A Missing Link from Data to Knowledge: Infrastructure that Curate the Meaning of Data

Markus Stocker (1,2)
http://orcid.org/0000-0001-5492-3212

Markus Fiebig (3)
http://orcid.org/0000-0002-3380-3470

Alex Hardisty (4)
http://orcid.org/0000-0002-0767-4310

(1) TIB Leibniz Information Centre for Science and Technology
Welfengarten 1 B, 30167 Hannover, Germany
markus.stocker@tib.eu
(2) MARUM Center for Marine Environmental Sciences
PANGAEA Data Publisher for Earth & Environmental Science
Leobener Strasse 8, 28359 Bremen, Germany
mstocker@marum.de
(3) NILU - Norsk Institutt for Luftforskning, Dept. Atmospheric and Climate Research
Instituttveien 18, 2007 Kjeller, Norway
markus.fiebig@nilu.no
(4) School of Computer Science and Informatics, Cardiff University
Queens Buildings, 5 The Parade, Cardiff CF24 3AA, United Kingdom
hardistyar@cardiff.ac.uk

With themes of "The critical role of university RDM infrastructure in transforming data to knowledge," "From data to knowledge," and "Translating ecological data into knowledge and decisions in a rapidly changing world", the Göttingen-CODATA RDM Symposium 2018[1], the RDA 11th Plenary Meeting[2], and the 10th International Conference on Ecological Informatics[3], respectively, have in common the notion "from data to knowledge". This is also the 2017 Motto of the Helmholtz Association of German Research Centres[4].

The idea of transforming data into knowledge is popular, indeed. Among others, research infrastructures emphasize there is knowledge to obtain through observation. For instance, the Integrated Carbon Observation System (ICOS) research infrastructure uses the tagline "Knowledge through observations"[5]. 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

[1] https://conference.codata.org/conference/2018_Goettingen_RDM/about/
[2] https://www.rd-alliance.org/plenaries/rda-eleventh-plenary-meeting-berlin-germany
[3] http://icei2018.uni-jena.de/
[4] https://www.helmholtz.de/ueber_uns/die_gemeinschaft/mission/motto_2017/
[5] https://twitter.com/ICOS_RI/status/803156982349729793

[...] to enter a new paradigm of knowledge in the XXI Century"[6]. The European Open Science Cloud (EOSC) is envisioned as an environment that enables turning ever increasing amounts of data "into knowledge as renewable, sustainable fuel for innovation in turn to meet global challenges"[7].

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 transforming data into 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" [1]. Indeed, information is represented as data in (computer) systems, which could suggest a "regression" from information to data. If we qualify data as observational, experimental or computational [2] - for simplicity, primary data - and information to be about the unit of analysis, the logical progression may be more defensible. Information about a natural phenomenon of analysis is thus obtained from primary data; i.e., data that results from the activity of observation carried out by 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 observational, experimental or computational primary data. 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 transformation of primary data into information about units of analysis, and possibly knowledge. According to its standard definition, information is *meaningful* well-structured data [3]. Aamodt and Nygård [4] proposed that interpretation is the activity that transforms data as uninterpreted symbols with "no meaning for the system concerned" into information, i.e. "data with meaning." 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 [5].

This essential role demands a conceptualization that unifies people and infrastructures, including research infrastructures and university research data management infrastructures. Knowledge infrastructure, described by Paul Edwards [6] 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 and include artifacts such as scientific equipment, scientific data and computer agents[8].

We argue that the concept of knowledge infrastructure 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. Surely among others, a key challenge faced by such infrastructures is the curation of information, i.e. caring for

---

[6] http://www.emsodev.eu/work_packages.html

[7] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

[8] This view is aligned with the EU Regulation No 1291/2013 [7] which describes research infrastructures as "facilities, resources and services that are used by the research communities to conduct research and foster innovation in their fields."

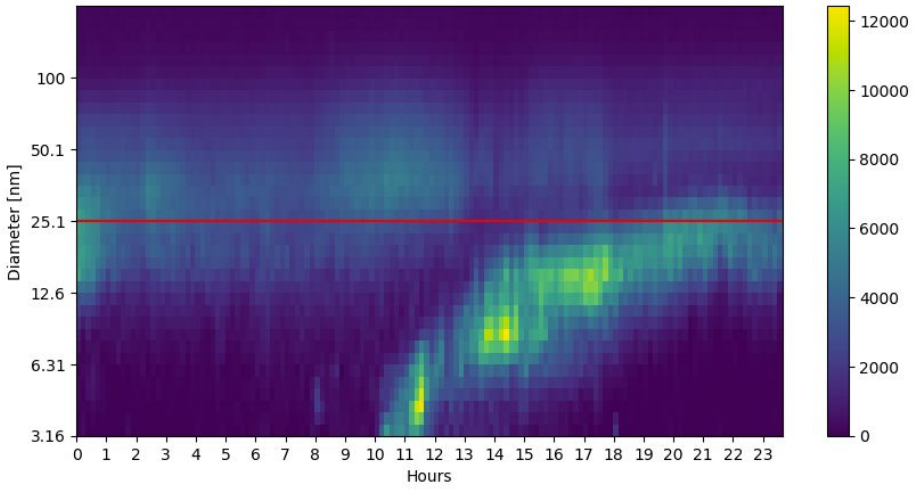
and presenting data *and* their meaning. Concretely, the meaning of data resulting from human-in-the-loop primary data interpretation activities should be explicit and formal, and thus, machine readable. This implies systematic acquisition of the meaning of data resulting from primary data interpretation activities by infrastructures. 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. For instance, in addition to curating a color coded world map as a matrix of pixel values, systems should also curate a corresponding digital object that represents the image's information content, explicitly and formally. Of course, it should also be possible to interrogate such an object to discover specific details / subsets of information of interest.

Incidentally, some aspects of the FAIR Guiding Principles [8] are relevant here. To be Interoperable, the Principles suggest that "(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation" (I1). The phrase '(meta)data' indicates that the principle should be applied to both metadata and data. Principle I1 is relevant here because formal languages for knowledge representation, such as the Web Ontology Language [9], enable systems to curate the meaning of data, explicitly and formally. The Principles further suggest that to be Interoperable "(meta)data use vocabularies that follow FAIR principles" (I2). Hence, the vocabularies of terms used to describe data should be specified using a formal language for knowledge representation, among other requirements.

```
In [1]: from pynpf.smear.datafetcher import fetchdata
        from pynpf.smear.dataplotter import plotdata
        from pynpf.factory import record, event

        day = '2013-04-04'
        place = 'Hyytiälä'

In [2]: # Fetch and plot particle size distribution data for the given day and place
        # Data fetched from SmartSMEAR, https://avaa.tdata.fi/web/smart
        plotdata(fetchdata(day, place))
```




```
In [3]: # Record information about particle formation
        record(event(day, place, beginning='11:00', end='19:00', classification='Class Ia'))
```

Figure 1: Jupyter based approach to expose the interpretation activity "as a service" to the research community. Particle formation occurred on April 4, 2013, in Hyytiälä, Finland.

To illustrate the challenge, we present a use case we have studied for some time [10]. Researchers of a community in aerosol science interpret observational data to obtain information about atmospheric particle formation processes. Observational data result in observing (measuring) atmospheric aerosol over

time, an activity carried out automatically by a sensing device. Daily and for specific locations, aerosol scientists interpret the corresponding observational data, visually. Visual data interpretation amounts to determining the presence of a shape that reflects the occurrence of particle formation on the given day and location (see plot in Figure 1). Here, data interpretation is a human-in-the-loop activity. As particle formation is the unit of analysis in this context, the result of data interpretation is information, i.e. well-formed data describing particle formation that are also meaningful in this context. Terms of various vocabularies are used in such descriptions, including: concepts for atmospheric particle formation; categories of classification schemes for particle formation; and concepts and relations for describing time and space.

Lacking infrastructure support, researchers curate information about particle formation as well-formed data but with little expressed explicit and formal meaning. Indeed, information about particle formation is typically reduced to well-formed data in tabular form with metadata determining the correct interpretation and meaning of values. Thus, metadata determines that the value 735328 is to be interpreted as MATLAB datenum and hence means the day April 4, 2013, during which particle formation occurred in Hyytiälä (see Figure 1). As a result of not using a formal language for knowledge representation, curated information is well-formed data with largely implicit meaning, especially to computer agents.

Furthermore, the various research groups in a community may not have agreed on form and meaning of data. This can be observed in the use case presented here, where two research groups of the community encode information about particle formation as data with heterogeneous form and meaning. Information is thus difficult to comprehend and integrate.

In the context of this use case, we offer a Jupyter[9] [11] based approach that aims to address some of these issues (Figure 1). Jupyter exposes the interpretation activity "as a service" to the research community while it enables infrastructure to ensure information about new particle formation is acquired and curated as well-formed data with explicit and formal meaning. Indeed, here infrastructure represents information about particle formation using the Web Ontology Language, a language for knowledge representation, and adopting vocabularies that meet domain-relevant community standards. We argue that the use case demonstrates how infrastructure can support transforming primary data into information about phenomena of the natural world, an essential subtask along the way to knowledge.

We have suggested that to support transforming data to knowledge, it is critical for Research and RDM infrastructures to acquire and curate information, and that such acquisition and curation relies on infrastructure to represent meaning explicitly and formally. Infrastructure should thus not merely acquire and curate the aerosol plot as a digital image but also the plot's information content, as well-formed data that are meaningful also to computer agents.

The kind of institution best suited to operate such infrastructure is an interesting aspect to discuss. Perhaps relatively uncontroversial is that e-Infrastructures / cyberinfrastructures are ideally suited to expose the interpretation activity as a service to the research community. In our example, it is the European Grid Infrastructure[10] (EGI) that provides the necessary resources. However, institutional e-Infrastructure, including those of universities, can provide the similar resources.

[9] http://jupyter.org
[10] https://www.egi.eu/

As curators of the primary data analysed by research communities, the data centers of research infrastructures are also ideally positioned to acquire and curate information derived from analysis. In the context of our use case, the European Aerosols, Clouds and Trace gases Research Infrastructure[11] (ACTRIS) and the Finnish Station for Measuring Ecosystem-Atmosphere Relations[12] (SMEAR) [12] are two research infrastructures that serve the research community with relevant observational data. While these infrastructures support the visualization of particle size distribution data (Figure 2), they do not currently deal with derived information about particle formation.

Figure 2: ACTRIS visualization of particle size distribution data for June 2006, in Birkenes, Norway.

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 such RDM infrastructure. 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 ownership, confidentiality (where necessary) and re-use. However, with their broader scope compared to specialized research infrastructures, university RDM infrastructures / digital libraries lack required domain expertise. A third alternative are specialized data curators / publishers, such as PANGAEA[13]. Notwithstanding which institution ultimately operates such RDM infrastructures, most if not all of them must develop their systems in order to acquire and curate information as suggested here.

Another aspect to consider is the streamed and possibly temporary nature of information. In our use case, observational data are interpreted regularly. Albeit with low frequency, information is thus created in a streamed manner. Furthermore, it may be temporary as it can be updated or deleted, e.g. in quality control. Moreover, it may only be partially utilized in scientific work. This raises the question of how RDM infrastructure should treat such information. Should it, at first, be acquired and curated by individual researchers with their personal infrastructure (e.g., workstation) and only the information relevant to scientific work be acquired and curated by institutional RDM infrastructure? Should all

[11] http://www.actris.eu/
[12] https://www.atm.helsinki.fi/SMEAR/
[13] https://www.pangaea.de/

information be acquired and curated (after all it may be valuable to other scientific work) but in batch rather than streamed mode? As with all aspects of research data management, infrastructures must be capable of supporting multiple policies catering to a wide range of scenarios arising from different community needs and practices.

Finally, there is the proposition that attaching formal meaning to data (i.e., creating interpretable information) leads directly to improved interoperability (one of the FAIR principles); 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 involves possession of a shared and congruent understanding of the context, including the important assumptions, principles, facts, notions and relations existing within that context. Alternatively, it involves possession of the capability to infer and build that understanding i.e., the context from (meta)data exchanged.

The notion of transforming primary data to information and knowledge is the focus of the (prospective) Research Data Alliance (RDA) Interest Group (IG) From Observational Data to Information (OD2I)[14]. The OD2I IG recognises that primary data are interpreted for their meaning along value chains in determinate contexts of scientific, industrial, or broadly societal relevance, and aims to analyse this constant in different contexts to determine common patterns and develop solutions for common problems.

Acknowledgement. This work has received support by the ENVRIplus project, which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182, as well as by ACTRIS, which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 654109.

References

[1] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge Journal of the American Society for Information Science and Technology, Wiley Subscription Services, Inc., A Wiley Company, 58, 479-493. https://doi.org/10.1002/asi.20508

[2] Borgman, C. L. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. MIT University Press.

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

[4] Aamodt, A., Nygård, M. (1995). Different roles and mutual dependencies of data, information, and knowledge – An AI perspective on their integration, Data & Knowledge Engineering, vol. 16, no. 3, pp. 191-222. https://doi.org/10.1016/0169-023X(95)00017-M

[5] Jennings, N.R., Sycara, K., Wooldridge, M. (1998). A Roadmap of Agent Research and Development. Autonomous Agents and Multi-Agent Systems, 1(1):7-38. https://doi.org/10.1023/A:1010090405266

[6] Edwards, P. (2010). A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. ISBN: 9780262013925. MIT Press.

[7] Regulation (EU) No 1291/2013 of the European Parliament and of the Council http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:347:0104:0173:EN:PDF

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

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

[9] W3C OWL Working Group (2012). OWL 2 Web Ontology Language Document Overview (Second Edition), W3C. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/
[10] Stocker, M., et al. (2014). Representing situational knowledge acquired from sensor data for atmospheric phenomena. Environmental Modelling & Software, 58:27-47. https://doi.org/10.1016/j.envsoft.2014.04.006
[11] Pérez, F. Granger, B. E. (2007) IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering, 9(3):21-29. https://doi.org/10.1109/MCSE.2007.53. URL: http://ipython.org
[12] Hari, P., et al. (2013). Station for Measuring Ecosystem-Atmosphere Relations: SMEAR. In: Hari P., Heliövaara K., Kulmala L. (eds) Physical and Physiological Forest Ecology. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-5603-8_9