2018

Reflections on the future of research curation and research reproducibility

To appear: April, 2018, *Proceedings of the IEEE*, V.106:4.

Reflections on the Future of Research Curation and Research Reproducibility[1]

John Baillieul,[2] Gerry Grenier,[3] Gianluca Setti[4]

The Web Has Changed Everything. In the years since the launch of the World Wide Web in 1993, there have been profoundly transformative changes to the entire concept of publishing – exceeding all the previous combined technical advances of the centuries following the introduction of movable type in medieval Asia around the year 1000[5] and the subsequent large-scale commercialization of printing several centuries later by J. Gutenberg (circa 1440). Periodicals in print – from daily newspapers to scholarly journals – are now quickly disappearing, never to return, and while no publishing sector has been unaffected, many scholarly journals are almost unrecognizable in comparison with their counterparts of two decades ago. To say that digital delivery of the written word is fundamentally different is a huge understatement. Online publishing permits inclusion of multi-media and interactive content that add new dimensions to what had been available in print-only renderings. As of this writing, the IEEE portfolio of journal titles comprises 59 online only[6] (31%) and 132 that are published in both print and online. The migration from print to online is more stark than these numbers indicate because of the 132 periodicals that are both print and online, the print runs are now quite small and continue to decline. In short, most readers prefer to have their subscriptions fulfilled by digital renderings only.

While modest economies accrue in moving from print to electronic dissemination of information, major transformations in the fundamental nature of research curation have made the disappearance of print a mere side effect of the larger changes in scholarship and scholarly communication. In the age of print, it was somewhere between infeasible and impossible to disseminate the large data sets that were the basis of much of the research that was published. Today's cloud services and information delivery apps have enabled researchers to release additional essential parts of their research findings, ranging from data sets, to code, to videos documenting experimental processes. As publishing has moved into the digital domain, the analytical tools of research have rapidly evolved as well. It is increasingly easy – with the

---

[1] The authors gratefully acknowledge the U.S. National Science Foundation through NSF Grant Number ECCS-1641014 that was awarded to support the First IEEE Workshop on the Future of Research Curation and Research Reproducibility.

[2] College of Engineering, Boston University, Boston, MA 02215, johnb@bu.edu

[3] IEEE, Inc., 445 Hoes Ln., Piscataway NJ 08854, g.grenier@ieee.org

[4] University of Ferrara, Via Saragat 1, 44100, Ferrara, Italy, gianluca.setti@unife.it

[5] The first printed book is generally thought to be the *Diamond Sutra*, 968AD, which was done by wood black printing. Printing by means of movable type has been attributed to the Chinese inventor Bi Sheng (990-1051). See Joseph Needham, *Science and Civilisation in China: Paper and Printing*. Cambridge: Cambridge University Press. Page 201-202.

[6] For the IEEE, *online only* periodicals may typically have a minuscule print run in order to meet contractual requirements of certain library customers.

proliferation of technologies like *Dockers* and *Vagrant* open-source virtual code development environments – to make software part of the research record.

The basic definitions of what constitutes a research product are thus in flux, and the entire process of conducting scientific inquiry is in the process of being fundamentally reshaped. The research missions of research institutions, the funders that support them, the publishers that have been custodians of the research record, and individual scholars and scientists themselves are changing accordingly. In what follows, we shall examine the changes, and share some concerns about safeguarding the integrity of rigorous, arms-length peer review. Publications are now only part of what science is expected to produce. Research notebooks, experimental protocols, data, and software are important artifacts that are increasingly seen as being equal in value to the peer reviewed articles that describe them. Both software and data are now prominently mentioned as essential research products by funding agencies, including the U.S. National Science Foundation (NSF)[1] and the European Commission (EC).[2] Since 2011, NSF has required that every research proposal includes a plan for data management and sharing of the products of proposed research. At the time of this writing, there are few agency-wide specific requirements for such plans, but the Engineering Directorate has additional detailed requirements that involve specific definitions of the digital data and metadata items that must be archived and made available to the research community.[3] Similarly, within Horizon 2020, the European Commission has launched the Open Research Data Pilot (ORD Pilot) to promote open access to and reuse of research data generated by Horizon 2020 projects. Design of this pilot followed the so-called FAIR data principles: all research data should be Findable, Accessible, Interoperable and Reusable (FAIR). At the beginning (2014-2016) the ORD pilot included only selected areas in the Horizon 2020 work program, but at this writing in 2018 it has been significantly extended to cover all its thematic areas.

Online Curation Is Really Different. For nearly two decades, research communities have worked to define appropriate policies for archiving and disseminating the products of publicly funded research. As part of these efforts, agencies in the U.S. and elsewhere have held public hearings[4,5] and sponsored numerous workshops[6,7] dealing with public access, data management requirements, and the formats needed to make research products maximally useful to the broad scientific community. There is a clear underlying premise that research products are useful only to the extent that both the research community and the general public have easy access to them and that they are easy to discover. Digital curation has led to new expectations and new metrics of quality, and funding agencies worldwide now want to evaluate research according to the extent of its being repeatable, replicable, reproducible, reusable, and validated. (See Table 1 for definitions of these terms.)

A little over a year ago, a number of the volunteers and staff who oversee IEEE publications began discussing the need to broaden the discussion to include greater participation by people from well known STEM publishing houses. A small group that included the authors of this article drafted a workshop proposal that was submitted to the National Science Foundation in April of 2016. The grant was awarded, and *The First IEEE Workshop on The Future of Research Curation and Research Reproducibility* was held in Washington, DC, over the weekend of November 5,6, 2016. The Workshop Report is now available (http://www.ieee.org/researchreproducibility), and in what follows we'll briefly describe some

highlights and discuss some rather large challenges (as well as opportunities) that now face the STEM publishing industry.

Table 1 – Research Reproducibility Terminology

Repeatability	Same team, same experimental setup
Replicability	Different team, same experimental setup
Reproducibility	Different team, different experimental setup
Reusability	Research that exceeds minimum expectations; this means there is good documentation and procedures can be repeated and repurposing is facilitated
Validation	A certificate is issued that the research has been replicated or reproduced

The Association for Computing Machinery (ACM) has led the effort to define terms associated with reproducible research. Badges that appear on published articles have been created to indicate levels of reproducibility as defined here.

The Open Science Movement Is Driving Change in Research Curation. A very broad summary of the Workshop report is that *open science* — and all that it implies for shared code and data — may involve new burdens on researchers unless protocols for sharing are carefully crafted and respected. Beyond protocols for sharing, a dominant theme of the Workshop was the challenge of making open science financially sustainable. There appears to be fairly broad agreement among all stakeholders that funding for open science and all the new dimensions in research curation will be an ongoing challenge. Researchers who have been supported by funding agencies both in the U.S. and abroad are well aware that funds for nontraditional line items in grant budgets are limited. In the U.S. especially, most segments of the research enterprise face the challenges of a zero-sum game. If grant funds are budgeted to cover the cost of open access article processing charges (APCs), they will typically be at the expense of other line items in the research budget. Libraries must similarly allocate funds among content, services, and infrastructure—enhancing any one area by reducing support to the others. Financial support of infrastructure for research curation will be a challenge that continues into the foreseeable future.

Curation Reimagined – Versioning and Distributed Archives. When the record of scholarly research was maintained in printed journals, the *version of record* of any article was whatever the publisher printed. Electronic dissemination quickly led publishers to rethink the entire concept. Recognizing the enormous potential benefits of linked digital archives, the Publishers International Linking Association (better known as Crossref) was created. By means of a Certificate of Incorporation that was signed by twelve leading publishers (including the IEEE) on January 18, 2000[7] Crossref was established to enable persistent cross-publisher citation linking in online academic journals through the assignment and maintenance of persistent Digital Object Identifers (DOIs). The mission of *Crossref* has grown to include forward reference linking (*Cited-By*), Similarity Checking (powered by *iThenticate*) and, most recently, the curation of a

[7] The number of voting members now participating in *Crossref* approaches 8000 including learned societies, hosting platforms, submission systems, libraries, publishing services companies, and metrics and analytics companies.

Funder Registry – a taxonomy of grant-giving organizations. An equally revolutionary service made possible by the move from print to online publishing is *Crossmark*.According to the *Crossref* website, https://www.crossref.org/services/crossmark/, *Crossmark* allows scholars to easily identify instances of documents that are being dynamically maintained and updated by their publishers. The appearance of a *Crossmark* logo on a PDF, HTML, or ePub document indicates that the publisher is taking care of or stewarding it through any updates, corrections, retractions, or other changes.

The concept of the version of record has thus changed from something completely static to something that can continually change and that needs to be authoritatively monitored. While the management of article versioning seems like a very natural undertaking for the *Crossref organization*, the concept becomes considerably more complex if it is extended to research artifacts beyond published articles – specifically to data, metadata, and software. The infrastructure that will support archives of the broad categories of research artifacts will in all likelihood be highly distributed and include a range host types including:

- Researchers' personal web sites,
- Enhanced publisher portals (such as http://www.portico.org/digital-preservation/, https://figshare.com/),
- University repositories (such as MIT's DSpace (https://dspace.mit.edu/), Boston University's OpenBU (https://open.bu.edu/), and the University of California's eScholarship (https://escholarship.org/), and
- Government supported repositories such as NSF's PAR (https://par.nsf.gov/), the U.S. Department of Energy PAGES (https://www.osti.gov/pages/), and the European Union's OpenAIRE (https://www.openaire.eu).

There is a clear challenge in designing the metadata that will be necessary to make scattered artifacts searchable and accessible – especially as they will evolve and change location and perhaps even disappear over time.

Workshop participants received updates on Portico (https://www.portico.org/) and the RMap Project (http://rmap-project.info/rmap/?page_id=98)) that is aimed at creating the technology needed for distributed research curation. Funded by the Sloan Foundation and carried out by a collaborative effort of the Data Conservancy, Portico, and IEEE, the goal is a well-designed, large-scale taxonomy that will tie together related content across multiple platforms and data types. The vision for RMap is that scholarly communication becomes not an object, not a journal article, but a network of heterodox objects, each consisting of multiple constituents, all of which can exist in multiple versions, and all complexly joined to the others. There are interlinked models, vocabularies, ontologies, and languages that support preservation and that ultimately ensure the reproducibility of scientific results.

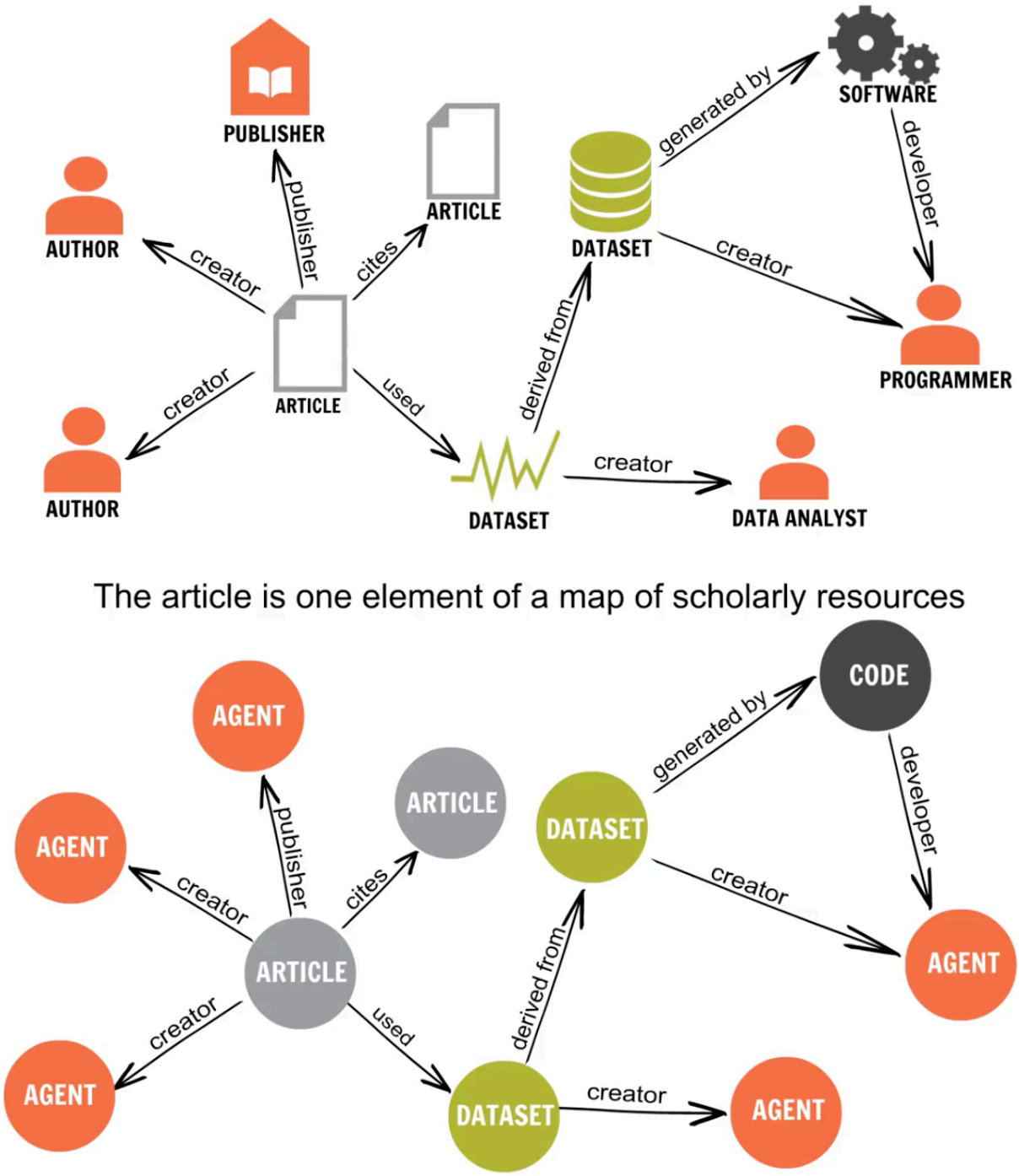

Figure: The RMap project describes the foundations needed for distributed curation of research objects. See https://demo.rmap-hub.org/app/

Finally we mention the Research Data Alliance (RDA) which is a non-governmental organization launched in 2013 with support from the European Commission, the U.S. National

Science Foundation, the U.S. National Institute of Standards and Technology, and the Australian Department of Innovation. The mission of the RDA is the creation of a "social and technical infrastructure to enable open sharing of data." (https://www.rd-alliance.org/) The RDA meets every six months, and RDA working groups focus on metadata standards, best practices for repositories, and repository sustainability. The work focuses on discipline-specific concerns, such as geology and polar climate data, and best practices for managing that content.

The future: Avoiding the Dark Side of Openness and Thinking Outside the Box. There is an almost giddy enthusiasm for "open science" on the part of some evangelists.[8] With lofty goals of greater scientific rigor, enhanced reproducibility of results, wider access to results and methods for both the research community and public at large – what's not to like? Certainly the well publicized efforts of the ACM[9] to include data, software, and other non-textual data as part of the research record are laudable, as are the efforts of organizations like the Center for Open Science.[10] Despite the energy, good will, and integrity of these representative groups, care and vigilance are called for in this era of *fake news* and *alternative facts*. Scientists and engineers have historically been trained to respect and support the validation process of peer review. They have relied on leading scholarly journals to carefully vet those invited to serve on editorial boards to ensure that the highest standards of technical competence are met. With the abundance of alternative information channels proliferating, and the tendency to rely on now ubiquitous new forms of media that are effectively un-moderated wiki's, there is considerable risk that some of what is disseminated will escape rigorous scrutiny by disinterested and unbiased peers. An interesting article[11] by Elizabeth Kolbert in *The New Yorker* discusses precisely this kind of dark side in the context of mass media. Kolbert notes that when newspapers and magazines, and by extension scholarly journals, are behind paywalls, the uber-media like Google, Facebook, and Twitter tend to bury them.

For scholarly journals, the movement toward enhanced free public access to scientific research[12] has come at the cost of placing new financial stresses on publishing houses that have been the historical guardians of peer validated scholarship. Open Access (OA) advocates and entrepreneurs have created many open access publishing alternatives as indicated in the growth statistics provided by the Directory of Open Access Journals, DOAJ, https://doaj.org/. While the DOAJ has taken steps to ensure the quality of journals that they list,[13] there remain reasons to be cautious about the science that comes for free on the Web. Opportunities are greater than at any earlier time in history to create and publish information that serves a business or political agenda.

An equally concerning modern phenomenon is the way social media is now being used in scientific discourse. While much of the discourse serves the purpose of enhancing and promoting research, there are also examples of blogs and Facebook posts making *ad hominem* attacks that have no place in scholarly discourse. At this writing, the New York Times has just run a front page Times Magazine article entitled "When the revolution came for Amy Cuddy."[14] The article describes a research paper, published by Amy Cuddy in the high profile journal *Psychological Science*, dealing with the effects of "power poses" on the people who strike the poses. The study reported that people who struck poses connoting power (legs astride or feet up on a desk) felt empowered. Moreover, Cuddy's paper reported that after striking a power pose, subjects had elevated levels of testosterone and decreased levels of cortisol. Unfortunately, a subsequent study failed to replicate the findings, and when Cuddy mounted a modest defense of

her study, she became the target of attacks in blog posts by other social scientists. It appeared that the critical civil discourse of traditional peer review had been replaced by cyber bullying in the blogosphere.

This of course does not mean that the goals of open access and more importantly open science are unworthy. On the contrary, if they are properly conceived and executed, the techniques being advocated for open science can help navigate the landscape of high quality reproducible research. They can also help suppress fake science.

Where Do We Go From Here. All things considered the enterprise of research curation is in an incredible state of flux. Driven both by evolving technology and public policy, there is widespread interest in questions of how best to achieve improved public access, including data storage and preservation, discoverability, and reuse with a particular focus on data underlying the conclusions of peer-reviewed scientific publications[15]. Over the past decade, there has been a proliferation of institutional repositories at leading research universities, enormous growth in the open access ePrint server arXiv.org, and the launch of funding agency access requirements together with public access repositories (NSF PAR, the EU's OpenAIRE, and many more). The published records of research are no longer the sole property of publishers, and the responsibility for preservation of the scholarly record is now shared among funders and their repositories, research libraries, and STEM publishers as well. Whether these organizations are unnecessarily duplicating each other's efforts, there is a more important question of whether re-aligning the roles of funders, publishers, and libraries will ultimately enhance or diminish the openness and quality of the research enterprise. Publishers have been – and perhaps remain – uniquely able to provide neutral refereeing of research. They have typically not been affiliated with research institutions, and since they do not fund research, they have no stake in the outcome of any particular experiment or project. If publishers are casualties of the push toward open science, who or what agency will assume the role of honest and neutral arbiter?

Endnotes

[1] https://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#IIC2j

[2]http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] https://nsf.gov/eng/general/ENG_DMP_Policy.pdf

[4] http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_083127.pdf

[5] http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_083128.pdf

[6] http://sites.nationalacademies.org/DEPS/BMSA/DEPS_153236

[7] "Robust Research in the Social, Behavioral, and Economic Sciences," https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

[8] "Knowledge is power. With key information openly available, power can be held to account, inequality challenged, and inefficiencies exposed. With the latest research openly available, everyone has the potential to understand our world, and the knowledge they need to tackle major challenges such as poverty and climate change. – From the vision statement of Open Knowledge International, https://okfn.org/about/vision-and-values/

[9] http://www.acm.org/data-software-reproducibility

[10] https://cos.io/blog/content-open-science/

[11] Elizabeth Kolbert, "The Content of No Content," *The New Yorker,* August 28, 2017.

[12] Holdren JP, 2013. "Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research." Executive Office of the President, Office of Science and Technology Policy. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[13] Monya Baker, "Open-access index delists thousands of journals," *Nature News*, May 9, 2016, doi:10.1038/nature.2016.19871

[14] Susan Dominus, "When the Revolution Came for Amy Cuddy," New York Times, October 22, 2017. https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-for-amy-cuddy.html

[15] NSF's Public Access Plan: Today's Data, Tomorrow's Discoveries, National Science Foundation, March 18, 2015, NSF 15-52. http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf.