Data & Policy (2022), 4: e15
doi:10.1017/dap.2022.7

COMMENTARY

Open data: The building block of 21st century (open) science

Corina Pascu[1] and Jean-Claude Burgelman[2,3,*]

[1]ENISA, Athens, Greece
[2]Open Science, Vrije Universiteit, Brussels, Belgium
[3]Frontiers Policy Labs
*Corresponding author. E-mail: jean-claude.burgelman@vub.be

Disclaimer: The information and views set out in this article are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Received: 07 December 2021; **Revised:** 21 March 2022; **Accepted:** 23 March 2022

Key words: artificial intelligence; open data; scientific knowledge

Abstract

This paper identifies the potential benefits of data sharing and open science, supported by artificial intelligence tools and services, and dives into the challenges to make data open and findable, accessible, interoperable, and reusable (FAIR).

Policy Significance Statement

Digital advancements—inter alia the Internet of (every)Thing whereby all material and immaterial acts of the universe become a data point—make it possible that anything gets researched.[1] It implies that almost all research and science to be data intensive and interconnected with researchers producing and sharing increasing volumes of data. Although digital technology enables this transition, what makes this drive to the datafication of science and research irreversible is the data version of Metcalfe's "law" of intangibles[2] that is the value of a data set increases with the number of other data sets being made available; allowing more correlations and cross linking. Hence the importance of open data: a completely digitized but "closed" science is indeed only incrementally better than its analogue versions. Just like with the FAANG's[3] of this world, will the benefit for science only exponentially increase if the data sets are openly available and reproducible. In the following, this paper identifies the potential benefits of data sharing and open science, supported by artificial intelligence tools and services, and dives into the challenges to make data open and findable, accessible, interoperable, and reusable (FAIR).

1. The Promise: Scientific Knowledge will be "Liquid"

Given this irreversibility of data driven and reproducible science and the role machines will play in that, it is foreseeable that the production of scientific knowledge will be more like a constant flow of updated data

[1] Tactile MIT initiative (https://innovation.mit.edu/pathway-post/tactile/); MIT spinout Endor has developed a predictive-analytics platform http://news.mit.edu/2017/endor-inventing-google-predictive-analytics-1220

[2] Facebook's data over the past 10 years show a good fit for Metcalfe's law https://ieeexplore.ieee.org/document/6636305

[3] FAANG' is an acronym of the five prominent US tech companies: Facebook, Amazon, Apple, Netflix, and Alphabet. Jim Cramer of CNBC's Mad Money coined the term in 2013. A variant of this acronym is "FANGAM" which includes Microsoft.

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/dap.2022.7 Published online by Cambridge University Press

driven outputs, rather than a unique publication/article of some sort. Indeed, the future of scholarly publishing will be more based on the publication of data/insights with the article as a narrative.

For open data to be valuable, reproducibility is a sine qua non (King, 2011; Piwowar et al., 2011) and—equally important as most of the societal grand challenges require several sciences to work together—essential for interdisciplinarity.

This trend correlates with the already ongoing observed epistemic shift in the rationale of science: from demonstrating the absolute truth via a unique narrative (article or publication), to the best possible understanding what at that moment is needed to move forward in the production of knowledge to address problem "X" (de Regt, 2017).

Science in the 21st century will be thus be more "liquid," enabled by open science and data practices and supported or even co-produced by artificial intelligence (AI) tools and services, and thus a continuous flow of knowledge produced and used by (mainly) machines and people. In this paradigm, an article will be the "atomic" entity and often the least important output of the knowledge stream and scholarship production. Publishing will offer in the first place a platform where all parts of the knowledge stream will be made available as such via peer review.

The new frontier in open science as well as where most of future revenue will be made, will be via value added data services (such as mining, intelligence, and networking) for people and machines. The use of AI is on the rise in society, but also on all aspects of research and science: what can be put in an algorithm will be put; the machines and deep learning add factor "X."

AI services for science[4] are already being made along the research process: data discovery and analysis and knowledge extraction out of research artefacts are accelerated with the use of AI. AI technologies also help to maximize the efficiency of the publishing process and make peer-review more objective[5] (Table 1).

Table 1. *Examples of AI services for science already being developed*

Research lifecycle	Examples of AI services/tools developed
Scientific discovery	IRIS.AI,[6] AI2's Semantic Scholar,[7] AlphaFold[8] project at DeepMind, Alan Turing Institute *New arrivals e.g. Yewno*[9]
Analysis	A.I.R.A[10]
New research methods	Alan Turing Institute's project "Living with machines"[11]
Unravelling unforeseen relationships or new models of the world	DeepMind AlphaGo[12]

(Continued)

[4] The AI revolution in scientific research (royalsociety.org)

[5] https://blog.frontiersin.org/2020/07/01/artificial-intelligence-to-help-meet-global-demand-for-high-quality-objective-peer-review-in-publishing/?utm_source=ad&utm_medium=lk&utm_campaign=ba_cco_corp_pr4

[6] https://iris.ai/

[7] https://www.semanticscholar.org/

[8] https://deepmind.com/research/case-studies/alphafold

[9] https://www.yewno.com/

[10] https://aira.io/

[11] https://www.turing.ac.uk/research/research-projects/living-machines

[12] See, for example: https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/, https://deepmind.com/blog/alphago-zero-learning-scratch

https://doi.org/10.1017/dap.2022.7 Published online by Cambridge University Press

Table 1. *Continued*

Research lifecycle	Examples of AI services/tools developed
Writing scientific manuscripts	SciNOte[13], A.I.R.A, Iris.AI
Research publishing including the optimization of peer-review workflows	Unsilo,[14] Springer,[15] Elsevier with Pending.AI[16]
Science outreach	Tl;dr

Abbreviation: AI, artificial intelligence.
Source: Authors' research based on public sources, 2021.

Ultimately, actionable knowledge and translation of its benefits to society will be handled by humans in the "machine era" for decades to come. But as computers are indispensable research assistants, we need to make what we publish understandable to them.

The availability of data that are "FAIR by design" and shared Application Programming Interfaces (APIs) will allow new ways of collaboration between scientists and machines to make the best use of research digital objects of any kind. The more findable, accessible, interoperable, and reusable (FAIR) data resources will become available, the more it will be possible to use AI to extract and analyze new valuable information. The main challenge is to master the interoperability and quality of research data.

2. FAIR Data are Essential … But It will not Happen Sui Generis

The opportunity costs of having non-FAIR data are estimated at least €10.2bn every year in Europe alone (European Commission, 2019).[17] In addition, there are also a number of consequences from not having FAIR which could not be reliably estimated, such as impact on research quality, economic turnover, or machine readability of research data. Even so, FAIR principles as such seem however to be relatively unknown to the community (Digital Science, 2019).

Moving to "FAIR-by-design" digital research outputs requires further efforts to develop, refine and adopt shared vocabularies, ontologies, metadata specifications, and standards, as well as increasing the supply and professionalization of data stewardship,[18] data repositories and data services in Europe and globally.

The European Open Science agenda contained the ambition to make FAIR data sharing the default for scientific research by 2020. To support as much as possible the proliferation of data that are FAIR, the emphasis has evolved from encouraging open access to research data for those projects funded by the EC (in Horizon 2020) to making research data open by default in Horizon Europe, following the principle "as open as possible, as closed as necessary" (Burgelman et al., 2019; Budroni et al., 2019) taking into account the need to balance openness and protection of scientific information, commercialization and Intellectual Property Rights, privacy concerns, and security.

[13] SciNote Manuscript Writer—using Artificial Intelligence

[14] https://unsilo.ai/

[15] Springer Nature advances its machine-generated tools and offers a new book format with AI-based literature overviews | Corporate Affairs Homepage | Springer Nature

[16] Elsevier and Pending.AI collaborate on AI-driven chemistry retrosynthesis tool

[17] Seven indicators were identified, defined and then quantified, such as time spent, cost of storage, licence costs, research retraction, double funding, interdisciplinarity, and potential economic growth.

The inefficiencies arising in research activities due to the absence of FAIR data were assessed to estimate the first five indicators and the time wasted due to no having FAIR was computed and the associated costs. Then the cost of extra licences that researchers have to pay to access data that would otherwise be open with the FAIR principles was assessed, as well as the additional storage costs linked to the absence of FAIR data. For the last two indicators, mostly qualitative considerations were provided.

[18] https://www.ausy.com/en/technical-news/where-are-alldata-stewards

https://doi.org/10.1017/dap.2022.7 Published online by Cambridge University Press

The move to open data means that researchers have to consider what data their research will produce and how the data will be made available. Practices with regard to data management, storage, and sharing differ widely across disciplines. A data management plan provides information on these issues, including metadata and standards, identifies suitable data repositories that will provide a unique and persistent identification of their data sets, curation and preservation and data sharing.

There are numerous legal issues in a research data environment, for example, regulation of copyright, ownership, and intellectual property for research data[19]; limitations to the sharing of research data that contains personal information by privacy requirements.[20] The "barriers to the free flow of data are caused by the legal uncertainty surrounding the emerging issues on "data ownership" or control, (re)usability and access to/transfer of data and liability arising from the use of data."[21]

The academic community is particularly concerned by these challenges (Aspesi et al., 2019), for example, ownership of data; (open) procurement of information tools and services, transparency of the algorithms, portability of the results, and sensitive data.

A broad range of changes (policy, cultural, and technical) would be needed to turn FAIR into reality in Europe (European Commission, 2018a). FAIR Digital Objects would be needed to enable discovery, citation, and reuse; data services to support FAIR; interoperability frameworks to incorporate research community practices; a distributed, federated infrastructure to unlock the potential of analysis and data integration; skills for data science and data stewardship; incentives for open science (metrics and indicators); and funding for FAIR to bring strong return on investment.

The use of trusted or certified data management environments like the European Open Science Cloud (EOSC)[22] (European Commission, 2018b) will be required for research data in some Horizon Europe work programs. Fostering the FAIR principles and data interoperability in the scientific community[23] is an important landmark. These ongoing actions are paving the way but the long tail of science still needs further support and coordination at national and European level.

Publicly funded science should be a "commons." The EOSC will enhance the possibilities for researchers to find, share and reuse publications, data, and software leading to new insights and innovations, higher research productivity and improved reproducibility in science. Europe is the first in the world to do that and EOSC the place of the science commons in Europe for data-intensive science and innovation (the "Web for FAIR data and services[24]).

3. The Bottleneck of Open Data: Data Stewardship

Currently, researchers—and the machines they use to crawl the data universe—spend significant time in the process of transforming and mapping data, for lack of standards, services, or culture. An open research labor force of data scientists is needed, with expertise in analytics, code and workflows, statistics, machine learning, data mining, and data management.

The *data steward*, with strong domain knowledge and the ability to apply this know-how within organizations to create value, has become an invaluable asset to manage data better. But data roles

[19] https://sciencebusiness.net/viewpoint/how-make-open-science-work)

[20] General Data Protection Regulation (GDPR) https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1595110568124&uri=CELEX:32016R0679; the EU Directive on Open Data and the Re-Use of Public Sector Information https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1595110473975&uri=CELEX:32019L1024; the EU Directive on Copyright in the Digital Single Market https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1595110637552&uri=CELEX:32019L0790

[21] COM (2014) 442 final on a "Towards a thriving data-driven economy"; the European "Free Flow of Data Initiative" within the Digital Single Market initiative delivered by the European Commission in 2016.

[22] The European Open Science Cloud (EOSC) will enable a trusted, virtual, federated environment in Europe to store, share and re-use digital output from research (publications, data, and software) across borders and scientific disciplines. The Partnership will bring together institutional, national and European initiatives and engage all relevant stakeholders to co-design and deploy a European Research Data Commons where data are Findable, Accessible, Interoperable, Reusable (FAIR).

[23] The 5b cluster projects are the connection of ESFRI Projects and Landmarks to the European Open Science Cloud (EOSC).

[24] https://www.eoscsecretariat.eu/sites/default/files/open_consultation_booklet_sria-eosc_20-july-2020.pdf

https://doi.org/10.1017/dap.2022.7 Published online by Cambridge University Press

continue to evolve. Whereas once it was expected that data scientists be responsible for every aspect of the data life cycle, *data engineers* will be needed to work alongside data scientists and analytics specialists that develop analytics tools to deliver that value and bring the data to life. *Open Science specialists* and data stewards would be needed to publish research in an open and FAIR way. Finally, a wider set of digital skills[25] are needed in a wide range of data-related profiles, but also skills to manage software, code and orchestrate data-intensive workflows.

Today it seems that the major hurdle for fully deploying open data is the time needed to acquire skills and expertise and handle the data. The transition to open data will not go sui generis but researchers will understand the benefits—and funders will invest in skills (at the lab level, the headcount and the costs incurred should also be considered)—for faster and better science (Mons, 2020) estimates the need for data stewards in Europe today to be around 500,000 (for every 20 people that generate data one data steward is needed).[26] If it is accepted that all science will very soon become data driven science than it follows a special effort is done to raise the level of data stewardship. And if even only 1% of the estimated 10 billion a year Europe invest in data infrastructures is allocated for that task, plenty of money will be available.

So money is not a problem, making the right priorities is.

How and what a business plan of open research data would look like and entail is a less straightforward issue. Not the least because up till now no one ever asked for a business plan for science either (and all intentions to quantify the impact of science have only given unsatisfactory responses).

Within the EOSC community this issue of what a business plan for EOSC could be has been discussed and it is suggested that several models could co-exist (European Commission, 2018b): a Direct Support Model, when an institute receives a grant from a funding entity to build/operate the resource and make it available to other grantees of the funding entity (however, the ability of certain researchers to access these resources may be restricted, i.e., nongrantees of the funding entity cannot access to the resources); a Cloud Coin Model—based on a certification program for commercial and noncommercial providers of scientifically useful services ("cloud coins"); or a Hybrid Model, i.e., combination of Direct Support Model and Cloud Coin Model.

Which of these will in the end make it, no one can know, because if so, that business plan would simply exist already.

4. The "Achilles Heel" of Open Data: Rewards and Incentives for Researchers to Make their Data Open

Scientific knowledge activity today is only incentivized by one metric: the impact factor of the author. This means that out of the whole activity of a scientist, only one product is rewarded: the article. Which means that all the work that is done before that and without which the article would never exist, is not taking into account into as a key performance indicator so to speak.

This single indicator incentive system worked for analogue science but is clearly not fit for a data driven science future where the production of open data set will be at least, if not more, important for the progress of science than the article.

It follows that the production of relevant open data sets should become a key indicator for measuring scientific performance.

Only if researchers are indeed rewarded for their data activity, will it make sense for them to invest time and resources into it.

[25] https://ec.europa.eu/digital-single-market/en/policies/digital-skills

[26] Cf. (Mons, 2020): The figure is calculated against an estimated 10 million serious data producers among 70 million science and technology professionals and 1.7 million researchers in Europe.

[27] cOAlition S funders have also committed to valuing the intrinsic merit of the work and not consider the publication channel and its impact factor when assessing research outputs during funding decisions.

[28] https://sfdora.org/

https://doi.org/10.1017/dap.2022.7 Published online by Cambridge University Press

First steps have been made, for example, by the EU, for example, Open Science Platform (European Commission, 2020) and cOAlition S,[27] as well as around the world.[28] The momentum is now building up to launch a coalition of research actors[29] to make a step change toward a science system that delivers higher quality and more impactful results. So, what could possibly go wrong?

The value of ORD but also the risks for ORD for 21st century science are unprecedented. The speed with which the scientific community was able to react to the CORONA challenge, based on open transfer of knowledge will most likely become the best user case for open data. However, minimal safeguards need to be put in place for open research data to stay open. The policy challenge is how to avoid misuse by public and private actors (Zuboff, 2019) and dependability of all kind of providers (like in OA). ORD should be as open as possible as closed as needed and unintended use should not be allowed or only with consent.

AI that is "accountable to society" (Smith, 2019) is one of the top technology policy issues that will shape the science of the 21st century. Machines must remain accountable to people and people who create AI technology must remain accountable to society as a whole Face recognition technology and its use by governments and law enforcement is controversial.[30] Private technology companies alone cannot be trusted to safely manage the data they collect.[31]

"We cannot just complain about how tech is transforming our world; we need to invent the transformation."[32] Solutions should be global in nature[33]: smart deals could be signed with publishers and platform providers[34]; compliance could be embedded in the design of the ORD services and standards (GDPR compliant) or the architecture of the sharing and access system; public data banks could be created hosting key data sets and trusted third parties could be envisaged to outsource independent data handling (like the credit card system). Eventually, people could gain power with the platforms through "mediators of individual data."[35]

But fundamentally the world should accept that open science data is a commons: to the benefit of all.

Funding Statement. This work received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing Interests. The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent licensing arrangements), or nonfinancial interest (such as personal or professional relationships, affiliations, knowledge, or beliefs) in the subject matter or materials discussed in this manuscript.

Author Contributions. All authors contributed equally to the article.

Data Availability Statement. Data availability is not applicable to this article as no new data were created or analyzed in this study.

References

Aspesi C, **Allen NS**, **Crow R**, **Daugherty S**, **Joseph H**, **McArthur JTW and Shockey N** (2019) SPARC landscape analysis: the changing academic publishing industry—implications for academic institutions. https://doi.org/10.31229/osf.io/58yhb

Budroni P, **Burgelman J-C**, **Schouppe M** (2019) Architectures of knowledge: the European Open Science Cloud. *ABI Technik 2019 39*(2), 130–141

Burgelman J-C, **Pascu C**, **Szkuta K**, **Von Schomberg R**, **Karalopoulos A**, **Repanas K and Schouppe M** (2019) Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *In Front. Big Data.* https://doi.org/10.3389/fdata.2019.00043

[29] https://osec2022.eu/ (European Commission, 2021).

[30] https://www.foxnews.com/tech/google-ceo-ban-facial-recognition-sundar

[31] https://www.theguardian.com/cities/2019/jun/06/toronto-smart-city-google-project-privacy-concerns

[32] https://www.wired.com/story/wired25-jaron-lanier-glen-weyl-radical-equality/

[33] Microsoft's Brad Smith called for a global AI alliance of democratic societies https://www.microsoft.com/en-us/research/podcast/an-interview-with-microsoft-president-brad-smith/

[34] https://sciencebusiness.net/news/elsevier-signs-open-access-agreement-netherlands

[35] https://www.wired.com/story/wired25-jaron-lanier-glen-weyl-radical-equality/

https://doi.org/10.1017/dap.2022.7 Published online by Cambridge University Press

de Regt HW (2017) *Understanding Scientific Understanding*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190652913.003.0001

Digital Science, **Fane B**, **Ayris P**, **Hahnel M**, **Hrynaszkiewicz I**, **Baynes G and Farell F** (2019) *The State of Open Data Report 2019*. Springer Nature. https://doi.org/10.6084/m9.figshare.9980783

European Commission (2018a) Turning FAIR into reality. Final report and action plan from the European Commission Expert Group on FAIR data. Available at https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283

European Commission (2018b) Prompting an EOSC in practice, Final report and recommendations on the European Open Science Cloud (EOSC), 2018. https://doi.org/10.2777/112658

European Commission (2019) Cost-benefit analysis for FAIR research data—cost of not having FAIR research data. Available at https://publications.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1

European Commission (2020) *Progress on Open Science: Towards a Shared Research Knowledge System: Final Report of the Open Science Policy Platform*, Lawrence R (ed.) Publications Office. https://data.europa.eu/doi/10.2777/00139

European Commission (2021) *Towards a Reform of the Research Assessment System*. Publications Office of the EU (europa.eu).

King G (2011) Ensuring the data-rich future of the social sciences. *Science 331*(6018), 719–721. https://doi.org/10.1126/science.1197872

Mons B (2020). Available at https://www.nature.com/articles/d41586-020-00505-7

Piwowar HA, **Vision TJ and Whitlock MC** (2011) Data archiving is a good investment. *Nature 473*, 285. https://doi.org/10.1038/473285a

Smith B (2019) *Tools and Weapons: The Promise and the Peril of the Digital Age*. Penguin Press.

Zuboff S (2019) *The Age of Surveillance capitalism: The Fight for a Human Future at the New Frontier of Power*. Public Affairs.

Cite this article: Pascu C and Burgelman J.-C (2022). Open data: The building block of 21st century (open) science. *Data & Policy*, 4: e15. doi:10.1017/dap.2022.7

https://doi.org/10.1017/dap.2022.7 Published online by Cambridge University Press