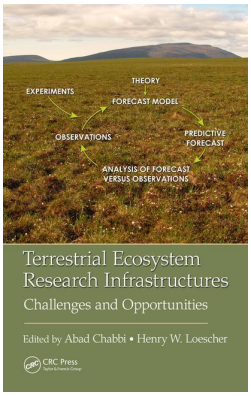

Terrestrial Ecosystem Research Infrastructures Challenges and Opportunities

Abad Chabbi, Henry W. Loescher

Advancing the Software Systems of Environmental Knowledge Infrastructures

Publication details
https://www.routledgehandbooks.com/doi/10.1201/9781315368252-16
Abad Chabbi, Henry W. Loescher, Markus Stocker
Published online on: 22 Feb 2017

15

Advancing the Software Systems of Environmental Knowledge Infrastructures

Markus Stocker

CONTENTS

15.1 Introduction

To acquire data about the environment is a core task of environmental infrastructures. Acquired data are for selected properties of certain elements of the environment, such as the temperature of air or the height of seedlings—where temperature and height are the properties of the elements air and seedlings, respectively. Data result in measurement, the "process of empirical, objective, assignment of numbers to properties" (Finkelstein, 1982). Measurement is repeated as the properties of elements are monitored in time and space (Meijers, 1986).

Environmental infrastructures often employ environmental sensor networks (Hart and Martinez, 2006) to automate data acquisition. Sensors automate monitoring, that is, automatically repeat measurement. A sensor generally monitors one property over time. A sensor system, with multiple sensors as its constituent parts, enables monitoring of multiple properties over time. Networked sensor systems, deployed at multiple locations, enable monitoring of properties over time and space.

Acquired data are processed to gain information about the environment, and information is transferred into knowledge. These tasks are typically performed manually by human agents. Environmental infrastructures thus consist of technical agents as hardware and software, for example, sensors and databases, and human agents, for example, technicians, engineers, and scientists. Environmental infrastructures are thus sociotechnical systems (Fox, 1995), consisting of technical and social subsystems. To increase human knowledge and understanding of the environment is arguably the primary aim of these sociotechnical systems. We thus speak of environmental *knowledge* infrastructures and of environmental knowledge *research* infrastructures if they serve primarily research. Edwards (2010) defined knowledge infrastructures as "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds." Environmental knowledge infrastructures focus on generating, sharing, and maintaining specific knowledge primarily about *natural* worlds. Being "networks of people, artifacts, and institutions," Edwards also underscores the sociotechnical character of knowledge infrastructures.

Especially in large-scale environmental knowledge infrastructures (Kratz et al., 2006; Keller et al., 2008; Michener et al., 2011), data acquisition, curation, access, and processing are increasingly often left to technical subsystems as they enable automation, for example, data acquisition by means of sensors or data management by means of databases. The technical subsystems of environmental knowledge infrastructures are thus data-based systems. In contrast, data analysis and interpretation, that is, the acquisition of information from data and transfer of information into knowledge, are carried out by social subsystems, often with little support from technical subsystems.

We envision that the technical subsystems, in particular software systems, of future environmental knowledge infrastructures will advance from data-based systems to knowledge-based systems. The technical subsystems of future environmental knowledge infrastructures may thus more actively support human agents in information acquisition and support the curation of machine interpretable knowledge and automated knowledge processing. We thus envision future environmental knowledge infrastructures with knowledge-based technical subsystems.

In this chapter, we present the environmental knowledge infrastructures of two case studies and underscore the sociotechnical character of the infrastructures. We discuss the kinds of information the infrastructures acquire from data; the agents and methods involved in information acquisition and transfer of information into knowledge; and the kinds of resulting knowledge. We then present methods and technologies that enable technical subsystems of environmental knowledge infrastructures to more actively support human agents in—or altogether automate—information acquisition and support the curation of machine interpretable knowledge and thus automated knowledge processing. The discussed approaches are presented as possible building blocks toward knowledge-based technical subsystems.

Information and knowledge are *about the environment* monitored by the environmental knowledge infrastructure. Information is acquired from data about the monitored environment. Other information and knowledge types are obviously relevant to environmental knowledge infrastructures, for example, information about technical agents, workflows, or data. These are however not of concern here.

Other authors have discussed the idea of knowledge-driven sociotechnical systems that learn from scientific data. For instance, Peters et al. (2014) present the architecture of a Knowledge Learning and Analysis System that aims at being "a knowledge-driven, open access system that 'learns' and becomes more efficient and easier to use as streams of data, and the number and types of user interactions, increase." Peters et al. discuss the integration of hypothesis-driven and data-intensive machine learning scientific approaches. Ganguly et al. (2007) propose a framework for knowledge discovery on environmental data in scientific applications. Naturally, knowledge infrastructures—in particular also infrastructures that employ environmental sensor networks—are not limited to scientific applications. For instance, Parmiggiani and Monteiro (2016) discuss a knowledge infrastructure developed at a Norwegian oil and gas company and note that the infrastructure based on environmental monitoring attempts "to abstract the datasets into general representations of environmental risk that make sense for the oil and gas professionals." In contrast to these works for high-level architectural and system descriptions and analysis, this chapter discusses the application of concrete software methods and technologies that enable the development and implementation of knowledge-based technical subsystems in environmental infrastructures.

15.2 Case Studies

Environmental knowledge infrastructures are sociotechnical systems consisting of technical and social subsystems. Hardware and software are agents of technical subsystems, while humans and communities are agents of social subsystems. While data acquisition and the curation and processing of data are fundamental to environmental knowledge infrastructures, data are merely intermediate products from which infrastructures acquire information and transfer information into knowledge.

We present the environmental knowledge infrastructures of two case studies that underscore the sociotechnical character of the infrastructures and highlight how their overall aim is to acquire information, transfer information into knowledge, and curate and process knowledge. For each case study, we describe the relevant data, information, and knowledge as well as the technical and social agents involved in data processing, information

acquisition, transfer of information into knowledge, and knowledge curation and processing. The case studies were originally developed in Stocker (2015).

15.2.1 Atmospheric New Particle Formation

The first case study is in aerosol science for the study of atmospheric new particle formation. The environmental knowledge infrastructure is thus a *research* infrastructure, and Stocker et al. (2014a) discuss the case study in more details.

Atmospheric new particle formation is an atmospheric phenomenon whereby new particles are formed and over time grow in size (Kulmala et al., 2004). The phenomenon has been documented in a wide variety of environments all over the world (Kulmala et al., 2004) and is studied because aerosol particles are known to scatter sunlight and influence quality of life, for instance, by affecting human health (Pope III et al., 2002). The scattering of radiation has a cooling effect on the climate (Solomon et al., 2007). The study of new particle formation is thus relevant to climate change research.

The environmental knowledge infrastructure involves the Finnish Station for Measuring Ecosystem–Atmosphere Relations (Hari and Kulmala, 2005, SMEAR), in particular the station located at the Puijo observation tower in Kuopio, Finland. This station is part of SMEAR IV, which is itself part of the wider SMEAR network with stations located in Eastern Lapland, Hyytiälä, Helsinki, and Kuopio.

The station consists of sensing devices for the monitoring of aerosols, weather, and atmospheric gases (Leskinen et al., 2009). Of interest here is the Differential Mobility Particle Sizer (DMPS) utilized to monitor the particle size distribution of polydisperse aerosols. A DMPS consists of a Differential Mobility Analyzer (DMA) and a Condensation Particle Counter (CPC). The particles of polydisperse aerosols are first classified according to diameter size by the DMA and then counted by the CPC (Kulkarni et al., 2011). The instrument measures the particle number concentration (cm^{-3}) for 40 discrete diameter sizes in the range of 7–800 nm, on average five times per hour.

In studying atmospheric new particle formation, a core task for the environmental knowledge infrastructure is to identify and classify individual events, that is, instances of the atmospheric phenomenon, as they occur in time and space. Different classification schemes have been proposed to characterize individual events (Dal Maso et al., 2005; Hamed et al., 2007; Vana et al., 2008). The identification and classification of events occur on processed DMPS data and are performed by human agents, in particular aerosol scientists.

Data acquisition in this environmental knowledge infrastructure, in particular measurement by the DMPS and collection over the network, is largely automated by the hardware and software agents of the technical subsystem. In contrast, the extraction of information about new particle formation

events, the transfer of such information into knowledge, and the curation and processing of knowledge are performed by aerosol scientists and are—with the exception of software for statistical computing and reporting, that is, MATLAB® and Excel—hardly supported by the technical subsystem of the infrastructure.

Acquired data undergo an inversion from sensor data in (V) to particle number concentration in (cm^{-3}) (Wiedensohler et al., 2012). Such data processing is implemented in MATLAB. The resulting data are curated as text files consisting of an $m \times n$ data matrix, where m is the number of measurements over 24 h (1 day) for 40 particle diameter sizes and $n = 41$ (includes the timestamp). The daily text files are stored on a file system accessible to researchers. The technical subsystem of the environmental knowledge infrastructure largely automates these steps.

Researchers access processed data and create data products that support them in the visual assessment of new particle formation events on a particular day and location (Hamed et al., 2007). MATLAB is the software agent used to create figures for visual assessment. The figures display time (24 h) on the x-axis and particle number concentration for the 40 measured particle diameter sizes of monitored polydisperse aerosol on the y-axis. A color gradient is used to represent low-to-high concentration. On a day during which a clearly visible event occurred, the figure displays a characteristic so-called banana shape, reflecting the high concentration of very small particles that grow in diameter size over time.

Having identified a new particle formation event, aerosol scientists characterize the event. Among the extracted features, scientists may classify the event, for instance, based on its visual clarity, obtain an estimate for event start and end times, and compute formation and growth rates (Hamed et al., 2007). Acquired information about events is recorded and Excel is the software agent used for the curation of information. At a minimum, information includes the day at which an event occurs and the event class.

By recording information about a particular event of atmospheric new particle formation in the columns of an Excel row, the scientist creates a knowledge object about the event. The knowledge object integrates contextual information about the event and is curated in Excel. In this environmental knowledge infrastructure, we may further specialize the object as a *situational* knowledge object. Barwise and Perry (1980) and Devlin (1991) suggested that a situation is a structured part of reality that an agent manages to individuate. New particle formation events are objects in parts of reality, that is, connected regions of space–time (Barwise and Perry, 1981). They are thus objects in situations. The environmental knowledge infrastructure individuates structured parts of reality and is thus the agent that individuates situations.

Scientists utilize the recorded knowledge objects in further analysis, for example, to compute the monthly frequency of event classes, seasonal differences in hourly mean total particle concentration between event and

nonevent days, or monthly mean event duration, formation, and growth rates (Hamed et al., 2007). Knowledge objects are thus processed. The results of such analysis are presented as figures and tables and are discussed in the natural language text of scientific journal articles. More abstractly, the results of such analysis are new information integrated into existing knowledge structures (Aamodt and Nygård, 1995).

Clearly, the infrastructure is an environmental *knowledge* infrastructure. Beyond acquiring, curating, and processing data about particle size distribution of polydisperse aerosols, the environmental infrastructure extracts, curates, and processes information and knowledge about events of atmospheric new particle formation. The infrastructure's aim is to increase human knowledge and understanding of atmospheric new particle formation.

The environmental knowledge infrastructure is furthermore a sociotechnical system as technical and social subsystems collaborate to attain the infrastructure's aim to increase human knowledge and understanding of atmospheric new particle formation. The technical subsystem consists of hardware and software agents. Hardware agents include the DMPS, communication links, and computers. Software agents include MATLAB, custom MATLAB scripts, and Excel. While hardware and software agents certainly do serve toward data processing and analysis, their role is primarily in data acquisition, curation, and access. The social subsystem consists of aerosol scientists and technicians. Human agents are involved in data acquisition, curation, and access, but the role of human agents, in particular scientists, is in data analysis, information acquisition, transfer of information into knowledge, and knowledge curation and processing. The technical subsystem of the environmental knowledge infrastructure is a data-based system and as such primarily concerned with tasks required prior to data analysis. The social subsystem builds on the data-based system and extends it with functionality for data analysis and interpretation. The social subsystem thus turns the infrastructure into a knowledge-based system.

15.2.2 Plant Disease Outbreaks

The second case study is in precision agriculture, for the assessment of (acute) disease outbreaks in plants. The environmental knowledge infrastructure serves agricultural advisors to farmers, and Stocker et al. (2016) discuss the case study in more details.

Plant disease is a threat to plant growth, quality, harvest, and thus economic return. Hence, farmers need to monitor disease progress to determine the right time when plants need to be protected, for example, by spraying chemical agents. Various factors other than disease progress influence decisions to protect plants, for example, regulations, utilized protective agent, or protection history. Decision-making thus depends on knowledge, that is, integrated information. Indeed, modern precision agriculture

is "intrinsically information intensive" (Fountas et al., 2006). Farmers are guided by agricultural advisors, and together they are part of the social subsystem of an environmental knowledge infrastructure designed to support decision-making.

Agricultural advisors rely on computer models and systems that support them in information acquisition. In this case study, agricultural advisors utilize a mechanistic model for estimating disease pressure. Disease pressure is computed as the cumulative value $AR_t = AR_{t-1} + DR$, where AR_t is the accumulated disease pressure value on day t and DR is the change on a given day. DR is constructed from a base risk value modified by daily modifiers. The base risk depends on the susceptibility of the selected crop and farming history. The daily modifiers are computed from data for the weather on the given day, specifically average temperature, humidity, wind, and the amount of rainfall. Diseases included in the model are *Pyrenophora teres*, *Pyrenophora tritici-repentis*, and *Stagonospora nodorum* and follow this general disease pressure model. How the base risk or the daily modifiers are used, however, depends on the disease, as all diseases react to changes in the environmental variables in a unique manner. The model updates disease pressure once per day.

Agricultural advisors operate weather stations as part of the environmental knowledge infrastructure to monitor a range of environmental properties, including temperature, relative humidity, wind speed, and cumulative precipitation. The weather stations are part of the SoilWeather Wireless Sensor Network (WSN) (Kotamäki et al., 2009). Each weather station is a sensor system and consists of several sensing devices. Observation data can be accessed via a Web service. Observation data are complemented with seasonal data for the agricultural parcels in the region observed by the infrastructure. Such data include the preceding crop, current crop, current crop susceptibility, tillage method, and seeding date.

The environmental knowledge infrastructure utilizes data to compute disease pressure for the region observed by the infrastructure. The results are (daily) maps that display disease pressure as color-coded spatial features. Given such maps, agricultural advisors can monitor the progress of disease pressure in the region and obtain information about disease pressure that exceeds the threshold for which outbreaks are expected. Knowledge about possible outbreaks is obviously of interest to farmers.

The infrastructure is clearly an environmental knowledge infrastructure. As in the previous case study, beyond acquiring, curating, and processing data about weather parameters and agricultural parcels, the infrastructure acquires, curates, and processes knowledge about disease outbreaks. The infrastructure's aim is to inform decision-making.

The environmental knowledge infrastructure is again a sociotechnical system as technical and social subsystem collaborate to attain the infrastructure's aim. Among other devices and systems, the SoilWeather WSN is an important hardware component of the technical subsystem. The technical

subsystem is primarily tasked with data acquisition, curation, and access; it is therefore a data-based system. The social subsystem consists, primarily, of agricultural advisors and farmers. Their primary role is, as in the previous case study, in information acquisition and transfer of information into knowledge, as well as decision-making. Their involvement in lower-level tasks such as data acquisition, curation, and access is minor. While data-based functionality is largely automated—in particular, the continuous acquisition of data via SoilWeather, the management of data by the database, and the access to data by the Web service—knowledge-based functionality is "implemented" by human agents. Most importantly, the technical subsystem of the infrastructure is not involved in knowledge curation. Indeed, knowledge about outbreaks is only implicit in the color-coded spatial features of maps for disease pressure. The technical subsystem does not have explicit representations of knowledge about outbreaks. Knowledge is thus not curated by the technical subsystem.

In the following section, we present software methods and technologies that can advance the data-based technical subsystems of state-of-the-art environmental knowledge infrastructures into knowledge-based systems. As a result, the technical subsystems of future environmental knowledge infrastructures will more actively support information acquisition, transfer of information into knowledge, as well as knowledge curation, access, and processing.

15.3 Approaches

We have presented the environmental knowledge infrastructures of two case studies to highlight how infrastructures acquire information about a monitored environment, transfer information into knowledge, and curate and process knowledge. We argued that the infrastructures are knowledge-based sociotechnical systems because social and technical subsystems collaborate to further human knowledge and understanding about the environment. We highlighted that in state-of-the-art environmental knowledge infrastructures it is because of social subsystems that the infrastructures are knowledge-based systems. Technical subsystems are prevalently data-based systems and provide the social subsystems with little support for information acquisition from data, transfer of information into knowledge, and knowledge curation and processing.

Our claim is that the technical subsystems of future environmental knowledge infrastructures will advance from data-based systems to knowledge-based systems. In other words, the technical subsystems will more actively support, and possibly largely automate, the execution of higher-level tasks currently mostly carried out by social subsystems.

As the data volumes acquired and processed by environmental knowledge infrastructures steadily increase—and as infrastructures become more interoperable, thus facilitating data fusion—this advancement is arguably a necessity. Daily visual extraction of situational knowledge about new particle formation events by an aerosol scientist is feasible for a single location. However, the task becomes increasingly expensive as new particle formation is to be identified at more locations. Furthermore, as the identification and characterization of new particle formation events are relatively straightforward and repetitive tasks, it makes good sense to automate them.

In this section, we present how information acquisition, transfer of information into knowledge, and knowledge curation and processing may be implemented by the technical subsystems of environmental knowledge infrastructures. We present relevant software methods and technologies.

We distinguish data, information, and knowledge and do so following the Data–Information–Knowledge model proposed by Aamodt and Nygård (1995). According to the model, data are syntactic entities. The syntactic entities resulting in the process of measurement, for example, sensor data for particle number concentration, are a kind of data relevant here. Data are input to an interpretation process. Information, according to the model, is interpreted data, that is, data with meaning or *semantic entities*, and is the output of an interpretation process. The semantic entities resulting in the process of information acquisition, for example, identified atmospheric new particle formation, are a kind of information relevant here. Finally, knowledge is learned information, that is, information incorporated into an existing body of knowledge. A situational knowledge object that integrates information about a new particle formation event (situation) is a kind of knowledge relevant here. Knowledge is itself a semantic entity, one that relates semantic entities.

To execute information acquisition, the technical subsystems of environmental knowledge infrastructures require one or more software agents designed to extract information from data. Technical subsystems must be able to control and execute the agents. Agents implement an interpretation process. Data are input to agents and information is the output.

Software agent design follows a model. Two model types are of particular interest: data driven and physically based. Data-driven models, also known as empirical models, may be *supervised*, that is, trained by labeled examples. Preconceived knowledge about the modeled phenomenon does not influence model development. In contrast, physically based models, also known as mechanistic models, are developed to include some degree of understanding about the processes underlying the modeled phenomenon. Mulligan and Wainwright (2004) provide an overview of (environmental) models and modeling and discuss in more depth the characteristics of various model types.

For the aerosol science case study, the technical subsystem of the environmental knowledge infrastructure could employ a supervised data-driven

software agent trained with labeled examples for processed daily particle number concentration data and corresponding information for whether or not new particle formation occurred during the day. Given a data-driven agent trained with such labeled examples, the technical subsystem can then automate the classification of new input data to output information for identified new particle formation. In other words, equipped with a trained data-driven agent, the technical subsystem of the environmental knowledge infrastructure can automate the task otherwise carried out visually by a scientist. The degree of confidence in the accuracy of automated classification can be estimated empirically and can hint at how well the technical subsystem will perform—or how carefully the automated assessment ought to be curated by the scientist.

For the case study in agriculture, the technical subsystem of the environmental knowledge infrastructure could employ a software agent that implements the presented mechanistic model. The agent uses preconceived expert knowledge about plant disease infection development, current observation data for weather parameters, and seasonal data about the crop, pathogen, and agricultural parcel in an equation that estimates daily disease pressure. The technical subsystem can execute the agent to generate information about disease outbreaks, that is, situations in which disease pressure exceeds defined thresholds.

In both environmental knowledge infrastructures, data-driven and physically based agents have the same purpose: to automate data interpretation. Software agents enable the technical subsystems to automate information extraction in environmental knowledge infrastructures. The automated assessment by technical subsystems is accurate to a certain degree. Extracted information thus needs to be curated by the social subsystem. Of particular interest is quality control of extracted information.

A description for what an environmental knowledge infrastructure observes, for example, a description for an event of new particle formation, generally involves different kinds of information. What infrastructures observe is located in space and time. Descriptions thus involve information for temporal and spatial locations, for example, timestamp, latitude, and longitude. A symbolic identifier for the observed phenomenon, for example, a character string for an instance of new particle formation, is information that enables reference to the observed phenomenon in a description. Descriptions also characterize the observed phenomenon, for example, describe the class of new particle formation and the duration of the event. Characterization results in additional information.

Descriptions for what an environmental knowledge infrastructure observes are thus structures that relate information. We call such structures knowledge objects. Knowledge objects integrate information in a body of knowledge. A particular type of knowledge object is the *situational* knowledge object. It is a description of a situation. Other knowledge object types, for example, for descriptions of processes, are of interest as well.

The integration of information in a knowledge object follows a pattern. Technical subsystems may implement such patterns to automate the integration of information in knowledge objects. Situation theory (Devlin, 1991) provides a pattern for integrating information about a situation in a situational knowledge object.

In addition to models for information extraction and patterns for information integration into knowledge objects, technical subsystems of environmental knowledge infrastructures also require a framework for the representation of knowledge objects. The framework provided by the Semantic Web (Berners-Lee et al., 2001) and its technologies is one approach to equip technical subsystems of environmental knowledge infrastructures with functionality for knowledge representation.

The Web Ontology Language (W3C OWL Working Group, 2012) is a core technology of the Semantic Web. In information science, ontology is classically defined by Gruber (1993) as "an explicit specification of a conceptualization." Guarino et al. (2009) provide a succinct analysis of Gruber's definition. Some authors have extended Gruber's definition (Borst, 1997; Studer et al., 1998) while others have provided alternatives (Neches et al., 1991; Swartout et al., 1996; Hendler, 2001). For the purpose here, an ontology is a document that specifies the concepts and relations of some domain so that the semantics of specified terms are interpretable by both software and human agents. To specify concepts and relations we need a language with formal semantics. Today, the Web Ontology Language (OWL) is arguably the *de facto* standard ontology language. It is also the language adopted here for knowledge representation in environmental knowledge infrastructures.

OWL language constructs support the formal specification of the semantics of concepts and relations as class axioms and property axioms, respectively. For instance, the language enables us to state that `C` and `D` are classes and that they are equivalent or that `Q` is an inverse property of the property `P`. The language also supports the specification of individuals. Concept assertions and role assertions specify the class membership and property values of individuals, respectively. For instance, we can state that `a` and `b` are individuals. The concept assertion `C(a)` states that the individual `a` is a member (instance) of the class `C`; `D(b)` thus states that `b` is a member of `D`. The role assertion `P(a,b)` states that the individuals `a` and `b` are related by property `P`. Given that `Q` is inverse of `P` it holds that `Q(b,a)`.

Ontologies are a means for the social subsystems of environmental knowledge infrastructures to convey the semantics of relevant concepts and relations to technical subsystems. As a result, the two subsystems share term semantics. For software agents that implement the language, `C(a)` is not merely a string. If the social subsystem states that `C` is a subclass of `B` (formally $\mathtt{C} \sqsubseteq \mathtt{B}$), then software agents automatically conclude `B(a)`. Ontologies are also a means for the technical subsystems of environmental knowledge infrastructures to convey social subsystems information

objects automatically acquired in data interpretation and knowledge objects resulting from automated information integration. Ontologies are thus a key component for the representation of knowledge communicated between the technical and social subsystems of environmental knowledge infrastructures.

Information objects extracted from data by technical subsystems are semantic entities and, specifically, entities of an OWL ontology. Given the class axiom `NewParticleFormation` $\sqsubseteq$ `AtmosphericPhenomenon`, the technical subsystem of our environmental knowledge infrastructure represents the symbolic identifier `f` for observed new particle formation as the class assertion `NewParticleFormation(f)`. The class assertion is an information object and a semantic entity of an OWL ontology. Information objects for locations in time and space are represented similarly as members of a class and with property values. For the representation of time and space, there exist ontologies that both the technical and the social subsystems can adopt. A candidate ontology for time is OWL-Time (Hobbs and Pan, 2006), which, among other terms, provides definitions for `Instant` and `Interval`. For space, an infrastructure may adopt GeoSPARQL (Perry and Herring, 2012), which provides definitions for spatial `Feature` and `Geometry`.

Knowledge objects resulting from information objects automatically integrated by technical subsystems following determined integration patterns are also ontological semantic entities. For the particular case of situational knowledge objects, describing situations observed by an environmental knowledge infrastructure, we may adopt the Situation Theory Ontology (STO) (Kokar et al., 2009). An event `s` for `NewParticleFormation(f)` may thus be represented as the assertion `Situation(s)`. The definition of the class `Situation` in the STO follows the Situation Theory developed by Barwise and Perry (1981) and Devlin (1991), which specifies how information about an observed situation is integrated in a situational knowledge object.

In the Semantic Web, the Resource Description Framework (RDF) (Cyganiak et al., 2014) is the data model utilized to encode the axioms and assertions of OWL ontologies. Originally conceived as a model of data about Web resources (Lassila and Swick, 1999), RDF can be utilized as a model of data about any resource, including physical objects, abstract concepts, or any entity that can be named by a Uniform Resource Identifier (URI) (Berners-Lee et al., 2005). The RDF *statement* is at the core of the framework and is a triple consisting of a resource, a property, and the value for the property of the resource. These are the subject, the predicate, and the object of the statement, respectively. According to the framework, the concept assertion `C(a)` is encoded as the triple `<a,type,C>` and the role assertion `R(a,c)` as the triple `<a,R,c>`, whereby `a`, `type`, `C`, and `R` are URIs and `c` may be a URI or a literal value, such as a string.

The subjects, predicates, and objects of two or more statements can share the same URI. Such statements *join* over the shared identifier. Common joins are subject–subject and object–subject. The former are statements about the

same subject resource. It can be easily seen that a set of statements form a directed labeled graph of subject and object nodes related by predicates, which act as vertexes directed from the subject to the object. Adding a new statement for an existing resource is akin to expanding the graph with an additional vertex. Two resources can be related by simply adding a new vertex to the graph. The flexibility of the graph data structure is arguably one of the interesting features of RDF.

Figure 15.1 is an example situational knowledge object for a new particle formation event, represented using the discussed Semantic Web technologies. Clearly visible is the directed labeled graph structure of the knowledge object. Object semantics are interpretable by the technical subsystem of environmental knowledge infrastructures. To facilitate machine readability of knowledge objects, infrastructures can adopt one of several syntaxes for RDF, such as RDF/XML (Gandon and Schreiber, 2014).

There exist several RDF database systems, which generally also implement the SPARQL Protocol and RDF Query Language (SPARQL; Harris and Seaborne, 2013) to support querying, updating, or deleting RDF statements managed by the database. For popular programming languages, libraries are available to support reading, processing, and writing RDF data (e.g., Beckett, 2002; Broekstra et al., 2002; Carroll et al., 2003). Libraries designed for programmatic interaction with OWL ontologies, specifically, are available for some programming languages (e.g., Horridge and Bechhofer, 2009). Some software packages for statistical computing also support loading RDF data.

SPARQL supports formulating queries with complex graph patterns. The language thus enables us to formulate queries for knowledge objects meeting certain criteria. For instance, an agent may interrogate an RDF database for events (situations) of strong new particle formation that occurred in 2015 in a particular region of Finland, with perimeter determined by the coordinates of a polygon geometry. As the subsystems commit to a shared ontology, agents of two or more subsystems understand how information about situations is represented; that "strong" is a category of the classification by Hamed et al. (2007); that "2015" is a date–time interval; and that we are constraining our search for situations that occurred within the given polygon geometry. In other words, agents involved in knowledge acquisition represent acquired knowledge objects according to the same ontology used by agents that access knowledge objects. The semantics of relevant terms are specified externally to agents.

Together, these technologies enable environmental knowledge infrastructures with technical subsystems that (1) acquire information from data using data-driven or physically based models, (2) integrate information into knowledge objects according to patterns, and (3) formally and explicitly represent knowledge objects. The technologies thus facilitate the curation, access, and processing of knowledge objects in technical subsystems of environmental knowledge infrastructures. Such technical subsystems are knowledge-based systems.

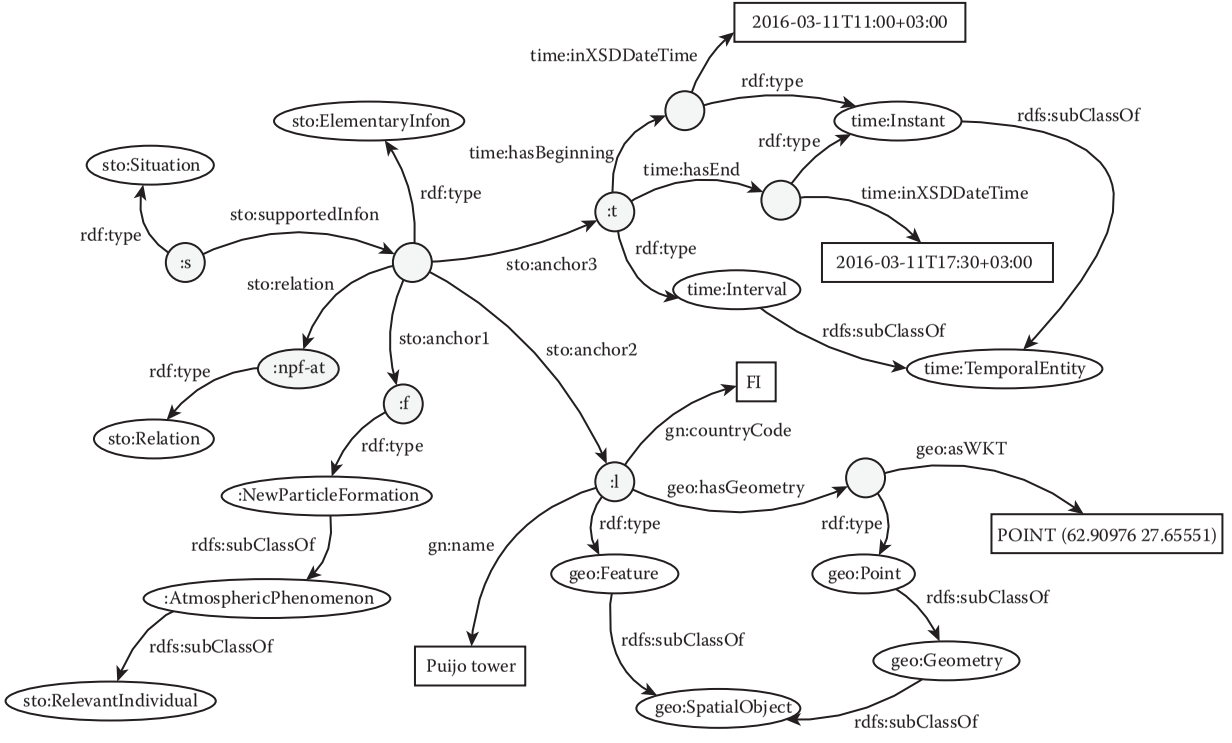

FIGURE 15.1
Example situational knowledge object `:s` represented in RDF for an event of new particle formation `:f` at spatial location `:l` and temporal location `:t`. In Situation Theory, a situation supports one or more elementary infons, that is, information about the situation, which are tuples consisting of a relation (`:npf-at`), a set of anchored objects, and a polarity (omitted here). The prefixes of terms, for example, `sto:`, reflect the ontologies.

15.4 Challenges

Earth and environmental science research communities have recently started to systematically study environmental research infrastructures. Based on an analysis of six infrastructures of the European Strategy Forum on Research Infrastructures (ESFRI)—that is, ICOS, EURO-Argo, EISCAT-3D, LifeWatch, EPOS, and EMSO*—Chen et al. (2013b) present a reference model for environmental research infrastructures. The reference model, called ENVRI-RM, structures identified and shared functionality of environmental research infrastructures into subsystems—namely, data acquisition, data curation, data access, data processing, and community support—and captures the requirements of the "archetypical" environmental research infrastructure from three viewpoints: science, information, and computational. The science viewpoint describes the requirements for "the people who perform their tasks and achieve their goals as mediated by the infrastructure" (Chen et al., 2013a). The information viewpoint describes the requirements for information handled by the infrastructure. The computational viewpoint describes the requirements for expected computational objects and the interfaces by which they interact.

The ENVRI-RM makes evident that state-of-the-art environmental research infrastructures are data-based systems and are thus modeled as such. Indeed, following the ENVRI-RM, technical subsystems are expected to support data acquisition, in particular with sensors; data curation, in databases and on storage systems; data access, for example, via Web portals; and data processing. This is despite that, beyond data, information and knowledge are arguably more important products of environmental research infrastructures. While data analysis and data mining are functionality of the ENVRI-RM data processing subsystem, it is unclear how the reference model and, thus, state-of-the-art environmental research infrastructures account for information and knowledge resulting from data analysis and mining. Accounting for knowledge objects, and their life cycle in environmental knowledge infrastructures, is important because the formal and explicit

*The Integrated Carbon Observation System (ICOS) aims at quantifying and understanding the greenhouse gas balance of Europe and neighboring regions (https://www.icos-ri.eu/); EURO-Argo is the European contribution to Argo, a global ocean observing system (http://www.euro-argo.eu/); EISCAT-3D primarily aims at investigating how the Earth's atmosphere is coupled to space (https://eiscat3d.se); LifeWatch aims at biodiversity and ecosystem research (http://www.lifewatch.eu/); the European Plate Observing System (EPOS) aims at developing a more holistic understanding of the processes underlying Earth's dynamics (https://www.eposip.org/); and the European Multidisciplinary Seafloor and water-column Observatory (EMSO) aims at long-term, high-resolution, (near) real-time monitoring of environmental processes including natural hazards, climate change, and marine ecosystems (http://www.emso-eu.org/).

representation of such objects enables automation in knowledge curation, access, and processing. Moreover, knowledge representation rests on methods and technologies different from those widely utilized for the representation of data objects, for example, tabular or relational data structures.

One of the expectations for the ENVRI-RM is to mature and serve as a blueprint for the implementation of environmental research infrastructures. The aim is harmonization and interoperability. The commitment by infrastructures toward the reference model is arguably a key requirement for its success. However, gaining such commitment is challenging. As experience shows, alone the commitment to a particular schema and format for data management and exchange is challenging to achieve (see, for instance, the brief account by Edwards et al. (2011) on the adoption of the Ecological Metadata Language in the Long-Term Ecological Research program).

Proposals for extending the ENVRI-RM with functionality for information acquisition, transfer of information into knowledge, knowledge representation, curation, access, and processing have recently been suggested in the literature (e.g., Stocker et al., 2015b). However, advancing the reference model and implementations to include such functionality presents major challenges. First, at the community level, a shift is needed toward understanding, modeling, and implementing environmental infrastructures as knowledge-based systems. Clearly, data are only an intermediary product in ICOS, EMSO, and other infrastructures—including those that serve purposes other than scientific research (e.g., Stocker et al., 2014b; Parmiggiani and Monteiro, 2016). For instance, ICOS acquires and processes gas flux observation data but is interested in information and knowledge about strong and weak carbon sinks and sources, for example, forests and cities. Using real-time data processing, EMSO is interested in early warning of tsunami (Best et al., 2014), a situated real-world phenomenon about which the technical subsystem of an early warning system ought to provide near real-time integrated information. As knowledge-based systems require first consolidated architectures and implementations for the lower-level data layers, it is expected that the advancement toward environmental knowledge infrastructures with knowledge-based technical subsystems will require time, significant resources, and commitment from interdisciplinary teams involving at least earth and environmental scientists and computer and information engineers and specialists.

Second, at the technical level, introducing new methods and technologies for knowledge-based systems further complicates already complex infrastructures. The successful design, implementation, and testing of software agents that implement data-driven or physically based models for information extraction are mostly a nontrivial task. The deployment of validated software agents into an environmental knowledge infrastructure typically comes with further technical challenges, such as near real-time execution of the agent, interface requirements, and performance issues. Ontology engineering and the implementation of knowledge-based systems with Semantic

Web technologies also require a specialized set of skills. For the case of research infrastructures, scientists familiar with these methods and technologies are arguably few.

Conversely, engineers capable of addressing technical challenges typically lack the science understanding. Such understanding is, however, required for information acquisition model development, ontology development, as well as for knowledge-based system application development, for example, applications for knowledge processing. Given an arbitrary environmental knowledge infrastructure, an important question engineers will typically have is what kind of information and knowledge are of interest to the infrastructure. The effort of building and maintaining environmental knowledge infrastructures is thus inherently an interdisciplinary endeavor where human agents in scientist and engineer roles need to collaborate. Unfortunately, interdisciplinary collaboration in science is plagued by what Edwards et al. (2011) call "science friction," "the difficulties encountered when two scientific disciplines working on related problems try to interoperate." Science friction "resists and impedes" and poses significant challenges to development.

Currently, a practical challenge is to build compelling case studies that demonstrate the architecture, implementation, and capabilities of environmental knowledge infrastructures with knowledge-based technical subsystems, in particular research infrastructures. Such case studies will highlight the significant opportunities in environmental knowledge infrastructures with advanced knowledge-based technical subsystems. We discuss some of the opportunities next.

15.5 Opportunities

We set forth the vision of future environmental knowledge infrastructures with knowledge-based technical subsystems that more actively support human agents in information acquisition, transfer of information into knowledge, as well as knowledge curation, access, and processing. We have also argued that, thanks to the employed technologies, technical subsystems may partially automate such tasks.

It is important to underscore that automation is *to some degree* and it is to support human agents in these tasks. The social subsystem of environmental knowledge infrastructure remains critical. Software agents for information acquisition need to be developed, for example, a supervised data-driven agent needs to be trained with labeled examples; acquired knowledge needs to be quality controlled; and knowledge serves toward decision-making. Labeling, quality control, and decision-making are generally performed, or at least supervised, by human agents. Hence, in environmental knowledge infrastructures, social and knowledge-based technical subsystems

collaboratively learn from the wealth of data acquired and curated by the infrastructures. Collaborative learning in environmental knowledge infrastructures using the presented methods and technologies comes with several interesting opportunities.

One of the most interesting aspects of environmental knowledge infrastructures with knowledge-based technical subsystems is that acquired knowledge objects are machine readable and interpretable. Consider our two case studies. Knowledge about new particle formation events is recorded in Excel, and the results of statistical analysis are described in scientific articles, using tables, figures, and natural language text. In our second case study, knowledge about plant disease outbreaks is equally implicit, in the images for regional maps and the color scheme used to inform human agents about the acuteness of outbreaks in the region. Knowledge encoded in these forms is hardly machine processable. It is implicit in higher-level data products. Presented with an intuitively designed map, a human agent can effortlessly extract information and knowledge conveyed by the image. Unfortunately, the same cannot be said for technical agents. As a consequence of encoding information and knowledge implicitly in higher-level data products, human agents need to manually extract information from articles, for example, to perform a meta-analysis. Another practice is to attempt to algorithmically extract the characteristics of spatial features by processing image pixels (e.g., Epitropou et al., 2015; Stocker et al., 2015a).

As noted earlier, to delegate information acquisition and the transfer of information into knowledge to the technical subsystems of environmental knowledge infrastructures is particularly useful when such processes are well defined and repeated. The more often they are repeated in space–time, the greater is arguably the benefit of automation as it frees human agents from carrying out the processes manually. Automation can also eliminate subjective bias by individual human agents in manual assessment.

The distinction between data, information, and knowledge is not clear-cut. The knowledge objects curated by an environmental knowledge infrastructure may arguably be data to other systems. More concretely, knowledge objects generally relate data of primitive types, for example, numbers, which an agent may want to access and process further. Hence, data related in knowledge objects may become elements of a dataset in another system. However, from a technical perspective, the distinction can be more obviously made based on data structure. Data in an environmental knowledge infrastructure are often structured as dataset, where the columns most commonly represent variables and rows are observations with values for the variables. In contrast, graphs are more suited to structure information in knowledge objects as graphs integrate by linking objects. Any node in a graph can be flexibly expanded with further vertexes to nodes. If new information is available for a particular knowledge object, it can be integrated by accordingly expanding the corresponding graph. If the class of new particle formation has not been assessed for a particular event, the graph corresponding to

the knowledge object simply has fewer vertexes compared to the one corresponding to a new particle formation event for which the class was assessed. Knowledge objects, in particular also those of the same type, can thus have varying types and counts of relations to objects without resulting into structures filled with `Null` values. New vertexes between existing nodes can also be added or removed flexibly. Hence, when an environmental knowledge infrastructure uncovers a new relation, the corresponding objects are simply linked by a new relationship.

RDF is a suitable data model to represent graphs and is thus arguably an interesting framework for the representation of knowledge objects curated by knowledge-based technical subsystems of environmental knowledge infrastructures. RDF addresses the syntactic interoperability of knowledge objects, while OWL addresses their semantic interoperability by formally restricting the meaning of terms to the one intended. Being a Web technology, RDF has further interesting aspects. Curated knowledge objects, as well as the information objects they integrate, are referred to by URI. They are thus globally identifiable and can be linked across distributed environmental knowledge infrastructures. Another potentially interesting aspect is the association of URIs with persistent identifiers (Hakala, 2010), such as Digital Object Identifiers, to enable location-independent reference to knowledge objects. Doing so could facilitate the citation of knowledge about environmental phenomena, for example, a hurricane, described by environmental research infrastructures. The human agents of the social subsystem responsible for the acquisition and curation of cited knowledge could be credited for their work.

Curated knowledge objects can be processed in various ways. An important type of processing is visualization. Environmental phenomena observed by environmental knowledge infrastructures are generally located in space–time. Knowledge about observed phenomena can thus be visualized along these two dimensions. For instance, situational knowledge for disease outbreaks in agriculture can be visualized for outbreak development over time and space. Given the commitment of situational knowledge objects to ontologies and underlying theories, such as Situation Theory, consumer applications that visualize situational knowledge in space–time can trivially extract spatial and temporal information from situational knowledge objects and utilize the information in processing for visualization. As a result, the environmental knowledge infrastructure visualizes knowledge for situations of disease outbreaks rather than data underlying situational knowledge acquisition, such as data for current weather, from which human agents have to draw knowledge about outbreaks manually. Furthermore, as knowledge is explicitly represented by knowledge-based technical subsystems, environmental knowledge infrastructures avoid having information and knowledge only implicit in higher-level data products. Explicitly represented knowledge can be reused for purposes other than those originally intended.

Knowledge processing forms other than visualization are also of interest. Consumers can fuse knowledge independently acquired and curated by two or more environmental knowledge infrastructures (e.g., Stocker et al., 2015a). The retrieval of knowledge from multiple infrastructures practically amounts to executing a federated SPARQL query (Prud'hommeaux and Buil-Aranda, 2013) over the distributed SPARQL services provided by the infrastructures. Automated reasoning, including rule-based reasoning, is a further possibility in knowledge processing. Interesting to evaluate is also the potential of curated knowledge for the empirical parameterization of high-level models, such as agent-based or Bayesian models.

15.6 Conclusion

For environmental infrastructures, in particular research infrastructures, we have highlighted that, beyond the acquisition and processing of data, they generate information and transfer information into knowledge. We thus argued that environmental (research) infrastructures are environmental (research) *knowledge* infrastructures. The emphasis on knowledge is important as it underscores that data are merely intermediate products in environmental infrastructures and that state-of-the-art architectural models, such as the ENVRI-RM, may want to reflect this aspect in order to represent environmental (research) infrastructures more holistically.

Discussing the environmental knowledge infrastructures of two case studies, we highlighted how the infrastructures are sociotechnical systems, that is, systems composed of technical and social subsystems. We highlighted how the technical subsystems are predominantly involved in lower-level functionality of the infrastructure, in particular data acquisition and curation, and how technical subsystems often to a great extent automate such functionality, for example, by means of environmental sensor networks and database systems. In contrast, the social subsystems are more actively involved in higher-level functionality of the infrastructure, that is, information acquisition, transfer of information into knowledge, and knowledge curation and processing. Functionality executed by social subsystems is not as automated as functionality executed by technical subsystems.

Of particular interest to well-defined information acquisition processes that are executed frequently over space–time, we discussed methods and technologies that could advance the technical subsystems of state-of-the-art environmental (research) infrastructures from data-based to knowledge-based systems. We argued that knowledge-based technical subsystems can better support the social subsystems of infrastructures in information acquisition, transfer of information into knowledge, and knowledge curation

and processing. In some cases, knowledge-based technical subsystems may largely automate such functionality.

The most important aspect of knowledge-based technical subsystems is their ability to represent acquired knowledge objects formally and explicitly. Knowledge about the environment observed by infrastructures is hence accessible and interpretable not just for social subsystems but for technical subsystems as well. This is in stark contrast to how knowledge is curated, accessed, and shared via higher-level data products such as digital maps, figures, tables, or natural language text typically generated by state-of-the-art environmental infrastructures. Knowledge conveyed via such higher-level data products is implicit and hardly accessible for technical subsystems.

We discussed some of the challenges and opportunities that lay on the path toward environmental (research) infrastructures with knowledge-based technical subsystems. An important challenge is the additional complexity of knowledge-based methods and technologies, added to already complex infrastructures. One of the most interesting opportunities for the research community may be the possibility of associating persistent identifiers to knowledge descriptions for discovered environmental phenomena—thus making such knowledge objects citable.

Acknowledgment

This work was supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182.

References

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.

Barwise, J. and Perry, J. (1980). The situation underground. In Barwise, J. and Sag, I., editors, *Stanford Working Papers in Semantics*, Vol. 1, pp. 1–55. Stanford Cognitive Science Group, Palo Alto, CA.

Barwise, J. and Perry, J. (1981). Situations and attitudes. *The Journal of Philosophy*, 78(11):668–691.

Beckett, D. (2002). The design and implementation of the Redland RDF application framework. *Computer Networks*, 39(5):577–588.

Berners-Lee, T., Fielding, R., and Masinter, L. (2005). Uniform Resource Identifier (URI): Generic syntax. Request for Comments 3986, IETF.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):29–37.

Best, M., Beranzoli, L., Chierici, F., Delaney, J. R., Embriaco, D., Galbraith, N., Huber, R., Orcutt, J. A., and Weller, R. A. (2014). CoopEUS EMSO-OOI case study: Tsunami modelling and early warning systems for near source areas (Mediterranean, Juan de Fuca). In *AGU Fall Meeting Abstracts*. San Francisco, CA. https://agu.confex.com/agu/fm14/webprogram/Paper18074.html.

Borst, W. N. (1997). Construction of engineering ontologies for knowledge sharing and reuse. PhD thesis, Centre for Telematics and Information Technology, University of Twente, Enschede, the Netherlands.

Broekstra, J., Kampman, A., and van Harmelen, F. (2002). Sesame: A generic architecture for storing and querying RDF and RDF schema. In Horrocks, I. and Hendler, J., editors, *The Semantic Web—ISWC 2002, Lecture Notes in Computer Science*, Vol. 2342, pp. 54–68. Springer, Berlin, Germany.

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K. (2003). Jena: Implementing the semantic web recommendations. Technical Report HPL-2003-146, HP Laboratories, Bristol, U.K.

Chen, Y., Hardisty, A., Preece, A., Martin, P., Atkinson, M., Zhao, Z., Magagna, B., Schentz, H., and Legré, Y. (2013a). Analysis of common requirements for environmental science research infrastructures. In *Proceedings of the International Symposium on Grids and Clouds (ISGC)*, Academia Sinica, Taipei, Taiwan. Proceedings of Science (SISSA).

Chen, Y., Martin, P., Magagna, B., Schentz, H., Zhao, Z., Hardisty, A., Preece, A., Atkinson, M., Huber, R., and Legré, Y. (2013b). A common reference model for environmental science research infrastructures. In Page, B., Fleischer, A. G., Göbel, J., and Wohlgemuth, V., editors, *Proceedings of the 27th International Conference on Environmental Informatics for Environmental Protection, Sustainable Development and Risk Management*, pp. 665–673, Hamburg, Germany.

Cyganiak, R., Wood, D., and Lanthaler, M. (2014). RDF 1.1 concepts and abstract syntax. Recommendation, W3C.

Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P., and Lehtinen, K. (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, Cambridge, UK.

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. MIT Press, Cambridge, MA.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., and Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41(5):667–690.

Epitropou, V., Bassoukos, T., Karatzas, K., Karppinen, A., Wanner, L., Vrochidis, S., Kompatsiaris, I., and Kukkonen, J. (2015). Environmental data extraction from heatmaps using the airmerge system. *Multimedia Tools and Applications*, 75(3):1589–1613.

Finkelstein, L. (1982). Theory and philosophy of measurement. In Sydenham, P. H., editor, *Handbook of Measurement Science, Theoretical Fundamentals*, Vol. 1, pp. 1–30. John Wiley & Sons, Hoboken, NJ.

Fountas, S., Wulfsohn, D., Blackmore, B., Jacobsen, H., and Pedersen, S. (2006). A model of decision-making and information flows for information-intensive agriculture. *Agricultural Systems*, 87(2):192–210.

Fox, W. M. (1995). Sociotechnical system principles and guidelines: Past and present. *The Journal of Applied Behavioral Science*, 31(1):91–105.

Gandon, F. and Schreiber, G. (2014). RDF 1.1 XML syntax. Recommendation, W3C.

Ganguly, A. R., Omitaomu, O. A., Fang, Y., Khan, S., and Bhaduri, B. (2007). Knowledge discovery from sensor data for scientific applications. In Gama, J. and Gaber, M. M., editors, *Learning from Data Streams*, pp. 205–229. Springer, Berlin, Germany.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In Staab, S. and Studer, R., editors, *Handbook on Ontologies, International Handbooks on Information Systems*, pp. 1–17. Springer, Berlin, Germany.

Hakala, J. (2010). Persistent identifiers—An overview. Technical report, The National Library of Finland, Helsinki, Finland.

Hamed, A., Joutsensaari, J., Mikkonen, S., Sogacheva, L., Dal Maso, M., Kulmala, M., Cavalli, F. et al. (2007). Nucleation and growth of new particles in Po Valley, Italy. *Atmospheric Chemistry and Physics*, 7(2):355–376.

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.

Hart, J. K. and Martinez, K. (2006). Environmental sensor networks: A revolution in the earth system science? *Earth-Science Reviews*, 78(3–4):177–191.

Hendler, J. (2001). Agents and the semantic web. *Intelligent Systems, IEEE*, 16(2):30–37.

Hobbs, J. R. and Pan, F. (2006). Time ontology in OWL. Working draft, W3C.

Horridge, M. and Bechhofer, S. (2009). The OWL API: A Java API for working with OWL 2 ontologies. In Hoekstra, R. and Patel-Schneider, P.F., editors, *Proceedings of the 6th International Workshop on OWL: Experiences and Directions (OWLED 2009)*, Chantilly, VA, Vol. 529, pp. 11–21. CEUR Workshop Proceedings. ISSN 1613-0073.

Keller, M., Schimel, D. S., Hargrove, W. W., and Hoffman, F. M. (2008). A continental strategy for the National Ecological Observatory Network. *Frontiers in Ecology and the Environment*, 6(5):282–284.

Kokar, M. M., Matheus, C. J., and Baclawski, K. (2009). Ontology-based situation awareness. *Information Fusion*, 10(1):83–98. Special Issue on High-Level Information Fusion and Situation Awareness.

Kotamäki, N., Thessler, S., Koskiaho, J., Hannukkala, A. O., Huitu, H., Huttula, T., Havento, J., and Järvenpää, M. (2009). Wireless in-situ sensor network for agriculture and water monitoring on a river basin scale in Southern Finland: Evaluation from a data user's perspective. *Sensors*, 9(4):2862–2883.

Kratz, T. K., Arzberger, P., Benson, B. J., Chiu, C.-Y., Chiu, K., Ding, L., Fountain, T. et al. (2006). Toward a global lake ecological observatory network. *Publications of the Karelian Institute*, 145:51–63.

Kulkarni, P., Baron, P. A., and Willeke, K. (2011). *Aerosol Measurement: Principles, Techniques, and Applications*. John Wiley & Sons, Hoboken, NJ.

Kulmala, M., Vehkamäki, H., Petäjä, T., Dal Maso, M., Lauri, A., Kerminen, V., Birmili, W., and McMurry, P. (2004). Formation and growth rates of ultrafine atmospheric particles: A review of observations. *Journal of Aerosol Science*, 35(2):143–176.

Lassila, O. and Swick, R. R. (1999). Resource Description Framework (RDF) model and syntax specification. *Recommendation*, W3C.

Leskinen, A., Portin, H., Komppula, M., Miettinen, P., Arola, A., Lihavainen, H., Hatakka, J., Laaksonen, A., and Lehtinen, K. E. J. (2009). Overview of the research activities and results at Puijo semi-urban measurement station. *Boreal Environment Research*, 14(4):576–590.

Meijers, E. (1986). Defining confusions—Confusing definitions. *Environmental Monitoring and Assessment*, 7(2):157–159.

Michener, W. K., Porter, J., Servilla, M., and Vanderbilt, K. (2011). Long term ecological research and information management. *Ecological Informatics*, 6(1):13–24.

Mulligan, M. and Wainwright, J. (2004). Modelling and model building. In Wainwright, J. and Mulligan, M., editors, *Environmental Modelling: Finding Simplicity in Complexity*, pp. 7–73. John Wiley & Sons, Hoboken, NJ.

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., and Swartout, W. R. (1991). Enabling technology for knowledge sharing. *AI Magazine*, 12(3):36–56.

Parmiggiani, E. and Monteiro, E. (2016). A measure of "environmental happiness": Infrastructuring environmental risk in oil and gas off shore operations. *Science & Technology Studies*, 29(1):30–51.

Perry, M. and Herring, J. (2012). OGC GeoSPARQL—A geographic query language for RDF data. *Technical Report OGC 11-052r4*, Open Geospatial Consortium Inc, Wayland, MA.

Peters, D. P. C., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., and Villanueva-Rosales, N. (2014). Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. *Ecosphere*, 5(6):1–15.

Pope III, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., and Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *The Journal of the American Medical Association*, 287(9):1132–1141.

Prud'hommeaux, E. and Buil-Aranda, C. (2013). SPARQL 1.1 federated query. Recommendation, W3C.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M., and Miller, H. L. (2007). *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, U.K.

Stocker, M. (2015). Situation awareness in environmental monitoring. PhD thesis. Publications of the University of Eastern Finland. Dissertations in Forestry and Natural Sciences No 192, University of Eastern Finland. ISBN: 978-952-61-1907-6.

Stocker, M., Baranizadeh, E., Portin, H., Komppula, M., Rönkkö, M., Hamed, A., Virtanen, A., Lehtinen, K., Laaksonen, A., and Kolehmainen, M. (2014a). Representing situational knowledge acquired from sensor data for atmospheric phenomena. *Environmental Modelling & Software*, 58:27–47.

Stocker, M., Kauhanen, O., Hiirsalmi, M., Saarela, J., Rossi, P., Rönkkö, M., Hytönen, H., Kotovirta, V., and Kolehmainen, M. (2015a). A software system for the discovery of situations involving drivers in storms. In Denzer, R., Argent, R. M., Schimak, G., and Hřebíček, J., editors, *Environmental Software Systems. Infrastructures, Services and Applications,* volume 448 of IFIP Advances in Information and Communication Technology, pp. 226–234. Springer International Publishing, Switzerland.

Stocker, M., Nikander, J., Huitu, H., Jalli, M., Koistinen, M., Rönkkö, M., and Kolehmainen, M. (2016). Representing situational knowledge for disease outbreaks in agriculture. *Journal of Agricultural Informatics*, 7(2):29–39.

Stocker, M., Rönkkö, M., and Kolehmainen, M. (2014b). Situational knowledge representation for traffic observed by a pavement vibration sensor network. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1441–1450.

Stocker, M., Rönkkö, M., and Kolehmainen, M. (2015b). Knowledge-based environmental research infrastructure: Moving beyond data. *Earth Science Informatics*, 9(1):47–65.

Studer, R., Benjamins, V., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1–2):161–197.

Swartout, B., Patil, R., Knight, K., and Russ, T. (1996). Toward distributed use of large-scale ontologies. In *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Wowrkshop*, Banff, Alberta, Canada.

Vana, M., Ehn, M., Petäjä, T., Vuollekoski, H., Aalto, P., de Leeuw, G., Ceburnis, D., O'Dowd, C. D., and Kulmala, M. (2008). Characteristic features of air ions at Mace Head on the west coast of Ireland. *Atmospheric Research*, 90(2–4):278–286.

W3C OWL Working Group (2012). OWL 2 web ontology language document overview, 2nd edn. Recommendation, W3C.

Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B. et al. (2012). Mobility particle size spectrometers: Harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions. *Atmospheric Measurement Techniques*, 5(3):657–685.