Jul 11th, 3:10 PM - 3:30 PM

A Scientific Data Management Infrastructure for Environmental Monitoring and Modelling

Daniel Henzen
Technische Universität Dresden, daniel.henzen@tu-dresden.de

Matthias Mueller
Technische Universität Dresden, matthias_mueller@tu-dresden.de

Simon Jirka
52° North Initiative for Geospatial Open Source Software GmbH, jirka@52north.org

Julia Senner
Fraunhofer Institute for Computer Graphics Research IGD, julia.senner@igd.fraunhofer.de

Thomas Kaeseberg
Technische Universität Dresden, thomas.kaeseberg@tu-dresden.de

See next page for additional authors

Presenter/Author Information

Daniel Henzen, Matthias Mueller, Simon Jirka, Julia Senner, Thomas Kaeseberg, Jin Zhang, Lars Bernard, and Peter Krebs

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

A Scientific Data Management Infrastructure for Environmental Monitoring and Modelling

Daniel Henzen[a], Matthias Mueller[a], Simon Jirka[b], Julia Senner[c], Thomas Kaeseberg[d], Jin Zhang[d], Lars Bernard[a], Peter Krebs[d]

[a] *Chair of Geoinformatics, Technische Universität Dresden, 01062 Dresden, Germany*
(daniel.henzen@tu-dresden.de, matthias_mueller@tu-dresden.de, lars.bernard@tu-dresden.de)

[b] *52° North Initiative for Geospatial Open Source Software GmbH, 48155 Muenster, Germany*
(jirka@52north.org)

[c] *Fraunhofer Institute for Computer Graphics Research IGD, 64283 Darmstadt, Germany*
(julia.senner@igd.fraunhofer.de)

[d] *Institute of Urban Water Management, Technische Universität Dresden, 01062 Dresden, Germany*
*(*thomas.kaeseberg@tu-dresden.de*,* jin.zhang@tu-dresden.de, peter.krebs@tu-dresden.de*)*

Abstract: Environmental projects often require the collaboration of researchers from different disciplines or domains in an interoperable context. With regard to data handling, most of these projects have an analogous workflow: phenomena are monitored, observation data are captured, (pre-) processed, exchanged, published, and finally disseminated among other scientists, practitioners and stakeholders. In many cases, each of the project partners implements these workflows separately and the integration of the distributed data sets happen in later project stages. In this paper, we present building blocks for research data infrastructure which covers the complete project cycle and supports data integration right from the beginning. Building upon open source software components, we created a software framework, which covers data capture and storage, semantic enrichment, publication and service-based dissemination thus fulfilling the typical needs and requirements of interdisciplinary research projects. A security layer authorizes access to private data sets thus protecting commercial or otherwise, licensed data that are required by the project partners but not intended for public release and display. Besides standards-based geospatial web services, the framework provides a lightweight RESTful API that suits the needs of web developers and facilitates application development for stakeholder engagement and outreach activities.

***Keywords**:* data dissemination, data management, research data infrastructures, open source software

1 INTRODUCTION

Research infrastructures facilitate environmental research. They provide a set of shared resources to the project teams and its collaborators such as probing equipment, data management solutions or publication archives. In data-driven research (Gray, 2009), scientific data infrastructures have become a vital tool that supports daily data management, as well as publication and outreach activities (Bernard et al., 2014). A major drawback of many existing systems is a technology disconnect between data management and publication. Quite frequently data management and curation components are not very well connected to data publishing and provisioning services (Mason et al., 2014, Jones et al., 2015, Samourkasidis and Athanasiadis, 2014). In fact, they are usually very different systems and the data preparation for publication is often a manual and tedious task.

This paper discusses the shortcomings of existing systems, and proposes an integrated software architecture that supports in-project data management, publication, and application development. Data management in many environmental research projects involves three typical phases. During the first stage, data are acquired, pre-processed and analysed to build, calibrate, and validate environmental models. In a second phase, later in the project, some of the project data are referenced in publications and needed to be published such that other research teams can verify them or use them for their own research. During the third phase, further projects develop applications for monitoring, forecasting, and decision support on this data. Even though the data are roughly the same in all three phases, it usually requires collecting and providing additional metadata for the subsequent publications and applications development.

The most prominent infrastructures for environmental data serve the data publication and dissemination tasks. Researchers can register their data in portals such as PANGAEA (PANGAEA - Data Publisher for Earth & Environmental Science), DataOne (DataONE - Data Observation Network for Earth) or GEOSS (GEOSS - Global Earth Observation System of Systems) which can be searched and queried by external parties. Projects like the European project EUDAT (EUDAT - Research Data Services Expertise & Technology Solutions) try to provide a generic research data publication platform to foster interdisciplinary data exchange. In addition, major publishers promote data publication offers along with research papers. All these platforms are widely recognised and used by the community. However, most of them support only the publication phase and assume that data is provided in a well-described common format so that other parties can readily use it.

Exceptions are data providing organisations such as remote sensing satellite operators, climate data centres, or meteorological offices that operate their own data centres and have a formalised publication and update process. Their services are either updated perpetually, or release periodically updated data products. However, a lot of valuable data is gathered in smaller projects with a stronger focus on empirical studies rather than data management, curation, and harmonisation. Such projects require a technology stack that facilitates their in-project data management first and provides extended support for data publication and application development. By using open standards, these data products may be subsequently integrated into broader networks such as research data infrastructures.

In the next section we will briefly introduce a case study on urban hydrology that illustrates typical data integration, management, and publication tasks in an interdisciplinary research project at a smaller scale. Subsequently, an abstract architecture is proposed that describes the required information system components and the required interfaces for interoperable data exchange with external data platforms and applications and introduces a software stack that integrates several open-source products to provide the required functionality. The paper closes with a set of concluding remarks and discusses plans for future developments.

2 COLABIS: A CASE STUDY IN URBAN HYDROLOGY

The project COLABIS (Collaborative Early Warning Information Systems for Urban Infrastructures) investigates local rainfall events, pollutant accumulation during dry weather periods and cascading events affecting urban water and sewage infrastructures. In the event of heavier rains, the capacities of the urban wastewater treatment plants of combined sewer networks - sewage and storm water discharged in the same system/pipe - are often insufficient to treat the total storm water influx. Here, to protect the treatment efficiency of the plant, water will be directly or indirectly (water storage volumes) discharged – diluted, but untreated – into the receiving waters (Figure 1). It is hypothesised that an improved understanding and modelling of pollutant deposition, precipitation, and surface runoff enables short-term projections on the expected wastewater influx. Operators of the sewer system and water treatment facilities could use this information for more efficient control of the sewage system and act anticipatorily in critical situations.

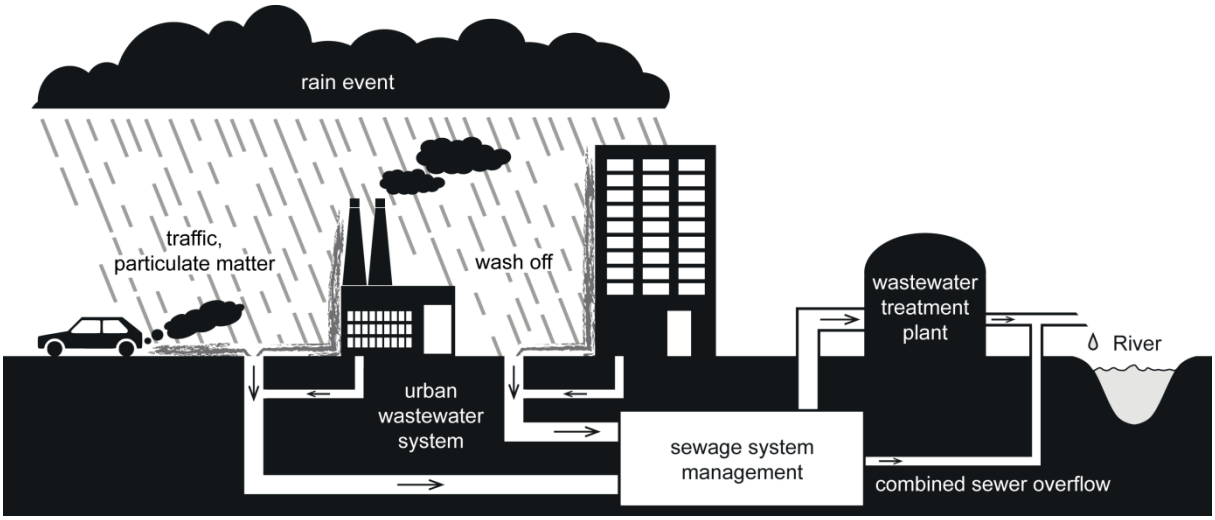

Figure 1. COLABIS case study overview.

To meet these requirements, a variety of data from different sources, such as in-situ sensors, crowdsourcing, environmental simulations, administrative and historical data, has to be collected and fused to enhance information. The spatio-temporal resolution of pollutants in an urban watershed (e.g. road surface in urban areas) differs by source (vehicular traffic, atmospheric deposition), and the knowledge about the extent and dynamics of the pollutant wash off (Zhang et al., 2016). To estimate retention and wash off effects of pollutant in a watershed, the influences of physical and chemical properties of that watershed (e.g. pavement material, slope, texture depth, surface coating of road surface) on the pollutants build-up, wash off and transportation are key factors to be investigated. Such data could be further used for storm water discharge modelling.

Traffic causes deposition of pollutants and other particulate matters on the road surface which will be eventually discharged into receiving waters. Some examples are copper from auto brake pad erosion, zinc from tyre debris (Zhang et al., 2015b) and polycyclic aromatic hydrocarbons from auto engine emission (Zhang et al., 2015c). Therefore, the aquatic environmental source control and storm water pollution mitigation of pollutants should be designed based on the in-depth understanding of spatial distribution and temporal accumulation of pollutants on urban impervious surfaces (Zhang et al., 2015a). Furthermore, sewage changes its composition throughout the day almost systematically and adds to the pollution of the combined wastewater. In addition to organic input and nutrient matter, sewage contains increasing amounts of anthropogenic micro pollutants (Marx et al., 2015).

There are three main objectives in this project:

1) An improved understanding and quantitative modelling of pollutant deposition, mobilizations and transportation in the urban water cycle,
2) The provision of information products to stakeholder in the municipal administration and water treatment companies, and
3) A long term enhancement of the data pool for hydrological modelling and simulation.

To accomplish these goals, a variety of data sources must be integrated (Figure 2). Maps of land use, built environment and cadastral data are provided by local authorities as well as community projects such as *OpenStreetMap* (http://www.openstreetmap.org). A synergetic use of administrative and crowd sourced data sets is beneficial in terms of completeness and currency in many regions (Haklay, 2010, Girres and Touya, 2010). The different spatio-temporal resolutions of pollutants in an urban watershed or meteorological parameters can be improved and condensed with crowd sourced data (Henzen et al., 2015). However, it ought to be noted, that crowd sourced data itself is mostly of lower quality regarding accuracy, resolution or completeness (Reis et al., 2015, Al-Bakri and Fairbairn, 2010). Nonetheless, the overall data quality can be increased when professionally collected data is fused with crowd sourced data to obtain densified information products with better spatio-temporal resolution. In-situ sensor networks deliver data for meteorological parameters, wastewater discharge, pollution measurements and traffic volume. This data can be assimilated by modelling and simulation tools, which, in turn, generate additional data products that must be stored and maintained.

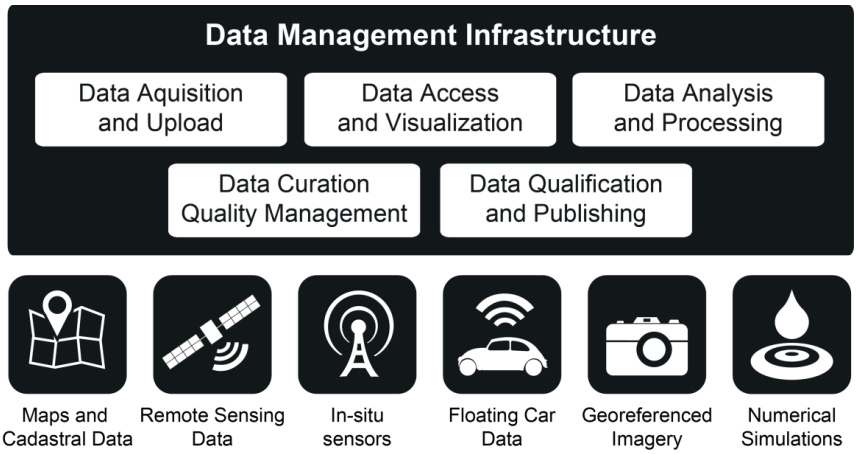

Figure 2. Services for data integration, management and provision in the COLABIS case study.

Prior to publication, data qualification is an important intermediate step in which the data is properly documented, described, validated and encoded. Then, the other users and researchers can read and evaluate the provided data (Hook et al., 2010). Data sets are usually transformed to a commonly accepted data interchange format (ISO, 2005) to facilitate their distribution. There is also an established body of interface standards available from the International Organization for Standardization (ISO), the Open Geospatial Consortium (OGC), the World Wide Web Consortium (W3C) and Internet Engineering Task Force (IETF) for interoperable data access and exchange, which should be considered for public data access. After qualification and publication, higher level services for data visualization and inspection may be offered to support users in conducting simple screening and analysis tasks. Finally, published measurements and survey data are an input to more complex analysis and modelling tasks and may be embedded into larger analysis and statistics workflows that derive composite monitoring and forecasting products. With this in mind, the subsequent section proposes a data management infrastructure based on these established workflows and standards.

3 SOFTWARE ARCHITECTURE AND COMPONENTS

The design of a data management infrastructure can be built on a wealth of open source software packages that can be assembled into a software stack that provides the required functionality. Since most products focus on particular tasks in the data management process, they can be organised in a layered architecture shown in Figure 3.

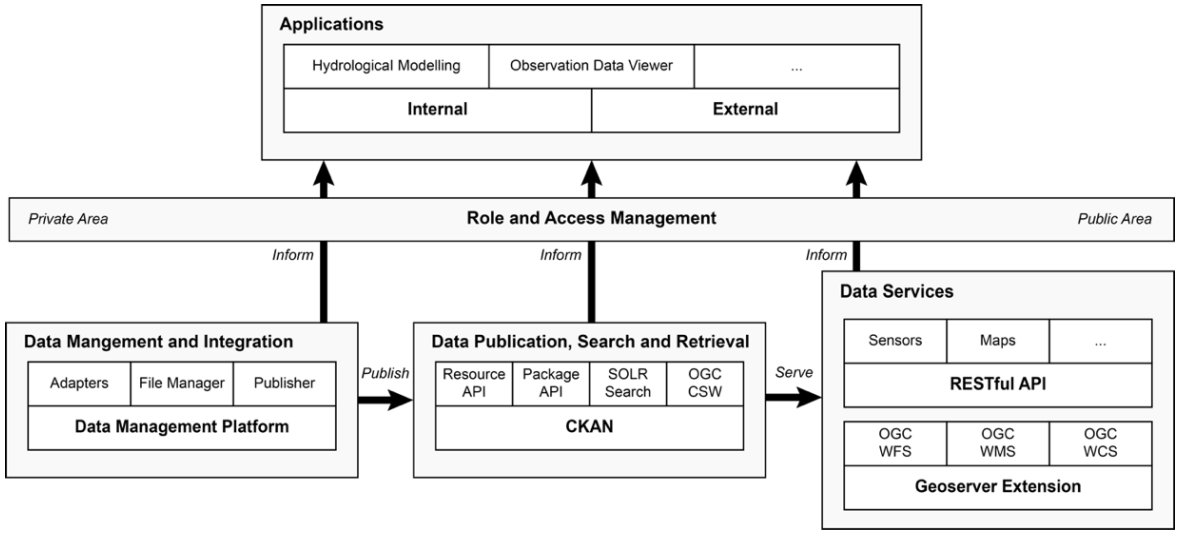

Figure 3. Software stack for the proposed data management infrastructure.

This architecture can easily be mapped to the regular project life cycle phases: initiation, planning, implementation, closure. During the first two phases, the project objective or need is identified, and a solution is further developed. Here, the *data management platform* (DMP), as the central part and initial component in the proposed workflow, can play a significant role as a data collection system for researchers and therefore as data basis for realisation ideas and as search portal of previous research and experiences.

The DMP is a scalable, multi-purpose file manager to collect, manage, and share project data internally. This platform is meant to be used by the researchers as their day-to-day-work data management system, thus being easy to use, not providing barriers and guaranteeing that all project research data gets stored in a common environment. It assists researchers and technical staff of various fields of research in building and managing a consolidated project data base and provides simple tools for semantic enrichment. Major intermediate datasets and research results, can be qualified and then published with a single click for public access. Methods to support the full qualification process are still under development and currently limited to format and field definitions as proposed by Hook et al. (2010). For flexibility, the DMP implements a generic tagging system that is flexible enough to incorporate future metadata fields. Furthermore, data-adapters are used to periodically harvest and integrate external data sources, such as atmospheric data from meteorological offices and data distributors. In brief, the DMP facilitates the following tasks:

- Upload, manage and share existing data,
- Integrate external data sources via adapters,
- Enrich and update the data, while keeping track of the changes,
- Apply a semi-automatic qualification process to semantically enrich data, and
- Publish qualified data to the CKAN catalogue component.

The DMP can consume various kinds of data: (a) semi-structured raw data from experiments and observation campaigns without semantics, (b) semi-structured and semantically enriched data formats like CSV, and (c) standard-based data products following a fully qualified metadata description. This fact eases the initial start into data management infrastructures and the follow-up publishing process.

All published data is stored in the CKAN component (Comprehensive Knowledge Archive Network; http://ckan.org/), a software solution for open data management and dissemination. CKAN has a plugin ecosystem that provides modules for spatial indexing, data preview, and service interfaces such as the OpenGIS Catalogue Service (CSW; OGC, 2007). Content indexing and search in CKAN are powered by Apache SOLR search engine. CKAN also has a powerful stable web API for convenient integration with other applications and services such as DMP and the data services (cf. Figure 3). The CKAN web application allows users to find collections of scientific data quickly and easily, irrespective of their origin, discipline, or community. Using standardizes facets allows allows quick overviews and easy browsing of available data. It is aimed at researchers and practitioners who intend to find useful data resources, which they can use for their research purposes. Hence, it largely supports the third goal in the COLABIS case study (long term enhancement of a data pool for hydrological modelling and simulation).

To summarize, CKAN facilitates the following tasks in the data management infrastructure:

- Publication and indexing of published data
- Faceted, keyword based search
- Support for geographic and temporal coordinate references
- Data preview as interactive charts, tables and maps, and
- Connection and integration with other components through a HTTP API.

At the next level, the data management infrastructure provides a set of higher level services for data inspection and visualisation. While the catalogue component is primarily focused on data search and retrieval, data service permits higher level operations on the data sets such as access of individual values within the whole data set, filtering of the data or image rendering. These services can serve appropriate pieces of data to thin clients or deliver visualised data for immediate display.

Since almost all data from the case study (cf. section 2) has a spatial reference and follows common geographical data models, it can be provided via OGC web services, such as Web Map Services (WMS), Web Feature Services (WFS), Web Coverage Services (WCS) and Sensor Observation

Services (SOS). These interfaces are defined in open international standards and hence frequently applied to establish interoperability in web based spatial data infrastructures(Bernard et al., 2005). Since CKAN's software ecosystem already provides connectors to the Geoserver project (http://geoserver.org/), an OGC service layer can be generated for many spatial datasets.

During the last years, RESTful service interfaces (Fielding, 2000) have become increasingly popular for application development. Hence, the data management infrastructure provides a RESTful API in addition to standardised owsOGC service interfaces. Similarly to the Geoserver connector, we are using a CKAN connector that pipes sensor data into a RESTful Sensor Web API (52° North Initiative for Geospatial Open Source Software GmbH, 2014). It gives JavaScript applications effortless access to these data sources so that they do not have to deal with these sometimes complex interfaces of SOS or WFS.

To conclude, the service layer shall support the following:

- Interoperability layer for data access and visualization (OGC Services)
- Easy-to-use API layer for stakeholder and decision support applications

At the top level of the software stack are for instance applications, which use the data access layer and create visual representations of the data. Especially in decision support scenarios like the proposed use case the visualisation of data and in particular interactive graphics successfully supports the involvement of stakeholders in a problem-solving process by offering them web-based tools for an individual analysis and reasoning (Andrienko et al., 2007). This comprises both a set of pre-developed applications (e.g. the observation data viewer as part of the stack of applications in an Urban Observatory) and, at a lower level, API components which facilitate the creation of applications. For example, the offered Sensor Web REST-API allows application developers to easily access different underlying data sources (e.g. the observation data available on the CKAN server or measurement data offered by OGC compliant SOS server) so that they do not have to deal with these sometimes complex interfaces. Hence, it facilitates the development of data driven applications and additionally supports our second goal, which is to provide information products to stakeholders in the municipal administration and water treatment companies.

The applications target different types of users. The observation data viewer for instance addresses scientists who want to discover and explore available observation data, but also operators of sensor networks who want to get an overview of the collected measurements. The observation data viewer is implemented in JavaScript, and it has been designed in a modular manner so that future enhancements and adjustments are easily possible. Furthermore, the design of the viewer is responsive such that it can be used on mobile phones, tablets, as well as desktop computers. The Sensor Web REST-API acts as a proxy to underlying data sources. This component is developed in Java and for instance able to interact with CKAN servers and SOS servers as data sources. The proxy regularly analyses the content offered by the data sources and stores the collected data and metadata in an internal data cache. Using this harvesting approach, the Sensor Web REST-API and therefore the application is capable of handling incoming queries at a high level of performance. All self-developed software components are published under open source licenses and free-to-use and to contribute to.

In brief, the initial development for our application stack – the observation data viewer - does have the following features:

- Diagram and table views for time series data
- Track display for visualising mobile sensors and their measurements
- Parameterised URL calls to start the viewer with pre-selected data sets
- Support of different device types (smart phones, tablet, desktop PC)

The authorization layer (see Figure 3 *Role and Access Management*), as a cross-cutting component, controls permission access to the various data sources. Currently, three roles are distinguished: (1) private usage, (2) world wide access, (3) internal usage, for specific groups of stakeholders, hydrological modellers, etc.

The described components contribute to a versatile software stack that supports data management in environmental projects. To prove the applicability of the described architecture and software stack to other types of projects we plan to test the described infrastructure within other running projects.

4 CONCLUSIONS AND OUTLOOK

We presented a data management infrastructure to support the research management and exchange in all phases of environmental research projects. It provides a simple to use and secure solution for the collection and management of all (newly) evolving research data within a project and a simplified inclusion of the external data sources within projects. Thus, our data management platform represents the central point of data storage and exchange and provides further offers tools to semantically enrich and transform the data for the subsequent publishing process. This publishing process includes the resource-based data dissemination process as well as the provision of standards-based OGC compliant web services. Together with the delivering of relevant web based data structures and interfaces we created a flexible and comprehensive infrastructure to support the complete workflow from the data collection, to preparation, to processing and finally to dissemination. The implementation of this infrastructure is still in a prototype phase and under intensive development. Individual components are gradually refined according to further requirements of the users. The expansion of the supported data formats portfolio in the semantic data enrichment process in the data management platform can serve here for instance. Although the development of the semantic enrichment is technically completed, it is necessary to inspect it closer, since every discipline – not necessarily only on an academic level – has a domain-specific vocabulary. To gain a widespread acceptance in particular communities and generally support a research process, it is mandatory to coordinate this semantic enrichment for important properties with all involved disciplines. Hence, in the next step we intend the design and integration of a more complex and complete qualification process for the data management and publishing process to better support a semantically and schematic transformation process and methods for data fusion. Furthermore, this includes a support for enhanced quality descriptions of the data to better distinguish between the fit for purpose of administrative data and, for instance, low-cost sensor data. Subsequently, more in-depth integration and user tests are mandatory to achieve the long-term goal of our data management: the transferability of the infrastructure beyond specific project boundaries.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the German Federal Ministry of Education and Research, programme Geotechnologien, under grant agreement no. 03G0852A. The project website is available at https://colabis.de.

REFERENCES

52° North Initiative for Geospatial Open Source Software GmbH, 2014. SensorWeb RESTful API Version 1.

Al-Bakri, M. & Fairbairn, D., 2010. Assessing the accuracy of crowdsourced data and its integration with official spatial datasets. Accuracy 2010 Symposium. Leicester, UK.

Andrienko, G., Andrienko, N., Jankowski, P., Keim, D., Kraak, M.J., MacEachren, A. & Wrobel, S., 2007. Geovisual analytics for spatial decision support: Setting the research agenda. International Journal of Geographical Information Science, 21, 839-857.

Bernard, L., Kanellopoulos, I., Annoni, A. & Smits, P., 2005. The European Geoportal - One step towards the Establishment of a European Spatial Data Infrastructure. Computers, Environment and Urban Systems, 29, 15-31.

Bernard, L., Mäs, S., Müller, M., Henzen, C. & Brauner, J., 2014. Scientific geodata infrastructures: challenges, approaches and directions. International Journal of Digital Earth, 7, 613-633.

DataONE - Data Observation Network for Earth. Available: https://www.dataone.org [Accessed 27/05/16].

EUDAT - Research Data Services Expertise & Technology Solutions. Available: https://www.eudat.eu [Accessed 27/05/16].

Fielding, T.R., 2000. Architectural Styles and the Design of Network-based Software Architectures. Dissertation, University of California.

GEOSS - Global Earth Observation System of Systems. Available: www.geoportal.org/ [Accessed 27/05/16].

Girres, J.-F. & Touya, G., 2010. Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS, 14, 435-459.

Gray, J., 2009. eScience: a transformed scientific method. In: Hey, A.J.G., Tansley, S. & Tolle, K.M. (eds.) The fourth paradigm: data-intensive scientific discovery. Redmond, WA: Microsoft Research.

Haklay, M., 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. Environment and Planning B: Planning and Design, 37, 682-703.

Henzen, D., Karrasch, P. & Bernard, L., 2015. Self-learning Virtual Sensor Networks using low-cost electronics in Urban Geosensor Networks. 18th AGILE International Conference on Geographic Information Science. Lissabon.

Hook, L.A., Vannan, S.K.S., Beaty, T.W., Cook, R.B. & Wilson, B.E., 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. In: Center, O.R.N.L.D.A.A. (ed.). Oak Ridge, Tennessee, USA.

ISO, 2005. Geographic information – Encoding. ISO 19118:2005.

Jones, A.S., Horsburgh, J.S., Reeder, S.L., Ramírez, M. & Caraballo, J., 2015. A data management and publication workflow for a large-scale, heterogeneous sensor network. Environmental Monitoring and Assessment, 187.

Marx, C., Mühlbauer, V., Schubert, S., Oertel, R., Ahnert, M., Krebs, P. & Kühn, V., 2015. Representative input load of antibiotics to WWTPs: Predictive accuracy and determination of a required sampling quantity. Water Research, 76, 19-32.

Mason, S.J.K., Cleveland, S.B., Llovet, P., Izurieta, C. & Poole, G.C., 2014. A centralized tool for managing, archiving, and serving point-in-time data in ecological research laboratories. Environmental Modelling & Software, 51, 59-69.

OGC, 2007. OpenGIS Catalogue Services Specification.

PANGAEA - Data Publisher for Earth & Environmental Science. Available: https://www.pangaea.de/ [Accessed 27/05/16].

Reis, S., Seto, E., Northcross, A., Quinn, N.W.T., Convertino, M., Jones, R.L., Maier, H.R., Schlink, U., Steinle, S., Vieno, M. & Wimberly, M.C., 2015. Integrating modelling and smart sensors for environmental and human health. Environmental Modelling & Software, 74, 238-246.

Samourkasidis, A. & Athanasiadis, I.N., 2014. Towards a low-cost, full-service air quality data archival system. Proc. 7th Intl. Congress on Environmental Modelling and Software, International Environmental Modelling and Software Society (iEMSs).

Zhang, J., Hua, P. & Krebs, P., 2015a. The build-up dynamic and chemical fractionation of Cu, Zn and Cd in road-deposited sediment. Science of The Total Environment, 532, 723-732.

Zhang, J., Hua, P. & Krebs, P., 2015b. The chemical fractionation and potential source identification of Cu, Zn and Cd on urban watershed. Water Science and Technology, 72, 1428-1436.

Zhang, J., Hua, P. & Krebs, P., 2016. The influences of dissolved organic matter and surfactant on the desorption of Cu and Zn from road-deposited sediment. Chemosphere, 150, 63-70.

Zhang, J., Wang, J., Hua, P. & Krebs, P., 2015c. The qualitative and quantitative source apportionments of polycyclic aromatic hydrocarbons in size dependent road deposited sediment. Science of The Total Environment, 505, 90-101.