7. Mobilising data

Environmental data, technical and governance issues

Lorenzo Bigagli and Stefano Nativi

Introduction

The argument so far has covered some of the general issues and trends in the mobilisation of open data. Chapters Three and Four show that there are a set of requirements including technological and governance ones that need to be addressed and implemented to support open data. To understand how these requirements can be realised, Chapters Five and Six explored the issues facing institutions and how making data open interacts with, and may shape, existing research practices across a range of disciplines. To address issues about data, technology and governance in making data open in detail, this chapter focuses on one context where open data is being mobilised: the environmental sector. This area of research has engaged with making data open and continues to do so; in so doing, it has revealed the issues of developing open access to data. It is also an area of research that has many characteristics of Mode 2 Knowledge production and thus it is an example of late modern science. The fact that it has gained experience about knowing how to make data open from core aspects of open data, namely the technical, governance and data aspects, and how these are related to each other, makes it a useful case to discuss. When this is combined with its Mode 2 Knowledge production characteristics, the environmental or earth sciences are an exemplar of the issues in making data open. Further, its work links closely with many of the challenges that society is facing and it therefore brings out the role of data in society to address contemporary societal grand challenges.

‘Environment’ is used here in its broadest terms, to include land, oceans, the atmosphere, polar regions, life, the planet’s natural cycles and deep earth processes, along with the mutual influences of these constituent parts on one another. This definition of environment also includes human society, which is an integral part of this enormously complex system, also known as the earth system. The following sections provide a brief overview of the most relevant initiatives promoting open access policies to environmental data, particularly from the global and European perspectives. Next, the chapter describes some of the key technological and infrastructural issues that stakeholders in the environmental sector are facing in attempting to
implement open access policies to data. Last, it discusses the all-encom-
passing issue of governance of this highly-heterogeneous landscape, where
geopolitical, economical, industrial, legal, and institutional issues must all
be harmonised in order to enable our society to gain a full picture of the
earth system.

Open access and the earth system

Scientific research over recent decades has led to the conclusion that the earth
system has been changing outside its range of natural variability, at least in the
last half million years, possibly under the influence of human activities (Steffen
et al. 2004). Such planetary-scale changes in the earth system, including large-
scale changes in society, are referred to as ‘global change’, so the environmental
sector illustrates the importance of open data to benefit society.

In fact, global change is posing unprecedented difficulties to decision
and policymakers, because strategic goals for globally-sustainable develop-
ment need to be agreed upon at a planetary level. To enable measuring,
monitoring and assessing of these strategies and policies, we need shared
indexes based on sound environmental and socio-economic indicators
that, in turn, originate from fundamental physical variables and hence,
ultimately, from data.

Nowadays, a fundamental enabler of environmental research practice is
space-based earth observation. Many international bodies (e.g. those par-
ticipating in the United Nations Conference on Sustainable Development in
2012) recognise that earth observation from space is key to addressing global
change, as also testified by the significant investments availed in this so far.
These have demonstrated a positive return across a wide range of societal
benefit areas akin to environmental monitoring, such as humanitarian
aid, increasing food security, crime prevention and disaster management.
However, the high costs involved mean that no single country, programme
or industry can undertake this daunting endeavour alone. Instead, the whole
of human society needs to be engaged and mobilised, in both developed and
developing countries. In fact, developing countries often face an extremely
difficult dilemma between preserving their natural resources and industrial-
isation; hence, their impact on the environment may be very significant.

Understandably, in this context, the sharing of data, resources and knowl-
dge is considered more of an opportunity than a liability. In fact, several
global initiatives are pursuing data sharing and capacity-building efforts,
including the United Nations Office for Outer Space Affairs (UNOOSA),
which aims to bring the benefits of space technologies for sustainable development in an equal way to the broadest spectrum of nations. As Director Simonetta Di Pippo notes:

although more and more satellites are launched into space almost weekly by a growing number of space-faring nations, and in spite of the rapid growth in cooperative efforts, only a small percentage of the over 200 countries in the world have adequate access or capacity to work with space-based technologies and data, due to technological or resource limitations and often a lack of capacity awareness. (Di Pippo 2014).

Open data could be a true enabler and a formidable asset for decision makers in developing countries, which typically lack the most sophisticated technology and expensive infrastructures required for space-based applications. Di Pippo praises the fact that ‘visionary programmes such as Landsat paved the way for a large amount of space-based data to be released into the public domain,’ and that ‘a few visionary private entities in particular in the last ten years have made space-based data accessible to the public through their significant investments into the accessibility of satellite imagery and applications, permitting greater familiarity with the availability and benefits of space-based data’ (Ibid.). Google Earth is a remarkable example of this. Di Pippo continues: ‘If there is a more equal playing field in terms of access to space-derived data, it is thanks to strategic and crucial decisions to make the data public. This has led to a growing interest and demand for training on how to work with such data and on how to derive information from it for more informed decision-making and for varied uses’ (Ibid.).

The environmental sector is also of particular interest because it is naturally interrelated with two important points of contact between science and technology and society: that of public sector information (PSI) and big data. The public sector is one of the main advocates of open access policies, and provides many examples of successfully implementing open data environmental management policies, in publications and research or government data. Meanwhile, the ‘big data revolution’ predicted by policymakers in Europe and beyond is having a significant impact on the governance of public resources, and is bringing about novel ways of addressing environmental challenges, as the European Commission’s (EC) Digital Agenda for Europe (DAE) recognises (EC 2016a).

Since the environment is part of the earth system, environmental data can be considered a subset of geospatial data, as they are both geographic and spatial in nature, and typically characterised by their position relative
to the earth. This chapter refers to geospatial data gathered from earth observation to illustrate the implications of accessing such data in conjunction with open policies.

The environmental data ecosystem

As the World Wide Web Consortium (W3C) and Open Geospatial Consortium noted in a recent collaboration agreement (W3C 2015), spatial data is ubiquitous and integral to many human endeavours. Therefore, making spatial data easier to access and use can be extremely valuable. For example, in the United States alone, geospatial data and services are estimated to generate $1.6 trillion annually (Henttu et al. 2012). At the level of the global environmental movement, scientists, governments, policymakers and activists widely support the use of open data.

This generalised sense of consensus has helped stakeholders to mobilise some level of open data within the wider environmental community. In addition, efforts to strengthen the political cohesion of geographical regions (e.g. the European Union), to digitise public administration, to better understand and mitigate global scale phenomena (e.g. climate change), and the growing interest in spatial exploration programmes, are all greatly contributing to the momentum of the open data movement in the environmental sector.

One of the main advocates of open data in the geospatial sector is the Group on Earth Observation (GEO no date), a global voluntary group comprising over 100 nations and more than 90 international participating organisations such as UNOOSA. GEO promotes information sharing across many different scientific disciplines and applications, by providing a coordinated and sustained observation framework with a global and flexible network of content providers (which currently interconnects more than 130 autonomous infrastructures) – the Global Earth Observation System of Systems (GEOSS) – which gives decision makers direct access to an extraordinary range of data and information.

The GEO explicitly acknowledges the importance of data sharing in achieving the societal benefits they anticipate from GEOSS: ‘The societal benefits of Earth observations cannot be achieved without data sharing’ (GEO 2005). Thus, the GEOSS implementation plan sets out a set of data sharing principles for exchanging data, metadata, and products:

- There will be full and open exchange of data, metadata and products shared within GEOSS, recognising relevant international instruments and national policies and legislation.
All shared data, metadata and products will be made available with minimum time delay and at minimum cost.

All shared data, metadata and products, being free of charge or no more than the cost of reproduction, will be encouraged for use in research and education (Ibid. p.8).

Conceding that these data sharing principles may remain an abstract goal until all parties (members, contributors, users) can appreciate how they will implement them, The GEO devised an action plan, which identifies some of the negative implications of open access for environmental data in the GEOSS context (GEO 2010). A primary negative implication is of a financial nature. Various data providers perceive that enabling a full and open exchange of data, metadata and products in GEOSS could pose challenges to their own development which would result in limited revenue, particularly as payments for reuse contradict the GEOSS data sharing principles. Furthermore, many providers cannot visualise a business model that would work if they adopted the principle of full and open exchange of data. Yet, in many cases, requiring users to pay for access to data impedes its use, especially if acquiring the necessary funding to purchase data is a long and arduous process. Hence, the data providers can only realise very limited societal benefits if the product is not attractive, and easily accessible, to the users. To rectify the above and to mitigate providers’ reluctance to share their data and products openly, the action plan suggests that the GEO community should demonstrate how the full and open exchange of data can lead to new applications, additional jobs and more open competition, in contrast to the old model of data protection. One action taken to achieve this goal is the GEO ‘Appathon’, a global app development competition that aims to develop new, exciting and (most importantly) useful applications using earth observation data (GEO 2014).

Another negative implication of open access is that different disciplines, sectors and countries have developed different socio-cultural approaches to open data in the environmental sector, resulting in language barriers and different rates of development in countries across the globe. The GEO recognises that a commonly-endorsed vision is needed to bridge these gaps and overcome such barriers. Incompatibilities between different countries’ legal frameworks are also seen as inhibitors that need to be adapted, in

1 In GEOSS terms, ‘full’ and ‘open’ are interpreted as ‘taking into account international instruments and national policies and legislation’, whereas ‘minimum cost’ is interpreted as ‘free or cost of reproduction’.
order to remove legal barriers that could slow the implementation of the GEOSS data sharing principles. In some cases, the principle of full and open exchange of data is inconsistent with current national policies. The GEO tries to address this issue by encouraging both national and international bodies to adopt the principle of full and open exchange of data.

The GEO appreciates that it is important to recognise such negative implications, in order to mitigate them. For example, paying for data may not only hinder their use because of the price, but also because the mechanisms for paying are too cumbersome. In fact, barriers to data access are not simply a matter of pricing policies, but also include the varying policies across data providers and countries, so that negotiating access with each provider is extremely complex and long, thus creating a de facto barrier.

At the European level, on 3 March 2010, the European Commission proposed the framework for the ‘Europe 2020’ initiative (EC 2016b), a ten-year strategy to advance the European Union’s economy. The first of seven Europe 2020 flagship initiatives, the Digital Agenda for Europe (DAE) (EC 2016a) contains a specific policy on open data (EC 2016c), including legislation on the reuse of government data (EC 2015), such as the Directive on Access to Public Sector Information (EC 2003a), which applies to any data held by public authorities.

There is also an emphasis on open government data policy (as discussed in Chapter Three) in this area, which is implemented by the Public Sector Information Directive. Open government data is at the centre of a cluster of initiatives within various EC policy domains, which build on and complement the open data policy. One example is the European Commission Communication on Marine Knowledge 2020 (EC 2010a), which aims, amongst other things, to make marine data easier and less costly to use. Other areas impacted by the open government data policy include transport systems, scientific research and cultural heritage. In addition, the gradual deployment of the EC open data policy is expected to have an impact on a number of domains that do not yet have concrete open policies, but which will undoubtedly profit from the benefits of opening up a wide range of public and business data across areas such as education, tourism, consumer protection and public health.

EC environmental policy initiatives are founded on the Directive on Public Access to Environmental Information (EC 2003b), based on the Aarhus Convention (UNECE 1998), and have the most solid links to the open data policy. A remarkable example of this is the INSPIRE Directive (EC 2007), which aims to achieve the widest possible harmonisation and sharing of environmental information throughout the European Member States.
Article 17(1) of the ‘INSPIRE Directive’ requires each member state to adopt measures to share spatial datasets and services between its public authorities, in relation to public tasks that may have an impact on the environment. Since most of these EC institutions and bodies have to integrate and assess spatial information from all the member states, INSPIRE acknowledges the need to be able to gain access to, and use spatial data and spatial data services in accordance with an agreed set of harmonised conditions.

The main points of the INSPIRE Regulation (EC, 2010b) are that:

– Metadata must include conditions which apply to access and use for EC institutions and bodies; this will facilitate their evaluation of the available specific conditions, even at the discovery stage.
– Member states are requested to provide access to spatial data sets and services without delay – within 20 days on receipt of a written request at the latest, although mutual agreements may allow an extension of this standard deadline.
– If data or services can be accessed under payment, EC institutions and bodies are entitled to ask member states to provide information on how these charges have been calculated.
– While fully safeguarding the right of member states to limit sharing – when this would compromise the course of justice, public security, national defence or international relations – member states are encouraged to find the means to give access to sensitive data under restricted conditions, (e.g. providing generalised data sets). Upon request, member states should give reasons for their limitations on sharing.

As regards public access to data and services supplied under INSPIRE, the EC guidance document states that, if no provisions are contained in the agreement between member states and the EC institutions and bodies, then access given should be guided by whether public access is already, or could be, allowed in the member state and under what conditions (EC 2013a). Public access should therefore be promoted as much as possible, while respecting any exemptions provided for by law. When this public access to spatial data sets or services cannot be allowed, due to an exemption provided for by law, data producers are encouraged to state the conditions under which such access would be possible, for example by removing sensitive information, downgrading the accuracy or restricting the size of the download. It also suggests that any such measures should be harmonised, as far as possible, within and between member states, so that they can effectively be applied to aggregated data sets that might, potentially, come from a large number of producers.
INSPIRE establishes a list of topics that it considers particularly critical for successful data and service sharing within and between member states, supplying criteria for good practice in each. For example, it defines public access as: ‘the ability of the public to discover, view and download information and data and to use available services and data. [...] The public authorities should make their data and services available in a way that makes it easy for the citizen to obtain access. It states that usage conditions and charges should be presented in an understandable way’ (INSPIRE 2013, p. 46). The following criteria characterise good practice in public access:

- Awareness by the public that data and services exist – the public knows where it can find data and services, i.e. there is a central portal with registries and search engines that allow citizens to find out where they should go to obtain access to data or services. Awareness-raising activities are also promoted through other means (e.g. flyers). Increasing public awareness will usually be reflected by the growth in use of such websites.
- A clear process for the public to access data and services – the public authorities provide clear and user-friendly information on how citizens can obtain access to data and services and under which conditions and charges. This information is also provided online, with contact details for obtaining more information.
- Online access wherever possible – citizens can also obtain access to data online rather than via paper or digital copies, on CD or consultation on site (Ibid.).

Data mobilisation at the European level is exemplified by Copernicus, previously known as Global Monitoring for Environment and Security (GMES), which is the main European programme for establishing capacity for earth observation (Copernicus 2016). Copernicus is a European system for monitoring the earth, which collects data from multiple sources, processes these data, and provides users with reliable and up-to-date information related to environmental and security issues. The main users of Copernicus’s services are policymakers and public authorities needing sufficient information to develop environmental legislation and policies, or to take critical decisions in the event of emergencies such as natural disasters or humanitarian crises.

Copernicus covers six thematic areas: land; marine; atmosphere; climate change; emergency management; and security. These support a wide range of applications, including environment protection; management of urban areas; regional and local planning; agriculture; forestry; fisheries; health; transport; climate change; sustainable development; civil protection;
and tourism. The architecture of Copernicus comprises in-situ stations (airborne, seaborne and ground-based sensors) and a space component that consists of both missions contributed by members (e.g. commercial or national satellites) and dedicated satellite missions, such as the Sentinel constellation. The space component also includes services to facilitate access to the massive amount of data and information expected from Copernicus, which will be many times more than the volume of data produced by the Sentinel-1, -2, -3 A-series, which is roughly equivalent to 25 Envisat missions (ESA 2012), or over 25 petabytes of data (Laur 2012).

Based on the Copernicus services and on the data collected through the Sentinels and contributing missions, many value-added services can be tailored to specific public or commercial needs, resulting in new business opportunities. In fact, several economic studies have already demonstrated a huge potential for job creation, innovation and growth. This is a major positive outcome expected from Copernicus, in terms of strengthening earth observation markets in Europe. In particular downstream actors, i.e. those developing products and services based on this data, should experience growth and job creation. European research, technology and innovation communities will also be supported in making the best use of these data to create innovative applications and services (Koch 2014).

As a strategic pan-European programme requiring significant resource investment, Copernicus is coordinated and managed by the European Commission, in cooperation with the European Space Agency (ESA) for the space element, and the European Environment Agency (EEA) and the member states for the in-situ component. The member states and the European Parliament have mandated the EC to define Copernicus’s overall data and information policy, whose basic principle is full and open access to all data and information produced by services and collected through Copernicus infrastructure, including the Sentinel missions.

The Sentinel data policy was jointly agreed by the EC and ESA (European Space Agency, Sentinel-2 Preparatory Symposium, April 2012, slide 9; cited in Desnos 2013), based on joint principles prepared in 2009 (ESA 2009). The policy has been implemented by the Copernicus regulation (European Parliament and the Council 2014), which replaces the previous regulation on the initial operations (2011 to 2013) of GMES (European Parliament and the Council 2010). The Copernicus regulation implies a commitment to follow the GEOSS data sharing principles. In fact, the Copernicus policy promotes the access, use and sharing of data and information on a completely full, free and open basis. To understand the extent of this freedom, it is interesting to highlight the key general provisions of Copernicus’ data and information policy:
– No restriction on use, including reproduction, redistribution, and adaptation, for commercial and non-commercial purposes; in particular, no difference is made between public, commercial and scientific use.
– All datasets, including the Sentinel data, are always available on the Copernicus dissemination platform free of charge (or at the minimum cost of fulfilling the user requests).
– Worldwide access to data for European and non-European users, without any limitation in time.

Security restrictions and licensing conditions, including registration, may limit these general principles. For example, access limitations are foreseen for conflict of rights, where the Copernicus open dissemination affects Intellectual Property Rights (IPR) from third parties and principles recognised by the Charter of Fundamental rights of the European Union. Other limitations may apply for security reasons, where the Copernicus open dissemination may affect the security of EU member states, or for urgency. In every case, the decision must be balanced between protecting security interests and the social benefits of open dissemination.

While no warranty is given on the data and information provided, the policy only imposes one obligation, which is an attribution clause, the need to cite the source of data and declare any modification that is made. Regarding user identification, the policy allows quasi-anonymous use, specifying that there should be no need for users to register to access and view services, and only a light registration for use of the download service. It is worth noting that data generated by missions contributed to by Copernicus members, such as commercial or national satellites, as well as in-situ data, are considered external to Copernicus, and therefore they are not covered by the policy. However, Copernicus follows or negotiates the rules set by the data providers for such external data.

Another fundamental contribution to the promotion of open data culture is the advent of data journals, a relatively recent addition to the panorama of scientific literature in the environment sector, and beyond. Although data journals are not open access per se, most of them adhere to the open access paradigm, since their main objective is to provide a formal way of publishing data as a citeable entity, similar to research articles in the scholarly literature. This is in contrast to simpler data-sharing approaches, where data producers make data available on a website. A data paper can be seen as an eloquent and readable version of metadata, which describes a dataset, including its purpose, scope, coverage, format, provenance and quality. Tools are already available to create data papers directly from
existing metadata, such as the GBIF Integrated Publishing Toolkit (Robertson et al. 2014). Importantly, a data paper has a unique persistent identifier (PI) assigned to it, which ensures that it can be identified and cited. This means that data producers can now receive credit for their work, which is expected to incentivise data publication. Some examples of environmental data journals are the Earth System Science Data (Copernicus), the Geoscience Data Journal (Wiley), the Biodiversity Data Journal (Pensoft), and the Data Papers of the Ecological Archives (ESA). More generic data journals include Scientific Data (Nature), Data Science Journal (CODATA/ICSU), and GigaScience (BioMed Central).

Open environmental data: Key technological and infrastructural issues

The main role of the technical infrastructure is to provide uniform and equal access to the broad variety of research outputs, i.e. making data understandable, searchable, retrievable, available, assessable and secure. Our work on the RECODE project highlighted five main key technological and infrastructural challenges that stakeholders in the environmental sector face in mobilising their data: heterogeneity, accessibility, sustainability, quality and security.

Heterogeneity relates to the different ways of formatting, storing and using the variety of data available from a growing number of disparate sources. It comprises low-level issues such as format encoding and interoperability of the communication protocols as well as higher-level matters such as application interoperability, semantics mismatches, cross-disciplinary usability, internationalisation, and discoverability – that is how easy it is for users to find the data they need.

Accessibility relates to the volume of data and its impact on the infrastructure’s capabilities and architecture. Data volume is a storage matter and becomes a processing issue when data must be analysed. Hence, this challenge is connected to the big data aspects of volume and velocity, in relation to data streaming, record structures, organisation of storage and processing resources, data indexing, filtering and delivery. Velocity concerns both how quickly data is produced, and how quickly data must be processed to meet demand, with related bandwidth issues arising from the huge amounts of data being stored and accessed.

The adoption of open standards helps to both mitigate heterogeneity and improve accessibility. Reinforcing the importance of metadata and data
standardisation (e.g. defining common models and encodings) promotes ease of deposit and retrieval for stakeholders such as researchers, universities, libraries and the general public. At the same time, data variety is inevitable, to some extent, so should be acknowledged and accommodated, using distributed architectures and interoperability solutions to fill the gaps between existing systems. Good practice advocated by the system-of-systems\(^2\) approach and brokering or mediation solutions should be considered, as adopted, for example, in GEOSS, where the infrastructure is able to provide harmonised discovery and access services to heterogeneous data by using a brokering approach (Nativi \textit{et al.} 2012). An infrastructure for open access to research data should be conceived as a system-of-systems, to leverage existing infrastructures, supplementing rather than supplanting them, to protect previous investment, guarantee sustainability, and ensure valuable participation from the whole research community.

Sustainability relates to the long-term impact of maintaining and operating an open infrastructure for research data, in relation to obsolescence, governance of updates and upgrades, data preservation and curation, persistence, scalability and energy footprint. Given the ever-growing amount of data, an increasingly pressing question is what data should be preserved, for how long and in what format (for example, online or offline). These decisions are context-specific and, in most cases, rather subjective. Some best practices that can improve sustainability include virtualisation technologies and periodical migrations to more recent technological solutions (through format conversion, transcoding, etc.) as well as using persistent identifiers. There is also the potential to outsource data curation and preservation to third-party archives (see, for example, DANS 2016), which suggests that the new professional roles and skills required to achieve open access to research data should be investigated. Data management culture is well established in some contexts, such as libraries, and some fields of science (e.g. physics or social sciences), but is almost absent in others, for instance in the wider administrative sector.

Quality denotes the technological support required to evaluate data suitability and appropriateness in relation to data accuracy, completeness, documentation (including metadata and other ancillary information), assessment, validation and peer review, usefulness and fitness for purpose. Quality is a crucial aspect throughout the whole data lifecycle. In the big data realm, it is typically referred to as veracity and is conceived as an indication of data integrity, including trustworthiness, provenance,

\(^2\) For a definition of ‘system-of-systems’ see Dersin 2015.
accuracy, and certainty. To address the quality challenge, it is necessary to enforce the presence of complete and accurate metadata, by requiring data producers and disseminators to provide and maintain appropriate ancillary information when publishing and curating their data. This should also comprise tools to auto-generate provenance information, manage versions, and enable data creators, providers and users to assess data quality. However, this is often perceived as being too onerous. To mitigate this sensitive issue, several data-sharing initiatives (e.g. GEOSS) advocate adopting the concept of fitness for use, which seems more neutral than quality. This could be implemented by supporting the collection of user feedback in data repositories, which could be integrated with the metadata to assist users in assessing the suitability of data for their specific purpose.

Security concerns restrictions on the usage, access and consultation of data and metadata as well as their enforcement from a technical viewpoint, e.g. protocols for authentication, authorisation and auditing or accounting, privacy issues and licensing. Technical challenges related to security mainly arise from the variety of data policies, licences, embargo periods, specific IPR, privacy and legal issues that need to be considered when building data infrastructures. For example, some disciplines deal with sensitive data that should be obfuscated (e.g. the location of endangered species), while others manage data under specific licences (e.g. academic programmes for commercial remote-sensing data vendors), where only derived data products can be shared. In other cases, data is withheld until the research project ends or an embargo period expires. Hence, a security framework for authentication, authorisation and auditing is a mandatory component for most data infrastructures in the environmental sector. A good practice is to enforce data policies automatically, where possible, using approaches like ‘privacy by design’, which designs privacy and data security protections into systems at the outset rather than relying on costly retro-fits. Furthermore, it is important to recognise that sharing does not necessarily mean unrestricted and free access. A common practice is to make metadata immediately available for discovery, with the underlying data only being published after a certain time.

All five of these aspects of the complex open access ecosystem are mutually interrelated. To effectively allow researchers to identify, evaluate, access and use relevant scientific information extracted from a variety of sources, in a variety of formats, it is necessary to recognise the importance of semantic and multidisciplinary interoperability and to adopt technical and infrastructural solutions that holistically address data harmonisation, preservation and technological obsolescence as well as data documentation and metadata, quality and relevance indicators and security aspects.
Despite this, there is a clear tendency in the open data debate to refer to science as a whole sector, ignoring the differences between disciplines in further policymaking. In reality, though, each discipline has different methods for gathering and analysing data, which may be visual, numerical, narrative or statistical, presented in small, medium or large data sets, discrete or interlinked. Moreover, the definition of research data includes public sector information. This implies an essential, inherent heterogeneity. Any policy for open access to research data should therefore take a flexible approach, applying adaptable technological and organisational solutions, and avoiding approaches that do not satisfy the specifics of different disciplinary communities, and thereby raise entry barriers that are already high, and growing.

RECODE research has also highlighted that technical barriers are considered to be more a concern in the environmental sector than cultural ones. This suggests that the acceptance of open access in environmental sciences could be limited more by technology than by stakeholders’ willingness to share their data. However, technological and infrastructural hindrances are not perceived as a key obstacle to achieving the adoption of open access to research data, when compared to financial, political, ethical and legal issues, and all of the administrative and process-oriented elements of data management, which is generally referred to as data governance.

Issues in open data governance

Although the environmental sector is witnessing a general push towards abandoning the traditional model of data protection, in favour of a full and open exchange of data, in the belief that this will lead to new applications, additional jobs and more open competition, the major obstacles in relation to open data governance are:

- Interoperability, due to the large heterogeneity of applications, languages, policies, and legal frameworks characterising the context.
- Financial, given the investments required for earth observation, and the industrial sector’s determination to protect their investments and competitiveness. This includes the need for one or more effective and sustainable business models characterising the open data process, and open science more generally.
- Curation and preservation, which are related to the previous two challenges, because a sustainable business model is required to guarantee data and related software interoperability over long periods, for instance, for longitudinal climate change research.
As already mentioned, many of the infrastructural and technological issues in open access to research data relate to those typical of big data. With the advances in satellite technology, the future ubiquity of sensors and the uptake of crowdsourced approaches, it is reasonable to expect a growing overlapping between open access and big data issues, particularly around the sharing, preservation and curation of research data. As big data and open access concerns coincide, and open data repositories become more immense, their governance will become an increasingly pressing concern. The current trend may lead to a bureaucratisation of governance, where critical decisions are delegated to politicians who are, typically, not fully aware of the related scientific implications. It is preferable that the governance of big open access repositories primarily involves scientists, who should eventually use them to advise policymakers about critical decisions.

One concern about open data governance is the way that the increasing momentum of open access is spurring on a significant number of volunteer efforts into data sharing in diverse contexts, which is resulting in data sharing solutions being implemented at very different scales, e.g. for a single community of practice or specific project, and in the fragmentation of data into a puzzle of individual pieces, which may be referred to as ‘semi-open data pools’, or ‘semi-commons’ (Reichman et al. forthcoming). Although, in principle, these are informed by the overall vision of data sharing, they actually work in isolation from each other. To mitigate this problem, funders may oblige publishers to publish data as open access, together with scientific articles, or force them to transfer their semi-commons data into open access repositories, when the projects cease to exist. Another issue is how to combine data that fall under different jurisdictions, e.g. EU and USA policies, especially when such a combination of data is suitable for commercial exploitation. Funders and policymakers should address this problem, for example by developing standard data transfer licences that may be automatically enforced at the infrastructural and technological level.

The Copernicus data policy itself provides an insight into the problems connected to open data governance. The fact that the Copernicus policy is supported by a regulation has both positive and negative implications: on the positive side, as a formal normative document, it could be aligned with other relevant directives, such as the EU INSPIRE Directive 2007/2/EC and the EU PSI Directive 2003/98/EC. This facilitates the consistent implementation of open access in the environmental sector throughout the whole EU. On the other hand, as a formal EU regulation, its provisions are legally binding for European entities, but they cannot have the same efficacy on foreign entities outside the EU. In particular, the principle of
worldwide (European and non-European users) access, without limitation in time, coupled with the absence of restrictions on the purpose of use (including commercial exploitation) has raised major concerns, particularly by the European industrial sector, about indirect negative implications for competitiveness. Among other concerns, industry has expressed the view that granting free access to Copernicus data and information to non-European entities, including those from countries such as China and India, which have less expensive cost structures, may result in them gaining a competitive advantage over European industry.

To counteract this problem, clear criteria to define targeted users, their legal status and origin in order are being determined, to ensure that implementing the Copernicus data policy will not reduce the European market share of the earth observation industry. Industry representatives have therefore asked the European Commission to review the current version of the Copernicus data policy and consider introducing limitations on data access for non-European entities, particularly for for-profit entities and their commercial use of Copernicus data and services.

The European Commission is analysing this request and the potential legal and policy impacts arising from measures that would restrict the principle of full, free and open access to Copernicus data and information for non-EU commercial entities. On 27 September 2013, the Committee on Industry, Research and Energy (ITRE) submitted a draft proposal amendment on the Copernicus regulation. In particular, the report submitted that Article 14 of the Copernicus regulation should be amended as follows:

Copernicus data and information shall be made available on a full, open and free-of-charge basis for all participating Member States, for emergency situations and for development aid purposes. In all other cases a policy of pay-for-data shall be adopted or a reciprocity principle shall be applied (ITRE 2013).

However, the issue is very complicated and it is likely that both industry and ITRE proposals on potential restrictions in accessing and using Copernicus data and information for non-EU entities could lead to a violation of EU obligations and commitments under the WTO General Agreement on Trade in Services (GATS) (Amedeo and Baumann 2013). This example illustrates the fact that addressing these problems requires all the parties involved to agree on mutual policies on the exchange, sharing, access and use of interoperable data and services across various levels of public authority and different sectors of society, at a global level.
The work of the Research Data Alliance (RDA) on middleware governance is of particular interest in understanding the most important challenges to governing and sustaining a digital infrastructure, and to share data sets that are not necessarily from a single community (RDA no date). Effective middleware governance has the potential to support longer-term development under a variety of funding models, to simplify and standardise access models, and to establish a basis to ensure sustainable, stable development and effectiveness in an operational environment. The RDA identifies the following main challenges for achieving effective and sustainable open data governance:

- Secure financial support – the efforts required to obtain and sustain funding. This may include proposal writing, obtaining venture capital, reporting, etc.
- Engage user communities – the efforts required to identify, target, engage and sustain a class of institutional and/or disciplinary and cross-disciplinary science users and their data facilities.
- Marketing – the effort required to understand requirements and then pursue the case(s) for commitment to using a shared capability. This is focused on individual users or facilities and moves beyond the general engagement of a user community.
- Human resources – the personnel support and expertise required for market development, management and achieving the technical capability for infrastructure evolution and sustainment.
- Software engineering – the effort, including formal and informal processes, required to provide development and ongoing maintenance, improvement and technology assessment of software assets.
- Product management – the management of documentation, versions, licensing, distribution, security and other administrative activities.

If necessary, the legal considerations of operating in a multinational or global environment must be added. The RDA study considers the following possible business models for achieving good open data governance:

- Government funding through assistance awards and contracts.
- Government funding through data facility guardianship.
- Software as a service (SaaS).
- Information and advertising sales.
- Corporate support and product or service sales.
- Consortia.

Naturally, each of these has advantages and weaknesses. Although business revenue and/or hybrid models for sustainability can be identified, these
approaches will be most successfully applied to the research community when they are amended to fit specific community cultures and practices. Therefore, the RDA study identifies the following community best practice which can be used to shape the most effective governance and business models:

- Low cost.
- Open source, which
  - promotes stakeholder engagement
  - reinforces community practices and standards
  - may promote interoperability standards across communities
  - may mitigate volunteer fatigue
- Community-driven development and evolution.

Conclusion

The environmental sector illustrates several implications of accessing data in conjunction with open policies, in particular regarding data stemming from earth observation, where a big data revolution is predicted. The increasing availability of multidisciplinary data from new observation platforms is expected to provide scientists and society with unprecedented resources through which to understand our planet and better control or mitigate the environmental dynamics. In turn, a better use of globally-available national and local data sets will enable policymakers to make informed and evidence-based decisions to address global change.

The examples discussed above show that the sector is experiencing a general drive to abandon the traditional model of data protection, in favour of full and open exchange of data, in the belief that this will lead to new applications, additional jobs and more open competition. However, despite this obvious mobilisation and the significant gains in achieving agreement and cooperation on key issues, key challenges remain. The major challenges of open environmental data sharing can be identified as interoperability issues, due to the significant heterogeneity of technologies, applications, languages and legal frameworks characterising the context as well as financial concerns, given the investments required for earth observation and the industrial sector’s determination to protect their investments and global competitiveness.

Addressing these problems requires mutually agreed policies on the exchange, sharing, access and use of interoperable data and services across various levels of public authority and different sectors of society,
at a global level. Such complexities may be reduced by following some recommendations:

- Build on the existing research infrastructures, to supplement, rather than supplant them, applying flexible and adaptable technological and organisational solutions. This is essential to guarantee sustainability and valuable participation from the research communities. The aim should be to fill the gaps and make the existing research infrastructure interoperable, by mediating, instead of imposing common solutions that may not satisfy each community’s specific needs, or even raise already-high barriers to entry.

- Distinguish between different approaches to open access, for instance, acknowledging that sharing does not always mean giving away for free.

- Discuss new business models that can sustain the open data approach (e.g. evolving governance and business models based on public-private partnerships).

- Leverage the experience and lessons learned from ongoing national, European and international initiatives such as INSPIRE, GEOSS and Copernicus.

- Discuss new professional roles and curricula which specialise in data science, and open data in particular.

In summary, to overcome the technological barriers of open research data access, there should be a particular focus on the problem of data discovery and access, of analytical search tools and techniques involving aspects such as the use of metadata, relevance indicators, key word searches and third-party recommendations, to help researchers and the public find their way through the mass of scientific information and research data, to identify the material that best fits their purpose. The problems of technological sustainability and obsolescence should also be considered, because these related issues have specific impacts on ensuring continued, sustained access to research data over time. Successful and emerging technologies that can be optimised to provide better access to scientific information and research data should be identified, including technological solutions that are being used in open access repositories, to identify which approaches might be replicated to increase interconnections between scientific information and research data repositories across Europe.

Metadata, particularly provenance information, are of paramount importance in ensuring the repeatability of processes, and good open standards would facilitate a culture of information sharing. The importance of metadata and data standardisation should be reinforced by, for example,
agreeing on common models and encodings, to promote the ease of deposit and retrieval for stakeholders including researchers, universities, libraries and members of the public.

Nonetheless, the approach to open research data in Europe should take account of the diverse attitudes in different fields of science towards the issue, as well as the specificities of different communities, which means that there is always an inherent heterogeneity. Therefore, data heterogeneity should be acknowledged and accommodated, by the use of distributed interoperability solutions between existing systems, to enable access to heterogeneous content via the usual platforms. To this end, system-of-systems and mediation solutions may be adopted, following the example of GEOSS, where the infrastructure is able to provide harmonised discovery and access to heterogeneous data by means of a brokering approach. In addition, the cultural changes needed to foster open access to environmental data are much bigger than the technical challenges. This holds true, in particular, for communities that do not require cutting-edge technology to perform their routine research tasks. In fact, communities that are limited by technology typically help to push technological boundaries and advance in terms of data mobilisation and sharing.

Providing open access to data is still at an early stage within Europe and internationally, and its impacts on the creation of a knowledge society are only beginning to be determined. Along with the lessons learned from more widespread open access in publications, the experience gained by global information sharing endeavours in the environmental sector, such as GEOSS, offer a useful insight into the challenges, suggest some possible solutions, and provide valuable experience and good practice to reflect on, when discussing strategies for the future. Nevertheless, this case study demonstrates that, even in a relatively bounded, although heterogeneous, discipline, where the benefits of information sharing are obvious and generally accepted by stakeholders, and where significant gains in working out issues related to institutions, legal frameworks and standards have been made, there are still many obstacles to be overcome to enable the sharing of data to significantly contribute to a knowledge society, even within earth observation itself. This indicates that much work remains to suitably leverage open access to data to achieve all of the potential benefits foreseen from this opening.