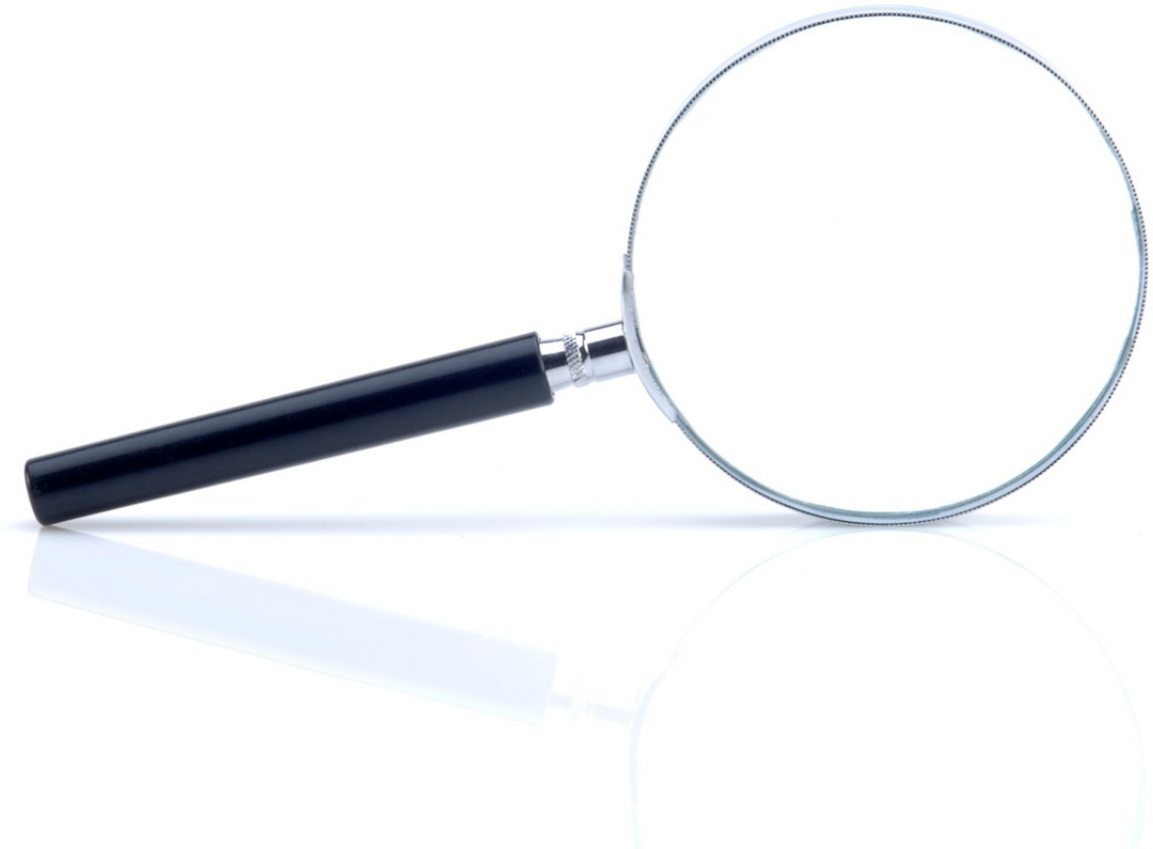
Open Data Access Policies and Strategies in the European Research Area and Beyond

August 2013



Science-Metrix

Aurore Nicol, Julie Caruso,
& Éric Archambault

Open Data Access Policies and Strategies in the European Research Area and Beyond

August 2013

Produced for the
European Commission DG Research & Innovation
RTD-B6-PP-2011-2: Study to develop a set of indicators to measure open access

by **Science-Metrix Inc.**
Brussels | Montreal | Washington
1335 Mont-Royal E., Montréal
Québec, Canada, H2J 1Y6
1.514.495.6505
info@science-metrix.com
www.science-metrix.com

Cover image: iStockphoto

Executive Summary

This report examines policies and strategies towards open access (OA) of scientific data in the European Research Area (ERA), Brazil, Canada, Japan and the US from 2000 onwards. The analysis examines strategies that aim to foster OA scientific data—such as the types of incentives given at the researcher and institutional levels and the level of compliance by researchers and funded organisations—and also examines how, and whether, these policies are monitored and enforced. The infrastructures developed to store and share OA scientific data are also examined. The analysis is supported by findings from the literature on the global progression of OA scientific data since 2000—including its growth as a segment of scholarly publishing—as well as some of the broader trends, themes and debates that have emerged from the movement.

Governmental OA Scientific Data Strategies

Governments produce and own large datasets. To date, most national open data policies primarily target these datasets—which are not necessarily generated through scientific research but may be used for research—rather than scientific data at large.

The importance of comprehensive OA policies was recognised in 2004 by the Ministers of Science and Technology of the then 30 OECD countries, and of China, Israel, Russia, and South Africa. Governments may reap important economic benefits from the release of OA scientific data, such as through economic growth and job creation deriving from innovation, and through better informed policy and research.

It has been estimated that, by unlocking the potential of big data, developed European economies could save between €150 billion and €300 billion annually in the form of operational efficiency gains and increased potential versus actual collection of tax revenue alone. A study has estimated the value of direct and indirect economic impacts of government-owned data across the EU-27 at €140 billion annually. It also estimated that, with lower barriers and improved infrastructure, this value could have been around €200 billion in 2008, representing 1.7% of the European GDP for that year.

In order to unlock the full economic potential of OA scientific data, machines must be able to access and mine this information without copyright infringement. The European Commission's Working Group on Text and Data Mining (TDM) *Licences for Europe* initiative has addressed this challenge through stakeholder dialogue. However, representatives from groups of researchers, science organisations, libraries, and small and medium enterprises have withdrawn from the process, arguing that the Working Group's decision to place licensing at the centre of discussions constitutes a bias against other solutions to adapt the legal framework to TDM applications.

The European Union issued a directive on the re-use of public sector information as early as 2003, while the US government was the first to establish its own open data portal, data.gov, in 2009. National public data repositories were launched in the succeeding years in 20 countries (Austria, Belgium, Brazil, Canada, Denmark, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Japan, the Netherlands, Norway, Portugal, Slovakia, Spain, Sweden and the United Kingdom).

Major international organisations have also launched open data portals, including the United Nations, OECD, European Commission, International Energy Agency (IEA), Nuclear Energy Agency (NEA), Programme for International Student Assessment (PISA), and International Transport Forum (ITF). The World Bank's repository, data.worldbank.org, launched in 2011, is a landmark repository in terms of its usability by machines and humans alike.

The bulk of currently available open governmental data repositories belong to states, provinces, and municipalities. However, a few national, which in many cases have taken international

proportions, now offer access to vast amounts of data. Notable governmental/funding bodies initiated OA scientific databases include the NIH's GenBank, the DNA DataBank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database, and the ClinicalTrials.gov database.

Funding Bodies' Strategies

Funding bodies may have an interest in archiving and disseminating data generated by their grantees to monitor the outcomes of their investments, increase the visibility of their contribution to research, and ensure long-term preservation of the data. Mining and further analysing the data collected from multiple projects could also highlight underrepresented or emerging areas of interest.

Funding bodies have far fewer OA policies for scientific data than policies for scientific publications. In an analysis conducted by Science-Metrix of 48 funders' OA policies listed in ROARMAP within ERA countries, Brazil, Canada, and the US, 23% explicitly excluded data from OA requirements and 38% did not mention data at all. Conversely, 29% of policies mandated open data archiving and 10% encouraged it without mandating it.

Although some rare funding bodies operate their own repositories, most recommend the use of institutional, disciplinary, or aggregating repositories. Institutions, governmental bodies, and research networks may be better able to define their needs in terms of standards, space, and functionality based on a more direct relationship with the researchers who produce the data or with the users of the data.

Research Institutions' Strategies

OA policies for scientific data are less common than OA policies for scholarly publications. In a survey of head librarians at universities and higher learning institutions conducted by Science-Metrix, 73% of respondents agreed that providing scholarly publications in open access form is a priority in their organisation, whereas only 45% agreed that providing scientific data in open access form is a priority in their organisation.

A similar trend is observed in the adoption of OA policies for scholarly publications and scientific data. Only 11% of respondents indicated that their institution has an OA policy for scientific data, while for scholarly publications, 42% indicated that their institution has an OA policy. Thus, establishing institutional frameworks for the diffusion of data in open access remains a secondary concern at this point.

OA scientific data has the potential to strengthen the credibility of scholarly publications and research institutions, as it opens peer review to the entire scientific community. A growing trend has been reported in the percentage of published articles retracted for fraud or suspected fraud. Easily accessible data could extend the peer-review process beyond a small group of reviewers acting before publication, to all readers after publication.

Few institutional repositories are dedicated exclusively to research data. Institutional repositories generally support datasets on repositories devoted to books, theses, peer-reviewed publications and multimedia objects. Moreover, the infrastructure required to host and share scientific data, while still uncommon, seems more developed than the associated policy frameworks. In the survey, 36% of respondents indicated that their organisation maintains one or more repositories for OA scientific data, whereas 11% indicated that their organisation had a policy to this effect.

The survey's results also suggest that a significant number of existing OA repositories for scientific data are not indexed in ROAR and OpenDOAR, the directories used in this study. This

impression is reinforced by the absence of known open data repositories, such as the NASA's AERONET and the European Molecular Biology Laboratory's databases. The repositories listed could merely be the tip of the iceberg.

Discussion

Open data is evolving rapidly in an environment where citizens, institutions, governments, non-profits, and private corporations loosely cooperate to develop infrastructure, standards, prototypes, and business models.

From a governmental point of view, open data confers a competitive advantage in an increasingly information-based economy. New products and services can be developed directly from the data or through extensive transformation that adds value to the information.

Additionally, the legal and economic implications for publishers are far fewer. In fact, the Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical & Medical Publishers (STM) have declared their support of OA scientific data initiatives.

Some organisations use their data as a source of revenue, by providing paid access to their datasets. In the case of government agencies within the EU, a review has estimated that the direct revenues generated from the sale of government-owned data generally represent less than 1% of each agency's revenues. However, for a few organisations, the sale of data may represent close to 20% of revenues, which is likely to hinder the development of open data policies even if the data was generated using public funds. Still, the economic benefits of open data would outweigh sales revenue by approximately two orders of magnitude within the EU-27.

The emergence of OA scientific data as a valid, citable form of reference is limited by the difficulties associated with the standardisation of data and metadata formats, poor indexation by internet browsers, as well as by the scarcity of directories or registries that could make data more visible.

Increasingly, disciplinary standards are being developed to index specific types of records in the face of an ever-rising tide of data. However, the heterogeneous nature of scientific data certainly is a challenge for the development of OA in this area. A few fields of research use highly standardized formats that facilitate the aggregation and reuse of data; genomics, proteomics, chemical crystallography, geography, astronomy and archaeology are among those fields. The archiving of other, less standardized types of data needs to be carefully thought out in order to generate datasets that will be usable by other researchers.

Whereas researchers have a 'natural incentive' to promote OA in the case of scholarly publications, as this makes their work more widely known, in some cases there might be negative incentives to make research data public. Researchers may indeed derive a competitive advantage by filling up their own data warehouse and widely diffusing these data may limit the growth and curtail the size of their research enterprise.

Contents

Tables

1 Introduction

Organised efforts to provide open access to scientific data date back to the 1950s and so predate the concept of open access as an organised movement (Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council, 2008). Yet open access to scientific data is less studied and documented than open access to scholarly publications and less lobbied for by the scientific community. This lag in the spread of open scientific data may be explained on one hand because it is further from the current model of scholarly communication and more complex to implement than open scholarly publications, and on the other hand because it does not directly threaten the scholarly publishers whose resistance has polarised the debate around OA publications, spurring OA communities to organise and lobby.

Open scientific data was mostly developed within the confines of specific disciplines where the pooling of highly standardized data was instrumental to the pursuit of research, such as genetics and earth sciences. Today, the convergence of rapidly developing technologies and the exponential power and market penetration of computing devices have made large datasets more manageable and their use more accessible to a wider array of research institutions, businesses, and individuals. The exploitation of large datasets or "big data" has shown considerable promise for innovation, productivity, and growth in commercial and industrial applications such as retail sales, transportation and energy. Increasing public awareness has also increased demand for transparency on the part of governments and scientists. This combination of factors has ushered in a new period of growth for open scientific data, with a burst in the development of infrastructure and adoption of new policies in the last decade.

This report presents an analysis of strategies for open access to scientific data in the European Research Area (ERA)[1], Brazil, Canada, Japan and the US from 2000 onwards. The analysis examines the strategies and infrastructure developed by governments, funding bodies, and universities for managing, enabling sharing of, and allowing open access to raw scientific data.

A first part covers governmental strategies and infrastructure for scientific data (Section 2.1) including funding bodies' strategies and infrastructure for scientific data including open access rules for grants' recipients (Section 2.2). This is followed by an examination of research institutions' strategies; in practice this means mostly the strategies of universities (Section 3). Section 4 discusses the costs and benefits of these strategies, explores emerging practices in the open data community, and mentions some of the barriers to the development of OA data.

[1] Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, the former Yugoslav Republic of Macedonia, Malta, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Turkey, United Kingdom.

2 Governmental strategies for OA to scientific data

Studying OA government *scientific* data is interesting as it is not necessarily a straightforward concept. Indeed, governments generate substantial amount of data, frequently in the form of statistics, and raw data. Not all of this data can be considered as 'scientific' but large parts of these can be used by scientists, and frequently by social scientists. OA policies by government, excluding the policies by or affecting research funding organisations, are presented in Section 2.1.

In general, funding bodies do not directly produce scientific data, but they may have an interest in archiving and disseminating data generated by their grantees. In doing so, funders may better monitor the outcomes of their investments, increase the visibility of their contribution to research, and ensure long term preservation of the data. If they have the means to mine and further analyse the data collected from multiple projects, funders may also use this information to highlight underrepresented or emerging areas of interest and adapt their policies accordingly. The policies of funding bodies are examined in Section 2.2.

Another important aspect of government policies has been at providing OA data infrastructures. Interestingly, though these may contain government data, they are frequently filled by data produced by university researchers. Government supported infrastructures are discussed in Section 2.3.

2.1 National policies and incentives

Governments produce large amounts of data through censuses, geographical surveys, public spending oversight and other large scale monitoring activities. These activities result in large datasets that are funded, produced, and owned by governments. Governmental data are not necessarily generated through scientific research, but may also be used for further research. To date, most national open data policies primarily target these types of datasets, rather than scientific data at large. Policies that directly address open scientific data are more frequent at the level of funding bodies and research institutions.

Although comprehensive open access policies may not be on the immediate agenda of governments, their importance was recognised in 2004 by the Ministers of Science and Technology of the then 30 OECD countries, and of China, Israel, [2] Russia, and South Africa (OECD, 2007). The Ministers asked the OECD to develop a set of principles and guidelines, which were published in 2007, to "facilitate cost-effective access to digital research data from public funding". The OECD's guidelines define openness as "access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination" (OECD, 2007). The guidelines further state that OA to research data from public funding should be easy, timely, user-friendly and preferably distributed through the internet, where the marginal costs of transmitting data are close to zero (OECD, 2007).

Governments may stand to reap the most economic benefits from the release of scientific data in open access form through economic growth and job creation deriving from innovation. The public release of Global Positioning System technology, developed by the US Department of Defence in the 1970s, is a powerful, albeit extreme, example of the potential for economic return

[2] Israel has joined the OECD on 7 September 2010.

of open scientific data. Civilian access to the technology was gradually granted under presidents Reagan, Clinton, and G.W. Bush (Pellerin, 2006). Today access to the service is entirely free, as is the data needed to develop programs that interface with the system, and legislation recognises that it is in the national interest to "encourage open access in all international markets to the Global Positioning System and supporting equipment, services, and techniques" (U.S. Code, 2010). A recent study has estimated that the use of GPS technology in the commercial sector alone directly contributed $96 billion per year to the American economy, that 3.3 million jobs relied heavily on it, and that the economic importance of the technology is expected to keep growing (Pham, 2011). Considering that the annual budget to maintain and improve the system is around one billion dollars and the total documented investment to date is $43 billion, the uptake of this innovation by citizens and the private sector illustrates the potential economic benefits of open data initially produced by or for governments (McNeff, 2002) (GPS Innovation Alliance, 2012).

In 2006, the MEPSIR study (Measuring European Public Sector Resources) estimated the direct re-use market of public data at €27 billion for the EU and Norway and predicted a rapid growth of this market segment, around 7% annually (Makx Dekkers, 2006). The study focused on domains where the principal activity was based on re-use of data, excluding other types of businesses as well as government and research. A later study aggregating direct and indirect economic impacts of government-owned data across the EU-27 estimated a value of €140 billion annually (Vickery, 2011). This later study also estimated that, with lower barriers and improved infrastructure, this value could have been around €200 billion in 2008, representing 1.7% of the European GDP for that year. A comparison of the two studies' results suggests that there is far greater value in secondary or indirect uses of data than in direct uses, as value is added at every successful transformation of the raw data. Such successive transformations may however elude estimation and the economic impacts of emerging data-intensive activities are difficult to foresee.

Opening up scientific data may also inform policy and further governmental research through the exchange of data generated by various agencies and ministries. The sharing of data between agencies of the same government is often limited by subscription fees or by the development of operational silos, leading to unnecessary expenses. Henri Verdier, director of Etalab, the French governmental organisation responsible for the development of open data, has declared that until recently, certain ministries, such as the Ministry of Agriculture and the Ministry of Ecology, had never coordinated their databases (Siméon, 2013).

The development of efficient means of data diffusion within a government or between levels of government within the same country can not only help to avoid duplication of research, but also generate richer data from the crossing of various datasets. Opening data to international users can also be a powerful lever to understand and solve multi-jurisdiction issues such as transportation, migrations and environmental issues. It has been estimated that, by unlocking the potential of big data, European developed economies could save between €150 billion and €300 billion annually in the form of operational efficiency gains and increased actual versus potential collection of tax revenue (James Manyika, 2011).

Opening scientific data for human use alone is likely insufficient to unlock the full economic potential of this information; machines must be able to access and mine data as well. The European Commission addressed this challenge through the Working Group on Text and Data Mining (TDM) of its *Licences for Europe* initiative, aiming to "work for a modern copyright framework that remains fit for purpose and seeks to foster innovative market practices" through stakeholder dialogue (European Commission, 2012). However, representatives of researchers, science organisations, libraries, and small and medium enterprises expressed concern about the

Working Group's approach, which places licensing at the centre of discussions and allegedly constitutes a bias against other solutions. This concern culminated in their withdrawal from the dialogue on May 22, 2013 (Association of European Research Libraries, 2013).

2.2 Funding bodies' policies and incentives

Funding bodies have far fewer OA policies for scientific data than policies for scientific publications. This may be due to the technical challenges in the implementation of suitable infrastructure and standards to manage the data. Funders may also perceive fewer advantages stemming from OA data than from OA peer-reviewed publications in terms of their visibility.

In an analysis conducted by Science-Metrix of 48 funders' OA policies listed in ROARMAP within ERA countries, Brazil, Canada, and the US, 23% explicitly excluded data from OA requirements and 38% did not mention data at all. Conversely, 29% of policies mandated open data archiving and 10% encouraged it without mandating it.

Although they are few, the funders who do mandate or encourage the publication of OA research data may have a noticeable impact considering their importance in the distribution of research funds.

Funding bodies could benefit from widespread open data to identify issues where funds from multiple sources are concentrated, as well as neglected issues and use this information to better target their investments. The international aid sector is a good example of the potential of data sharing among funders and recipients. The International Aid Transparency Initiative (IATI)[3] was launched to track aid-related spending from multiple donor governments to multiple recipient governments to better inform all stakeholders involved. Currently, IATI collects and pools raw data from 148 organisations into its data registry with the aim of offering timely, comprehensive, accessible, and comparable data. A common technical publishing framework allows data to be accessed and compared using the Extensible Markup Language (XML) format. Although IATI focuses on financial data rather than scientific data, it also collects data relative to an array of project outcomes and, as such, demonstrates that it is possible and beneficial for funders, recipients, and other stakeholders to collect data outside of highly technical fields. A pilot project covering 5 recipient countries revealed that, even if some discrepancies remain in data from various sources, considerable benefits could be expected from shared data (International Aid Transparency Initiative, 2010). Standard reporting and open publication of raw aid data may extend the breadth of stored data and improve its consistency, timeliness, regularity, and accuracy, which would result in significant time savings for users and reduced parallel reporting.

2.3 National infrastructure

A number of governments are already developing the infrastructure required to collect, manage, and disseminate scientific data. As with national open data policies, the focus of these efforts has been governmental data rather than scientific data at large. In 2009, a communication by the European Commission stated that "the landscape of data repositories across Europe is fairly heterogeneous, but there is a solid basis to develop a coherent strategy to overcome the fragmentation and enable research communities to better manage, use, share, and preserve data"

[3] http://www.aidtransparency.net/

(European Commission, 2009). The bulk of currently available open governmental data repositories belong to smaller governmental entities, such as states, provinces, and municipalities. However, a few national repositories, such as the US based data.gov, now offer access to staggering amounts of data. As the acceptability and economic benefits of open data grows, more countries can be expected to develop similar infrastructure.

OA databases for scientific data are a relatively new concern for funding bodies in general. Although there are instances of funding bodies that operate their own repositories, many more recommend the use of institutional, disciplinary, or aggregating repositories. Indeed, setting up a repository is a complex task involving not only the financial and technical aspects of the infrastructure, but also the adoption of appropriate standards. Institutions, governmental bodies, and research networks may be better able to define their needs in terms of standards, space, and functionality based on a more direct relationship with the researchers who produce the data or with the users of the data.

Funding bodies that require data archiving on their own repositories fall into two categories. The first type is large organisations that have a clear disciplinary profile that involves generating a critical mass of data to attract users. The second category is composed of smaller organisations that use their data repository mostly for verification and transparency purposes.

Perhaps the most notable OA scientific databases include the NIH's GenBank, the DNA DataBank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (funded by public research funds from 20 member countries). These databases are directly linked and seamlessly share data on a daily basis. This level of integration is facilitated by the uniform nature of their highly specialised content. This success might also be attributable to the maturity of these databases. GenBank has been active since 1982, the DDBJ since 1986, and the EMBL Nucleotide Sequence Database since 1982.

Public registries of clinical trials were first set up in the 1990s in Europe and in the US to provide general information about certain types of clinical trials while maintaining confidentiality of the results. In the US, the legislation governing the activities of the ClinicalTrials.gov database was amended to include data concerning the flow and baseline characteristics of participants and summaries of outcome measures, statistical analyses and adverse events (FDAAA 801, 2007). As of 2012, the European Union Clinical Trials Register still limited public access to a description of the design of the trial, the name of the substance investigated, therapeutic areas of application, the status of the trial, and its sponsors (Sec. 57 (EC) Regulation 726, 2004). However, the European Medicines Agency (EMA) has allowed release of clinical trials data on the basis of requests for access to documents since 2010. In 2012, the agency introduced a new policy of public access to clinical trial data which will take effect by January 2014 (Rabesandratana, 2013). The 2010 policy has been challenged as part of lawsuits by two pharmaceutical companies, in which interim rulings were made in favour of the companies (EMA, 2013).

Although the European Union issued a directive on the re-use of public sector information as early as 2003, the US government was the first to establish its own open data portal, data.gov, in 2009. At its onset, the repository contained only 47 datasets. As of May 2013, this number has grown to nearly 400 000 datasets including raw datasets, tools, and geodata (Data.gov, 2013). Several national public data repositories were launched in the following years and are listed in Table I. However, not all data accessible through these portals truly are OA, as access fees or restrictions may apply to certain datasets. Germany's public service information portal, govdata.de, uses a custom licence model which may preclude re-use of the data, especially in

combination with other datasets. This format has been widely criticised by German OA advocates who participated in the portal's development, and they have launched an online petition asking, among other demands, to revise the license model (Not your Govdata, 2013). Other national portals included in Table I have been criticised, but Germany's case is unique as its licence model may be a more permanent hurdle than low numbers of datasets or access restrictions.

Table I National Open Data portals within the ERA and selected countries

Country	Official Open Data portal
Austria	http://data.gv.at/
Belgium	http://data.gov.be/
Brazil	http://dados.gov.br/
Canada	http://www.data.gc.ca/
Denmark	http://digitaliser.dk
Estonia	http://pub.stat.ee/px-web.2001/Dialog/statfile1.asp
Finland	http://data.suomi.fi/
France	http://data.gouv.fr/
Germany	https://www.govdata.de/
Greece	http://geodata.gov.gr/geodata/
Ireland	http://www.statcentral.ie/
Italy	http://www.dati.gov.it/
Japan	http://datameti.go.jp/data/ (test version)
Netherlands	http://data.overheid.nl/
Norway	http://data.norge.no/
Portugal	http://www.dados.gov.pt/pt/inicio/inicio.aspx
Slovakia	http://data.gov.sk/
Spain	http://datos.gob.es/
Sweden	http://öppnadata.se/
United Kingdom	http://data.gov.uk/
United States of America	http://www.data.gov/

Notes: Countries had to have a functional, government owned open data portal to be included.
Source: Data compiled by Science-Metrix.

Major international organisations have also launched open data portals. The United Nations' portal, data.un.org, was launched in 2008, earlier than the US government's data.gov. It is the single entry point for 34 databases developed by 17 UN organisations, and currently contains close to 60 million records accumulated since the UN's formation. The OECD's portal, OECDiLibrary, launched in 2010, offers open access to publications and data emanating from the OECD, International Energy Agency (IEA), Nuclear Energy Agency (NEA), Programme for International Student Assessment (PISA), and International Transport Forum (ITF). However, organisations can purchase a subscription to access an undisclosed quantity of items excluded from the open access content. The World Bank's repository, data.worldbank.org, launched in 2011 and updated in 2013, is a landmark repository in terms of its usability by machines and humans alike. The European Commission's portal, open-data.europa.eu, was launched in 2012 and now includes 5911 datasets from European organisations, including Eurostat.

3 Research institutions' strategies for OA to scientific data

One of the largest incentives for OA data is to augment the capacity of the scientific community to detect frauds and to correct erroneous interpretations. However, in universities, there seem to a void of champion for that cause as the advocates of OA in scholarly publications are not engaged in this cause (Section 3.1).

3.1 Institutional incentives, and dearth of policies

Open scientific data has the potential to strengthen the credibility of scholarly publications, as it opens peer-review to the entire scientific community. Currently, papers submitted for publications are vetted by a limited number of reviewers, typically two or three, who are appointed by the editor as experts in the author's field. Peer-review is a fundamental aspect of scholarship, acting as a safeguard for the quality of published studies. Although the overwhelming majority of scholars agree that peer-review is essential, the current process has been criticised for its failings. Reviewers scrutinize the authors' work and may ask for precisions or supplemental data for their evaluation before giving their *imprimatur*. Once the article is accepted for publication, it is assumed to be correct and individual readers who want to examine data omitted from the paper must reach out to the authors, who have no obligation to reply. Hence, the validity of a publication is vetted by a handful of people outside of the research team and the discovery of mistakes or fraud after publication is unlikely unless subsequent studies attempt to duplicate the results.

It is impossible to know the true extent of peer-review failure, but the retraction of articles can be used as a proxy measure (Michael L. Griesen, 2012). PubMed's records contain over 25 million articles published in life sciences and medicine. Since the 1940s, more than 2000 of these articles have been identified as retracted, with the first occurrence of retraction in 1977 for a paper published in 1973. An examination of these retractions revealed that two-thirds of them were attributable to misconduct, which includes fraud or suspected fraud (43%), duplicate publication (14%), and plagiarism (10%), while the remaining retractions were attributable to error (Ferric C. Fang, 2012). The study also found a growing trend in the percentage of published articles retracted for fraud or suspected fraud by year of publication, which may be explained by a growth in the prevalence of scientific misconduct or to increased detection of faulty articles. Although retractions are rare, they demonstrate that traditional peer review is not infallible and suggest that errors and misconduct may frequently remain undetected.

In 1998, Wakefield et al. published a paper in The Lancet, establishing a link between the Measles-Mumps-Rubella vaccine and autism in children. The article was cited abundantly and launched a widespread scare, leading parents to shun immunisation. In 2010, Wakefield faced accusations of misconduct, was stripped of his licence to practice medicine, and the paper was retracted. Unfortunately, by the time the article was retracted, spikes in measles cases had been observed in the UK (Deer, 2011). Despite its far-reaching implications, Wakefield's misconduct directly involved only one publication, with a sample size of 12 children.

Other researchers have manipulated or fabricated evidence on a much larger scale. Critical care specialist Joachim Boldt specialised in the study of hydroxyethyl starches (HES), synthetic colloids widely used to replace blood transfusions to maintain blood pressure during surgeries. The body of literature on the subject found no evidence of harm in patients treated with HES compared to blood transfusions or other substitutes. However a meta-analysis of research on HES covering 38 trials and 10 880 patients found that, if Boldt's contributions were excluded from the analysis,

HES was associated with a significantly higher risk of mortality and renal failure (Ryan Zarychanski, 2013). Around the same time, a medical board inquiry revealed sweeping malpractice in Boldt's research, from failing to obtain approval by the ethics committee to fabricating entire studies and forging co-authors' signatures, prompting the retraction of 89 publications (Marcus, 2013). Boldt now faces criminal charges. Regulatory bodies must now review their guidelines regarding the transfusion of blood substitutes (Gever, 2011).

Biological sciences and medicine have been studied more closely in this respect, possibly because of the obvious impacts of faulty research. They are by no means the only fields of research where errors and fraud occur. A study of 4449 scholarly publications retracted from 1928-2011 across 42 bibliographic databases in all fields found that 47% of retractions were attributable to publishing misconduct, 20% to alleged research misconduct, and 42% to questionable data and interpretations, with some retractions attributed to more than one cause (Michael L. Griesen, 2012). The study did find that, as of 2010, medicine, chemistry, the life sciences, and multidisciplinary sciences were over represented in the retracted sample with respect to their representation in Web of Science records, while mathematics, physics, engineering, agriculture, earth and space sciences, ecology, natural resources, humanities, and social sciences were underrepresented. The study also found an increasing trend in retractions over time and that the "top" 15 offenders collectively accounted for 52% of the 725 retractions due to alleged research misconduct. Nine out of fifteen worked in medical fields.

Cases of repeat offenders clearly exist in other disciplines, but their low representation might be due to differences in monitoring mechanisms, less palpable negative impacts, *bona fide* lower occurrence or lesser pressure to publish "glamorous" research. Recently, social psychologist Diederik Stapel saw 53 of his publications retracted after an inquiry stemming from students' complaints demonstrated that he had manipulated and fabricated data and entire studies.

Other recent events demonstrated that the effects of fraud in social sciences could be as destructive, if not more so, than in medical research, when a University of Massachusetts team found flaws in the data behind a paper by Harvard economists Carmen M. Reinhart and Kenneth Rogoff. The paper was repeatedly cited by politicians to justify austerity measures. The Excel spreadsheet used to compile the study's statistics contained errors and, intentionally or not, excluded countries where results were not consistent with the studies' conclusions.

Cases of retractions and flawed studies, whether the flaws are due to error or scientific misconduct, should be a major argument in favour of open access to scientific data as a complement to traditional peer review. Easily accessible data would extend the peer review process beyond a small group of reviewers acting before publication to all readers after publication.

One of the responses to this potentially important challenge would be for universities and research hospitals to make experimental data available to all. However, as mentioned earlier, OA policies for scientific data are less common than OA policies for scholarly publications. Librarians and researchers, two communities that advocate for OA to peer-reviewed publications, are far less vocal about data. Arguably, the effort required to publish datasets in usable forms is greater than the effort required to archive papers or to publish in gold journals. From the librarians' point of view it may also seem less urgent to push for OA data than for OA peer-reviewed publications, considering the latter's bearing on their budget and the fact that managing publications is the traditional purview of librarians.

In a survey of head librarians at universities and higher learning institutions conducted by Science-Metrix, 73% of respondents agreed or strongly agreed that providing scholarly

publications in open access form is a priority in their organisation whereas only 45% agreed or strongly agreed that providing scientific data in open access form is a priority in their organisation (Table II). A similar trend is observed in the adoption of OA policies for scholarly publications and scientific data. Only 11% of respondents to the survey indicated that their institution has an OA policy for scientific data, while in terms of scholarly publications, 42% indicated that their institution has an OA policy (Table III). Thus, establishing institutional frameworks for the diffusion of data in open access remains a secondary concern at this point.

Table II Perceived priority given to Open Access scientific data and scholarly publications at the national and organisational level

	Head librarians (n=162)	
To what extent do you agree or disagree with the following statements:	Agree	Disagree
Providing scientific data in open access form is a priority in my organisation	45%	42%
Providing scholarly publications in open access form is a priority in my organisation.	73%	25%

Notes: Ratings of "agree" and "strongly agree" were combined, as were ratings of "disagree" and "strongly disagree". The percentage includes all respondents, including those who did not provide a rating.
Source: Science-Metrix survey of head librarians at universities and higher learning institutions.

Table III Prevalence of Open Access policies for scientific data and scholarly publications in universities and higher learning institutions

	Head librarians (n=162)	
	Yes	No
Does your organisation have an open access policy regarding scientific data produced as part of research?	11%	72%
Does your organisation have an open access policy regarding peer-reviewed scholarly publications (e.g., self-archiving policy; publications in open access journals)?	42%	52%

Notes: The percentages include all respondents, including those who did not provide an answer. Only respondents who declared that their organisation has a policy were asked if it was publicly available.
Source: Science-Metrix survey of head librarians at universities and higher learning institutions.

3.2 Infrastructure developed by research institutions

There are a handful of institutional repositories dedicated to research data. It is more common for institutions to support datasets on repositories otherwise devoted to books, theses, peer-reviewed publications and multimedia objects. In the survey of head librarians at universities and higher learning institutions conducted by Science-Metrix, 36% of respondents indicated that their organisation maintains one or more repositories for open access scientific data (Table IV), whereas 11% of respondents indicated that their organisation had a policy to this effect. Thus, the infrastructure required to host and share scientific data, while still uncommon, seems more developed than the associated policy frameworks.

The survey's results also suggest that a significant number of existing OA repositories for scientific data are not indexed in the directories and registries used in this study. For the countries covered in this study, ROAR lists 7 repositories for scientific data and Open DOAR lists 77 repositories that support the deposit of datasets. The ROAR and OpenDOAR figures encompass institutional repositories as well as disciplinary, governmental, and aggregating repositories. Considering the small sample size of the survey, it seems unlikely that our respondents could

account for 59 such repositories and it can therefore be surmised that ROAR and OpenDOAR fail to provide anywhere near a complete inventory of scientific data. This impression is reinforced by the absence of known open data repositories, such as the NASA's AERONET and the European Molecular Biology Laboratory's databases, from the directories and registries. OpenDOAR is the most comprehensive directory of OA repositories, but organisations must register for their repositories to be listed (Directory of Open Access Repositories, 2013). It is therefore by no means exhaustive and the repositories listed could merely be the tip of the iceberg.

Table IV Prevalence and type of repositories used to archive open access scientific data

Does your organisation maintain a repository (or several) for open access scientific data? Check all that apply.	Frequency	%
Yes, Centralised	47	28%
Yes, Decentralised	12	7%
No	78	47%
Don't know/Not applicable	16	10%
Other (please specify)	11	7%
NR	1	<1%

Notes: Respondents could select more than one answer.
Source: Science-Metrix survey of head librarians at universities and higher learning institutions.

4 Discussion and conclusion

The concept of open access to scientific data was first implemented more than 50 years ago to share data collected as part of the 1957-1958 International Geophysical Year by the International Council of Science (ICSU) (Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council, 2008). Since then, its growth has followed a complex pattern in terms of its spread across disciplines, countries, and types of organisations. Today, a few mature databases have grown to become inevitable references, while an unknown number of databases are still in their infancy. Importantly, however, compared to open access repositories of theses and scientific articles, institutional repositories that support the archiving of scientific datasets remain marginal.

Open data is evolving rapidly in an environment where citizens, institutions, governments, non-profits, and private corporations loosely cooperate to develop infrastructures, standards, prototypes, and business models. Hack-a-thons and social media are testing grounds for novel approaches. Repository overlay journals are experimenting with new forms of scholarly communication involving raw data.

Several disciplinary standards have been developed to index specific types of records in the face of an ever-rising tide of data. Although the development of some of these standards predates widespread internet access, all have evolved to adapt to it and most use XML, which is readable by machines and humans alike. A non-extensive list of standards currently used is presented in Table V. Many of these standards share characteristics which allow researchers to cross datasets emanating from different fields. For example, ecological data can be crossed with genomic data using taxonomic information or with meteorological data using geographic coordinates.

Though this report has mainly discussed governmental and institutional efforts, there are also noteworthy efforts in the private sector. For instance, Figshare is a platform where researchers can archive and share files including datasets, figures, audio and video files, posters and articles. Figshare encourages the diffusion of smaller datasets and negative results that are of little interest to publishers but can inform other researchers' process while conducting related research. A DOI is attributed to each object on Figshare to ensure traceability and citability. All figures, media, texts and filesets are registered under CC-BY[4] licence while datasets are registered under CC0[5] (Figshare, 2013). The platform's social media component allows for viral dissemination of content within the community and generates metrics based on sharing and uploads.

[4] CC-BY. Attribution 3.0 Unported: You are free: to Share — to copy, distribute and transmit the work; to Remix — to adapt the work; to make commercial use of the work. Under the following conditions: Attribution — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). See complete license at: http://creativecommons.org/licenses/by/3.0/

[5] CC0. Universal. Public Domain Dedication: 'The person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission'. See complete license at: http://creativecommons.org/publicdomain/zero/1.0/

Table V Disciplinary data archiving standards

Discipline	Standard	Types of data
Biology	Access to Biological Collections Data (ABCD)	Primary biodiversity data
	Darwin Core	Biodiversity data
	Ecological Metadata Language (EML)	Ecological data
	ISA-Tab	Combined 'omics'-based data
	Minimum Information for Biological and Biomedical Investigations-MIBBI	Minimum information from 40 biological research domains
	Open Microscopy Environment XML	Light microscopy data
Earth Sciences	Agricultural Metadata Element Set (AgMES)	Agricultural data
	Climate and Forecast (CF) Metadata Conventions	Meteorological data with space-time coordinates
	Directory Interchange Format (DIF)	Earth sciences datasets with information concerning instruments and space-time coordinates
	Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC/CSDGM)	Obsolete format for geospatial data. Replaced by ISO 19115
	ISO 19115	Comprehensive standard for geospatial data
	Metaphor Common Information Model CIM	Climate data, models, simulations, and software
Physical Science	Astronomy Visualization Metadata (AVM)	Astronomical imagery
	Crystallographic Information Framework (CIF)	Chemical structure data
	Council for the Central Laboratory of the Research Councils Core Scientific Metadata Model (CSMD-CCLRC)	High-level, generic information about scientific studies and associated data
	International Virtual Observatory Alliance Technical Specifications	Integrates 9 astronomical data types into a single virtual observatory
	Space Physics Archive Search and Extract (SPASE) Data Model	Heliophysics data
Social Sciences and Humanities	Data Documentation Initiative (DDI)	Data derived from the social, behavioral, and economic sciences
	MIDAS-Heritage	Cultural heritage data (buildings, archeological sites, shipwrecks, battlefields, parks, gardens, etc.)
	Statistical Data and Metadata Exchange (SDMX)	Statistical data

Source: Compiled by Science-Metrix based on data from the Digital Curation Centre

Importantly, whereas scholarly journal publishers often considered OA publications to be a threat to their traditional markets, the legal and economic implications for publishers are far fewer than in the case of scholarly publications. In fact, the Association of Learned and Professional Society Publishers (ALPSP) and the International Association of Scientific, Technical & Medical Publishers (STM) declared in a joint statement that "as a general principle, data sets [...] should wherever possible be made freely accessible to other scholars. We believe that the best practice for scholarly journal publishers is to separate supporting data from the article itself, and not to require any transfer of or ownership in such data or data sets as a condition of publication of the article in question" (ALSP and STM, 2006). This points to a potential facilitator for the development of OA data.

From a governmental point of view, "access to research data increases the returns from public investment in this area; reinforces open scientific inquiry; encourages diversity of studies and opinion; promotes new areas of work and enables the exploration of topics not envisioned by the initial investigators" (OECD, 2007). Overall, open data may spur economic development in an increasingly information-based economy. New products and services can be developed directly from the data or through extensive transformation adding value to the information. Crossing information from various sources can create opportunities for innovation.

The costs of setting up and maintaining repositories and databases are not recovered easily and this is obviously even more a cause for concern with OA data. Some organisations use their data as a source of revenue, by providing paid access to their datasets. Giving free, immediate and unlimited access would eliminate this source of funds. In the case of governmental agencies within the EU, a review has shown that the direct revenues generated from the sale of government-owned data are relatively low, with upper estimates ranging between €1.4–3.4 billion (Vickery, 2011). These estimates are based on the revenues generated by the sale of data in the UK and in the Netherlands, countries which have been particularly efficient at collecting these revenues. The actual amount generated across the EU-27 is probably lower, and in most cases it represents less than 1% of each agency's revenues. However, for a few organisations, the sale of data may represent close to 20% of revenues, which is likely to hinder the development of open data policies even if the data was generated using public funds. Still, the economic benefits of open data would outweigh sales revenue by approximately two orders of magnitude within the EU-27.

The heterogeneous nature of scientific data certainly is a challenge for the development of OA data. The emergence of OA scientific data as a valid, citable form of reference is limited by the difficulties associated with the standardisation of data and metadata formats, poor indexation by internet browsers, as well as by the scarcity of directories or registries that could make data more visible. Initiatives from academia and from the non-profit and private sectors seek to address these limitations. A few fields of research use highly standardised formats that facilitate the aggregation and reuse of data; genomics, proteomics, chemical crystallography, geography, astronomy and archaeology are among those fields. The archiving of other, less standardized types of data needs to be carefully thought out in order to generate datasets that will be usable by other researchers. The proliferation of data archiving standards indicates that this issue is addressed by communities of researchers, librarians, and database administrators but probably will not be settled in the near future.

One response may be provided by DataCite which is a non-profit organisation formed in London in 2009 and currently managed by the German National Library of Science aiming to provide easier access to research data online, increase the acceptance of research data as legitimate, citable contributions to science, and support data archiving in order for it to be verified and reused for

further study. The organisation has members in 15 countries, including Canada, Denmark, France, Germany, Italy, the Netherlands, Sweden, Switzerland, the UK, and US. DataCite signed a memorandum of understanding with the Registry of Research Data Repositories (re3data) in 2012 to collaborate on the elaboration of norms for the purposes of indexing research data repositories. Formed by the Berlin School of Information Science, the German Research Centre for Geosciences, and the central library of the Karlsruhe Institute of Technology, re3data aims to build a registry of research data repositories, to define selection criteria, and to formulate a metadata schema to describe research data repositories. If it were adopted, the extensive XML metadata schema proposed would greatly facilitate the discovery and reuse of scientific data. Although its complexity might hinder its uptake by the research community at the archiving level (input), its usability would be superior to pre-existing schemas, such as the Dublin Core, in terms of clear documentation of important characteristics of datasets and repositories.

Confidentiality certainly is a concern in the dissemination of data generated by studies of human subjects. Anonymisation of data is crucial but may pose a challenge if the sample population comes from a small group where subjects could nevertheless be identified.

The relatively slow progression of OA data repositories may be due in part to the lack of champions, such as been the case with OA scientific papers repositories which might have developed faster due to the role played by librarians. Scientists such as Harnad, Suber and others certainly had traction and helped the OA scholarly publication move along at relatively great speed. There seem to be far fewer visible vocal champions for OA data. Whereas open up scholarly publications was welcomed by many in the scientific community, giving away one's data may be harder. Scientists generally want the results of their research to be as widely known as possible and this may makes the OA scholarly paper movement a natural evolution of the scientific system. However, many scientists spend painstaking time collecting hard to get data which can then be used progressively to build a career and a large research team and to derive a competitive advantage vis-à-vis colleagues who do not have access to these data. Having hard to get data, publishing, and getting more grant money to continue to build up this system is commonly seen in academia. It might be more difficult for scientists who have built their career on this model to espouse communalism for themselves that it is for them to ask seemingly rich publishers to give up their own business models. Researchers are frequently seen with an idealistic model whereby they uninterested but many of the most successful ones are in fact running small businesses, giving up the data on which these businesses are built may not be easy and this can pose a very tangible barrier to greater data openness.

References

Anon. (2010). Global Positioning System, *U.S Code Sec*. 2281. Available at: http://www.gpo.gov/fdsys/pkg/USCODE-2011-title10/pdf/USCODE-2011-title10-subtitleA-partIV-chap136-sec2281.pdf

Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council (2008). *Earth Observations from Space: The First 50 Years of Scientific Achievements*. Washington D.C.: The National Academies Press.

Deer, B. (2011). How the case against MMR vaccine was fixed, *BMJ*, 342(c5347), 77-82.

Directory of Open Access Repositories. (2013). *OpenDOAR frequently asked questions*. Available at: http://www.opendoar.org/faq.html

EMA. (2013). *European Medicines Agency receives interim decisions of the General Court of the EU on access to clinical and non-clinical information*. Available at: http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2013/04/news_detail_001779.jsp&mid=WC0b01ac058004d5c1

European Commission. (2009). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee of the Regions - ICT infrastructures for e-science*. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52009DC0108:EN:NOT

FDAAA 801. (2007). *Food and Drug Adeministration Amendments Act, Section 801*. Available at: http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ085.110.pdf#page=82

Fenner, M. (2012). *Figshare: Interview with Mark Hahnel*. Available at: http://blogs.plos.org/mfenner/2012/02/16/figshare-interview-with-mark-hahnel/

Fang, F.C., Steen, R.G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, 109(42), 17028-17033.

Figshare. (2013). *How is your uploaded data licensed?* Available at: http://figshare.com/cc_license

Gever, J. (2011). New Twist in Colloids-Crystalloids Tussle. *MedPage Today*, 4 March.

GPS Innovation Alliance. (2012). *The U.S. Governement is the Biggest Single User of GPS; Has Invested at leat $43 Billion in IT*. Available at: http://www.gpsalliance.org/gps-key-documents.aspx

Marcus, A. (2013). German medical board issues sweeping findings in Boldt case. *Anesthesiology News*, 4 February.

McNeff, J. G. (2002). The Global Positioning System. *IEEE Transactions on Microwave Theory and Techniques*, 50(3), 645-652.

Grieneisen ML, Zhang M (2012) A Comprehensive Survey of Retracted Articles from the Scholarly Literature. *PLoS ONE*. 7(10), e44118. doi:10.1371/journal.pone.0044118

OECD. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Available at: http://www.oecd.org/science/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm

Pellerin, C. (2006). *United States Updates Global Positioning System Technology, New GPS satellite ushers in a range of future improvements*. Available at: http://www.america.gov/st/washfile-english/2006/February/20060203125928lcnirellep0.5061609.html

Pham, N.D. (2011). *The Benefits of Commercial GPS Use in the U.S. and The Costs of Potential Disruption*, s.l.: ndp consulting.

Rabesandratana, T. (2013). Drug Watchdog Ponders How to Open Clinical Trial Data Vault. *Science Translational Medicine*, 339(6126), 1369-1370.

Sec. 57 (EC) *Regulation 726, 2004*. REgulation (EC) No 1901/2006 of the European Parliement and of the Council of December 2006 on medicinal products for paediatric use and amending Regulation (EEC) No 1768/92, directive 2001/20/EC, Directive 2001/83/EC and Regulation (EC) No 726/2004. Available at: http://ec.europa.eu/health/files/eudralex/vol-1/reg_2004_726_cons/reg_2004_726_cons_en.pdf#zoom=100,0,0

STM and ALPSP. (2006). *Databases, data sets, and data accessibility - views and practices of scholarly publishers*. Available at: http://www.google.ca/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CDQQFjAA&url=http%3A%2F%2Fwww.stm-assoc.org%2F2006_06_01_STM_ALPSP_Data_Statement.pdf&ei=TJ96UaTIC83j4APSz4DADQ&usg=AFQjCNH83OkH0Gung69Cz7E9DIwki22wPQ&bvm=bv.45645796,d.dmg

Zarychanski, R. et al., 2013. Association of hydroxyethyl strarch administration with mortality and acute kidney injury in critically ill patients requiring volume resuscitation, a systematic review and analysis. *Journal of the American Medical Association*, 309(7), 678-688.