Mind the Gap: Big Data vs. Interoperability and Reproducibility of Science

Max Craglia and Stefano Nativi

The global landscape in the supply, creation and use of geospatial data is changing very rapidly with new satellites, sensors and mobile devices reconfiguring the traditional lines of demand and supply, and the number of actors involved. As the volume, heterogeneity and rapidity of change of the data increases many organisations worldwide are reflecting on how to manage and exploit Big Data. The opportunities are many for business, science and policy but so are the challenges at technical, methodological, organisational, legal and ethical levels. In this chapter, we situate the discussion of Big Data in the context of the increasing challenges of the scientific method in a world of contested politics, in which science can no longer be seen as "neutral". We argue for a more open and participative science starting from the shared framing of problems across multiple stakeholders. In this context, the reproducibility of science is not just about the ability to repeat an experiment but also about the transparency of the process leading to a shared outcome. Opening up science to make it truly participative will need a major paradigm shift. It also needs an underpinning information infrastructure geared towards sharing data, information and knowledge across multidisciplinary and transdisciplinary boundaries. We use the development of the Global Earth Observation System of System (GEOSS) as a case study, because it highlights well the nature of these challenges when handling multidisciplinary Big Data across more than 80 countries and 90 international organisations. As we show, there is an increasing gap between the rapidity of technological progress and the slow pace of the organisational and

M. Craglia (✉)
European Commission Joint Research Centre, Ispra, Italy
e-mail: massimo.craglia@ec.europa.eu

S. Nativi
Italian National Research Council, Institute of Atmospheric Pollution Research (CNR - IIA), Rome, Italy
e-mail: Stefano.nativi@cnr.it

P.-P. Mathieu, C. Aubrecht (eds.), *Earth Observation Open Science and Innovation*, ISSI Scientific Report Series 15, https://doi.org/10.1007/978-3-319-65633-5_6

cultural change needed to address interoperability, reproducibility and legitimacy challenges effectively.

Introduction: the Big Data Paradox

We are living a paradox: at one level we have quantities of data at our disposal like never before to support scientific research, with claims of a new dawn or paradigmatic shift towards data science, the so-called Fourth Paradigm (Hey et al. 2009). At the other, we have an increasing mistrust in scientists, and a "crisis" in science with evidence of increased malpractice, irreproducible evidence and faked results, which Big Data are only likely to exacerbate (Benissa et al. 2016, Economist 2013, Nature 2014, 2015). In this chapter we explore this paradox, and look at the evolution of scientific approaches towards more participative, open and shared knowledge creation. We then look at the implications this has for information systems and Big Data handling using the development of the Global Earth Observation System of System (GEOSS) as a case-study, because it highlights well the nature of these challenges when handling multidisciplinary Big Data across more than 80 countries and 90 international organisations. We conclude with a reflection on the role of Big Data in the new world of participative, post-normal science.

It's All in the Framing!

The reality and rhetoric, of Big Data, or the Data Deluge, are difficult to grasp. SINTEF[1] (2013) for example indicated that 90% of all data in the world had been generated in the previous 2 years, while Turner et al. (2014) suggested that the "digital universe" will grow at 40% a year for the next decade, reaching some 44 trillion gigabytes. This abundance of data leads the Research Data Alliance, an international initiative led by the USA, Australia and the European Commission to facilitate the opening up of scientific data, to use the metaphor of the "data harvest" (RDA 2014), and claim that a bright new future is around the corner:

> when data volumes rise so high, something strange and marvellous happens: the nature of science changes. Problems that were previously not even recognised suddenly become tractable. Researchers who never met, at different institutions and in divergent fields, find themselves working on related topics. Work that previously plodded along from one experiment or hypothesis to another can accelerate .... Why should we care? Because, just as the World Wide Web has transformed our lives and economies, so this new data wave

[1] http://www.sintef.no/en/news/big-data--for-better-or-worse/.

> will matter eventually to every one of us, scientist or not. In the first instance, developing the tools, systems and businesses required for this will create jobs, revenues and economic growth (RDA 2014, p. 5).

To reap the benefits of this abundance, there are issues to be addressed with respect to data management, incentives to data sharing, tools and methods, and data skills (ibid.) but these are tractable problems if there is sufficient political will, and investment, as advocated by the Research Data Alliance (RDA): "Europe's leaders, [. . .] must act—or go down in history as the politicians who missed the Next Big Thing" (RDA 2014, p. 6).

Against this optimistic outlook, confidence in science is being shaken by increasing reports of malpractice and lack of reproducibility, which is at the basis of the "scientific method". For example, Begley and Ellis (2013) reported that 47 out of 53 seminal publications in hematology and oncology could not be reproduced. Similarly, Robert-Jan Smits, Director General of the European Commission Directorate for Research and Innovation, reported at the Fourth Plenary of the Research Data Alliance in 2014 that the reproducibility of scientific research was often as low at 10–30%, thus arguing for greater transparency of methods and access to data (Smits 2014). The lack of reproducibility, and thus accountability, may also hide deliberate bias or manipulation as indicated by the increasing number of papers retracted and the developments of studies and tools to uncover fraudulent behavior. For example, Markovitz and Hancock (2015) analysed a corpus of 253 retracted papers to find language patters that signaled fraudulent data reporting, Newman (2013) reports on two initiatives to detect data and image manipulation in scientific articles, while Springer and the University Joseph Fourier in Grenoble launched SciDetect in 2015, an open source software that "discovers text that has been generated with the SCIgen computer program and other fake-paper generators" (http://scidetect.forge.imag.fr/).

The concept of the reproducibility of scientific results was set in the context of the experimental sciences, in which the scientist had control over experiments, methods, and the generation, and "ownership", of the data. In this sense, scientific enquiry based on Big Data, i.e. on vast volumes of rapidly changing, highly heterogeneous, and distributed data not "owned by the scientist" faces many additional challenges because of loss of control over the data, as well as the algorithms that generate the data that may be proprietary, not accessible, and also changing frequently like the APIs of popular search engines or micro-blogging companies (Mei-Po Kwan 2016).

Ostermann and Granell (2015) make a useful distinction between *reproducibility* and *replicability*:

> Reproducibility is . . . concerned with the validity of the results of that particular study, i.e. the possibility of readers to check whether results have been manipulated, by reproducing exactly the same study using the same data and methods. Replicability is more concerned with the overall advancement of our body of knowledge, as it enables other researchers to conduct an independent study with different data and similar but not identical methods, yet arriving at results that confirm the original study's hypothesis. This would be strong evidence that the original study has uncovered a general principle through inductive research, which now another study has proven in deductive research design.

Therefore, reproducibility requires full access to both data and methods used. Replicability is more modest, but not less useful, and requires access to a description of the method or pseudo-code and access to metadata describing how the data was collected and its context, even if the original data set is not accessible. The many open data initiatives around the world (see for example http://www.opendataenterprise.org) and efforts of the Research Data Alliance, CODATA, GEO, and other international organisations are important to increase both reproducibility and replicability, and thus transparency of the scientific process.

Important as they are, these initiatives still frame reproducibility and replicability in the traditional ("modern") scientific discourse in which science is separated from society and decision-making, facts are separated from values, and there is one single reality (truth) that the scientist can discover to then advise decision-makers with neutral evidence.

This "positivist" model, still prevalent in the physical sciences, is of course based on an abstraction that falls rapidly apart at the interface between science, policy and society in our increasingly complex and globalized world. Here, there are no facts of nature, but only socially-constructed objects (Latour 1993) in which disciplines play an important role in framing the production of knowledge through discursive practices (Foucault 1980). Using urban planning as an example of a field at the intersection between policy-making and social science, we see the transition from "modern" to "post-modern" interpretations on the role of science and knowledge.

Up until the 1970s, urban planning was characterised by engineering approaches underpinned by management science, and neoclassical economics in which individuals make rational decisions based on perfect information. This "positivist" style of planning echoed the scientific approach of natural sciences, and assumed that it was possible to "objectively" understand reality, develop and test hypotheses, and develop universal laws of cause and effect on which to base predictions. Complex transport and city models were thus developed on the assumption that it was possible to predict the future and provide resources accordingly. From a socio-political point of view, this approach worked until there was strong economic growth and a post-war consensus on society's goals (Silva et al. 2015).

With the economic crisis of the 1970s, this social consensus broke down whilst several environmental and civil right movements pointed to the raising environmental costs of our model of development and the widening inequalities in society. Post-modernisms emerged as the new intellectual paradigm with a stinging critique of "positivist" science when applied to the social realms. Post-modernists would argue that we can never grasp reality in an "objective" fashion, but only interpret it based on our own experiences, values, and cultures. This has given rise to a reflective planning approach (Healey 2006) in which practitioners and researchers seek to expose the assumptions underpinning their work and confront them openly with the value systems of other stakeholders. The analysis of spaces, typical of quantitative methods was combined the analysis of places, which are defined by cultural identities and dynamic relationships in the physical and social

environments. This "interpretative" planning approach does not assume that there is a single reality, but accepts that there are multiple, equally valid, realities held by different groups in society.

The trajectory of the planning discipline reflects current debates on the crisis of science (Benissa et al. 2016) and the emergence of a "post-normal science" (PNS), which is an approach designed to apply "where facts are uncertain, values in dispute, stakes high, and decision urgent" (Funtowicz and Ravetz 1993, 744).

> In PNS the focus is on participation, legitimacy, transparency and accountability. In the "extended participation model" (Funtowicz 2006) deliberation (on what to do) is extended across disciplines ... and across communities of experts and stakeholders (Saltelli et al. 2016b, p. 20).

In other words, both post-modern and post-normal science analysts alert us that the traditional model of science is no longer adequate in a globalized world with hotly contested social, political, and environmental issues, in which science is not "neutral" but agent of different economic and commercial interests. There is not a single problem space in which to search for answers, but multiple spaces with competing values and views. This is an absolutely crucial point: if we believe in the primacy of our view of the world, then the problems of complexity and disagreement are addressed by more data, more processing, and more tools (e.g. Big data, Internet of Things, High Performance Computing) on the one hand, and "educating" those which "do not get it" on the other. To note that in this context, the calls for increased public participation, open data and citizens science to open up and help address the "crisis of science" (Saltelli et al. 2016b) are not enough if these resources are co-opted to contribute to a pre-defined problem space.

By contrast, the acknowledgement that there are multiple, and legitimate, different problem spaces and perspectives calls for humility and reflexivity, for openness and participation in finding a shared "framing" of the problem first, i.e. a collective understanding of what are the important questions to ask, and only then defining the methods, data, and tools to address them. Multidisciplinarity, transdisciplinarity (i.e. involving not just other disciplines but also non-academic stakeholders, see Vaccari et al. 2012), public participation and citizen science are crucial but only if already involved from the beginning, at the initial stage of framing the problem space (i.e. *frame first*, *compute later*).

Towards Open (Shared) Knowledge

Developing shared framing of what is the problem typically requires sharing of data, information knowledge, often tacit assumptions, values among key stakeholders. An initial crucial challenge is identifying and getting all relevant stakeholders around a table, nurturing and then harnessing the necessary willingness to talk and

compromise, having a mediator with the right set of skills (e.g. Time 2015)[2]. A supportive information infrastructure is needed throughout the iterative stages of framing (what are the questions to ask), exploring alternatives, assessing through the lenses of different stakeholders with respect to:

- Feasibility: compatible with boundary conditions beyond human control;
- Viability: compatible with internal structures and their control system;
- Desirability: compatible with the values of the different stakeholders (Saltelli and Gianpietro 2016a)

The information infrastructure supporting these processes needs to enable the sharing of data but also, and crucially in a multidisciplinary and transdisciplinary context, also the context in which the data was collected/produced, the methods, workflows, and models that are often associated with the data to generate information and knowledge, and the outcomes/results that need then to contribute to the generation of the alternatives to be assessed. Open Data is just a first stage, Open (Shared) Knowledge is the goal requiring many steps in between, particularly to make explicit much of the tacit assumptions, values and knowledge that are necessary to understand and meaningfully use the data. The conceptual framework can be represented as in Fig. 1, where the Reproducibility of Science needs to be understood in the context not just of reproducing an experiment, but also of a transparent shared process of reaching an outcome.

An example of what is required to build the necessary semantic bridges across disciplines is provided in Europe by the INSPIRE Directive (2007/2/EC) which is a legal framework to share the data provided by the infrastructures for spatial information established and operated by the 28 member states of the European Union (http://inspire.jrc.ec.europa.eu/). INSPIRE was set up to support environmental policy in the EU, and addresses 34 spatial data themes needed for this purpose (Table 1).

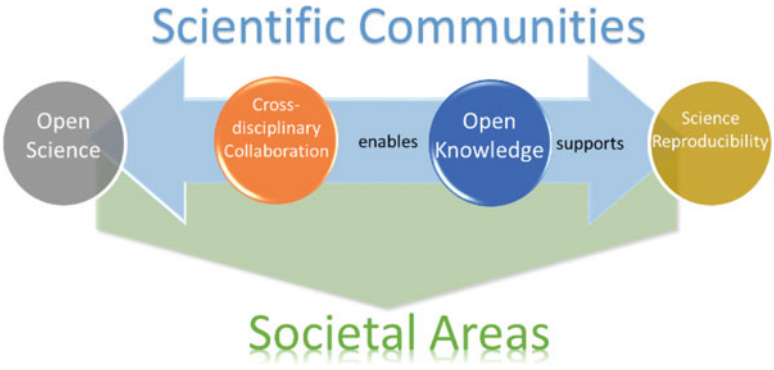

Fig. 1 Conceptual framework and rationale for knowledge sharing and open science

[2]http://time.com/3859497/communication-negotiation-basics/.

Table 1 Key data themes addressed by INSPIRE

Annex I	**Annex III**
• Addresses • Cadastral parcels • Transport networks • Hydrography • Protected sites • Coordinate reference systems • Geographical grid systems • Geographical names • Administrative units	• Statistical units • Buildings • Soil • Land use • Human health and safety • Utility and governmental services • Environmental monitoring facilities • Production and industrial facilities • Agricultural and aquaculture facilities • Population distribution—demography • Area management/restriction/regulation zones and reporting units • Natural risk zones • Atmospheric conditions • Meteorological geographical features • Oceanographic geographical features • Sea regions • Bio-geographical regions • Habitats and biotopes • Species distribution • Energy Resources • Mineral resources
Annex II	
• Elevation • Land cover • Ortho-imagery • Geology	

Each of these themes involves communities of scientists and practitioners in public administrations across 28 countries and 24 languages. To share data across these different communities has already the best part of 10 years of work to document them through metadata, making the data searchable, viewable and accessible through catalogues and related services. To date the INSPIRE geoportal[3] contains some 120,000 datasets documented with agreed metadata. This is an important first step because INSPIRE was the first Directive to introduce harmonized rules to document datasets and make them searchable in Europe. Nevertheless, the most difficult challenge in INSPIRE is to achieve the interoperability of datasets, i.e. arriving at a shared understanding of the structure of the datasets and the meaning of the variables they contain. To do so, it was necessary to identify and mobilise the relevant multidisciplinary communities in each of the 34 data themes, and through a patient process of reviewing, refining, and agreeing arrive at shared (generalized) data models that define the structure, content, and meaning of the data needed to support environmental policy. It took some 6–7 years to reach these agreements across hundreds of stakeholder organisations in the member states, and it will take another 10 years to "translate" the existing data in the Member States to the new European models.

[3] http://inspire-geoportal.ec.europa.eu/discovery/.

A visit to the INSPIRE website[4] in the data specifications' sections demonstrates the huge amount of work involved. There are thousands of pages of specifications and not to be forgotten, tens of thousands of comments that had to be addressed individually during the stakeholders' consultations. The INSPIRE Registry (http://inspire.ec.europa.eu/registry/) is a repository of the agreed definitions, codelists, and dictionaries necessary to underpin the interoperability across thematic layers. The process was long because there were few agreed standards to draw on, and reaching agreement is a slow process when the financial and organizational stakes are high. Although the implementation of INSPIRE takes a long time with variable degree of progress (Ansorge et al. 2014), it is underpinned by European legislation, which requires its implementation across the EU. In the next Section, we use the case-study of the Global Earth Observation System of Systems (GEOSS) to look at the technical and organizational issues to be faced when scaling up to the global context, with millions of datasets from multiple disciplines, and a voluntary, rather than legal framework.

The GEOSS Case

The Group on Earth Observation (GEO)[5] is a voluntary partnership of governments and international organizations launched in response to calls for action by the 2002 World Summit on Sustainable Development and by the G8 (Group of Eight) leading industrialized countries. These high-level meetings recognized that international collaboration is essential for exploiting the growing potential of Earth observations to support decision making in an increasingly complex and environmentally stressed world. To this aim, GEO is coordinating efforts to build a Global Earth Observation System of Systems, or GEOSS (GEO 2005). GEOSS is intended as a global and flexible network of content providers allowing decision makers to access an extraordinary range of data and information at their desk.

GEO is developing GEOSS based on cycles of 10-Year Implementation Plans (the first period went from 2005 to 2015 and the new one will end in 2025) (GEO 2012, 2016). The Plans define a vision statement for GEOSS, its purpose and scope, expected benefits, and the targeted "Societal Benefit Areas" (SBAs) (i.e. *Biodiversity and Ecosystem Sustainability*, *Disaster Resilience*, *Energy and Mineral Resources Management*, *Food Security and Sustainable Agriculture, Infrastructure and Transportation Management*, *Public Health Surveillance*, *Sustainable Urban Development*, *Water Resources Management*).

[4] http://inspire.jrc.ec.europa.eu/index.cfm.

[5] http://www.earthobservations.org/.

A key achievement of GEO has been to agree on common Data Sharing Principles:

- Data, metadata and products will be shared as Open Data by default, by making them available as part of the GEOSS Data Collection of Open Resources for Everyone (Data-CORE) without charge or restrictions on reuse, subject to the conditions of registration and attribution when the data are reused;
- Where international instruments, national policies or legislation preclude the sharing of data as Open Data, data should be made available with minimal restrictions on use and at no more than the cost of reproduction and distribution; and
- All shared data, products and metadata will be made available with minimum time delay.

The implementation of these principles takes the form of the GEOSS Data-CORE which now contains more than five million datasets. This is a very significant success given the heterogeneity of the organisations participating in GEO. As awareness of this pool of open data increases, and applications are built using these datasets, it is increasingly important that the data is well managed and dependable. For this reason, GEO has recently adopted also a set of Data Management Principles addressing discoverability, accessibility, usability, preservation and curation (see http://earthobservations.org/dswg.php). Making the data easy to share and well managed is clearly important to underpin transparency, accountability, and reproducibility. They need however also to be embodied into an information infrastructure, as described below.

Realizing a System-of-Systems, GEOSS is composed of contributed Earth Observation systems, ranging from systems collecting primary data, to systems concerned with the creation and distribution of information products. Although all GEOSS systems continue to operate within their own mandates and will evolve, GEOSS systems can leverage each other so that the overall GEOSS becomes much more than the sum of its component systems (GEO 2007). This is achieved by implementing a digital infrastructure (e-infrastructure) that coordinates access to these systems, interconnecting and harmonizing their data, applications, models, and products: the *GEOSS Common Infrastructure* (GCI).

The GCI is an instrument—realized as a third-party service layer—that interconnects—in a transparent way—the heterogeneous GEOSS data supplier systems and the applications developed (by public and private bodies) to serve GEOSS users (Nativi et al. 2012a, 2013). It provides a set of core services supporting the integration of Earth Observation resources available in the framework of GEO with the goal of setting up GEOSS as an operational System-of-Systems. It is also aimed at allowing GEOSS end-users to search, discover, evaluate and access the resources (e.g. data, information, tools and services) made available by the GEO Members (e.g. institutions, agencies, private industry) via their shared supply systems.

As depicted in Fig. 2, the GCI consists of three main components: the GEOSS Web Portal, the GEO Discovery and Access Broker (DAB), and the contributed

Fig. 2 High level GEOSS architecture and GCI

Resources Registry. The GCI relies upon a set of interoperability agreements that ultimately aim at defining rules for tackling existing incompatibilities with the goal to facilitate integration and interaction of heterogeneous components and systems in GEO (GEO 2012).

Big Data Infrastructure Services: The GEOSS Common Infrastructure (GCI) Big Data Strategy

GEOSS provides access to more than 200 million single datasets (i.e. single files), as of January 2016, characterized by a large heterogeneity—in terms of content, ranging from individual sensor observations to high-level environmental indicators and indexes. Therefore, the development of GEOSS and the GCI poses challenges along all the Big Data dimensions—in particular, volume, variety, velocity and veracity These challenges and the way they have been addressed is discussed by Nativi et al. (2015) and are summarized in Table 2. The strategy has centred on the continuous development of the GEOSS Discovery and Access Broker (DAB), and the use of cloud services to develop a public cloud-based software ecosystem that characterizes the present GEOSS Information Systems. We highlight below some key aspects as they provide a practical implementation of the effort to share not just data but also knowledge and the way it is generated.

Table 2 GCI and GEOSS strategies and solutions to address Big Data challenges (source: Nativi et al. 2015)

Big Data challenges		Solutions adopted to address the challenges
Volume	Discovery Challenges high number of catalogs, inventory, listing services to be brokered; Large number of metadata records; Large number of Users' discovery requests	Reduce the number of matching results, by supporting advanced constraints in addition to the more traditional "what", "where", "when". Design and apply a ranking metrics and related paging strategy. Support distributed queries, along with harvesting approach, to reduce the number of large metadata records to be stored and managed by the DAB. Use of load balancing and auto-scaling clusters to support large number of queries.
	Access Challenges high number of data services to be brokered; large amount of datasets; big data volume; Large number of Users' access requests	Use of server-side transformation functionalities to limit downloaded data. Supplement missing transformation functionalities (not supported by data servers). Support data caching and map tiling. Use of load balancing and auto-scaling clusters.
Variety	Discovery Challenges Support of highly heterogeneous metadata models and discovery service interfaces; Publication of the set of metadata models and discovery interfaces implemented by GEOSS Users' applications; Long-term data access sustainability in a multidisciplinary environment	Introduction of a brokering tier dedicated to mediation of service interfaces and metadata models harmonization in a transparent way for both Users and data providers. Design and implementation of a brokering semantic and metadata model used. Extensible architecture of brokering to support new service interfaces and metadata models.
	Access Challenges Support of highly heterogeneous data models, encoding formats, and access service interfaces; Publication of the set of data models, encoding format, and access interfaces implemented by GEOSS Users' applications; Long-term data access sustainability in a multidisciplinary environment	Introduction of a brokering tier dedicated to mediation of access service interfaces and data formats harmonization in a transparent way for both Users and data providers. Design and implementation of a brokering data model used to: (1) harmonize and integrate the heterogeneous data formats brokered by GEOSS; (2) expose the data formats well-supported by GEOSS Users. Extensible architecture of brokering to support new access service interfaces and data formats. Transformations facilitating re-use.

(continued)

Table 2 (continued)

Big Data challenges		Solutions adopted to address the challenges
Velocity	Discovery Challenges To manage the increasing rate at which metadata flows; Fast metadata processing to satisfy Users' needs	Operational data store that periodically extracts, integrates and re-organizes brokered metadata records for operational inquire and ranking generation. Caches that provide instant access to the results of distributed queries while buffering data provider systems from additional load and performance degradation. Design of the DAB architecture that balances metadata latencies with GEOSS Users' requirements, avoiding assuming that all data must be near-real time. Incremental harvesting strategy. Live query distribution combined with caching of results. Load balancing to route incoming requests to machines with lowest workload. Use of auto-scaling clusters to increase computing capacity in response of rapid workload growth.
	Access Challenges To manage the increasing rate at which data flows; Fast data processing to satisfy Users' needs	Operational data store that periodically generates and stores preview tiled maps of brokered data for operational data preview. Caches that provide instant access to the results of previous access requests. Supplementing missing transformations allows limiting the local processing time. For extremely large processing requests, Users are allowed to opt for an asynchronous version of the access functionality.
Veracity, Value, and Validity	Challenges Reduce the "information noise"; Retrieved data comparison; Data trustiness for GEOSS decision makers; Effective data re-use; Data meaningfulness for User requests; Data accuracy for intended use	The brokering data model includes a specific multidisciplinary quality extension. Implementation of a flexible ranking metrics including quality of service and metadata completeness as valuable indexes. The brokering metadata model supports a harmonized presentation of retrieved metadata facilitating their comparison. Use of GEOSS Essential Variables as an additional parameter for improving the existing ranking metrics. The prototyped "fit-for-purpose" and Users' feedback extensions aim to provide Users with quality-aware results.
Visualization	Challenges Visualization speed; Contextualized visualization	Support Community Portals and Applications publishing DAB APIs for client development. Support the following visualization strategy: (1) provide an overview (trying to keep that simple and show important elements), (2) allow zoom and filter unnecessary clutter, (3) provide more details if requested by Users. Provide fast previews by generating preview tiles in batch.

Content Harmonization and Information/Knowledge Generation: The Brokering Framework

The Brokering framework

In a complex ecosystem of domain infrastructures like GEOSS, multidisciplinary interoperability has been traditionally pursued on a one-to-one basis or by asking the stakeholders (i.e. resource providers and consumers) to be able to utilize the plethora of interoperability standards (both international and Community-based) characterizing the different disciplinary systems. Clearly, this has represented a high entry barrier for developing cross-disciplinary science and applications (Nativi et al. 2013). For this reason, a new solution was proposed first by a European FP7 project (Vaccari et al. 2012) and then by a US-NSF initiative (Nativi et al. 2011), namely: the Brokering approach.

The Brokering approach follows these principles to make existing infrastructures and data systems interoperable, in a System-of-Systems (SoS) framework (Nativi et al. 2012a; 2013):

1. To keep the existing capacities as autonomous as possible by interconnecting and mediating between standard-based and non-standard-based capacities.
2. To supplement, without supplanting, the individual systems' mandates and governance arrangements.
3. To assure a low entry barrier for both the resource providers and the end users.
4. To be flexible enough so as to accommodate the existing systems as well as future ones.
5. To build in an incremental fashion upon the existing infrastructures (information systems) and incorporate heterogeneous resources by introducing distribution and mediation functionalities.
6. To specify interoperability arrangements focusing on the modularity of interdisciplinary concepts rather than just on the technical interoperability of systems.

The Brokering approach relaxes the requirement for implementing a common data model and exchange protocol, providing the necessary mediation and transformation functionalities in a transparent way to the SoS components. It builds on existing data systems and federation systems, complementing the federation approach: for SoS Engineering, brokering architecture addresses those challenges that are not solved by federated systems. In brokered systems, interoperability is then in charge of dedicated interconnection (i.e. mediation and transformation) components—the *brokers*—managed by a third-party and deployed in the SoS common infrastructure (Nativi et al. 2013). End systems need only to formally agree to the participation in the SoS and just document the interfaces and data models that they already adopt.

This makes the brokered approach flexible and applicable in very heterogeneous and distributed environments when the overarching organization has not any possibility to enforce the adoption of a common model for information sharing.

Moreover, since the information technology complexity (required to interoperate) lies in the brokers, a brokered SoS does not require strong information technology skills to participants (e.g. data suppliers and application developers).

The DAB

One of the key components to achieve multidisciplinary interoperability of the GCI is the GEO Discovery and Access Broker (DAB)[6]. This component stems from work done in the EuroGEOSS project[7] funded by the European Commission Seventh Framework Programme, and implements the Brokering approach for multidisciplinary interoperability in GEOSS.

Any request received by the GEOSS Web Portal (see Fig. 2) is forwarded to the DAB, which connects user requests to an ever-increasing number of databases and information systems around the world—i.e. the GEOSS resources supply system provided by the SoS enterprise systems. DAB applies the brokering principles to interconnect the many enterprise systems constituting GEOSS, the global SoS managed by GEO. Through the DAB services, GCI relaxes the requirement for implementing a common data model and exchange protocol, providing the necessary mediation and transformation functionalities in a transparent way to the SoS components—see Nativi et al. (2006) and Nativi and Bigagli (2009).

The DAB supports more than 50 well-used and standard protocols, commonly implemented by the GEOSS data and information and service suppliers to share their resources, harmonizing them to provide a unique and consistent response to the GEOSS user requests.

The DAB exposes a set of well-used standard Internet interfaces and high-level JavaScript APIs[8] (Application Program Interfaces) enabling the Developers stakeholders to implement applications and sophisticated downstream services for the end Users. The APIs implements discoverability, accessibility and simple transformations (i.e. data encoding transformation, coordinate reference systems mapping, data subsetting and data resolution change) functionalities.

More (Value) knowledge to reduce Volume

GEOSS is required to be able to manage any Earth Observation resource considered useful to study Global Changes. Usually, the amount of datasets and their individual size decrease moving from low-level observations (such as sensor raw data) to high-level data and information—like Essential Variables and primary indicators. In other words, it is important to recognize and manage the right level of data required by Users.

[6] http://www.geodab.net/.

[7] http://www.eurogeoss.eu.

[8] http://api.eurogeoss-broker.eu/docs/index.html.

In Big data terminology, this is related to the "Value" feature: a system should focus on information that is more relevant for its Users, and preliminary select the right sources. This is especially true for those systems that are not general-purpose (e.g. Web search engines) but have a clear definition of Users and use scenarios. GEOSS objective is expressed as: "exploiting the growing potential of Earth observations to support decision making". This means that only the information useful for decision-making should be delivered by GEOSS. The on-going GEO activity on the identification of "essential variables" and primary indicators goes in that direction. Focusing on the delivery of relevant content, representing essential variables and indicators for the eight GEO SBAs, would reduce the Big Data requirements for the GCI and GEOSS.

Further considering users' requirements and feedbacks (e.g. dataset fit-for-purpose) would also help to solve other issues concerning the Volume aspects, such as dataset granularity. Datasets could be (virtually) aggregated according to users' needs (e.g. time series) instead of providers' convenience. While focusing on users' requirements would probably reduce the amount of datasets made accessible through the GCI, on the other hand it would require implementing smarter functionalities. For instance, Users' needs might guide smart ranking strategies for the presentation of query results. The system might also be tailored to and provide functionalities for refined queries, based on users' needs, instead of explicit requests. The typical GEOSS User (e.g. decision-maker or scientific expert) should be allowed to express queries in terms of scientific or societal challenges instead of data parameters.

The identification, by the different scientific communities, of Essential Variables for the GEO SBAs and Communities of Practice will presumably help to relieve the big volume aspects in GEOSS. However, this, and more generally the shift from data to information and knowledge, will introduce new kind of resources to be managed by the GCI—such as knowledge bases, ontologies, environmental and ecological models for the generation of significant indicators. The GEO Model Web initiative has started envisioning a possible architecture and discussing technological and non-technological issues (Nativi et al. 2012b).

Recently, GEOSS recognized the need to try addressing Big Data challenges by identifying and satisfying high-level Users' needs. From the GEO Community, there is a consensus on auspicating the GEOSS evolution from a data infrastructure to an information and knowledge system.

From Data to Knowledge: the GEOSS Knowledge Base and High-Performance Data Analytics

The DIKW pattern

Evolving GEOSS from a data infrastructure to an information system entails to understand and connect shared resources: information is an added-value product

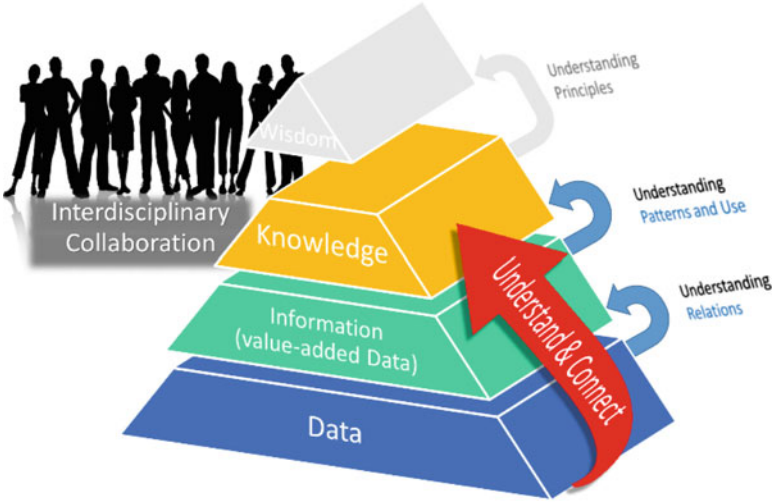

Fig. 3 The DIKW pattern

generated by understanding data and working out relations among them and with physical and/or social phenomena; while, understanding information and working out valuable patterns generates knowledge. GEOSS is required to gradually implement the DIKW (Data, Information, Knowledge, Wisdom) model (Zins 2007), as depicted in Fig. 3.

To apply the DIKW pattern in a transparent and open way is extremely important in order to enable Open Science and allow results reproducibility. The understanding and connection principles and rules, applied to generate information first and then knowledge, must be accessible and re-usable in order to allow science reproducibility.

The GEOSS Knowledge Base

An important objective of the GEOSS Knowledge Base is to collect and share the relations, patterns, principles, rules and implementation instruments that the GEOSS SBAs commonly use to generate information and knowledge from the Earth Observations.

GEOSS Knowledge Base must closely interoperate with the brokering framework (e.g. the DAB) that provides harmonized and consistent documentation on available Earth Observations—solving the Big Data variety challenge. On the other hand, the DAB must leverage the Knowledge Base content to advance the present discoverability capability by understanding and formalizing meaningful links among the available Earth Observations and with other related resources—documents, models, etc.

It is envisioned that GEOSS (intermediate and final) users can access such a Knowledge Base to understand the provenance of information accessed via GEOSS and, where meaningful, to reproduce results. Another important use case considers users getting the necessary knowledge for generating more information from the shared resources. Users should be able to run (complex) workflows, collected and shared by the Knowledge Base and discovered and accessed via the DAB. This requires to advance the DAB and evolve GEOSS and the GCI to improve data accessibility and allow data processing—in other words, to further address the Big Data analytics challenges: velocity, volume and variety.

High-Performance Analytics and GEOSS

To discover patterns and generate useful information from its shared resources, GEOSS has to face an important and new challenge: to keep its SoS nature while evolving to leverage the High-Performance analytic capabilities offered by the innovative infrastructures—i.e. Clouds, HPC, Grids, etc.

Considering the GEO scope and organizational structure, GEOSS is a "System of Systems" and its success depends on building interoperability among the different and autonomous systems shared by GEO members, presently and in the next future. This makes of GEOSS a significant framework to advocate the feasibility and benefits of Open Science.

In keeping with its SoS nature, GEOSS introduced a set of architectural principles as the basis for evolution and ensure interoperability with relevant research and policy-driven (data) infrastructures:

- Openness;
- Effectiveness;
- Flexibility;
- Sustainability;
- Reliability;
- Support the implementation of quality principles—i.e. the GEO Data Management principles.

These principles were considered to design and implement the present GCI building on the existing Data Systems and being flexible enough to support the next coming ones. The same should be done for implementing a GEOSS High-Performance Analytic capability by building on existing high-performance computing infrastructures and being flexible enough to include the next ones.

Cloud/Infrastructure brokering solutions play an important role (as the DAB does for the data systems). This is a third-party technology that acts as an intermediary between the consumer of a cloud/infrastructure storage/computing service and the provider of that service. In general, it is an intermediary between the GCI and the many available cloud/infrastructures providing storage and computing services. Figure 4 shows a possible System-of-Systems architecture to leverage the Big Data

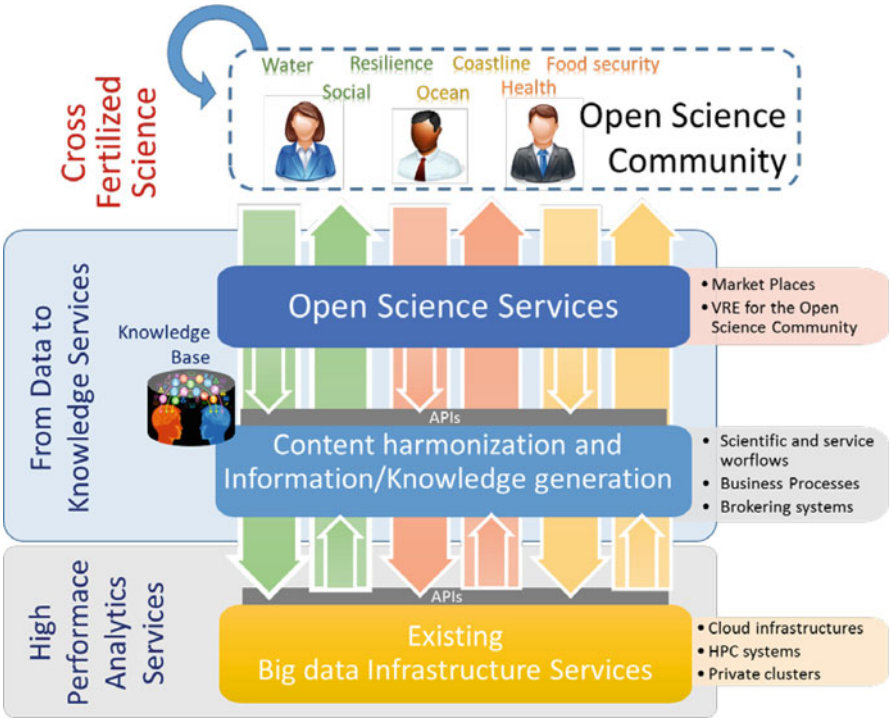

Fig. 4 System-of-Systems architecture leveraging Big Data Analytics to move from Data to Knowledge

Analytics and generate Knowledge from Data. Table 2 summarises the key Big Data Challenges facing GEOSS and how they are being addressed, as discussed in Nativi et al. (2015).

From the point of view of openness and shared opportunity to participate in framing decision spaces and contributing meaningfully to debates, access to distributed processing and cloud services is particularly interesting because it means that even in a Big Data world it is not necessary to have your own high-cost infrastructure for data processing, but it is sufficient to use existing services when needed. This in principle, democratizes access to processing and sense-making from the vast amount of data available.

Conclusions

In this chapter we have situated the discussion on Big Data into the broader framework of the challenges faced by science today when advising policy, or more generally when addressing topics that have social, economic, and environmental

implications. The increasing lack of trust in policy (and politicians), and science (and scientists) by civil society needs to be addressed with greater humility and reflexivity and engage into meaningful forms of participation and dialogue at the very early stages of the process, when the problems are framed, not after the direction is set and decisions are taken. Open participation that recognizes the legitimacy of different viewpoints and perspectives, needs to be underpinned by a shared information infrastructure enabling access and "meaningful" use of the data needed to support one's position in the initial framing and debate. By "meaningful" we mean here ability to access not just the data, but also the context giving meaning to the data (how it was collected by whom, for what purpose, with what methods, definitions, classifications ...), and the methods used to extract information from the data (e.g. algorithms, models, analytical steps), which in turn are underpinned by theories and often tacit assumptions that need also to be made explicit to avoid misunderstanding. From open data, we need to move to open knowledge and shared infrastructures and tools accessible and usable by the different interests. We introduced GEOSS and its common infrastructure (the GCI) as an example of this move from a data infrastructure to a knowledge-base one. Given the complexity of building a global multidisciplinary system of systems and the voluntary nature of this initiative, we do not claim that GEOSS has succeeded in addressing all the challenges. We are for sure a long way from that. Nevertheless, it provides a good example of a strategy to address the issues, in which the ethos of mediation, or brokering, across multiple disciplines and stakeholders in a global setting is not just a technical approach but a philosophical one that recognizes the legitimacy of the many "others", and draws strength from openness and diversity.

Acknowledgements This research was partially funded by the European Commission, grants number: 641538 (ConnectinGEO project), 620400 (ENERGIC-OD Project), 641762 (ECOPOTENTIAL project) and 620400 (ENERGIC-OD project). The authors would also like to thank the ESA, USGS, OGC, and IEEE for their cooperation within the GCI.

References

Ansorge C, Craglia M et al. (2014) Mid-term evaluation of INSPIRE, EEA: Copenhagen. http://www.eea.europa.eu/publications/midterm-evaluation-report-on-inspire-implementation/at_download/file. Accessed Mar 31 2016

Begley CG, Ellis LM (2013) Drug development: raise standards for preclinical cancer research. Nature 483:531–533. http://www.nature.com/nature/journal/v483/n7391/full/483531a.html. Accessed Apr 4 2016

Benissa A, Funtowicz S et al. (2016) The rightful place of science: science on the verge. Consortium for Science, Policy & Outcomes, Arizona State University

Economist (2013) How science geos wrong. October 19th. http://www.economist.com/news/leaders/21588069-scientific-research-has-changed-world-now-it-needs-change-itself-how-science-goes-wrong. Accessed Apr 4 2016

Foucault M (1980) Power/knowledge: selected interviews and other writings, 1972-1977. Pantheon, New York, NY

Funtowicz S (2006) Why knowledge assessment? Chapter 8. In: Guimarães Pereira A, Guedes Vaz S, Tognetti S (eds) Interfaces between science and society. Greenleaf Publishers, Sheffield

Funtowicz S, Ravetz J (1993) Science for the post-normal age. Futures 31:739–755. http://www.uu.nl/wetfilos/wetfil10/sprekers/Funtowicz_Ravetz_Futures_1993.pdf. Accessed Apr 4 2016

GEO (2005) GEOSS: 10-year implementation plan reference document, ESA Publications Division. http://www.earthobservations.org/documents/10-Year%20Plan% 20Reference%20Document.pdf. Accessed Apr 4 2016

GEO (2007) Strategic Guidance for Current and Potential Contributors to GEOSS, printed by JAXA on behalf of GEO Architecture and Data Committee, Oct 2007. https://www.earthobservations.org/documents/portal/25_Strategic%20Guidance%20Document.pdf. Accessed Apr 4 2016

GEO (2012) GEO 2012–2015 Work Plan, GEO publication. http://www.earthobservations.org/documents/work%20plan/GEO%202012-2015%20Work%20Plan_Rev2.pdf. Accessed Apr 4 2016

GEO (2016) GEO strategic plan 2016–2025: implementing GEOSS. https://www. earthobservations.org/documents/GEO_Strategic_Plan_2016_2025_Implementing_GEOSS.pdf. Accessed Apr 4 2016

Healey P (2006) Collaborative planning: shaping places in fragmented societies. Palgrave McMillan, Basingstoke

Hey T, Tansley S, Tolle K (eds) (2009) The fourth paradigm: data-intensive scientific discovery. Springer, Berlin, p 252

Latour B (1993) We have never been modern. Harvard University Press, Cambridge, MA, p 4. isbn:978-0-674-94839-6

Markovitz DM, Hancock JT (2015) Linguistic obfuscation in fraudulent science. J Lang Soc Psychol 35:435. https://doi.org/10.1177/0261927X15614605

Kwan M-P (2016) Algorithmic geographies: big data, algorithmic uncertainty, and the production of geographic knowledge. Ann Am Assoc Geogr 106:274. https://doi.org/10.1080/00045608.2015.1117937

Nativi S, Domenico B et al (2006) Extending THREDDS middleware to serve OGC community. Adv Geosci 8(8):57–62

Nativi S, Bigagli L (2009) Discovery, mediation, and access services for earth observation data. Select Top IEEE J Appl Earth Observ Rem Sens 2(4):233–240

Nativi S, Khalsa SJ et al. (2011) The brokering approach for Earth Science Cyberinfrastructure. EarthCube white paper, US NSF. http://semanticommunity.info/@api/deki/files/13798/=010_Domenic. Accessed Apr 5 2016

Nativi S, Craglia M, Pearlman J (2012a) The brokering approach for multidisciplinary interoperability: a position paper. Int J Spat Data Infrastruct Res 7:1–15. http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/281/319. Accessed Apr 5 2016

Nativi S, Mazzetti P, Geller G (2012b) Environmental model access and interoperability: the GEO model web initiative. Environ Model Software 39:214–228

Nativi S, Craglia M, Pearlman J (2013) Earth science infrastructures interoperability: the brokering approach. IEEE J Select Top Appl Earth Observ Remote Sens 6(3):1118–1129

Nativi S, Mazzetti P et al (2015) Big Data challenges in building the global earth observation system of systems. Environ Model Software 68:1–26

Nature (2015) Challenges in irreproducible research, Nature Special feature. http://www.nature.com/news/reproducibility-1.17552#/Editorial. Accessed Apr 5 2016

Nature (2014) Journals unite for reproducibility, Nature editorial. vol 512, Nov 2014. http://www.nature.com/news/journals-unite-for-reproducibility-1.16259. Accessed Apr 5 2016

Newman A (2013) The art of detecting data and image manipulation. https://www.elsevier.com/editors-update/story/publishing-ethics/the-art-of-detecting-data-and-image-manipulation. Accessed Apr 4 2016

Ostermann F, Granell C (2015) Advancing science with VGI: reproducibility and replicability of recent studies using VGI. Transactions in GIS. http://onlinelibrary.wiley.com/doi/10.1111/tgis.12195/full. Accessed Apr 5 2016

RDA (2014) The data harvest report – how sharing research data can yield knowledge, jobs and growth. A RDA Europe report, December 2014. https://europe.rd-alliance.org/sites/default/files/repository/files/TheDataHarvestReport_%20Final.pdf. Accessed Apr 4 2016

Saltelli A. Gianpietro M. 2016a. The fallacy of evidence-based policy. Benissa A et al. The rightful place of science: science on the verge. Tempe, AZ: Consortium for Science, Policy & Outcomes, 31-70.

Saltelli A., Ravetz J. Funtowicz S. 2016b. Who will solve the crisis in science? Benissa A et al. The rightful place of science: science on the verge. Tempe, AZ: Consortium for Science, Policy & Outcomes, 1-30.

Silva EA, Healey P, Harris N, Van den Broek P (eds) (2015) The Routledge handbook of planning methods. Routledge, New York, NY

Smits RJ (2014) Keynote at 4th RDA Plenary, Amsterdam. https://collegerama.tudelft.nl/Mediasite/Play/0844aefac5bb49ca9032069c6edc668f1d?catalog=3984a02f-bf33-4c70-a080-94a04d3e8112 (minute 16:07). Accessed Sept 2017

Turner V et al (2014). The digital universe of opportunities http://www.emc.com/leadership/digital-universe/2014iview/internet-of-things.htm. Accessed Apr 4 2016

Vaccari L, Craglia M, Fugazza C, Nativi S, Santoro M (2012) Integrative research: the EuroGEOSS experience. IEEE J Select Top Appl Earth Observ Remote Sens 5(6):1603–1611

Zins C (2007) Conceptual approaches for defining data, information, and knowledge. J Am Soc Inform Sci Technol 58(4):479–493