The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-7/W3, 2015
36th International Symposium on Remote Sensing of Environment, 11–15 May 2015, Berlin, Germany

BIG DATA BREAKING BARRIERS – FIRST STEPS ON A LONG TRAIL

S. Schade

European Commission – Joint Research Centre, Institute for Environment and Sustainability, Via Enrico Fermi 2749, 21027 Ispra, Italy

sven.schade@jrc.ec.europa.eu

KEY WORDS: Big Data, Digital Earth, Breaking Barriers, Visual Analytics, Multidisciplinary

ABSTRACT:

Most data sets and streams have a geospatial component. Some people even claim that about 80% of all data is related to location. In the era of Big Data this number might even be underestimated, as data sets interrelate and initially non-spatial data becomes indirectly geo-referenced. The optimal treatment of Big Data thus requires advanced methods and technologies for handling the geospatial aspects in data storage, processing, pattern recognition, prediction, visualisation and exploration. On the one hand, our work exploits earth and environmental sciences for existing interoperability standards, and the foundational data structures, algorithms and software that are required to meet these geospatial information handling tasks. On the other hand, we are concerned with the arising needs to combine human analysis capacities (intelligence augmentation) with machine power (artificial intelligence). This paper provides an overview of the emerging landscape and outlines our (Digital Earth) vision for addressing the upcoming issues. We particularly request the projection and re-use of the existing environmental, earth observation and remote sensing expertise in other sectors, i.e. to break the barriers of all of these silos by investigating integrated applications.

1. INTRODUCTION

Today, more and more data becomes available (discoverable and accessible) – on purpose, or unintended. In this era of "Big Data" – i.e. in a situation in where the volume, variety, velocity and veracity (3+1 Vs) in which data sets and streams become available challenges current management and processing capabilities (Hey et al, 2009) – we undergo a paradigm shift from the mentality to ask for all images related to theme X in region Y at time Z, to requests such as: "give me all that you have that is related to this area" or "give me all that you have that is related to that object". Potentially relevant data does not any more come from a known (small) community, but from everywhere. This naturally leads to a clash of working practices and cultures.

With this in mind, our work exploits earth and environmental sciences for existing interoperability standards, and the foundational data structures, algorithms and software that are required to meet the geospatial information handling tasks in Big Data research. Furthermore, we are concerned with the arising needs to combine human analysis capacities (intelligence augmentation) with machine power (artificial intelligence) in order to advance knowledge discovery across data sets and streams from open, commercial and civic sources.

This paper provides an overview of the emerging landscape and outlines our (Digital Earth) vision for addressing the upcoming issues. We particularly request the projection and re-use of the existing environmental, earth observation and remote sensing expertise in other sectors, i.e. to break the barriers of all of these silos by investigating integrated applications.

The remainder of this paper is structured as follows. The next section presents the emerging Big Data landscape both on the general level, as well as in relation to geospatial information, earth and environmental sciences. Thereafter, in Section 3, we briefly outline our recent activities that explore multiple facets of geospatial data analysis and visualisation, particularly considering new data sources, novel technologies, and means for integration. Section 4 discusses the findings from these activities and sets them into the technical, semantic and organisational context, just before we draw our conclusions and derive future work items in Section 5.

2. BIG DATA LANDSCAPE

It only requires a quick look at www.bigdatalandscape.com to understand that the landscape of Big Data technologies, architectures and applications is complicated. While a few prominent players could already establish themselves, many specialised products are equally available. Below, we report on our impressions of the mainstream technologies, as well as of dedicated geospatial information handling tools.

2.1 Overall landscape

Although the overall functional requirements and system components have been identified (see e.g. the work of the US National Institute of Standards and Technology – NIST (bigdatawg.nist.gov)) on the general level and (Lee and Kang, 2015) for a representative example that is particularly related to Big Geospatial Data), the underlying technologies are still evolving, and their landscape remains dynamic. We might expect stabilization only in the medium term.

The required ecosystem of technologies and infrastructures demands contributions from a wide community and it is difficult to provide full-fledged solutions off-the-shelf. Looking into the technologies and infrastructures, some commercial tools tend to become open source, and many are even undergoing the incubation process of the Apache Software Foundation (www.apache.org). Several products/components both in the Apache Hadoop stack (hadoop.apache.org) and in commercial products – even within the same company – have

overlapping functionality. It appears that, if an organisation envisages a wide range of Big Data applications, then it is most likely best served with an open source solution. The specialised commercial products might - on the one hand - not adopt to all needs, and - on the other hand - many of the provided more generic capabilities might remain unused. Still, the rich set of application areas lead to case-dependent adaptations of the available solutions, for example, many of the Big Data analytics platforms criticize the pure MapReduce (Dean and Ghema,2004) and provide their optimized versions or simply replace it, e.g. using Drill (http://drill.apache.org).

Nevertheless, we witness some consistency across current approaches when voluminous data has to be handled quickly. Here, incoming data is channelled into processing pipes. The control of the data flow (together with resource allocations) is often separated from the specific algorithms that are required in each processing step. These might even be realized with diverse programming language, such as R (www.r-project.org) or Python (www.python.org). It might be possibly to consider a generalization over all approaches, but it remains to be seen if the most common denominators are still meaningful or resolve in common sense. Abstractions of workflows as interconnected functions might prove useful.

All in all, Big Data as such does not necessarily imply the need for huge computing power, or (for the geospatial information science, earth observation and remote sensing communities) to focus on computational capacities. Storage and computing facilities are advancing while they become increasingly requested. Undoubtedly, this area requires dedicated and coordinated action, but the related concerns should be addressed by computer scientists and system engineers so that the required supporting technologies are provided across application domains. In the end this should lead to the optimal e-Infrastructure that offers cross cutting support to (Research Infrastructures of) multiple application areas – including remote sensing, but also helping to breach out into other fields.

2.2 Landscape on geospatial capabilities for Big Data handling

Ultimately, most data sets and streams have a geospatial component. Years ago, some people claimed that about 80% of all data is related to location. In the era of Big Data this number might even be underestimated, as data sets interrelate and thus initially non-spatial data becomes indirectly geo-referenced (by associating it to some spatial data set). Consequently, any ignorance of expert geospatial solutions for Big Data challenges unavoidably limits knowledge extraction and thus fails to exploit hidden potentials.

Indeed, geospatial intelligence increasingly finds applications across sectors, not only within earth and environmental sciences - where it is traditionally applied. Some of the many examples include health care, utilities, transport and retail (Buchholtz et al, 2014). All of these sectors – and many more – currently investigate possible benefits from the use of the spatially-enabled Internet of Things (IoT), geo-located social media, and more general aspects of geospatial information handling.

When talking about Big Data in geospatial intelligence, we particularly see the following match with the 3+1 Vs:

- *Volume*: large data volumes primarily appear from remote sensing (usually 2D images, such as those delivered by the Sentinels of the European Space program (Copernicus), or point clouds in 3D, such as LIDAR), or from intense modelling as mostly done for immediate and medium range weather forecasting (see e.g. www.ecmwf.int) and climate modelling (see e.g. www.noaa.gov). Array data bases and (for point clouds) columns stores are applied (Baumann et al, 2014).
- *Velocity*: high throughput appears while transmitting and processing large single volumes or continuous inputs - of the same type but from massive amounts of sources, e.g. in the context of the IoT. In stream processing and distributed computing (e.g. cloud or HPC) are applied. Parallelization algorithms depends on the applied tools (such as STORM/trident (storm.apache.org) or Kafka (kafka.apache.org)).
- *Variety*: given any place on earth (or elsewhere), we already today receive spatially-related data sets and streams for multiple sources. These are expected to grow and accumulate over time. Using spatial co-occurrence, classical geospatial technology and the possibility to ground information in physical space already provide huge asserts for data integration. However, - as in many other domains - data integration from multiple sources still poses huge organizational, legal, semantic and technical interoperability challenges. Some of the promising approaches that require further investigations include: brokering (Nativi et al, 2012), linked data (Auer et al, 2009) and semantic integration/fusion (Mau é and Schade, 2009).
- *Veracity*: the question of reference data and differentiation between ‘authoritative’ sources and user-contributed contend (sometimes Volunteered Geographic Information (Goodchild, 2007) is still heavily discussed in the geospatial community (a, for example, during the Geospatial Information Observatories workshop at last year’s GIScience conference), and – closely related - also in statistics research (see for example the latest conference on New Techniques and Technologies for Statistics).

In terms of available geospatial information handling capabilities, we see mature support in the area of gridded data sets (and streams), which usually represent field-like phenomena in space time – including point clouds from radar, images, grids of all kinds – these mostly support earth observation and climate sciences, but also some areas of hydrology and hydrography. Native support for vector formats (except grids) remains in its infancy. Apart from some support of (2D) geospatial indexing and simple geospatial filters for data queries, we currently do not see much sophisticated Big Data capabilities for the processing of geospatial objects. This observation follows the overall support of geospatial data handling by mainstream Information and Communication Technology (ICT), which usually does not expand beyond point data (latitude/longitude). We see room for extended research and innovation relating to the spatial-temporal processing of object related data, including 3D, such as trajectories of all sorts of entities (e.g. from RFID, GPS, Galileo or mobile phone data) and data streams from the IoT.

Together with new modes of immersive and collaborative visual analytics for use in education and science, these capabilities will enable the implementation of a next-generation Digital Earth (Goodchild et al, 2012). Accordingly, within the context of this paper, we call the solutions for Big Data analysis and visualisation in the environmental and earth sciences, which can

be used for small as well as large heterogeneous datasets, *Digital Earth platforms*.

3. OVERVIEW OF CASE STUDIES

Last year (2014), we carried out ten case studies in order to (i) examine components of such Digital Earth platforms; (ii) identify some new possibilities; and (iii) gain hands-on experiences from both, the examination of new data sources, as well as the projection of already ongoing work into a Big Data context. In order to illustrate parts of the arising capabilities and to provide a basis for discussion, we briefly present all ten case studies below, grouped by the underlying motivation, and including pointers to further readings – where available.

3.1 3D platform for geospatial data handling

In order to investigate the current opportunities for handling 3D geospatial visualizations and thereby identify promising ways to provide a core baseline for any Digital Earth platform, we investigated two technology options for the potential of the "Core003" data set, a Very High Resolution (VHR) optical coverage over the member and cooperating countries of the European Environment Agency (EEA) that was generated from SPOT-5 data through multi-spectral 2.5 meters resolution data ortho-rectified with a geo-location accuracy of less than 5 meters Root Mean Square Error (RMSE):

- *3D browser based viewer*: We initiated an experiment to advance the current Core003 viewer (JRC, 2014) that has been developed by our colleagues. The new activity aims to develop a web viewer showing a detailed 3D representation of the European land, using the 2.5 meters Core003 true colour mosaic as a raster overlay. The implementation is realized with Cesium and WebGL (cesiumjs.org), and allows to overlay visualisations from standard conform Web Map Services (WMSs) (OGC, 2006).
- *Advanced 3D application*: In parallel to the viewer, and in order to compare the potentials of a browser-based solution with the capabilities of a stand-alone application, we started the development of a second platform. This should provide a powerful 3D desktop application with advanced functionality. Furthermore, the experiment has the goal to also work on large touch screen systems, allowing a direct interaction with the "hands on" the 3D model representations.

Considering 3D visualisation, and particularly the application of Cesium and WebGL, we found a mature, highly customizable and open software solution that is able to deal with voluminous geospatial datasets. The Cesium platform is open to all affirmed and emerging standards in the 3D visualisation field to design and integrate detailed models into the virtual globe and is a highly powerful spatio-temporal platform. However, if we desire to apply advanced visualisation technologies efficiently and effectively, then these investigations teach us that – at least currently – we have to rely on desktop applications.

3.2 Investigating usage potential of social media platforms

Social media provide potential now data sources that might complement traditional remote sensing and earth observation with "social sensing" in future Digital Earth platforms. We focus our examinations on the-re-use of existing software libraries and applications of novel data handling technologies with the following initial activities:

- *Using new database technologies to store and query social media data*: In order to investigate the particular capabilities of the NoSQL database MongoDB (www.mongodb.org), we ran a case study that investigated its potential use for social media analysis, here especially focusing on data feeds from the microblogging site Twitter in Dorset, a small region in the UK (Juhász, 2014).
- *Using social network analysis to sense social behaviour*: We conducted a short-term case study that focuses on communication patterns in Twitter before and during the United Nations climate summit in September of 2014. This activity was conceived as a didactic example of how to make scientific processes more transparent (and reproducible) and re-used tools of the first mentioned experiment.
- *Using social media platforms to complement authoritative vector data*: In this case study we investigated the suitability of social media data (especially from Foursquare) as a data source for determining building use. A case study has been conducted in Amsterdam, in an area of 72.12 km2, where 112,567 buildings are located (Spyratos et al, submitted).

All three experiments together reveal possibilities and limitations when extracting knowledge from these relatively new data sources. Any social media analysis has to face linguistic issues – not only across languages, but also in respect to stop words or modifications of terms, e.g. to express sentiments. Issues of geo-location remain, as still approximately only 1% of Tweets are geo-located. The extraction of place names is of limited success. It has to be particularly considered that – due to the usage conditions of most social media Application Programming Interfaces (APIs) – we always retrieve (unknown) subsets/samples. To this sense, social media is a fragile source and results are thus rather indicative than conclusive. More operational activities would benefit from full access, which mostly would mean purchase of the full data set.

If social media data is used for a new purpose, it cannot be expected to fully replace a targeted method that is already in place. However, it might provide useful complementing insights. The use of social media data in combination with other sources (e.g. coming from the public sector) remains a promising research direction. Still, if should also be noted that such applications are highly case dependent, i.e. each application area requires a dedicated set-up, calibration and evaluation mechanism. In any case, data from social media always cover only a non-representative part of society.

3.3 Sensing technologies and the Internet of Things

As sensor networks remain to flourish with the IoT paradigm (Kortuem et al, 2010), we also began to investigate potential processing mechanisms and tools. Here, applications exceed way beyond current (environmental) monitoring networks, due to the increased integration of industrial sensing devices into all sorts of manufactured goods, but also the use of low-cost sensors by layman (as a form of Citizen Science (Haklay, 2012)). We particularly investigated two mechanisms:

- *Real-time event detection from sensor networks*: Each sensor in each network produces a stream of data and has the capacity to send a large number of observations. It

becomes difficult to analyse all of these observations in the moment that the raw values are obtained. This case study was targeted to this this particular Big Data challenge and investigated mechanism to analyse the arising flood of monitoring data. We provided a proof of concept implementation based on the Storm framework and tested in with a regional environmental sensor network (Trilles et al, 2015, Trilles et al, submitted).

- *Service-Enabled Sensing Platform for the Environment*: Considering the particular challenges of handling information from sensor networks and taking an approach of reducing the data transfer and storage needs outside the originally data producing agents (the sensors), we investigated a novel methodology for handling data within networks of distributed sensors. This case study concentrated on the underlying architectural considerations and possibilities of deploying sensor-near processing facilities in order to deal with Big Data originating from the IoT. The first developments focused on a test set-up in the area of air quality monitoring (Kotsev et al, 2015).

These solutions do not only address (big) data velocity. Each of them also covers aspects of scalability and flexibility, two essential requirements when dealing with large data volumes in future Digital Earth platforms. Last but not least, although both case studies were carried out with specific environmental data sources, they offer topic independent designs and could also be applied to data integration, i.e. resolving issues of (big) data variety. The surrounding methodology and essential software components could be identified and facilitated. These more general capabilities provide room for continues and extended testing.

3.4 Handling the complexity of data integration

With a more integrative view on Digital Earth platforms, and following our earlier work on the Infrastructure for Spatial Information in Europe (INSPIRE) (Schade, 2013), integrated modelling (Granell et al, 2013), as well as a Digital Earth Nervous System (De Longueville et al, 2010), the concept of the Observation Web (Havlik et al, 2011) we continued our investigations of integration mechanisms across multiple types of data sources. Those included:

- *Visualisations of complex metadata*: As connecting (linking) rich metadata is increasingly recommended on top of Big Data, we begin to investigate possibilities to explore rich metadata and to highlight relevant aspects in given practical context. Hence, we selected the INSPIRE metadata that is available from the official INSPIRE geoportal (inspire-geoportal.ec.europa.eu). It provides an interesting case for the visual analysis of environment-related data in an EU policy-making context. With more than 300.000 metadata records of largely unstructured data, a web of relationships emerges when visualised properly. We particularly used Gephi (www.gephi.org) as a tool to highlight health related themes in the available INSPIRE metadata about data, services and applications.
- *Model transparency*: In our digital age, model transparency, i.e. the access to models, platforms, frameworks and systems, together with their descriptions, related input and output data, impact assessments as well as related documentation of any kind, is one of the holy grails across all sciences. Addressing transparency often requires institutional, cultural and technical challenges. Especially the challenge of complexity closely relates to the visualisation of varying types of (big) data. Accordingly, we related ongoing work of the management of models and related access services within the Joint Research Centre (JRC) to Big Data challenges on integration and visualization (Ostländer et al, submitted).
- *New modes for multi-sensory integration*: We began to exploit the potentials of multi-sensory integration to further develop the surrounding concept of a Digital Earth Nervous System, thereby not processing different data streams in parallel but together. We found that particularly promising research objectives include the assessment of a sensor's observations' validity through possibility methods and the use of crowd-sourcing to supervise machine-learning of algorithms and rules to filter, sort and organized stimuli into coherent perceptions. (Ostermann and Schade, 2014).

As already indicated in relation to the case studies on social media, but equally true when also considering the IoT, any investigation, especially if a combination of data sources is considered, required – at least in parts - dedicated data flows and particular calibrations considering the targeted questions. A generic detection of anomalies for initiating more detailed (and specialised) investigations that also consult other data sources might be desirable. It also became obvious that not only the facilitated data sources will have to be well described, but also the used software tools, models, algorithms and underlying assumptions. The resulting flood of meta-data requests now forms of visual analytics, so that potential users can identify potentially relevant information and judge their fitness for purpose in respect to their particular contexts – that are largely unknown at the time of data gathering.

4. DISCUSSION

In the light of investigating possible Digital Earth platform(s) of the future, we focus our discussion not on particularities of the case studies that were just presented above, but reflect on the overall experiences and impressions gained from the numerous investigations. Taking the standpoint that underlying issues of infrastructure and hardware should be addressed by computer science and software engineering (see Section 2.1), we found three major barriers that should be overcome in order to fully address the challenges that the ever growing volume, variety and velocity of data are posing to earth and environmental sciences (with strong dependencies between each other, as we will see in the conclusions).

4.1 Technical barrier

Alongside the ten case studies that we carried out and by reviewing the current Big Data landscape we saw a wide range of architectural solutions, partial implementations and software components, which each addressed some issue of geospatial information handling and were usually specialized for a particular use. Having a rich choice for implementations is useful on the one hand because existing resources might be re-used, but on the other hand puts not only a burden in the identification of an appropriate solution to a particular problem and potential implications of a technological choice (and investment) on the capacities to solve future scientific challenges. It also introduces a barrier in sharing experiences between any two parties that follow different approaches. This might result in a diversification of methodologies and tools, already within advanced environmental and earth analytics. We see a danger to provide solutions for on-demand knowledge

extraction from highly assimilated structures, semi-structured and unstructured data (including in-situ measurements, ex-situ observations, and remotely sensed images and point clouds) that do not work together. Such technical heterogeneities put barriers to data integration, one of the main assets promoted by the Big Data movement.
Having said this, one might assume that the joint development of one technical platform for all environmental and earth analytics would be the optimal solution. However, apart from cultural and political issues, this is highly unlikely to happen because of reasons of complexity and complication. The use of earth observation and environmental data is already so diverse that the optimisation of a solution for one particular field of expertise would not fit the others. Furthermore, why should some researcher be forced to use a universal tool that by definition would be hard to learn when (s)he has only a very dedicated small data handling task to complete? In other words, a Digital Earth platform should never be the next-generation Geospatial Information System (GIS) or Spatial Data Infrastructure (SDI).

In essence, although present solution, such as NASA World Wind (worldwind.arc.nasa.gov) or Google Earth Engine (earthengine.google.org) exist, we should not expect one single solution for all possible purposes. On the contrary, we should not focus on the one fits all technical solution, i.e. to develop yet another platform that is supports all possibly required analysis tasks, but consider an un-platform, in the sense that allows to re-use and connect already existing pieces and captures the meta level of describing performed experiments and lessons learned.

4.2 Semantic barrier

The wide range of architectural solutions, partial implementations and software components, can equally be witness beyond the earth and environmental sciences. Accordingly, we see a strong requirement to provide easier access and means for connecting 'foreign' (i.e. non- geospatial, environmental, remote sensing and earth observation) domains.

While a few trends, such as the move from Hadoop (and thus MapReduce) to Scala (scala-lang.org) or the separation between the handling of data flows and the execution of algorithms to the content emerge, portability between knowledge communities remains difficult. This is particularly a problem for Big Data, because – as already mentioned in the first paragraph of this paper – potentially useful data does not reside inside a well-known community any more, but might be offered by any third-party. Work across currently existing communities, including the earth observation and remote sensing communities, can only be established large scale if we do not bind the use of the community data sets and related tools to one (or few) community specific tools. We have to find a way to easily and quickly understand third-party data sets, their fit for purpose, and the required processing capabilities.

Consequently, proposed solutions should not (only) work together, but it should be easy to transfer generated knowledge between specific institutions, infrastructures and technological components, as well as to replace and customize parts of the methodology with another implementation or even architecture, especially across knowledge domains.

4.3 Organisational barrier

The interdisciplinary work that we just argued for obviously has to overcome organizational issues, including not only the crossing of scientific cultures such as the collaboration between the natural sciences and the social sciences. It also has to address the relation between science, industry and the public. Issues of privacy and ethics obviously arise when dealing with data from as many sources as possible and by deriving new findings out of their combination. These items have to be addressed in any serious Big Data research and citizen engagement seems a promising (if not the only) pathway.

Some of the already existing platforms allow to reduce complexity that far, that also stakeholders without any scientific background can be involved in the analysis activities. Until some years ago, the use of complex algorithms and analysis methodologies were only available to scientists. Now, advanced visual analytics also allow citizen participation to integrative research in a trans-disciplinary way which foresees tightly integrated research, the latter involving participants without any academic background.

This finally moves us into an era that breaks barriers for Citizen Science. Here, GalaxyZoo or the many other projects of Zooniverse (www.zooniverse.org) provide impressive examples of successful storytelling and the use of gamification techniques –many of which largely benefit from the increased resolution in earth observation, as people become able to identify objects in pictures and can get engaged because they see their house or local neighbourhood from a birds-eye-perspective. These developments open a whole new range of applications driven by earth observation products, way beyond the traditional use in expert systems or as pure background imagery. In this way, latest visualisation and visual analytics technologies empower us to move beyond social sensing - in which laymen collect data - to (social) co-delivery of scientific evidence. With solutions such as Geo-Wiki (www.geo-wiki.org), and follow-up activities, everybody gets empowered to also analyse his/her own data, information collected by others, and much more. The convergence of (a) increasing data volumes from earth observation (and space) technologies, which pose data processing challenges and excel the limits of automated feature detection from imagery; and (b) enabling essentially everybody who can use a web-based application, is a huge chance for massive social engagement. This provides immense new possibilities in developing the "social machine" (de Roure, 2014), i.e. the optimised combination of human analysis capacities (intelligence augmentation) with machine power (artificial intelligence) in which simplification, pattern recognition and ground truthing by humans feeds into self-learning algorithms and vice versa.

With this we reach a state in which it becomes clear that we should not (only) address the technocratic dimension of Big Data, but increase investigations of the social and behavioural dimension, i.e. real stakeholder engagements, community building, and possibly before all other, citizen participation.

5. CONCLUSION

In this paper we investigated the ongoing work around the notion of Big Data from an exploratory point of view and introduced the notion of Digital Earth platforms which might integrate traditional remote sensing and earth observation tools with newly arising knowledge sources powered by concepts

such as social sensing and the IoT. We our view on the emerging landscape, explained the way we try to learn about this arising field ourselves, while deriving findings that might be valuable for the wider community. This last point seems particularly timely and valuable because many research organisations are currently testing and taking first steps in the Big Data landscape, while large scale roll outs and operational deployments are still rare. We particularly underline the requirement to exchange knowledge between communities and grow together.

On the technical level, we promote advanced environmental and earth analytics services to provide on-demand knowledge extraction from highly assimilated structures, semi-structured and unstructured data (including in-situ measurements, ex-situ observations, and remotely sensed images and point clouds). Considering semantic interoperability we stress the requirement to provide easier access and means for connectivity 'foreign' – i.e. non- geospatial, environmental, remote sensing and earth observation – domains. Organisationally, we argue for a social and behavioural approach that crosses scientific cultures (including thematic practices, open research and citizen science) and thus fostering inter-disciplinary work.

Considering the three barriers that we identified for the earth and environmental sciences – do not develop (yet) a(nother) platform, do not (only) work together, and do not (only) be technocratic – we might conclude that the real elephant in the room, which so many are searching for, is real openness; or in other words: do not (only) address the easy part of Open Data. As Adams and Gahegan (2014) pointed out recently, we have to extend the data producers view to also include the data consumers perspective. Here, data should not only be understood in the narrow sense, but it should also include generated code, methodologies, description of experiments, and much more. The description and sharing of such contextual information in a way that can be perceived and unambiguously understood by potentially interested users is the major future challenge when aiming at optimal data re-use.

In order to improve the joint understanding of the real potential and feasibility of Big Data analysis capabilities, we will have to include investigations on the potential requirements for large scale operations and set those into relation with the gained benefits (and threads). In our future work, we will further investigate a structured methodology to derive these findings from the many existing case studies and use it for the planning of new activities. We are currently investigating the use of RM-ODP (www.rm-odp.net) for this purpose. It seems promising to use this standard methodology to describe information systems that is already widely used in the geospatial information domain to develop a high-level view on Digital Earth platforms.

With this we hope that we could illustrate some of the most eminent challenges of our data-driven age, particularly in relation to earth and environmental science and geospatial information handling. Many more case studies have been developed across the globe and we certainly took our first steps on the long trail of successful and useful knowledge and thereby value extraction from small and big data – and their combinations. We hope that the required barriers will be overcome and remind all of us to take to time to occasionally check if we are still on the right track.

ACKNOWLEDGEMENTS

Most of the work reflects the 2014 activities of a whole group of people working at the Digital Earth and Reference Unit of the JRC, especially including (in alphabetical order of surnames) Massimo Craglia, Davide De Marchi, Irene Eleta, Jacopo Grazzini, Jiří Hradec, Alexander Kotsev, Frank Ostermann, Nicole Ostländer, Francesco Pantisano, Elena Roglia, Cristina Sanchez, Sven Schade, Spyridon Spyratos, Chrisa Tsinaraki, Lorenzino Vaccari, as well as the two visiting scientists (Levente Juhász and Sergi Trilles).

REFERENCES

Adams, B., Gahegan, M., 2014. Emerging data challenges for next-generation spatial data infrastructure. Paper presented at Research@Locate'14, Canberra, Australia. 7 - 9 April 2014.

Auer, S., Lehmann, J., Hellmann, S., 2009. LinkedGeoData: Adding a Spatial Dimension to the Web of Data. Proceedings of the International Semantic Web Conference 2009, pp. 731-746.

Baumann, P., Yu, J., Misev, D., Lipskoch, K., Beccati, A., Campalani, P., Owonibi, M. (2014). Preparing Array Analytics for the Data Tsunami. *I*n Pourabbas, E. (ed): *Geographical Information Systems: Trends and Technologies*, CRC Press, 2014, pp. 1 – 19.

Buchholtz, S., Bukowski, A., Śniegocki, M., 2014. *Big and open data in Europe - A growth engine or a missed opportunity?* Report commissioned by demosEUROPA.

De Longueville, B., Annoni, A., Schade, S., et al., 2010. Digital Earth's nervous system for crisis events: real-time Sensor Web enablement of Volunteered Geographic Information. *International Journal of Digital Earth 3(3)*, pp. 242-259.

De Roure, D., 2014. The Emerging Paradigm of Social Machines. *In O'Hara et al. (Eds.) Digital Enlightment Yearbook 2014*, IOS Press.

Dean, J., Ghemawat, S., 2008.. MapReduce: simplified data processing on large clusters. *Communications of the ACM 51(1)*, pp. 107-113.

Goodchild, M. F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal 69*, pp. 211–221.

Goodchild, M. F., Guo, H., Annoni, A., et al., 2012. Next-Generation Digital Earth. *Proceedings of the National Academy of Sciences - PNAS 109 (28)*, pp. 11088-11094.

Granell, G., Schade, S., Ostländer, N., 2013. Seeing the forest through the trees: A review of integrated environmental modelling tools. *Computers, Environment and Urban Systems 41*, pp. 136–150.

Haklay, M., 2012. Overview and Typology of Participation. Crowdsourcing Geographic Knowledge. *In: Citizen Science and Volunteered Geographic Information*, pp 105-122.

Havlik, D., Schade, S., Sabeur, Z., et al., 2011. From Sensor to Observation Web with Environmental Enablers in the Future Internet. *Sensors 11*, pp 3874-3907.

Hey T., et al. (Eds.), 2009. *The Fourth Paradigm: Data Intensive Scientific Discovery*. Microsoft.

JRC, 2014. Web viewer for Copernicus Core_003 Tile Service. Web service of the Joint Research Centre http://cidportal.jrc.ec.europa.eu/copernicus/services/webviewer/core003/ (25 March 2015).

Juhász, L. 2014, Documentation of the case study on using new database technologies to store and query social media data: http://blog.jlevente.com/twitter-data-analysis-from-mongodb-part-1-introduction, http://blog.jlevente.com/twitter-data-analysis-from-mongodb-part-2-exploring-data, http://blog.jlevente.com/twitter-data-analysis-from-mongodb-part-3-basic-spatial-and-temporal-content, http://blog.jlevente.com/twitter-data-analysis-from-mongodb-part-4-visualizing-tweets, and http://blog.jlevente.com/mongodb-postgresql-speed-comparison (25 March 2015)

Kortuem, G., Kawsar, F., Fitton, D., Sundramoorthy, V., 2010. Smart objects as building blocks for the internet of things. *IEEE Internet Computing 14 (1)*.

Kotsev, A., Pantisano, F., Schade, S., Jirka, S. 2015. Service-enabled sensing platform for the environment. *Sensors scientific journal – special issue on Wireless Sensor Networks and the Internet of Things*.

Lee, J.G., Kang, M., 2015. Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, in press.

Maué, P., Schade, S., 2009. Data Integration in the Geospatial Semantic Web. *Journal of Cases on Information Technology 11(4)*, pp. 100-122.

Nativi, S., Craglia, M., Pearlman, J., 2012. The brokering approach for multidisciplinary interoperability: A position paper, *International Journal of Spatial Data Infrastructures Research 7*, pp. 1-15.

OGC, 2006 OpenGIS Web Map Service (WMS) Implementation Specification – Version 1.3.0.

Ostermann, F., Schade, S., 2014. Multi-sensory Integration for a Digital Earth Nervous System. AGILE Conference 2014, Castellon, Spain, 3.6 June 2014.

Ostländer, N., et al, submitted Developing a Model Inventory and access services in order to increase the transparency of policy making.

Schade, S., 2013. INSPIRE Monitoring and Reporting – a First Glimpse into 2013. INSPIRE Conference 2013, Florence, Italy, 24-27 June 2013.

Spyratos, S., Stathakis, D., Lutz, M., Tsinaraki, C., submitted. Using Foursquare place data for estimating building block use: A case study in Amsterdam, the Netherlands.

Trilles, S., Schade, S., Belmonte, O., Huerta, J., (in press). *Real-time anomalies detection from environmental data streams*. Full-paper contribution (Springer proceedings) to AGILE 2015.

Trilles, S., Schade, S., Belmonte, O., Huerta, J., submitted. A methodology to analyze sensor data streams in real time - with a proof of concept implementation for anomaly detection from environmental data.