# A Path to Big Data Readiness

Claire C. Austin, PhD
Science and Technology Strategies Directorate*
Environment and Climate Change Canada
Gatineau, Canada
claire.austin@canada.ca, https://orcid.org/0000-0001-9138-5986

***Abstract*—"*Big Data readiness*" begins at the source where data are first created and extends along a path through an organization to the outside world. This paper focuses on practical solutions to common problems experienced when integrating diverse datasets from disparate sources. Following the Introduction, Section 2 situates *Big Data* in the larger context of open government, open science, science integrity, and Standards, internationally and in Canada. Section 3 analyses the *Big Data* problem space, while Section 4 proposes a *Big Data* solution space. Section 5 proposes eight data checklist modules and suggests implementation strategies to effectively meet a variety of organizational needs. Section 6 summarizes conclusions and describes future work.**

***Keywords*—Big Data, data quality, data checklist, data repository, open science, open government, data science.**

* All opinions author's own.

## I. INTRODUCTION

*Big Data*[1] has the potential to answer questions, provide new insights previously inaccessible, and strengthen evidence-informed decision making. However, the harnessing of data into the *Big Data* net can also very easily overwhelm existing resources and approaches, keeping those answers and insights out of reach. For the purposes of this paper, data that can potentially end up in the *Big Data* net are research data, defined as: "*Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data*" [1].

Environment and Climate Change Canada, like many organizations, has important data assets for various uses, including confidential and sensitive data, and multiple data flow pathways. This heterogeneity presents people at the working level and upper management alike with enormous challenges in developing and implementing solutions that will enable *Big Data* and *Big Data Analytics*.

The purpose of this paper is to contribute to the development of innovative thinking transferable to a wide range of organizations and domains with the goal of effecting changes needed to achieve *Big Data*. To support corporate governance and data management planning and strategies that may not yet be fully developed, this paper offers suggestions for a PATH TO "*BIG DATA READINESS*" based on Open Science, FAIR[2,3,4,5,6] data and an "*It's good enough*" approach. FAIR data, endorsed by the G20 in 2016 means that the data are Findable, Accessible, Interoperable, and Reusable. "*It's good enough*" means doing what can be done now to make things work with the tools and the people currently in place. A "*Big Data readiness*" approach will support long-term planning and enable short-term solutions.

This paper proposes a generic strategy and actions directed primarily at the working level that can be anticipated to have significantly positive short-term impacts without overwhelming workers, managers, or stakeholders, and to increase the chances of success of a *Big Data* project and implementation of a future data strategy. It will take some time to realize the business value of data strategies that may be under development in an organization and for some scenarios, an organization cannot afford to wait until implementation.

*Big Data* transformation does not need to happen all at once; nor does the organization or its base need to wait for the development of a *Big Data* Framework, governance model, data policy, data strategy, master data management, or Open Science plan before taking action to help accelerate the implementation of *Big Data*. The approach presented in this paper is proposed as an effective first step for what can be done now in the present (taking into account current organizational maturity, capabilities, and data flow realities) to position an organization to meet opportunities provided by the *Big Data* revolution.

## II. OPEN SCIENCE AND OPEN GOVERNMENT

### A. Open Science

*Big Data* is one of the components of open science described in FOSTER's open science taxonomy (Facilitate Open Science Training for European Research, 2015, a European Commission funded project), cited also by the National Academies of Science, Engineering, and Medicine (NASEM) in 2018 in,

[1] Big Data consists of extensive datasets – primarily in the characteristics of volume, variety, velocity, and/or variability - that require a scalable architecture for efficient storage, manipulation, and analysis (National Institute for Standards and Interoperability (NIST), Big Data technology Roadmap, Big Data Public Working Group, Volume 1, Definitions, Draft version 2). International Standards Organization - ISO (2018), "Information technology - Big data - Overview and vocabulary," Draft International Standard, ISO/IEC DIS 20546).

[2] http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm

[3] https://www.nature.com/articles/sdata201618

[4] https://github.com/FAIR-Data-EG/Action-plan

[5] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

[6] https://rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases.html

"Open science by design" [2,3]. Open science includes: (1) open science ***policies***; (2) open science ***guidelines***; (3) open access ***publications***; (4) open science ***tools*** (open services, open workflow tools, curated repositories); (5) ***open data*** (Big Data, data use and reuse, data journals, data standards, FAIR data); (6) ***reproducible research*** (tools, workflows, open source, open code, open lab/notebooks); (7) open science ***projects***; and, (8) open science ***evaluation*** (metrics, indicators, impact).

Governments have significant data assets and their commitment to open data and open science is critical for *Big Data* to become a reality. It is informative to consult the European Union's 2016 Amsterdam Call for Action on Open Science [4], the European Open Science Agenda [5], and the April 22, 2018 European Commission's Open Science Policy Platform (OSPP) Recommendations for achieving open science [6,7]. There is a recommendation that all researchers be required to deposit their research outcomes in infrastructures compliant with the European Open Science Cloud (EOSC) to ensure interoperability and the free movement of information across all national and international boundaries and between disciplines. On April 25, 2018, the European Commission published 12 recommendations on Access to and Preservation of Scientific Information [8]. In June 2018, the U.S. National Institutes of Health (NIH) released its first Strategic Plan for Data Science that provides a roadmap for modernizing the NIH-funded data science ecosystem.[7] On July 4, 2018, France adopted a National Plan for Open Science that aligns with the French 2018-2020 OGP National Action Plan on Open Government commitment to open science: "*Developing an open science ecosystem*" [9-10]. The Science Europe September 4, 2018, Plan S proposal would require Open Access publication with penalties for noncompliance.[8] Open science is gaining momentum.

Foundations, non-profits, and academia are also important players in advancing open science. Examples of the former include the Wellcome-Trust policy on sharing research outputs,[9] and the Center for Open Science (COS) that provides valuable tools and guidelines.[10] Journals increasingly encourage deposit of data and computer code associated with published articles, new peer-reviewed data journals are in existence, there is a rise in institutional data repositories, and the need for reproducibility and openness is leading to a re-examination of the traditional reward system and criteria for assessing scientists [11-14].

### *B. Government investment*

A European Commission impact study on re-use of public sector information estimated that total direct economic value of public sector information (i.e. government data) in the EU is expected to increase from 34B€ in 2018 to 195B€ in 2030 [15]. Governments are investing in open science as an economic stimulus. The French National Plan for Open Science, for example, includes 5.4M€ the first year and 3.4M€ the following years for open access and open data.

In February 2018, the Government of Canada (GoC) Federal Budget included a focus on harnessing *Big Data*. The government proposed a $4 billion investment in Canada's research system to support the work of researchers and to provide them access to the state-of-the-art tools and facilities. This includes $572.5 million over five years, with $52 million per year on going, to implement a Digital Research Infrastructure Strategy that will deliver more open and equitable access to advanced computing and *Big Data* resources to researchers across Canada. The Minister of Science is working with interested stakeholders, including provinces, territories and universities, to develop the digital strategy, including how to incorporate the roles currently played by the Canada Foundation for Innovation, [11] Compute Canada, [12] and CANARIE, [13] to provide for more streamlined access for Canadian researchers.

### *C. Open government in the Canadian federal government*

In 2006, pro-active disclosure was introduced by the Federal Accountability Act.[14] In 2011, Canada joined the international Open Government Partnership (OGP)[15] and, in 2013, endorsed the G8 Charter on Open Data [16]. In October 2014, the GoC Directive on Open Government[16] was issued under the authority of the Financial Administration Act.[17] In 2017, Canada was elected to the OGP Steering Committee for a 3-year term (2017-2020) and, in October 2018, assumed the role of Lead Government Chair for a period of one year.

The position of Canada's Chief Science Advisor (CSA), created in 2017, reports to both the Prime Minister and to the Minister of Science. The CSA's mandate includes development and implementation of guidelines to ensure that government science is fully available to the public. One of the priorities is development of an open science framework. In 2018, the position of Chief Information Officer (CIO) of the Government of Canada was elevated to the level of Deputy Minister reporting to the newly created post, Minister of Digital Government. The CIO is a Public Officer of the committee of the Queen's Privy Council for Canada, known as Treasury Board Secretariat, established in 1985 by the Financial Administration Act. In 2018, Canada joined the Digital 7, a network of the world's most advanced digital nations, and committed to the D7 Charter [17].

Open Science was first included in Canada's 2nd OGP National Action Plan on Open Government in 2014 and carried over to the 3rd Plan in 2016. Draft 4th Plan[18,19] activities relating to Open Science span five of 10 OGP commitments and 11

[7] https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf
[8] https://www.scienceeurope.org/coalition-s/
[9] https://wellcome.ac.uk/funding/guidance/policy-data-software-materials-management-and-sharing
[10] https://cos.io/our-products/osf/
[11] https://www.innovation.ca/about
[12] https://www.computecanada.ca/about/
[13] https://www.canarie.ca/about-us/
[14] http://laws-lois.justice.gc.ca/eng/acts/F-5.5/
[15] https://www.opengovpartnership.org/about/about-ogp
[16] https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=28108
[17] http://laws-lois.justice.gc.ca/eng/acts/f-11/FullText.html
[18] CSA (Chief Science Advisor, Prime Minister's Office); CSPS (Canada School of Public Service); ECCC (Environment and Climate Change Canada); IDRC (International Development Research Centre, a Crown Corporation); LAC (Library and Archives Canada); NRC (National Research Council); NRCan/FGP (Natural Resources Canada/Federal Geospatial Platform); OD4D (Open Data for Development under IDRC); PCO (Privy council Office); StatCan (Statistics Canada); and, TBS (Treasury Board Secretariat).
[19] https://open.canada.ca/en/4plan/creating-canadas-4th-plan-open-government-2018-20

departmental leads. These include development of data quality criteria, creation of a Digital Policy and Data Strategy, promotion of open science and wide consultation on stakeholder and federal scientists' needs with respect to open data and open science, and implementation of a pilot project for cross-jurisdictional common data standards. Canada's participation in the OGP serves as a vehicle for promoting Open Science.

### *D. Science integrity*

In 2018, Canada's CSA released a model Science Integrity Policy (mSIP) acknowledging the importance of openness and transparency about all elements of the research and scientific process as well as timely release of scientific and research information in keeping with the Directive on Open Government [18]. Indeed, the policy's definition of scientific integrity specifically includes adherence to the concepts of transparency and openness, mirroring the definition of Open Science. See, also, NASEM's 2017 "Fostering integrity in research" [19].

### *E. Big Data Standards*

The 2018 European Commission impact study on re-use of public sector information concluded that the use of standards is the most underestimated factor of data re-use [15]. That *Big Data* is a rapidly evolving area is evidenced by global efforts to develop new international Standards to provide guidance as organizations collectively move forward. Several international working groups (WG) are focusing on this area, including the National Institute of Standards and Technology (NIST) *Big Data Interoperability Framework,* [20] International Standards Organization (ISO) *Big Data – Artificial Intelligence,* [21] and IEEE *Big Data Governance and Metadata Management.* [22] These efforts, results from Research Data Alliance (RDA) workgroups, [23] and others will profoundly affect overall data management practices at all levels and for all types of data.

## III. THE BIG DATA PROBLEM SPACE

### *A. Barriers to Big Data*

#### *1) Legacy systems*

Data management gaps at the working level and lack of data governance at the corporate level have been identified in organizations in private and public sectors dealing with decades old systems and procedures. Legacy systems do not only refer to the dark data buried in printed output, on CDs, in notebooks, on external hard drives, and on personal computers, etc. Legacy systems also refers to hardware and software that are still in use in the organization but are no longer supported by either the original vendor or by the organization's IT department, and to in-house computer code that may be poorly documented or developed without a well-structured approach. Additional challenges include more recent hardware and software that fail to meet the demands of *Big Data* and modern analytics, and people who experience challenges in adapting to new ways of doing things. New organizations may have a competitive advantage in that they have the opportunity to build state of the art systems from scratch relatively inexpensively, unencumbered by legacy systems or by other technical and non-technical barriers that are a function of an organization's overall readiness for *Big Data* measured by organizational maturity, organizational capability, and organizational alignment.

#### *2) Organizational maturity*

Increased organizational maturity has been observed in the public sector, where there is more structural collaboration between organizations [20]. Organizational capabilities for *Big Data* can be described in terms of internal attitude, external attitude, legal compliance, IT resources, data science expertise, IT governance, and data governance, with the last three leading to the greatest improvements in organizational capability. Organizational alignment (i.e. whether or not *Big Data* applications are suited for the organization in question) was found to be vital for the success of *Big Data*. When evaluating organizational alignment, it is not surprising that the intensity of data *use* was found to be a determinant of the readiness for *Big Data*. Paradoxically, intensity of data *collection* was not necessarily associated with data quality or readiness for *Big Data*. This is important to keep in mind in the case of monitoring networks (writ large), for example, where the intensity of data collection is high, but the intensity of data use is low because the data users are found elsewhere within the organization or externally. It may well be that the greatest barrier to *Big Data* is not organizational maturity or capability, but organizational alignment with the data provider's or program priorities.

#### *3) Breaking out of 'Lock-in'*

Organizations can be locked into old ways of thinking and old ways of doing things that impede *Big Data*. Best practices in data management have not kept up with changes in technology that resulted in a rapid increase in the speed of generation, quantity, variety, complexity, variability and new uses for the data collected. In addition, there is uncertainty regarding data accuracy, inconsistency in vocabulary, and confusion over the meaning of *Big Data*, data mining, and artificial intelligence. Meanwhile, many organizations are still struggling to emerge from a paper-based world governed in siloed organizations to a digitally literate and interconnected world. This is a very difficult transition. It requires the transformation of longstanding, well-adapted thinking processes that no longer work well, to new thinking processes adapted to a new world.

#### *4) Culture change*

*Big Data* is being propelled from an emerging area to the fore of open data and open science. However, data that may be "locked in" traditional approaches are largely inaccessible to *Big Data* end users. This limits an organization's ability to use *Big Data* approaches for knowledge acquisition and innovation. Changes in thinking across organizations are needed to achieve a coordinated and harmonized system that is simple, effective and geared to meet organizational needs.

Organizations and various groups within them have developed data management processes that work for them internally. They tend to be project- or client-centric to meet their

[20] https://bigdatawg.nist.gov/

[21] https://iecetech.org/Technical-Committees/2018-03/First-International-Standards-committee-for-entire-AI-ecosystem

[22] https://standards.ieee.org/industry-connections/BDGMM-index.html

[23] https://www.rd-alliance.org/

specific mandate and needs, but not necessarily user-centric in the context of open science and *Big Data* where the user is unknown. A paradigm shift in thinking and culture is needed in many organizations to achieve agile delivery of "analysis-ready" data that can be incorporated seamlessly into a *Big Data* workflow. The underlying principle for success is a "*Big Data readiness*" approach from the bottom up at the working level, in operations, research, and business lines. Targeted generic actions will help create the necessary conditions on the ground. Culture change will follow.

This bottom up change in thinking and culture must work hand-in-hand with top down culture change that needs to happen if data are to become a strategic asset. Resources assigned to data life-cycle management must become a priority for program areas, supported appropriately by senior managers. Ultimately, sustainable culture change needs to work in both directions.

*5) Data standards*

There is a need for common data standards for the preparation and updating of FAIR data. Previous approaches to data governance may have led to uncontrolled data flows, data fragmentation, variation in data quality, and incomplete information concerning the data (Fig. 1). Where this may be satisfactory within specific mandates, it is problematic for open science, reproducible research and *Big Data*.

*6) Data quality*

Gartner estimates that poor data quality costs an average organization $13.5 million per year and that data governance problems are worsening [21]. There are seven levels of data quality: (1) Quality of the observations or measurements; (2) Quality of the recording of the observations and measurements; (3) Quality of the descriptors associated with the observations and measurements; (4) Quality of the information needed for an end user to completely understand the data and their limitations; (5) Organization of the observations/measurements/descriptors in a dataset or collection; (6) Compliance with recognized consensus Standards; and, (7) Quality of the management of the data and information, including sharing. While there is a need for shared responsibilities across all six levels, the first two are primarily the realm of domain expertise, the fourth requires domain and information management expertise, and the last two are primarily data management expertise.

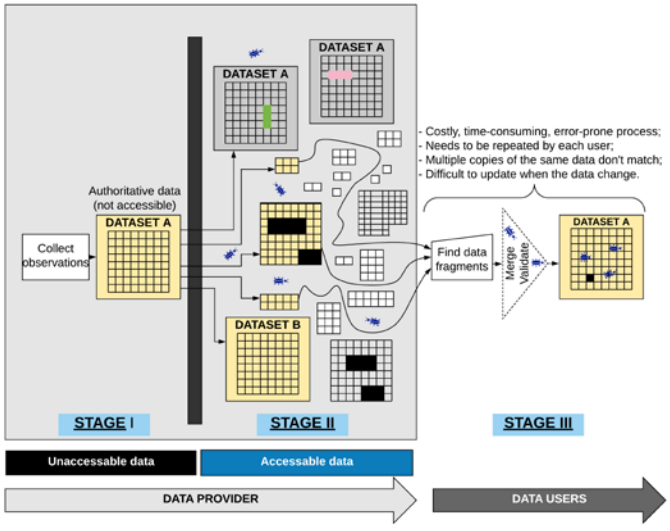


Fig. 1. Dataset fragmentation. In Stage I, the data provider produces high quality observations and measurements. In Stage II, the data are published to various platforms and portals, during which data fragmentation and duplication may occur and the data lineage lost. During Stage III, the data user must find all of the data fragments and reassemble them into something resembling the original dataset in Stage I.

A very high-quality dataset produced under strict quality assurance/quality control (QA/QC) protocols can become fragmented in the absence of data governance encompassing the complete data life cycle (Fig. 1). From the viewpoint of the data providers, they have produced extremely high quality data. From the viewpoint of the data users, they see poor quality data that are difficult or impossible to use. In order to use such data, each user inherits the task of reassembling the data before being able to use them yet lacks all the information needed to perform the task reliably. This is an error-prone, costly, time consuming, and inefficient use of resources. Furthermore, it is unlikely that data reassembled by different end-users will result in matching datasets. The problem compounds exponentially when trying to integrate these data into *Big Data*.

*7) Merging datasets from diverse sources*

A commonly seen workflow is illustrated in Fig. 2 where multiple datasets from different sources somehow have to be merged. In addition to the problem of dataset fragmentation and simply finding the data, there is confusion about which one is the approved copy, lack of version control, absent or incomplete metadata, lack of common fields, variety in nomenclature and measurement units, inconsistent data structures, etc.

Before the analyst can use the data, there may be unavoidable manual work involved in collecting and cleaning each of the data streams before they can be used (Stage III in Fig. 1), and in integrating these disparate data from diverse sources (Fig. 2). All of these data would be lost to *Big Data* where reliance on manual processes is no longer possible, or an inordinate amount of time would need to be spent on data preparation.

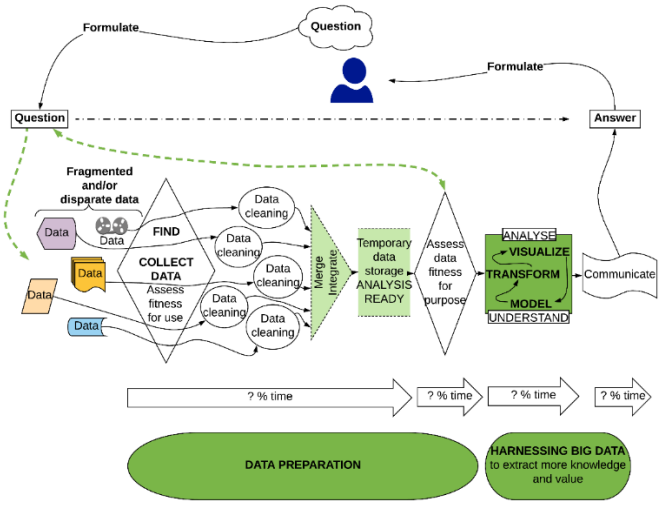


Fig. 2. Integration of data from diverse sources. If data providers published FAIR data that are analysis ready, data users would not need to spend 70-80% of their time on data preparation.

*Data preparation*

A major hurdle for the researcher or data scientist is data cleaning which can take up to 70% or more of the total time spent for the analysis [22], essentially performing tasks left undone when data providers release data that are not FAIR (Fig. 1,2). It takes enormous time, effort, and money to output small datasets to meet a variety of requests in Stage II of Fig. 1, and an even greater amount of time, effort and money for an analyst to reassemble the data before they can be used (Fig. 1, Stage III). Elimination of Stages II and III would eliminate the associated costs and wasted time, and result in more reliable analyses and stronger insights. Long-term data governance is the solution to these dataset, data flow, and metadata problems and to eliminating the hidden costs that result from them.

Short term targeted actions that address gaps in Data Governance and data management will improve the ability to integrate data from multiple sources and to reliably extract new knowledge and insights from large and complex collections of digital data. Adopting a "*Big Data readiness*" approach within an organization will help enable *Big Data analytics*, machine learning, and Artificial Intelligence (AI).

## IV. A BIG DATA SOLUTION SPACE

*1) A "Big Data readiness" approach*

The respective roles of data providers and data users require clarification. Data providers in the field, laboratory, and other organizational levels need to recognize at the outset that there will be unknown data users and that it is an integral part of their job to prepare their data to a standard that meets the requirements of these unknown users. Data providers also need to accept that how the data will be used and for what purpose will remain unknown to them. It is not the role of the data provider to assess if their data are fit for the purpose envisaged by some unknown user. That is the responsibility of the data user. However, to implement Open Data and *Big Data* it must be part of the data provider's role to make sure that data transmitted from one person or group to the next throughout the data life cycle are FAIR and tidy (organized for ease of use).

FAIR data include all related metadata and documentation so that an unknown end-user can completely understand the data and the data quality without having to contact the data provider. FAIR data have been verified by the data provider to be "fit for use" by any future unknown user who is then in a position to assess whether or not the data are "fit for purpose" in some specific context. FAIR, tidy, analysis ready data can be easily integrated into a *Big Data* workflow.

Best practices, standards, and training are key to data providers being able to prepare data appropriately. The organization must take on the responsibility of defining those practices and standards so that data can be integrated easily. A "*Big Data readiness*" approach should be included in organizational data strategies for short-term success in *Big Data* projects. For example, defining data quality and data standards strategies to support a Data Management Operational Plan could also include components of a "*Big Data readiness*" approach.

A "*Big Data readiness*" approach at the working level will concomitantly help solve existing data flow and data quality issues irrespective of whether or not the data will eventually enter a *Big Data* workflow. A "*Big Data readiness*" approach will improve an organization's overall data stewardship and governance, help make open data and open science a reality, and improve the chances of success of future corporate solutions such as a *Big Data* interoperability framework and Reference Architecture that support *Big Data* and analytics.

*2) Disrupting the status quo*

Implementation of a "*Big Data readiness*" approach at the working level may be easier to implement than imagined. The person best equipped to prepare "analysis-ready" data is the data provider – the person at the data source who knows the data best. Success in implementation of "*Big Data readiness*" requires inclusion of data providers – especially those who are experiencing the greatest challenges – in developing solutions. Inclusion means going beyond providing support. It means saying not only, "What can we do for you?" but also, "This is what we need from you." It means disrupting the status quo. "*Big Data readiness*" requires a paradigm shift in thinking at the working levels that is revolutionary, not evolutionary.

*3) It's "good enough"*

People are easily overwhelmed by disruption of the status quo. This can be mitigated by developing well thought out, "It's good enough" modular checklists that will result in what is needed now to move forward on the pathway to *Big Data*. It is unrealistic to expect that people at the working level, in the field and in the laboratories, have or can acquire the necessary skills and tools to design and maintain databases or to output their data in unfamiliar formats. However, it is realistic and necessary to expect that they can output their data in a form that can be easily understood and used by other people and systems. If this is achieved, it will be good enough.

*4) Data governance*

*Big Data* will not improve data quality, solve data management problems, reduce the need for good quality, well-managed data, or obviate requirements for competent statistical analysis. Grappling with poor quality data (Fig. 1,2) is not the essence of what it means to "harness" *Big Data*. Harnessing *Big Data* refers to analysts and systems extracting more knowledge from existing data. Data governance that also includes "*Big Data readiness*" is a fundamental and essential piece of the solution to extensive data preparation time and eliminating hidden costs.

Fig. 3 is a solution diagram for an organization. Data governance and improved data management frees up time for analysts to do analysis instead of data cleaning and preparation. The onus needs to be put on the data provider to provide FAIR data that are ready for analysis. Time thus freed-up can then be used for the harnessing of *Big Data* in the continuum of reproducible science.

Good data governance and FAIR data will result in reduction or elimination of inefficiencies and costly errors. Improved data quality, usability and discoverability will increase the value of data products thereby providing a bigger return on investment. *Big Data* can then reduce costs by reusing existing data instead of collecting more data unnecessarily. *Big Data* can also reduce costs by getting better answers more quickly.

*5) Open data science practices and tools*

While a stepwise move toward "*Big Data readiness*" and reproducible science means changing the way things are done with the tools currently in place, it also means adopting new tools and new competencies. For example, the EU funded Education for Data Intensive Science to Open New science frontiers (EDISON) aims to increase the number of competent data scientists [23]. Meanwhile, Lowndes et al. have published a refreshingly candid account of their path to adoption and implementation of open data science tools and reproducible science in a complex environmental sciences framework [24].

*6) Proposed strategy*

This proposal focuses on structured digital scientific data and the identification of a pathway from *Small Data* to *Big Data*, providing a rational stepwise approach to harnessing *Big Data*. Implementing actions that are generic and independent of systems currently in place means that they can be implemented "now":

1. Create awareness of "*Big Data readiness*" from the bottom up in operations and research contexts via communications such as newsletters, bulletins, and a dedicated website or wiki.
2. Provide online training modules to increase digital literacy across the organization.
3. Deploy "*It's good enough*" checklists for data Findability, Accessibility, Interoperability, and Re-usability (FAIR data) to help data providers produce data that are ready for *Big Data* workflows.
4. Implement a "*user-centric*," approach to data preparation to replace project- and client-centric approaches.[24]
5. Create linear data pathways to authoritative data sources to eliminate data fragmentation, duplication, and to preserve data lineage.
6. Develop and pilot test models of data-intensive scientific workflows for the preparation of FAIR, tidy, and analysis ready data and "reproducible science" in line with national and international best practices.
7. Encourage the use of open data science practices and tools.
8. Implement semi-automated data verification and feedback loops to ensure that data are ready for integration into *Big Data* workflows.
9. Maximize chances of success of Actions 1-8 by including data providers in the development of solutions.

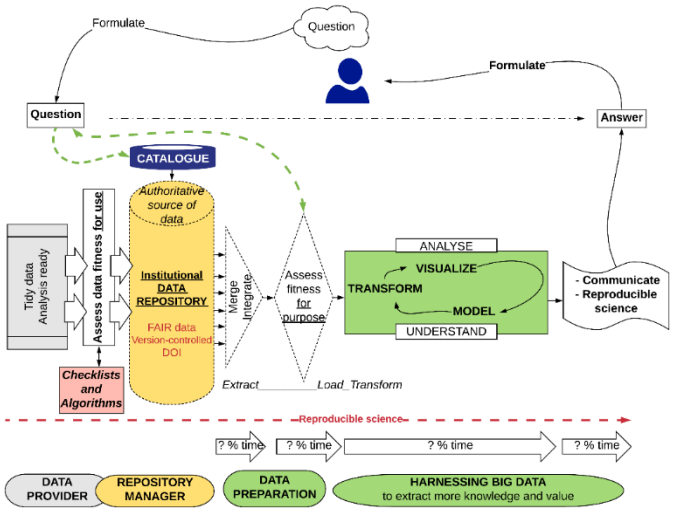


Fig. 3. Improved data governance. The data repository (yellow container) is the input/output control point and source of authoritative data. FAIR and analysis ready data on the left side of the diagram are released by the data providers. Semi-automated checklists implemented on the input side of the data repository are a critical component to ensure that the data are, in fact, FAIR.

## V. MULTI-FUNCTIONAL DATA CHECKLISTS

Data checklists can be a useful data management tool for data providers and data repositories, as well as for data stewards and managers who need to approve data without having been involved in their production. The use of checklists will help promote consistency, awareness, understanding, and efficiency in data governance. Implemented on the input side of the data repository in Fig. 3, they are a critical component to help ensure FAIR data and to maintain data quality, consistency, and transparency.

*1) Multiple uses for data checklists*

Well-designed checklists can serve multiple functions, for example:

1. The data provider can use the checklists as a data auto-evaluation tool.
2. The checklists can be used as a learning tool.
3. Checklist results can be submitted to data stewards and/or management along with or in lieu of the actual data for the purpose of data approval.
4. The institutional digital repository can use the checklists to identify datasets for acceptance into the repository, and to return to the provider for correction datasets that fail to meet all the criteria.
5. Management can easily merge checklist results received from across the organization to get a snapshot of the overall state of data quality and data management.

[24] A pivotal turning point is the release of all data in human readable and machine-readable format. For example, CSV files in tabular form can be understood by humans and can be read by statistical or database software (other than Excel, Word, or Acrobat) without the need to write extensive computer code to extract information and put it in a machine useable form.

6. Management can quickly scan the results to identify areas that may require closer attention within a project, identify gaps and areas in general need of improvement across the organization, or identify special cases that legitimately depart from general guidelines.

*2) Data checklist design*

Model data checklists were developed based on insights from the literature [25-32], and lessons learned from downloading and using a wide range of research, monitoring, and crowd-sourced data. The checklists developed for this paper comprise eight thematic modules with a total of 23 modules or sub-modules (Table 1).

**Table 1. Model data checklist modules**

| Module | Sub-modules |
|---|---|
| 1. Metadata | a) Metadata management. b) Provenance. c) Multilingualism. d) Accessibility. |
| 2. Data | a) Raw data. b) Data format/structure. c) Data collection. d) Data preparation. e) Geospatial data – additional considerations. f) Data management. g) Data fitness for use. |
| 3. Source | a) Data repository. b) Website. |
| 4. Visualization | a) Graphics. b) Cartography |
| 5. Software | a) Computer code. b) Project organization. c) File organization. d) Computer code changes |
| 6. Reproducibility | |
| 7. Manuscripts | |
| 8. Standards | |
| 9. Confidentiality | |

In order to keep implementation of the modules manageable, each module or submodule comprises no more than 10-20 questions each. The model checklists are not meant to be prescriptive, nor are they exhaustive. There is no "one-size-fits-all" or "off-the-shelf" solution. Organizations should adapt the modules and questions to their particular needs – modifying, removing, or adding new ones where necessary – and, implementation should be incremental.

Questions are formulated such that the '*preferred*' answer is "yes." The items are a mix of general questions (e.g., are the data FAIR? are the data accurate?) and detailed sentinel or "*canary-in-the-mine*" questions (e.g., are dates consistently formatted as YYYY-MM-DD?). When results are compiled across an organization this structure makes it easy to scan and zero in on areas that may require closer attention within a project, identify gaps and areas in general need of improvement, identify training needs, or identify special cases where a "*no*" response is in fact acceptable. Controlled responses are: "*yes*", "*no*", "*I don't know*", or "*not applicable*".

The complete set of 23 modules and submodules and the detailed questions included in each of them is available in GitHub.[25] Eight of these (1a, 1c, 2b, 3a, 4a, 4b, and 8) are presented in detail in Table 2.

*3) Implementation of checklists*

If the checklists are to achieve their intended goal, how they are used is as important as their content. The aim is to improve the organization's data quality and enable *Big Data*. This should be done in a context of a process of modernization. It requires a phased in approach and a supportive environment, including training both at the working level and for managers. The checklists and the way they are used must have strong support at the highest level of upper management.

An implementation plan should be developed to roll the checklists out in a manner that will ensure effective uptake. The model checklists, offered as a starting point, may not all apply in all situations. They should be pilot tested within the organization prior to implementation, and implementation should be iterative.

1. *Iterative implementation of the checklists.* Create a working group (WG) to adapt the checklists to the needs and realities of the organization. Each item should be assigned a level of priority, with approximately one third of the items tagged as either essential, valuable, or desirable for the first round of implementation.
2. *Pilot test the checklist module and sub-module subject headings and adjust as necessary.*
3. *Pilot test the module and sub-module checklist questions and adjust as necessary.*
4. *Round 1 should be implemented uniformly across the organization in conjunction with a data inventory in order to give management a good sense of the overall state of the data*. This should provide the organization with a good understanding of the state of the data in the organization as a whole and in each work unit. This exercise will yield valuable information for long term planning and for identification of where priorities need to be placed in the short term.
5. *In Round 2, the levels of importance should be adjusted to establish "Round 2" goals.* Round 2 goals could target low hanging fruit and what can be done in the short term without increasing resources, as well as a few of the most pressing needs to maximize short term impact and valuable outcomes. Since data management and data quality vary across the organization, Round 2 checklists should be adapted to the needs and realities of each work unit.
6. *Round 3 and successive iterations in each work unit should modify the importance levels of the various items, adding items as necessary, until the final round of implementation when all items would achieve the level of importance, "essential," and all data will be compliant.* At this point, the checklists will have evolved from "It's good enough" to "Best practices," and will have achieved uniformity across the organization. Thereafter, checklist modules should be revised on a regular basis to keep them relevant to evolving realities.

## VI. Conclusion and Future Work

Although each organization will need to develop its own path to "*Big Data readiness*", these paths will have a number of similarities: effecting culture change, treating even small data as an organizational and inter-organizational asset, and adopting

[25] Complete set of data checklist modules on GitHub: http://doi.org/cvs8

common standards so that data are useable beyond their original purpose by unimagined systems. This will require commitment both on the part of the individual data creator and on the part of the organization. As Open Government and Open Science continue to evolve world-wide, Open Science may be the key in providing a necessary unifying framework that will better enable *Big Data* and support science integrity in these systems.

Next steps in the present work will include further development to automate (or semi-automate) the data checklists in order to reduce the amount of manual labor needed to implement them.

**Table 2. Data checklists.**
Showing eight out of 23 modules listed in Table 1. Please see the full set of modules on GitHub (http://doi.org/cvs8)

| ID | Priority [a] | Data checklist questions [b] | Answer |
|---|---|---|---|
| | | ***Module 1a. Metadata management*** | |
| 1a-1 | 1. essential | Do the metadata include a description of the dataset? | yes |
| 1a-2 | 3. desirable | Do the metadata include a dataset creation date? | yes |
| 1a-3 | 1. essential | Do the metadata include a dataset update date? | yes |
| 1a-4 | 3. desirable | Do the metadata include a link to related publications? | yes |
| 1a-5 | 3. desirable | Do the metadata include a link to related data products? | yes |
| 1a-6 | 2. valuable | Are all metadata provided in a machine-readable format? | yes |
| 1a-7 | 2. valuable | Are all metadata provided in a human-readable format? | yes |
| 1a-8 | 3. desirable | Are the terms used in the metadata compliant with relevant metadata standards or ontologies? | yes |
| 1a-9 | 3. desirable | Do the metadata include a citation that is compliant with JDDCP (Joint Declaration of Data Citation Principles)? | yes |
| 1a-10 | 2. valuable | Do the metadata include a description of the methods used for data collection? | yes |
| 1a-11 | 3. desirable | Do the metadata include a description of the experimental set-up, if applicable? | yes |
| 1a-12 | 2. valuable | Is this dataset part of a data collection and, if so, is this described in the metadata? | yes |
| 1a-13 | 2. valuable | Is there a data dictionnary (describing content, format, structure of the data collection, relationships between tables, etc.)? | yes |
| 1a-14 | 2. valuable | Do the metadata include all concepts, definitions and descriptions of all of the variables? | yes |
| 1a-15 | 2. valuable | Do the metadata include descriptions of methods, procedures and quality assurance/quality control (QA/QC) practices followed during production of the data? | yes |
| 1a-16 | 1. essential | Are the metadata accurate, complete, up to date, and free of contradictions? | yes |
| 1a-17 | 1. essential | Does the documentation match the data files received? | yes |
| 1a-18 | 3. desirable | Do the metadata contain keywords selected from a controlled vocabulary, and is the controlled vocabulary properly cited? | yes |
| 1a-19 | 2. valuable | Do the metadata distinguish between types of data (primary or original, derived, dynamic, raw, aggregated data, etc)? | yes |
| 1a-20 | 2. valuable | Are the metadata registered or indexed in a searchable resource? | yes |
| 1a-21 | 2. valuable | Are the metadata assigned a globally unique and eternally persistent identifier? | yes |
| | | ***Module 1c – Multilingualism*** | |
| 1c-1 | 2. valuable | Are all elements available in English? (i.e. filename, metadata, associated resources, exposed elements in Web services). | yes |
| 1c-2 | 3. desirable | Are all elements available in an official language other than English? (i.e. filename, metadata, associated resources, exposed elements in Web services). | yes |
| 1c-3 | 2. valuable | In the case where multilingual column names are a requirement, are separate rows used for column names in the different languages? (e.g., French in row 1, Spanish in row 2, Cree in row 3, English in row 4? | yes |
| 1c-4 | 2. valuable | In the case of multilingualism, do the metadata include a translation of all the column names into the relevant languages? (e.g., French, Spanish, Cree, English) | yes |
| 1c-5 | 3. desirable | In the case of multilingualism and datasets containing text fields in English, do the metadata include translation(s) of all the possible text entries for each variable? | yes |
| | | ***Module 2b – Data format/structure*** | |
| 2b-1 | 2. valuable | For self-describing digital datasets, is the format either JSON (preferred) or XML-based using a well-known schema? | yes |
| 2b-2 | 2. valuable | In the case where the data reside in a relational database, is the database in 3rd normal form? | yes |
| 2b-3 | 1. essential | In the case where the data do not reside in a relational database, are the data files tabular? i.e. There is one rectangular table per file, systematically arranged in rows and columns with the headers (column names) in the 1st row. Every record (row) has the same column name. Every column contains the same type of data, and only one type of data. | yes |
| 2b-4 | 1. essential | Are field types (colums) used appropriate? (e.g., date field for dates, alphanumeric for text, numerical for numbers). | yes |
| 2b-5 | 2. valuable | Was a logical, documented naming convention used for variables (column names)? | yes |
| 2b-6 | 1. essential | Are the column names in the first row of the data file? | yes |
| 2b-7 | 1. essential | If these data have undergone analysis and/or visualization, do these results appear in a separate file from the data file? | yes |
| 2b-8 | 1. essential | Are the data organized so that both humans and machines can easily read it? | yes |
| 2b-9 | 1. essential | Has the data file been examined for the presence of hidden information which, if found, has been either: made visible, moved somewhere else, or deleted? | yes |
| 2b-10 | 1. essential | Do all the columns have a column name? (i.e. variable name) | yes |
| 2b-11 | 1. essential | Are the column names consistent with the documentation? | yes |
| 2b-12 | 2. valuable | Where possible, is human understable information preferred over coded information (e.g., "cat", "dog" instead of "1", "2"). | yes |
| 2b-13 | 1. essential | Does each record (row) have a unique identifier? | yes |
| 2b-14 | 1. essential | Can the tables in a data collection be linked via common fields (columns)? | yes |
| 2b-15 | 1. essential | Can the data tables be linked to the metadata via common fields (columns)? | |
| 2b-16 | 2. valuable | Are the filenames consistent, descriptive, and informative (clearly indicates content) to humans? | yes |
| 2b-17 | 3. desirable | Is filename: <70 characters; most unique content at start of filename; no acronyms; no jargon; no organization name? | yes |
| 2b-18 | 2. valuable | Was a logical, documented naming convention used for file names? | yes |
| 2b-19 | 3. desirable | Are standard/controlled vocabularies used within the data? | yes |

a. In column 2, “Priority” is based on organizational maturity (It's "good enough" for now). b. In column 4, “Answers” must be one of: ‘yes’, ‘no’, ‘I don’t know’, or ‘not applicable.

**Table 2 (cont'd)**

| ID | Priority | Data checklist questions | Answer |
|---|---|---|---|
| | | ***Module 2e – Geospatial data – Additional condsiderations*** | |
| 2e-1 | 1. essential | If the dataset contains latitude/longitude, is the datum provided? | yes |
| 2e-2 | 3. desirable | Do the metadata include a description of the geospatial coverage? | yes |
| 2e-3 | 1. essential | Do the metadata include a description of the map projection? | yes |
| 2e-4 | 1. essential | Do the latitude/longitude match the data description? (e.g., land/water, mountain/valley, northern/southern hemisphere) | yes |
| 2e-5 | 1. essential | In the case of geospatial data, is the most complete data (all layers, appendices) provided, even if proprietary? | yes |
| 2e-6 | 1. essential | In the case of geospatial data, is the format compatible with widely adopted GIS (e.g., ArcGIS)? | yes |
| 2e-7 | 1. essential | In the case of geospatial data, is the format developed or endorsed by the Open Geospatial Consortium (OGC)? (e.g., GML)? | yes |
| | | ***Module 3a – Data repository [31]*** | |
| 3a-1 | 1. essential | Does the repository perform basic curation? (e.g., checking, addition of basic metadata or documentation)? | yes |
| 3a-2 | 1. essential | Does the repository have an explicit mission to provide access to and preserve data? | yes |
| 3a-3 | 3. desirable | Does the repository maintain all applicable licenses covering data access and use and monitor compliance? | yes |
| 3a-4 | 3. desirable | Does the repository have a written continuity plan to ensure ongoing access to and preservation of its holdings? | yes |
| 3a-5 | 3. desirable | Does the repository ensure that data are created, curated, accessed, and used in compliance with disciplinary & ethical norms? | yes |
| 3a-6 | 1. essential | Does the repository have adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission? | yes |
| 3a-7 | 1. essential | Does the repository have written mechanisms to secure ongoing expert guidance and feedback, including scientific guidance? | yes |
| 3a-8 | 2. valuable | Does the repository guarantee the integrity and authenticity of the data? | yes |
| 3a-9 | 1. essential | Does the repository accept only data and metadata meeting defined criteria to ensure relevance & understandability for users? | yes |
| 3a-10 | 2. valuable | Does the repository apply documented processes and procedures in managing archival storage of the data? | yes |
| 3a-11 | 2. valuable | Does the repository assume responsibility for long-term preservation and manage this in a planned and documented way? | yes |
| 3a-12 | 1. essential | Does the repository have appropriate expertise to address technical data and metadata quality and ensure that sufficient information is available for end users to make quality-related evaluations? | yes |
| 3a-13 | 2. valuable | Does repository archiving takes place according to defined workflows from ingest to dissemination? | yes |
| 3a-14 | 2. valuable | Does the repository enable users to discover the data and refer to them in a persistent way through proper citation? | yes |
| 3a-15 | 3. desirable | Does the repository enable reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data? | yes |
| 3a-16 | 2. valuable | Does the repository function on well-supported operating systems and other core infrastructural software and is it using hardware and software technologies appropriate to the services it provides to its Designated Community? | yes |
| 3a-17 | 1. essential | Does the technical infrastructure provide for protection of the facility and its data, products, services, and users? | yes |
| 3a-18 | 3. desirable | Does the repository meet all "Core Trustworthy Data Repositories" requirements? | yes |
| | | ***Module 4a – Graphics*** | |
| 4a-1 | 1. essential | In the case of time series data, do the time series display as expected? | yes |
| 4a-2 | 3. desirable | Are the symbols effective and appropriate to content; do they display well and contribute to ease of understanding? | yes |
| 4a-3 | 3. desirable | Are standard or standardized symbols used? (e.g., thematically standardized symbols for hazards, resources, etc.) | yes |
| 4a-4 | 3. desirable | Do the symbols convey attribute information (i.e. information about the thing represented by the symbol)? | yes |
| 4a-5 | 3. desirable | Is a clearly legible legend present? | yes |
| 4a-6 | 3. desirable | Is the legend meaningful (i.e. informative and clearly indicating the content) | yes |
| 4a-7 | 3. desirable | Does the legend include measurement units, where applicable? | yes |
| 4a-8 | 2. valuable | Does the visualization load in a reasonable time period? | yes |
| 4a-9 | 2. valuable | Is the color palette effective? | yes |
| 4a-10 | 2. valuable | Is the colour palette perceivable by most forms of colour blindness? | yes |
| 4a-11 | 2. valuable | Is visualization clearly rendered (i.e. quality of the visualization is high, quickly and easily understood at appropriate scale) | yes |
| | | ***Module 4b – Cartography*** | |
| 4b-1 | 2. valuable | In the case of digital maps, is the format GeoTIFF, GeoPDF, GeoJPEG2000, or shapefile? | yes |
| 4b-2 | 1. essential | Is the map title unique and specific? | yes |
| 4b-3 | 1. essential | Does the map display what the title says? | yes |
| 4b-4 | 3. desirable | Are Web mapping services abailable? | yes |
| 4b-5 | 3. desirable | Are the contents of the Web Map Service visible at all scales? | yes |
| 4b-6 | 3. desirable | Is the Web Map Service visible at appropriate scales for the level of detail fo the datasets(s)? | yes |
| 4b-7 | 3. desirable | Are the contents of the Web Map Service consistent between scales? | yes |
| 4b-8 | 3. desirable | Are the symbols effective and appropriate to content; does it display well and contribute to ease of understanding? | yes |
| 4b-9 | 3. desirable | Are standard or standardized symbols used? (e.g., thematically standardized symbols for hazards, resources, etc.) | yes |
| 4b-10 | 3. desirable | Do the symbols convey attribute information (i.e. information about the thing represented by the symbol)? | yes |
| 4b-11 | 1. essential | Is a clearly legible legend present? | yes |
| 4b-12 | 1. essential | Is the legend meaningful (i.e. informative and clearly indicating the content) | yes |
| 4b-13 | 3. desirable | Does the legend include measurement units, where applicable? | yes |
| 4b-14 | 1. essential | Is the map scale shown? | yes |
| 4b-15 | 1. essential | Is the orientation (north/south) shown? | yes |
| 4b-16 | 1. essential | Is the map projection shown? | yes |
| 4b-17 | 1. essential | Are the map credits shown? (e.g., date of the map data, source of the map data, name of the map creator) | yes |
| | | ***Module 8 – Standards (partial list)*** | |
| 8a-1 | 1. essential | Are date and time compliant with ISO 8601? | yes |
| 8a-2 | 1. essential | Are time zones compliant with the most recent version of the IANA (Internet Assigned Numbers Authority) time zone database? | yes |
| 8a-3 | 2. valuable | In the case of geospatial data, are the metadata compliant with ISO 19115-NAP? | yes |
| 8a-4 | 2. valuable | Are measurement units compliant with the unified code for units of measure? http://unitsofmeasure.org/trac/ | yes |

a. In column 2, "Priority" is based on organizational maturity (It's "good enough" for now). b. In column 4, "Answers" must be one of: 'yes', 'no', 'I don't know', or 'not applicable.

## ACKNOWLEDGMENTS

This work was conducted as part of the development of the National Institute of Standards and Technology (NIST) *Big Data* Public Working Group (NBD-PWG) on NIST *Big Data* Interoperability Framework (NBDIF) volume 9 (Adoption and Modernization), the IEEE *Big Data* Governance and Metadata Management (BDGMM), and the Research Data Alliance - Assessment of data fitness for use (RDA-ADFU). The author gratefully acknowledges Wo Chang (Chair, NBD-PWG and BDGMM), Russell Reinsch (Chair, NBD-PWG Standards Roadmap Subgroup, Editor of NBDIF vol. 9), and Michael Diepenbroek, Jonathan Petters, Helena Cousijn and Marina Soares e Silva (Co-chairs, RDA-ADFU) for their leadership in this endeavor, and the many work group members for their discussions that encouraged and helped guide this work. Special thanks to Jessie Gaylord at Lawrence Livermore Laboratory (LLNL) and Mark Conrad at the National Archives and Records Administration (NARA) for their comments that helped improve the data checklists. The author expresses grateful appreciation to Dr. Doris Fortin and Lindsay Copland (STSD-Science and Technology Strategies Directorate) and to Cathy Cormier (DMS-Data Management Services) at Environment and Climate Change Canada (ECCC) for their support of this work and their comments that helped improve the paper, and to Dr. Monica Granados (STSD) for her gracious assistance. The author declares no conflict of interest. This work was unfunded.

## REFERENCES

[1] CASRAI-CODATA (2018). "IRiDiuM – International Research Data Management glossary." Consortia Advancing Standards in Research Administration - International Council for Science Committee on Data https://dictionary.casrai.org/Category:Research_Data_Domain

[2] Pontika N, Knoth P, Cancellieri M, Pearce S (2015). "Fostering Open Science to Research using a Taxonomy and an eLearning Portal," In: iKnow: 15th International Conference on Knowledge Technologies and Data Driven Business, Graz, Austria, October 21-22, 2015. http://oro.open.ac.uk/44719/2/kmi_foster_iknow.pdf

[3] National Academies of Science, Engineering, and Medicine (2018). "Open Science by Design." https://www.nap.edu/download/25116

[4] European Union (2016). "Amsterdam Call for Action on Open Science." Amsterdam Conference 'Open Science – From Vision to Action', hosted by the Netherlands' EU Presidency on 4 and 5 April, 2016. https://www.government.nl/binaries/government/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science/amsterdam-call-for-action-on-open-science.pdf

[5] European Commission (2016). "European Open Science Agenda (draft)." https://ec.europa.eu/research/openscience/pdf/draft_european_open_science_agenda.pdf

[6] European Commission (2018). "Open Science Policy Platform (OSPP) Recommendations." Directorate General for Research and Innovation, Directorate A - Policy Development and Coordination, Unit A.2 - Open Data Policy and Science Cloud, B-1049. https://ec.europa.eu/research/openscience/pdf/integrated_advice_opspp_recommendations.pdf

[7] ESFRI (2018). "Strategy report on research infrastructures Roadmap 2018." European Strategy Forum on Research Infrastructures. https://www.evropskyvyzkum.cz/cs/storage/5119545dd72676f39380f523fd60d54d210de6a9?uid=5119545dd72676f39380f523fd60d54d210de6a9

[8] European Commission (2018). "Recommendation on Access to and Preservation of Scientific Information." Recommendation 2018/790 of April 25, 2018 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018H0790

[9] France (2018). "National Plan for Open Science." https://libereurope.eu/blog/2018/07/05/frenchopenscienceplan/

[10] France (2018). "National Action Plan on Open Government." https://www.opengovpartnership.org/sites/default/files/France%20Action%20Plan%202018-2020%20%28English%29.pdf

[11] Hesse BW (2018) "Can psychology walk the walk of Open Science?" American Psychologist, 73(2), 126-137.

[12] Alter G, Gonzalez R (2018). "Responsible practices for data sharing." American Psychologist, 73(2), 146-156.

[13] Austin CC, Bloom T, Dallmeier-Tiessen S, Khodiyar VK, Murphy* F, Nurnberger A, Raymond L, Stockhause M , Tedds T, Vardigan M, Whyte A (2017). "Key components of data publishing: using current best practices to develop a reference model for data publishing." International Journal of Digital Libraries, 18(2) 77-92. https://doi.org/10.1007/s00799-016-0178-2.

[14] Moher D, Naudet F, CristeaI A, Miedema F, Ioannidis JPA, Goodman SN (2018). "Assessing scientists for hiring,promotion,and tenure." PLoSBiol 16(3): e2004089. https://doi.org/10.1371/journal.pbio.2004089

[15] European Commission (2018). "Study to support the review of Directive 2003/98/EC on the re-use of public sector information." Final report. Prepared by Deloitte. http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51491

[16] United Kingdom (2015). "G8 Open Data Charter and Technical Annex. Policy paper." https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex

[17] New Zealand (2018). "D7 Charter." https://www.digital.govt.nz/dmsdocument/28-d7-charter

[18] Canada (2018). "Model Science Integrity Policy (mSIP)." Office of the Chief Science Advisor. https://www.ic.gc.ca/eic/site/052.nsf/eng/00010.html

[19] National Academies of Science, Engineering, and Medicine (2017). "Fostering integrity in research." https://www.nap.edu/download/21896

[20] Klievink B, Romijn B-J, Cunningham S, de Bruijn H (2016). "Big data in the public sector: Uncertainties and readiness." Information Systems and Frontiers, 19(2), 267-283. https://link.springer.com/content/pdf/10.1007%2Fs10796-016-9686-2.pdf

[21] Rivera J, van der Meulen R (2015). "Gartner Says CIOs and CDOs Must 'Digitally Remaster' Their Organizations." Gartner. https://www.gartner.com/newsroom/id/2975018

[22] Delgado R (2016). "Why Your Data Scientist Isn't Being More Inventive. Dataconomy." http://dataconomy.com/2016/03/why-your-datascientist-isnt-being-more-inventive/

[23] EDISON (2018). "Education for Data Intensive Science to Open New science frontiers. https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome

[24] Lowndes JSS, Best BD, Scarborough C, Afflerback JC, Frazier MR, O'Hara CC, Jiang N, Halpern BS (2017). "Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1(160), 1-10. https://doi.org/10.1038/s41559-017-0160

[25] Broman KW, Woo KH (2017). "Data organization in spreadsheets." The American Statistician, Special Issue on Data Science, 72(1), 2-10. https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

[26] Kitzes J (2016). "Reproducible workflows." http://datasci.kitzes.com/lessons/python/reproducible_workflow.html

[27] Statistics Canada (2018). "Data quality toolkit." https://www.statcan.gc.ca/eng/data-quality-toolkit

[28] Wickham H (2014). "Tidy data." Journal of Statistical Software. 59(40), 1-23. https://www.jstatsoft.org/article/view/v059i10

[29] Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Tea TK (2017). "Good enough practices in scientific computing." PLOS Computational Biology, 13(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510

[30] Google (2018). "Google data search - Guidelines for developers." https://developers.google.com/search/docs/data-types/dataset#approach

[31] International Standards Organization (2018). "Geographic information – Metadata - Part 1: Fundamentals." ISO 19115-1:2014/Amd 1:2018. https://www.iso.org/standard/53798.ht

[32] DSA-WDS (2017). "Core trust seal." Data Seal of Approval and ICSU World Data System. https://www.coretrustseal.org/