Thoughts on Starting the CODATA Data Science Journal

COLLECTION: 20 YEARS OF DATA SCIENCE

ESSAY

JOHN RUMBLE

ABSTRACT

This essay discusses some of the considerations that led to the founding of the [CODATA] Data Science Journal. Three factors were most relevant to the founding. First, there was a need to have a more formal publication mechanism for the papers given at the biennial CODATA International Conferences. Second, there was a pressing need for data science advancements made in one area of scientific data work to be shared with other scientific disciplines. Lastly the increasing number of scientists interested in data, throughout science, and throughout the world, required a more convenient publication outlet. Thus arose the Data Science Journal.

CORRESPONDING AUTHOR:

John Rumble

R&R Data Services, US

john.rumble@randrdata.com

KEYWORDS:

CODATA; CODATA Data Science Journal; Scientific Data; Data Science; Data Standards

TO CITE THIS ARTICLE:

Rumble, J. 2023. Thoughts on Starting the CODATA Data Science Journal. *Data Science Journal*, 22: 13, pp. 1–4. DOI: https://doi.org/10.5334/dsj-2023-013

When the editors first approached me about writing about the founding of the *Data Science Journal* (DSJ), my first reaction was 'I am not a historian' and my memories and recollections about the beginning were going to be inaccurate and warped by the passage of time. As I thought more about the subject, I realized my initial thoughts were correct, and I would not be able to accurately depict the DSJ's start. Over the summer I was fortunate to visit F. Jack Smith, DSJ's first editor and the person who really provided its initial philosophy and structure. During my two-week visit (the length was dictated by my wife and I having contracted COVID), Jack and I had the opportunity to talk about the start-up, and his essay captures history much better than I ever could.

Looking back to when I became President of CODATA in 1998, these were my (in-brief) thoughts about how we were communicating our work:

- The future of the Proceedings of the Biennial CODATA International Conferences, which contained many valuable insights on scientific data (or shall we say, scientific data science), was unclear.
- After a decade or more of advances in computerized scientific data, it had become clear that there was immense benefit to sharing the lessons and progress in one discipline with practitioners in other disciplines.
- Scientific data activities were expanding into every discipline and every country, and this new generation of scientific data scientists really had no convenient publication outlet.

I will discuss each of these points in more detail below, but first I would like to mention a recent insight I had after listening this week to a LinkedIn seminar on monetizing data products. As I listened to the discussion about tools that could identify correlation between economic/business data sets and business opportunities, I realized that the defining difference between that activity and scientific data science is that scientists are primarily looking beyond *correlation* and trying to identify *causation*. Specifically, we scientists want our data activities – from collection to management to analysis to dissemination to knowledge discovery – to support the discovery of causation. In general layman terms, does reported property data have enough supporting information (metadata) to enable a new measurement?

This key difference is the driving force, I believe, behind the success of CODATA, many society-based data initiatives, government funded data activities, and ultimately the CODATA *Data Science Journal*. The scientific data community requires more from data science than other communities, and this unique dedication to causation, or perhaps we should call it scientifically valid data, helps explains the explosive growth of data science within the scientific community over the last couple of decades. As I like to point out, when CODATA was started, those scientists interested in data were called zealots and were usually outside mainstream science. Today every scientist is a data scientist.

PUBLISHING PAPERS FROM THE CODATA INTERNATIONAL CONFERENCES

The CODATA biennial conferences began in 1968 and have been instrumental in fostering sharing of ideas about scientific data activities across disciplines and geographical regions. The conferences covered a wide range of subjects in what we now call data science as well as many different scientific disciplines. Topics such as database management, metadata handing, user interfaces, dissemination strategies, data and computer networks, and data displays routinely introduced participants to new concepts and emerging computerized data capabilities.

For many years, conference proceedings were published, as edited by the CODATA Executive Director Phyllis Glaeser and other CODATA leaders. By 1998, it was clear that the Proceedings faced two challenges: the amount of work necessary for editing and publishing and the imminent retirement of Phyllis Glaeser. A CODATA journal would allow for a more controlled workflow process as well as providing greater access to conference papers, as a journal would be indexed as part of the scientific publication indexing process.

While this was an important motivation for starting the DSJ, the reality was that soliciting, receiving, and editing papers from conference presenters was a greater challenge than anticipated. Linkage between acceptance by the conference and requesting a DSJ paper was difficult given that conference sessions were organized by different individuals and groups. Also, presenters perceived that a DSJ submission was more labor-intensive than the previous Proceedings requirements. Therefore, this goal was only partially met.

SHARING THE LESSONS AND PROGRESS ON ADVANCES IN COMPUTERIZED SCIENTIFIC DATA

The period from 1970 to the late 1990s saw an explosion in computerized scientific data activity. Three distinct waves of computer power fed this progress: first, diversity in mainframe computers, with both larger and smaller computers becoming accessible; second, the PC innovation of the 1980s; and third, the development of the internet and the World Wide Web. Each of these waves was accompanied by similar innovation in data-related software, which is ubiquitous today. It would be great if I could write today that these advances were optimized for use by the scientific data communities, but this has not happened. Straightforward scientific information such as superscripts, subscripts, significant figures, trailing zeroes, and Greek letters still present challenges to popular database management systems, spreadsheets, and other common data science software.

The CODATA community has long been a significant factor in spreading information on how to take advantage of advanced computing, including challenges as mentioned above. Through its task groups, workshops, conferences, publications, and handbooks, CODATA has been a major source of new developments for scientific data. In starting the DSJ, CODATA not only extended that tradition, but also provided more timely access to advances. As with most scientific journals, the DSJ also provided authors with peer review for their work and scholarly credit for their new ideas. Today DSJ remains a premier journal for scientific data science.

A VOICE FOR THE BROAD CODATA COMMUNITY

As the world recovered from two World Wars and as Colonialism was overcome, new and emerging nations around the world began developing vibrant scientific activities, and scientific data has become an important component of that work. CODATA was an early responder to these developments and gained both new formal country members as well as engagement from individuals from everywhere. Many of these scientific data workers were new to formal scientific publishing, and from the beginning, the DSJ has recognized its important role in introducing their work to the larger scientific community.

This effort required considerable resources in terms of editing and manuscript solicitation with positive results. The breadth of DSJ's authorship reflects the breadth of scientific data work being done internationally.

SOME CONCLUDING REMARKS

One of the greatest challenges facing DSJ at the beginning was the tension between the need for publication in discipline-specific journals and the interdisciplinary nature of scientific data science. I continue to advocate that most scientific data science articles have applicability beyond their discipline and have personally benefited from ideas far from my own discipline. That tension has probably grown.

I would be remiss if I did not mention several people who were critical to starting DSJ and helping it through its first few years. F. Jack Smith, then at Queens' University Belfast as the first editor, made it happen. Kathleen Cass, as CODATA Executive Director during DSJ's first few years, ensured smooth transitions in publishing partners. Diane Smith Rumble was the editorial assistant for many years and provided editing and administrative services for several editors. Finally, colleagues in Japan added a needed level of professionalism as DSJ matured.

FINAL COMMENTS

In looking over the contents over the first ten years of DSJ, I am struck again by the diversity and prescience of its articles, especially by how many of the articles, especially outside my physical sciences disciplines, were influential in providing ideas for my own data work. The most vivid illustration of this cross-cultural feedback related to the development of data reporting standards, and how difficult it is to take into account the subtle differences that arise because of discipline-specificity and language. DSJ plays an important role in this cross-disciplinary discussion because data science articles published in a discipline-specific journal are not noticed by people outside that particular field.

As for the future of the DSJ, I continue to encourage the journal to keep an emphasis on the major topic of interest for CODATA – that is, to continue to emphasize the practice of data science as related to scientific and technical data. Issues that to me continue to be important include:

- data quality, and how it is assessed (evaluated),
- development of data reporting standards that are usable by scientists not involved in their development,
- long-term preservation of scientific data – including economic and sociological discussions of what should be preserved, the implications of constantly changing storage technologies, who should pay and why (for examples, how could/should future data users provide support) – and it should be noted that long-term means decades or beyond, and
- role of AI in generating routine or difficult-to-measure scientific data.

In addition, DSJ should consider how it can help CODATA create a post-archival journal, or journal subsection, that encourages less formal communication among scientific data enthusiasts that remains a powerful *raison d'être* for CODATA itself. This might include virtual workshops and conferences, where contributions are more conversational than formal papers, or discussions (moderated or not) of topics. Perhaps I am a bit old-fashioned, but I believe there still is a need for serious conversations that lie between 140 and 280 characters and a 5000-word formally reviewed paper. The topics that I highlighted immediately above have been of importance for many decades and are even more important than ever in this era of data. The CODATA DSJ is well-positioned to play a major role in coming up with new ideas that reflect the diversity of science and the growing importance of data.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

John Rumble orcid.org/0000-0001-6705-5768
R&R Data Services, US

TO CITE THIS ARTICLE:
Rumble, J. 2023. Thoughts on Starting the CODATA Data Science Journal. *Data Science Journal*, 22: 13, pp. 1–4. DOI: https://doi.org/10.5334/dsj-2023-013

Submitted: 06 December 2022
Accepted: 24 April 2023
Published: 08 June 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.