Challenging the print paradigm: Web-powered scholarship is set to advance the creation and distribution of research.

Our containers for scholarly works – papers, monographs, PDFs – are anachronistic. **Marcus A. Banks** *argues the Web is flexible enough to facilitate far more opportunities for scholarship in a way that print could never do. A print piece is necessarily reductive, while Web-oriented scholarship can be as capacious as required. He highlights three innovations in particular that are set to transform the scholarly environment: data papers, scholarly HTML and sophisticated data mining tools like Content Mine.*

"Big data" is a pernicious buzz word of our age, responsible for scores of hastily assembled PowerPoints in board rooms and classrooms throughout the world. That haste reveals shallowness of thought, an over-reliance on supposedly empirical and objective "data" to elucidate the mysteries of human existence. That such an expectation is impossible is not lost on humanists, which I still fancy myself to be despite a career spent in health sciences libraries.

That career has placed me within computationally intensive academic health care enterprises. My employers have sought to mine patient care notes and health profiles to improve the delivery of care across a system as well as individual patient outcomes. These efforts are big data without the marketing spin, and as such they have earned my support. They represent an intelligent application of data mining tools, in order to address real-world challenges. Although "big data" as an intellectual construct is fatuous, data mining as a practical skill is valuable.

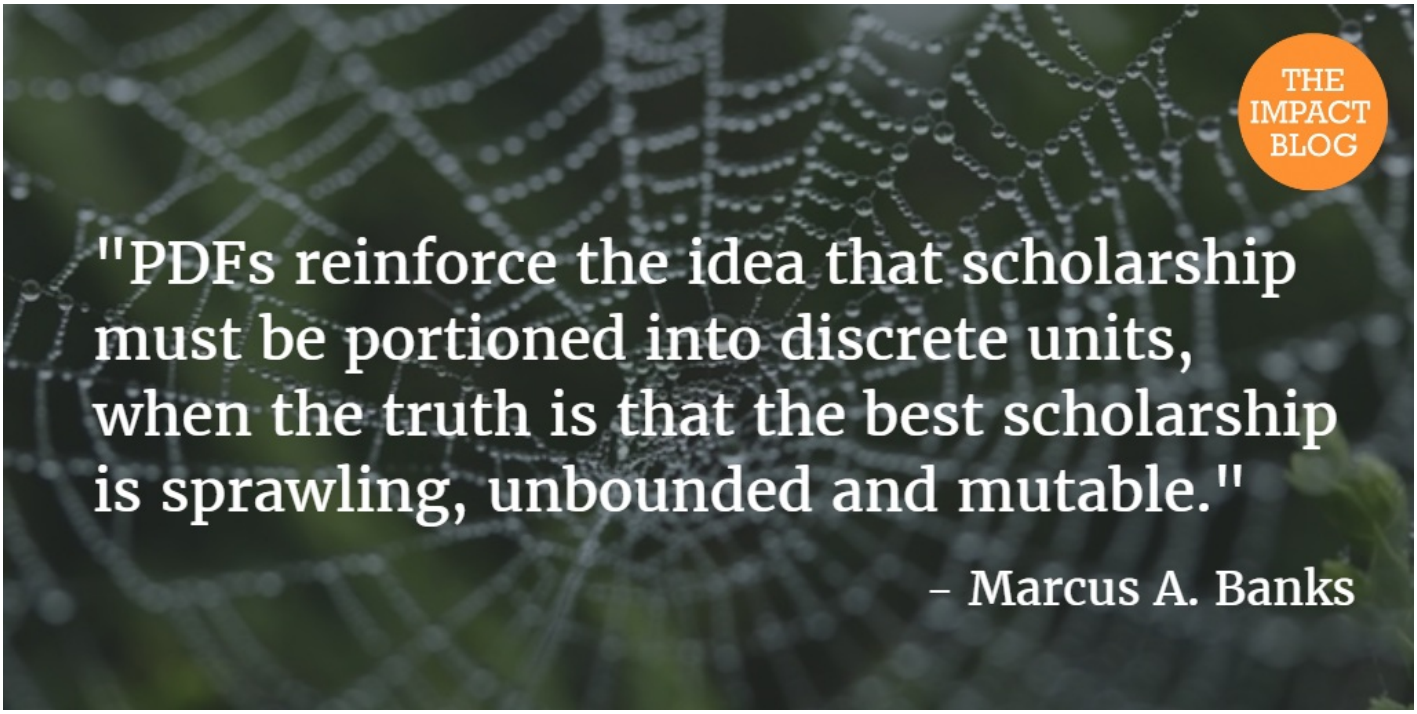
Image credit: HTML5 by ErickDimas 001 CC BY-SA

The scholarly literature is rife with potential for data mining, across all fields. Just for starters: data mining could be used to better understand how perceptions of key historical events have shifted with time, as well as how

philosophical conceptions of consciousness have evolved. These are just two of countless potential applications. Although data mining may feel like the province of the hard sciences, there is no epistemological reason why this is the case. Any field of scholarship could benefit from tools that detect patterns within the scholarly literature that are not apparent to the naked eye. A human being cannot easily digest 20,000 papers or monographs, but data mining software can do so with no trouble.

Despite my confidence in data mining I worry that our containers for scholarly works — "papers," "monographs" — are anachronistic. When scholarship could only be expressed in print, on paper, these vessels made perfect sense. Today we have PDFs, which are surely a more efficient distribution mechanism than mailing print volumes to be placed onto library shelves. Nonetheless PDFs reinforce the idea that scholarship must be portioned into discrete units, when the truth is that the best scholarship is sprawling, unbounded and mutable. The Web is flexible enough to facilitate this, in a way that print could never do. A print piece is necessarily reductive, while Web-oriented scholarship can be as capacious as required.

To date, though, we still think in terms of print antecedents. This is not surprising, given that the Web is the merest of infants in historical terms. So we find that most advocacy surrounding open access publishing has been about increasing access to the PDFs of research articles. I am in complete support of this cause, especially when these articles report upon publicly or philanthropically funded research. Nonetheless this feels narrow, quite modest. Text mining across a large swath of PDFs would yield useful insights, for sure. But this is not "data mining" in the maximal sense of analyzing every aspect of a scholarly endeavor, even those that cannot easily be captured in print.



One way to challenge the print paradigm is to evolve it cautiously. Hence the introduction of "data papers," which Chavas and Penew defined in 2011 as "a scholarly publication of a searchable metadata document describing a particular online dataset, or a group of datasets, published in accordance to the standard academic practices." Sure — using "papers" to describe non-textual items feels odd (not to mention contradictory to my previous point about the need to move away from print antecedents.) On the other hand, data papers utilize a format that has wide familiarity to offer something new and innovative.

Another way to challenge the print paradigm is head-on, at the conceptual foundations. Scholarly HTML is an ongoing initiative, with hopes of evolving into a formal W3C standard for Web communication. The creators note that, "Scholarly articles are still primarily encoded as unstructured graphics formats in which most of the information initially created by research, or even just in the text, is lost…Information cannot be disseminated if it is destroyed

before even having left its creator's laptop." This is an effort to thwart that information loss, by utilizing the native properties of the Web to present a more complete record of the results of research than is possible in a traditional PDF.

Yet another approach is to use sophisticated data mining tools to surface the numerous facts that are hidden within published research papers. Although papers themselves are generally subject to copyright protection — a vexing reality in its own right — the particular facts within them are not. The Content Mine seeks to extract "100 million facts from the scientific literature" and make them available for re-use and creating new work. While Scholarly HTML aims to build a new communication platform in which such facts are more easily discoverable, the Content Mine aims to maximize re-use within the constraints of current scholarly communication infrastructure.

The common thread between data papers, Scholarly HTML and the Content Mine is a desire to utilize the Web to advance both the creation and distribution of scholarship. Open access in its established sense is useful, but Web-powered scholarship promises to be much more transformational.

Note: This article gives the views of the author(s), and not the position of the LSE Impact blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author:

Marcus A. Banks *(@mab992) is Head of the Blaisdell Medical Library at the University of California, Davis. He has also worked at Samuel Merritt University, the University of California, San Francisco, and the US National Library of Medicine. Increasing the utility of scholarly work is his principal professional concern, although that focus is tested during baseball season.*