metadata, managed by different researchers and thus typically residing on multiple systems. Furthermore, across the three research groups, there is great syntactic and semantic heterogeneity in classification data. While syntactic heterogeneity can be overcome, the semantic differences of classification schemes is of substantial hindrance to interoperability of information about NPF events in the research community. As a result, to reconstruct integrated and interoperable information about NPF events is a daunting challenge. Machines cannot tackle the task without specialized software that knows where to locate data and metadata, how to read the metadata and apply them to interpret data. Worse, being classified with different schemes, reconstructing interoperable information in practice means that primary data need to be manually reclassified.

Lacking infrastructure support, researchers thus curate information about NPF events as well-formed data in tabular form but with little expressed explicit and formal meaning. Metadata does inform the correct interpretation and meaning of values but without a formal language for knowledge representation meaning remains implicit and inaccessible, especially to computer agents.

Such minimal standardization of methods is typical of "little science" projects. Borgman et al. (2015) note that in such projects "each scientist may use different tools and techniques to generate datasets similar in form and intent." The authors also highlight that since the scientist (data producer) is responsible for data management, "data tend to be managed by localized, ad hoc practices for the immediate purpose of the scientists." This reflects the state of affairs in this use case.

We now discuss the selected example NPF event that occurred at Hyytiälä, Finland, on April 4, 2013 within our Jupyter based prototype infrastructure.

5 Implementation

The presented prototypical infrastructure builds on the EGI federated e-Infrastructure of computing services for research and innovation using a virtual machine equipped with an instance of JupyterHub.[9] The interpretation of observational data and processing of resulting information is implemented as Jupyter notebooks (Kluyver et al., 2016). JupyterHub is a multiuser server for Jupyter notebooks.

We adopt Jupyter because it enables moving data analysis from local computing environments (in particular the researcher's workstation) into virtual research environments (see Section 6.3). Furthermore, it is straightforward to extend the environment with functionality we need to support the novel features presented here.

[8] https://www.mathworks.com/help/matlab/ref/datenum.html
[9] https://jupyterhub.readthedocs.io

The infrastructure utilizes semantic technologies for formal and explicit representation and curation of information. Of specific relevance are the Resource Description Framework (RDF) (Schreiber & Raimond, 2014), the Web Ontology Language (OWL 2) (Hitzler et al., 2012), and the SPARQL Protocol and RDF Query Language (SPARQL) (Harris & Seaborne, 2013). The virtual machine is equipped with an instance of Apache Jena Fuseki[10] for the acquisition, curation, and publishing of information.

The infrastructure includes a newly written and tested Python library with specialized functions that implement the program logic required to fetch and plot observational data and to represent, acquire and process information. Following PROV-O (Lebo et al., 2013), the functions also record provenance information relating to entities, involved agents and performed activities. Finally, the library implements a machine learning model trained to classify observational data to support automatic extraction of information.

Figure 2 provides a schematic overview of the conventional infrastructure for NPF event detection and description contrasted with the proposed infrastructure. Executing the data interpretation workflow in a virtual research environment, accessible to all research groups ensures that information about NPF events is generated with harmonized syntax and semantics across research groups in the community.

5.1 Data Interpretation

We have implemented a workflow to visually interpret observational data into information about NPF events as a Jupyter notebook. **Figure 3** illustrates the fetching and plotting of observational data from SmartSMEAR as well as the visualization used to interpret data for the purpose of detecting and describing the NPF event. Information, specifically the start and end times as well as the classification, about the event on April 4, 2013 at Hyytiälä is then recorded (**Figure 4**).

Listing 4 illustrates the machine readable information about the NPF event. For the sake of brevity and clarity, the listing omits prefixes and a few other details. While the representation in **Listing 4** includes more information than the conventional representations in **Figure 1**, it is clear that the proposed representation

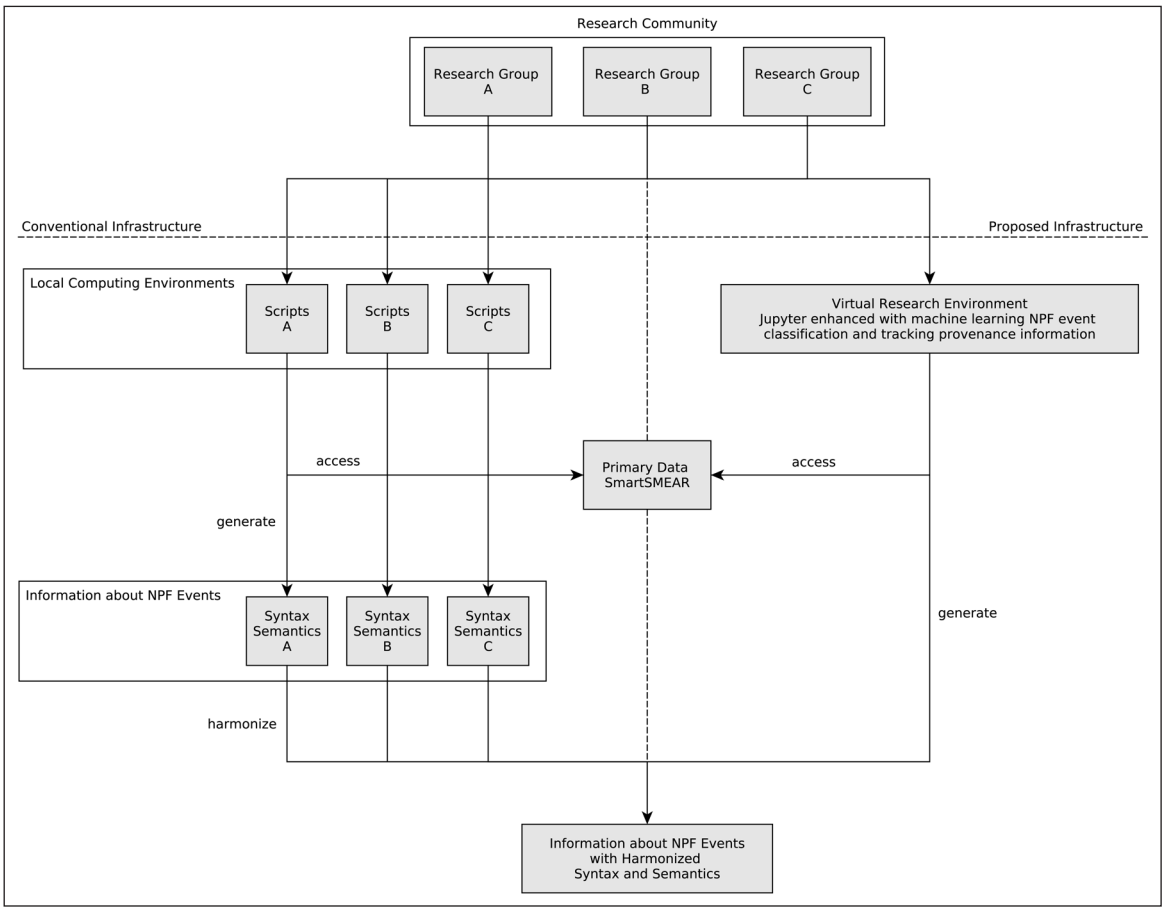

Figure 2: Schematic overview of the conventional infrastructure for NPF event detection and description contrasted with the proposed infrastructure.

[10] https://jena.apache.org/documentation/fuseki2/

```
plotdata(fetchdata('2013-04-04', 'Hyytiälä'))
```

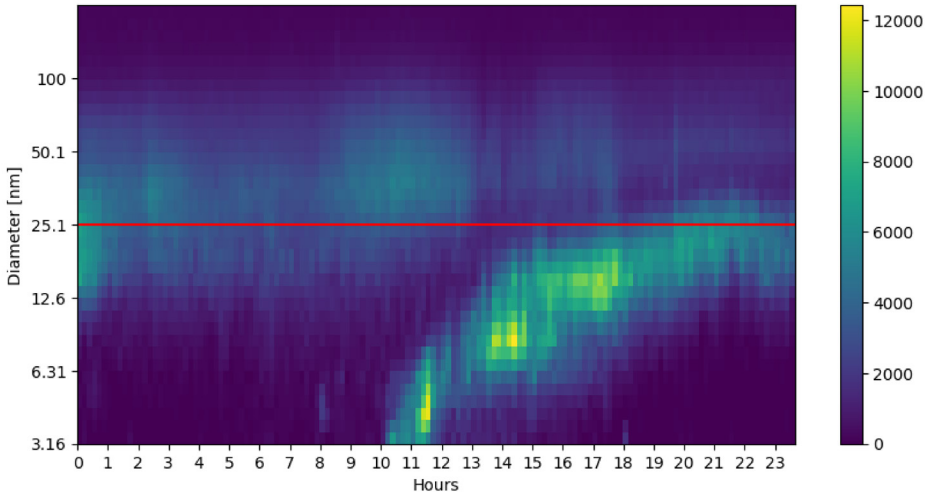

Figure 3: Fetching and visualizing particle size distribution data in Jupyter notebook for April 4, 2013 at Hyytiälä. The colour indicates the particle concentration level (cm^{-3}). The experts are interested in the yellow (white in grayscale) shape that reflects a NPF event for aerosol with particles of initially small but growing in diameter size.

```
record(event('2013-04-04', 'Hyytiälä', '10:00', '12:00', 'Class Ia'))
```

Figure 4: Recording information about the event that occurred on April 4, 2013 at Hyytiälä. It is a very clear and strong event (Class Ia) during which new particle formation was observed to start at 10 am and end at 12 pm.

```
[] a lode:Event ;
  smear:hasClassification [
    rdfs:label "Class Ia"^^xsd:string ;
    rdfs:comment "Very clear and strong event"^^xsd:string
  ] ;
  lode:atPlace [
    a gn:Feature, DUL:Place ;
    gn:countryCode "FI"^^xsd:string ;
    gn:locationMap <http://www.geonames.org/656888/hyytiaelae.html> ;
    gn:name "Hyytiälä"^^xsd:string
  ] ;
  lode:inSpace [
    a sf:Point, wgs84:SpatialThing ;
    geosparql:asWKT "POINT (24.29077 61.84562)"^^geosparql:wktLiteral ;
  ] ;
  lode:atTime [
    a time:Interval ;
    time:hasBeginning [
      a time:Instant ;
      time:inXSDDateTime "2013-04-04T10:00:00+03:00"^^xsd:dateTime
    ] ;
    time:hasEnd [
      a time:Instant ;
      time:inXSDDateTime "2013-04-04T12:00:00+03:00"^^xsd:dateTime
    ]
  ] .
```

Listing 4: Machine readable information about the NPF event at Hyytiälä on the 4th of April 2013.

is more expressive than the conventional representations. It captures more of the meaning acquired through interpretation of observational data carried out by researchers. Note that this information object is created automatically as a result of executing the statement shown in **Figure 4**.

As our experiments suggest, extracting information about NPF events can be automated by the infrastructure, at least to some extent. Such automated machine assessment can be integrated in the presented Jupyter based workflow, specifically following the visualization of observational data (**Figure 3**) and before the recording of NPF event information (**Figure 4**). We have implemented a specialized function `assess (day, location)` that uses a multilayer perceptron (MLP) artificial neural network (ANN) (Mitchell, 1997) to automatically detect NPF events (see also Stocker et al., 2014; Zaidan et al., 2017). The ANN has been trained using a labelled dataset consisting of 2938 samples for the period 1996–2016 at Hyytiälä. Each sample is a feature vector computed from daily observational data. Such data are: i) sampled to diameters smaller than 25.1 nm; ii) sampled to hours between 6 am and 6 pm to form a "daytime dataset"; iii) sampled to hours outside 6 am and 6 pm to form a "night time dataset". The feature vector consists of the computed values for sum, maximum, standard deviation, and variance for the daytime and night time datasets. The resulting training dataset is sampled to include the same number of positive and negative examples. Furthermore, it is min-max scaled. The performance of a MLP ANN with two hidden layers consisting of 5 and 3 neurons, respectively, has been evaluated using 10-fold cross validation. We obtained a mean accuracy of 88.73%, with standard deviation of 2.85%.

The infrastructure can thus support automatic extraction of information, which is subsequently reviewed by scientists. Following Floridi, automated information extraction provides factual semantic *content* while the result of review by scientists is factual semantic *information* i.e., well-formed, meaningful and truthful data. Note that the visual classification cannot be considered as absolute and objective truth as accuracies below 90% can result also from classification conducted by different individual researchers.

5.2 Information Processing

Given a database populated with information about NPF events, we can process information to, for example map events, compute statistics such as average duration of events at a particular location, or describe events both in natural language text and in machine readable format.

The analysis of information about NPF events, here "secondary data," results in further derivative information and data, here "tertiary data." An example is the computation of the average duration of events. Such computation takes a set of NPF events ε as input, in particular the values for the duration of events $e \in \varepsilon$, and returns a value $\bar{d}$ for the computed average duration. The value $\bar{d}$ can be acquired and curated by infrastructure, for instance because it is going to be published in literature. Such values can be represented using the same technologies as used for representing information about NPF events. Specifically, the Ontology for Biomedical Investigations (Bandrowski et al., 2016) provides a concept for average value defined as a "data item that is produced as the output of an averaging data transformation and represents the average value of the input data." Here, the input data are values for the duration of events $e \in \varepsilon$ and $\bar{d}$ is the output. The output is represented as `scalar value specification`, specifically a numeric duration with unit type hour. Hence, $\bar{d}$ is not merely a number, such as 7.5. Rather, it is information i.e., well-formed data and their meaning.

Since the interpretation of observational data and the analysis of derivative information about NPF events occurs *on* research infrastructure rather than on local computing environments, it is possible for the infrastructure to acquire and curate provenance information. **Figure 5** demonstrates how provenance information acquired and curated by the infrastructure can be interrogated. The first query is for derived entities generated in activities and attributed to a contributor identified by ORCID iD. The results show, in particular, that three entities (specifically, information about NPF events) are derived from observational data files in data visualization activities. The second query details an entity, namely information about the NPF event that occurred on April 4, 2013 at Hyytiälä. The result is a different representation of the information shown in **Listing 4** extended with provenance information. As we can see, information about the NPF event is a provenance entity (`prov:Entity`) that relates to the entity from which it was derived, to the agent (here identified by ORCID iD) the entity was attributed, and the activity (`obo:OBI_0200111` or 'data visualization') in which the entity was generated.

6 Discussion

Conventional approaches to represent information about NPF events (**Figure 1**) produce compact, structured data that are optimized, at the expense of human and machine readability for processing in specific computational environments (for example, MATLAB). Experts in NPF event studies intimately familiar with

```
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)
""")

query("""
    select ?p ?o where {
      smear:eb1ad69f11aecd2449f6d5741c3b8ac3 ?p ?o .
    }
""")
```

	entity2	entity1	activity
0	smear:eb1ad69f11aecd2449f6d5741c3b8ac3	file:2013-04-04-hyytiaelae.csv	data visualization
1	smear:280520d6022e59195bc2cab2946d2325	file:2015-03-10-vaerriö.csv	data visualization
2	smear:6d65ba1a7fe156b64361d2eb758c983c	file:2016-06-06-hyytiaelae.csv	data visualization
3	smear:b369975d2f5740fd6e32bf2fe3be3711	smear:39c190c629389e9cfe0df745179883f9	averaging data transformation

	p	o
0	prov:wasAttributedTo	https://orcid.org/0000-0001-5492-3212
1	smear:hasClassification	smear:ClassIa
2	prov:wasGeneratedBy	obo:OBI_0200111
3	rdf:type	lode:Event
4	rdf:type	prov:Entity
5	lode:atTime	smear:92be5465a05cc56156422d6cdb4603e1
6	prov:wasDerivedFrom	file:2013-04-04-hyytiaelae.csv
7	lode:inSpace	smear:7f885190eb43154e01c97f814b287a4b
8	lode:atPlace	http://sws.geonames.org/656888/

Figure 5: Interrogating provenance information acquired and curated by the infrastructure while data visualization and averaging data transformation activities are performed.

the SMEAR research infrastructure know the meaning of the produced data and they imply this into the computational environment by the way they write the processing statements. In contrast, the NPF event information produced by the approach proposed here, based on Floridi's distinction between data and information and Aamodt & Nygård's notion of data interpretation that derives the latter from the former, are less compact and not immediately processable in the specific computational environment. However, in addition to being structured, NPF event information is more meaningful to both humans and machines. We argue that our proposed approach is more effective at curating meaning acquired in data interpretation as well as ensuring that relevant provenance information is generated. Thus, the resulting NPF event information potentially has a greater fitness for purpose e.g., to share, integrate, or process information about NPF events.

6.1 Curating Information

The research data lifecycle, as described by the ENVRI Reference Model (Nieva de la Hidalga et al., 2017) for the 'archetypical' environmental research infrastructure, suggests that data produced by researchers in the data use phase can be acquired by a research infrastructure. In contrast to the links between other phases of the lifecycle, specifically between acquisition, curation, and publishing, the link between data use and further acquisition of new, derivative data (i.e., information as we have explained it in the present article) is relatively unexplored. Data produced by researchers in data use are not acquired and curated by research infrastructures as systematically as primary data are acquired and curated. We argue that the presented approach, illustrated as an integration with Jupyter, demonstrates one way in which the link between data use and acquisition could be strengthened.

As curators of the primary data analysed by research communities, the data centres of research infrastructures are ideally positioned to acquire and curate derivative information. In the context of our use case, the European Aerosols, Clouds and Trace gases Research Infrastructure[11] (ACTRIS) and SMEAR are two research infrastructures that serve the research community with relevant observational data. While these research

[11] http://www.actris.eu/

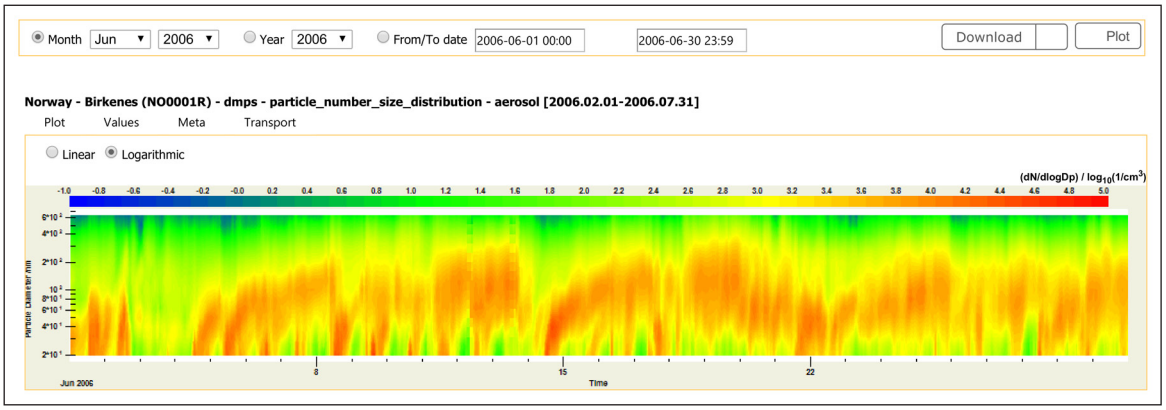

Figure 6: ACTRIS visualization of particle size distribution data for June 2006, in Birkenes, Norway. NPF events may be seen. However, information about events is implicit in the visualization. (Source: http://ebas.nilu.no/).

infrastructures support the visualization of observational data (e.g., **Figure 6**), they do not currently support 'Interpretation as a Service' and deal with derivative information about NPF events.

Since information is typically a result of activities carried out by researchers at their desks and in their laboratories, the institutions they are affiliated with e.g., universities, and their digital libraries may be better strategic choices to operate infrastructure that curate derivative information. Indeed, driven by requirements of funding agencies, research institutions of all kinds increasingly find themselves having to deploy infrastructure for digital research/scholarly data management and librarianship. The fact that data remain within the institution mitigates many legal issues e.g., regarding data control, confidentiality (where necessary) and re-use. However, with their broader scope compared to specialized research infrastructures, university infrastructures/digital libraries often lack the required domain expertise. An alternative are specialized data curators/publishers, such as PANGAEA[12] (Diepenbroek et al., 2002) or other members of the International Council for Science World Data System. Notwithstanding which institution ultimately operates such infrastructure, most if not all of them must develop their systems to acquire and curate information as suggested here.

6.2 Reproducibility and Interoperability

A stronger link between data use, acquisition and curation of derivative data, and thus further processing, is critical for reproducibility in science. While the systematic acquisition, curation and publishing of primary data ensures that such data are accessible, the same cannot be said for derivative data if they are not systematically acquired. A weak link between data use and data acquisition in research infrastructures not only means that information is not formal; it also means that derivative data are not readily accessible. Returning momentarily to our use case, information about individual NPF events is generally further processed into statistical indicators, such as statistical difference in mean event duration between seasons. Such figures are eventually published in literature. However, to reproduce the study, researchers not only require access to primary data but also to any derivative data (here information about NPF events).

Attaching formal meaning to data (i.e., creating interpretable information) leads directly to improved interoperability; not only between different persons and communities but, increasingly importantly, between different machines and computing systems. Machines work together better based on shared understanding of the meaning of what is exchanged between them. The key point here is that achieving semantic interoperability (Heiler, 1995) involves possession of a shared and congruent understanding of the context, including the important assumptions, principles, facts, notions and relations existing within that context. In the future, it involves possession of the capability to infer and build that understanding (i.e., the context) from (meta)data exchanged.

In discussing changes needed to the traditional von Neumann principles of programming computers (von Neumann, 1993) to support today's sophisticated resource and data intensive applications in clouds, Schubert & Jeffery (2015) enunciate the role of information arising from a needed shift away from

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

Turing/von Neumann approaches towards what they describe as "a new ICT of distributed parallel information valorization" (ICT enhancing the value of information). In new programming paradigms arising from their 3-pronged "Triple-I" Information-Incentive-Intention model, information (as opposed to data) assumes a role as a first-class object to be manipulated. The results we describe above, illustrating the utility of NPF event information as combination of data, structure and meaning sit well with structures Schubert & Jeffery propose and with reasons given to represent data as information i.e., "to guarantee its transformability into different views and usages, and to enable its composition and decomposition, such as for information mining or to communicate and act on partial data." Our work represents a contribution towards this future scenario.

6.3 Virtual Research Environment

Jupyter and the presented approach can be understood as elements of a larger Virtual Research Environment (VRE) (Candela et al., 2013). Standalone or as a component of a larger VRE, Jupyter and other comparable systems are gaining popularity to document and share executable software workflows and support reproducibility.

Discussing near real-time data processing in ICOS, Hellström et al. (2016) note that the ICOS Carbon Portal includes a Jupyter based VRE for user-initiated data processing. Goor et al. (2016) present a platform designed to support user exploitation of Earth Observation data and propose the possibility of using Jupyter to provide users a means for interactive data analytics. Characterizing the "Geoscience Paper of the Future", Gil et al. (2016) cite electronic notebooks and their ability to capture computational provenance as important tools to increase transparency and reproducibility. The authors note that commonly used frameworks e.g., spreadsheet software and programming environments, are unable to capture computational provenance i.e., capture "what functions were executed and with what parameters and data."

Capturing of such provenance information in notebooks has received some attention. For instance, Pimentel et al. (2015) present the integration of noWorkflow (Murta et al., 2015) in IPython (from which Jupyter originates) to support collecting and analyzing provenance in notebooks. The approach is different from the one pursued here since Pimentel et al. collect detailed provenance at the function, parameter, etc. granularity while our focus is on the provenance of information objects of primary interest to the domain. Hence, rather than collecting information about the function `event()` being called with determinate parameters (**Figure 4**) we collect information about NPF event information (**Listing 4**) being derived from primary data (**Figure 3**).

Cohen-Boulakia et al. (2017) discuss electronic notebooks for *in silico* experiments in relation to scientific workflow design and note that "bridging the gap between the use of scripts and workflows is of paramount importance and would have huge impact on reuse." Beaulieu et al. (2017) discuss the use of Jupyter in environmental research infrastructure (cyberinfrastructure) to generate a multi-disciplinary report for integrated assessment of a marine ecosystem. Notebooks are used to calculate, analyse, and plot indicators relevant for the report. Jupyter "acted as a lightweight, flexible, re-usable, scientific workflow technology to document data processing, analyses, visualization, and reporting." Whole Tale (Brinckman et al., 2018) is a further project that employs Jupyter in VREs that aim to support researchers in all phases of the research data lifecyle, from data acquisition to publication of results and the cross-linking of data, software, workflows and manuscripts to enable reproducibility and reuse.

The key distinctive feature of our work is the integration of semantic technologies to capture, in machine readable form, information acquired in data interpretation i.e., the semantics of derivative data in addition to the derivative data themselves, and the acquisition of such information as (RDF) data in research infrastructures—for further curation, publishing, processing, and use.

6.4 Limitations and Future Work

The prototype infrastructure presented here is intended to serve in discussions with research infrastructures and research communities. We are discussing the presented approach with ACTRIS and SMEAR to explore the possibility of integrating and expanding the presented concepts and approaches. Furthermore, we are collaborating with representatives of the research community to establish a concept for "new particle formation event" as part of the Environment Ontology (ENVO) (Buttigieg et al., 2016). This could enable distributed particle size distribution data interpretation for NPF event studies across Europe and globally as well as the systematic acquisition, curation, and publishing of information about NPF events. We are also interested in scaling the approach out to other kinds of phenomena, especially phenomena related to NPF events, to demonstrate integrated processing of information about phenomena of different kind.

The conceptualization of NPF event in collaboration with the research community and (ENVO) ontology engineers largely unfolds on GitHub and has been laborious. This is known from ontology engineering in general. While of obvious interest to this use case, the practicability of developing conceptualizations on a large scale for concepts as specialized as that of NPF event in collaboration with research communities and ontology engineers remains a significant challenge and practical hindrance to adopting these technologies in infrastructures. The right set of tools could help supporting such activities.

The Python library with specialized functions developed to support this use case is domain specific. This poses an issue in scaling the implementation out to other use cases, in environmental sciences or other disciplines. The program logic required to represent NPF events as presented here is implemented by the specialized `event()` function, which returns an object representing an event. The library handles such objects. For instance, specialized functions compute average durations on sets of event objects. We implemented these to hide complexity and keep the notebook lean and focused on the primary task. However, it is costly to write specialized functions, since it is time consuming and requires a fair amount of technical expertise (e.g., in semantic technologies) which researchers mostly lack.

This concern may be addressed with an approach for bidirectional translation between semantic (RDF) data and data frames e.g., of the Python Data Analysis Library[13] (`pandas`). Data frames enable flexible analysis of the data that constitute information but the semantics of data are implicit in such data structures. An approach for bidirectional translation could be a portable solution that ensures data semantics are curated without interfering with data analysis e.g., the computation of descriptive statistics. Preliminary attempts for such an approach exist e.g., the `pandasrdf` project that integrates `pandas` and RDF.[14]

Building machine learning classifiers also remains a difficult task that requires specialized technical skills. While programming libraries continue to simplify the training, testing, and use of such classifiers, building training data is costly and developing good models is generally not trivial. Furthermore, it is unclear to what extent classifiers are portable and can be reused to classify data from e.g., other locations that are thus acquired by different infrastructures. However, these issues are not unique to this use case and are being addressed in use of machine learning generally.

More foundational work is also needed regarding wider aspects. For instance, we need to understand if researchers are willing to share information acquired in data interpretation and the incentives for infrastructures to extend their support for the acquisition of information, as well as the curation and publishing of the derivative data.

The notion of transforming primary data into information and knowledge is the focus of the Research Data Alliance (RDA) Interest Group (IG) From Observational Data to Information (OD2I).[15] As data interpretation occurs within some sort of VRE, the OD2I IG plans to begin joint activities with the RDA Virtual Research Environment IG. We aim to better understand how (observational) data evolve to information in VREs. Furthermore, the joint activities aim to catalyse the adoption of the concepts presented here in existing and future VREs.

7 Conclusion

For a use case in aerosol science, we have presented how Jupyter and semantic technologies can facilitate the curation in research infrastructures of information acquired in workflows for data interpretation and represented in machine readable format combining structured data and their meaning.

Critical in this work is a coherent conceptual framework for data, information, and perhaps knowledge as well as relevant processes, in particular data interpretation. We build on existing work that provides a solid foundation, conscious of the fact that such conceptual frameworks are subject to continued debate, in particular also regarding their application in research infrastructures.

Our work is a contribution to this debate. We suggest the adoption of technologies in research infrastructures that are capable of representing machine readable information. Furthermore, we suggest the link between the data use and data acquisition phases in the research data lifecycle of research infrastructures can be strengthened to enable the systematic curation, publishing, processing and reuse of information.

We think a stronger link between use and acquisition phases could be a milestone in advancing the computing systems of current research infrastructures from data-based to information and later, knowledge-based. This would be a major step for the computational artefacts of research infrastructures as elements of knowledge infrastructures.

[13] http://pandas.pydata.org/

[14] https://github.com/westurner/pandasrdf

[15] https://www.rd-alliance.org/groups/observational-data-information

Acknowledgements

We thank Dr. Michael Pikridas (The Cyprus Institute) for insight on their NPF event classification scheme. We thank Giuseppe La Rocca, Dr. Yin Chen and EGI Federated Cloud providers for the use of resources. This work has received support from the ENVRIplus and GLOBIS-B projects, which are funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182 and No 654003, respectively; ACTRIS, which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654109; the Academy of Finland under project No 307331; and the European Commission under project No 742206.

Competing Interests

The authors have no competing interests to declare.

References

Aalto, P, Hämeri, K, Becker, E, Weber, R, Salm, J, Mäkelä, JM, Hoell, C, O'dowd, CD, Hansson, H-C, Väkevä, M, Koponen, IK, Buzorius, G and **Kulmala, M.** 2001. Physical characterization of aerosol particles during nucleation events. *Tellus B: Chemical and Physical Meteorology*, 53(4): 344–358. DOI: https://doi.org/10.3402/tellusb.v53i4.17127

Aamodt, A and **Nygård, M.** 1995. Different roles and mutual dependencies of data, information, and knowledge – An AI perspective on their integration. *Data & Knowledge Engineering*, 16(3): 191–222. DOI: https://doi.org/10.1016/0169-023X(95)00017-M

Bandrowski, A, Brinkman, R, Brochhausen, M, Brush, MH, Bug, B, Chibucos, MC, Clancy, K, Courtot, M, Derom, D, Dumontier, M, Fan, L, Fostel, J, Fragoso, G, Gibson, F, Gonzalez-Beltran, A, Haendel, MA, He, Y, Heiskanen, M, Hernandez-Boussard, T, Jensen, M, Lin, Y, Lister, AL, Lord, P, Malone, J, Manduchi, E, McGee, M, Morrison, N, Overton, JA, Parkinson, H, Peters, B, Rocca-Serra, P, Ruttenberg, A, Sansone, S-A, Scheuermann, RH, Schober, D, Smith, B, Soldatova, LN, Stoeckert, CJ, Taylor, CF, Torniai, C, Turner, JA, Vita, R, Whetzel, PL and **Zheng, J.** 2016. The Ontology for Biomedical Investigations. *PLOS ONE*, 11(4). e0154556. DOI: https://doi.org/10.1371/journal.pone.0154556

Barwise, J and **Perry, J.** 1981. Situations and Attitudes. *The Journal of Philosophy*, 78(11): 668–691. DOI: https://doi.org/10.2307/2026578

Beaulieu, SE, Fox, PA, Di Stefano, M, Maffei, A, West, P, Hare, JA and **Fogarty, M.** 2017. Toward cyberinfrastructure to facilitate collaboration and reproducibility for marine integrated ecosystem assessments. *Earth Science Informatics*, 10(1): 85–97. DOI: https://doi.org/10.1007/s12145-016-0280-4

Borgman, CL. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. MIT University Press.

Borgman, CL, Darch, PT, Sands, AE, Pasquetto, IV, Golshan, MS, Wallis, JC and **Traweek, S.** 2015. Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries* 16(3). 207–227. DOI: https://doi.org/10.1007/s00799-015-0157-z

Brinckman, A, Chard, K, Gaffney, N, Hategan, M, Jones, MB, Kowalik, K, Kulasekaran, S, Ludäscher, B, Mecum, BD, Nabrzyski, J, Stodden, V, Taylor, IJ, Turk, MJ and **Turner, K.** 2018. Computing environments for reproducibility: Capturing the "Whole Tale". *Future Generation Computer Systems*. DOI: https://doi.org/10.1016/j.future.2017.12.029

Buttigieg, PL, Pafilis, E, Lewis, SE, Schildhauer, MP, Walls, RL and **Mungall, CJ.** 2016. The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, 7(1): 57. DOI: https://doi.org/10.1186/s13326-016-0097-6

Candela, L, Castelli, D and **Pagano, P.** 2013. Virtual Research Environments: An Overview and a Research Agenda. *Data Science Journal*, 12(0): GRDI75–GRDI81. DOI: https://doi.org/10.2481/dsj.grdi-013

Cohen-Boulakia, S, Belhajjame, K, Collin, O, Chopard, J, Froidevaux, C, Gaignard, A, Hinsen, K, Larmande, P, Le Bras, Y, Lemoine, F, Mareuil, F, Ménager, H, Pradal, C and **Blanchet, C.** 2017. Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, 75: 284–298. DOI: https://doi.org/10.1016/j.future.2017.01.012

Dal Maso, M, Kulmala, M, Riipinen, I, Wagner, R, Hussein, T, Aalto, PP and **Lehtinen, KEJ.** 2005. Formation and growth of fresh atmospheric aerosols: Eight years of aerosol size distribution data from SMEAR II, Hyytiälä, Finland. *Boreal Environment Research*, 10(5): 323–336.

Devlin, K. 1991. *Logic and Information*. Cambridge University Press.

Diepenbroek, M, Grobe, H, Reinke, M, Schindler, U, Schlitzer, R, Sieger, R and **Wefer, G.** 2002. PANGAEA—an information system for environmental sciences. *Computers & Geosciences*, 28(10): 1201–1210. Shareware and freeware in the Geosciences II. A special issue in honour of John Butler. DOI: https://doi.org/10.1016/S0098-3004(02)00039-0

Edwards, PN. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. MIT Press.

Edwards, PN, Jackson, SJ, Chalmers, MK, Bowker, GC, Borgman, CL, Ribes, D, Burton, M and **Calvert, S.** 2013. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges. *Tech. rep.* Ann Arbor, MI: University of Michigan.

Floridi, L. 2010. *Information—A Very Short Introduction*. Oxford University Press. DOI: https://doi.org/10.1093/actrade/9780199551378.001.0001

Floridi, L. 2011. *The Philosophy of Information*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199232383.001.0001

Gil, Y, David, CH, Demir, I, Essawy, BT, Fulweiler, RW, Goodall, JL, Karlstrom, L, Lee, H, Mills, HJ, Oh, J-H, Pierce, SA, Pope, A, Tzeng, MW, Villamizar, SR and **Yu, X.** 2016. Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, 3(10): 388–415. DOI: https://doi.org/10.1002/2015EA000136

Goor, E, Dries, J, Daems, D, Paepen, M, Niro, F, Goryl, P, Mougnaud, P and **Vecchia, AD.** 2016. PROBA-V Mission Exploitation Platform. *Remote Sensing*, 8(7). DOI: https://doi.org/10.3390/rs8070564

Hamed, A, Joutsensaari, J, Mikkonen, S, Sogacheva, L, Dal Maso, M, Kulmala, M, Cavalli, F, Fuzzi, S, Facchini, MC, Decesari, S, Mircea, M, Lehtinen, KEJ and **Laaksonen, A.** 2007. Nucleation and growth of new particles in Po Valley, Italy. *Atmospheric Chemistry and Physics*, 7(2): 355–376. DOI: https://doi.org/10.5194/acp-7-355-2007

Hari, P and **Kulmala, M.** 2005. Station for Measuring Ecosystem-Atmosphere Relations (SMEAR II). *Boreal Environment Research*, 10(5): 315–322.

Harris, S and **Seaborne, A.** 2013. SPARQL 1.1 Query Language. *Recommendation W3C*.

Heiler, S. 1995. Semantic Interoperability. *ACM Comput. Surv*, 27(2): 271–273. DOI: https://doi.org/10.1145/210376.210392

Hellström, M, Vermeulen, A, Mirzov, O, Sabbatini, S, Vitale, D, Papale, D, Tarniewicz, J, Hazan, L, Rivier, L, Jones, SD, Pfeil, B and **Johannessen, T.** 2016. Near Real Time Data Processing In ICOS RI. In *2nd international workshop on interoperable infrastructures for interdisciplinary big data sciences (it4ris 16) in the context of ieee real-time system symposium (rtss)*. DOI: https://doi.org/10.5281/zenodo.204817

Hitzler, P, Krötzsch, M, Parsia, B, Patel-Schneider, PF and **Rudolph, S.** 2012. OWL 2 Web Ontology Language Primer (Second Edition). *Recommendation W3C*.

Jennings, NR, Sycara, K and **Michael, W.** 1998. A Roadmap of Agent Research and Development. *Autonomous Agents and Multi-Agent Systems*, 1(1): 7–38. DOI: https://doi.org/10.1023/A:1010090405266

Junninen, H, Lauri, A, Keronen, P and **Aalto, P.** 2009. Smart-SMEAR: On-line data exploration and visualization tool for SMEAR stations. *Boreal Environment Research*, 14(4): 447–457.

Karasti, H, Millerand, F, Hine, CM and **Bowker, GC.** (eds.) 2016. *Special issue on knowledge infrastructures*. Science & Technology Studies.

Keller, M, Schimel, DS, Hargrove, WW and **Hoffman, FM.** 2008. A continental strategy for the National Ecological Observatory Network. *Frontiers in Ecology and the Environment*, 6(5): 282–284. DOI: https://doi.org/10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2

Kissling, WD, Ahumada, JA, Bowser, A, Fernandez, M, Fernández, N, García, EA, Guralnick, RP, Isaac, NJP, Kelling, S, Los, W, McRae, L, Mihoub, J-B, Obst, M, Santamaria, M, Skidmore, AK, Williams, KJ, Agosti, D, Amariles, D, Arvanitidis, C, Bastin, L, De Leo, F, Egloff, W, Elith, J, Hobern, D, Martin, D, Pereira, HM, Pesole, G, Peterseil, J, Saarenmaa, H, Schigel, D, Schmeller, DS, Segata, N, Turak, E, Uhlir, PF, Wee, B and **Hardisty, AR.** 2017. Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews*, 93(1). DOI: https://doi.org/10.1111/brv.12359

Kluyver, T, Ragan-Kelley, B, Pérez, F, Granger, B, Bussonnier, M, Frederic, J, Kelley, K, Hamrick, J, Grout, J, Corlay, S, Ivanov, P, Avila, D, Abdalla, S, Willing, C and **Jupyter development team.** 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Loizides, F and Schmidt, B (eds.), *Positioning and power in academic publishing: Players, agents and agendas*, 87–90. IOS Press. DOI: https://doi.org/10.3233/978-1-61499-649-1-87

Kulmala, M, Vehkamäki, H, Petäjä, T, Dal Maso, M, Lauri, A, Kerminen, VM, Birmili, W and **McMurry, PH.** 2004. Formation and growth rates of ultrafine atmospheric particles: A review of observations. *Journal of Aerosol Science*, 35(2): 143–176. DOI: https://doi.org/10.1016/j.jaerosci.2003.10.003

Lebo, T, Sahoo, S and **McGuinness, D.** 2013. PROV-O: The PROV Ontology. *Recommendation W3C*.

Lindstrom, E, Gunn, J, Fischer, A, McCurdy, A and **Glover, LK.** 2012. A Framework for Ocean Observing. *Task Team for an Integrated Framework for Sustained Ocean Observing IOC/INF-1284 UNESCO*.

Mitchell, TM. 1997. *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc. 1st edn.

Murta, L, Braganholo, V, Chirigati, F, Koop, D and **Freire, J.** 2015. noWorkflow: Capturing and Analyzing Provenance of Scripts. In: Ludäscher, B and Plale, B (eds.), *Provenance and annotation of data and processes*, 71–83. Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-16462-5_6

Nieva de la Hidalga, A, Magagna, B, Stocker, M, Hardisty, A, Martin, P, Zhao, Z, Atkinson, M and **Jeffery, K.** 2017. The ENVRI Reference Model (ENVRI RM) version 2.2, 30th October 2017. DOI: https://doi.org/10.5281/zenodo.1050349

Pearlman, J, Schaap, D and **Glaves, H.** 2016. Ocean Data Interoperability Platform (ODIP): addressing key challenges for marine data management on a global scale. In: *Oceans 2016 mts/ieee monterey*, 1–7.

Pimentel, J, Nicolaci, F, Braganholo, V, Murta, L and **Freire, J.** 2015. Collecting and Analyzing Provenance on Interactive Notebooks: When IPython Meets noWorkflow. In *Proceedings of the 7th usenix conference on theory and practice of provenance TaPP'*, 15: 10–10. Berkeley, CA, USA: USENIX Association.

Roemmich, D, Johnson, G, Riser, S, Davis, R, Gilson, J, Owens, WB, Garzoli, S, Schmid, C and **Ignaszewski, M.** 2009. The Argo Program: Observing the Global Oceans with Profiling Floats. *Oceanography*, 22(2): 34–43. DOI: https://doi.org/10.5670/oceanog.2009.36

Schreiber, G and **Raimond, Y.** 2014. RDF 1.1 Primer. *Working Group Note W3C*.

Schubert, L and **Jeffery, K.** 2015. New Software Engineering Requirements in Clouds and Large-Scale Systems. *IEEE Cloud Computing*, 2(1): 48–58. DOI: https://doi.org/10.1109/MCC.2015.13

Stocker, M. 2015. *Situation Awareness in Environmental Monitoring*. University of Eastern Finland dissertation.

Stocker, M. 2017. *Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities.* chap. Advancing the Software Systems of Environmental Knowledge Infrastructures, 399–423. Taylor & Francis Group, CRC Press. DOI: https://doi.org/10.1201/9781315368252-16

Stocker, M, Baranizadeh, E, Hamed, A, Rönkkö, M, Virtanen, A, Laaksonen, A, Portin, H, Komppula, M and **Kolehmainen, M.** 2013. Acquisition and Representation of Knowledge for Atmospheric New Particle Formation. In: Hřebíček, J, Schimak, G, Kubásek, M and Rizzoli, AE (eds.), *Environmental software systems. Fostering information sharing*, vol. 413 IFIP Advances in Information and Communication Technology, 98–108. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-41151-9_10

Stocker, M, Baranizadeh, E, Portin, H, Komppula, M, Rönkkö, M, Hamed, A, Virtanen, A, Lehtinen, K, Laaksonen, A and **Kolehmainen, M.** 2014. Representing situational knowledge acquired from sensor data for atmospheric phenomena. *Environmental Modelling & Software*, 58: 27–47. DOI: https://doi.org/10.1016/j.envsoft.2014.04.006

Stocker, M, Rönkkö, M and **Kolehmainen, M.** 2015. Knowledge-based environmental research infrastructure: moving beyond data. *Earth Science Informatics*, 9(1): 47–65. DOI: https://doi.org/10.1007/s12145-015-0230-6

Von Neumann, J. 1993. First draft of a report on the EDVAC. *IEEE Annals of the History of Computing*, 15(4): 27–75. DOI: https://doi.org/10.1109/85.238389

Vossepoel, S and **Murray, MS.** 2016. Advancing Knowledge for a Changing North: Open, Interoperable, and Collaborative Pan-Arctic Data and Information Management at the Arctic Institute of North America. In: *Scidatacon 2016—advancing the frontiers of data in research*, Ubiquity Press.

Zaidan, MA, Haapsilta, V, Relan, R, Junninen, H, Aalto, P, Canova, FF, Laurson, L and **Foster, A.** 2017. Neural network classifier on time series features for predicting atmospheric particle formation days. In: Halonen, R, Nikandrova, A, Kontkanen, J, Enroth, J and Vehkamäki, H (eds.), *The 20th international conference on nucleation and atmospheric aerosols (Report Series in Aerosol Science 200)*, 687–690.

Zins, C. 2007. Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, 58(4): 479–493. DOI: https://doi.org/10.1002/asi.20508

How to cite this article: Stocker, M, Paasonen, P, Fiebig, M, Zaidan, MA and Hardisty, A. 2018. Curating Scientific Information in Knowledge Infrastructures. *Data Science Journal,* 17: 21, pp. 1–16, DOI: https://doi.org/10.5334/dsj-2018-021

Submitted: 31 May 2018 **Accepted:** 31 August 2018 **Published:** 20 September 2018

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS