Editorial

Artisanal and Industrial: The Different Methods of Data Creation

Last month, I talked about how one defines data and, perhaps more importantly for a cross-disciplinary journal like Patterns, how readers can be confident that they have the same understanding of what the data reported in the article are as the authors do. As I mentioned then, definitions and ontologies are fields of study in their own right, and editorial can only ever really scratch the surface of those topics. But editorial do provide the opportunity get conversations going about these more philosophical themes of interest to the community.

This month, I’d like to talk about how we can categorize data—not by subject area or file type, but by how the data are produced and who they are produced by. Looking at data from this point of view shines an interesting light on decisions that were made in the past about the data, especially when users come back to it after a time away or when other new users want to make use of the data for the first time.

A common phrase used by the data management and digital curation communities is “the long tail of research data.” This long tail consists of datasets that are produced by single researchers, or groups, where the data do not fall within the scope of discipline-based repositories. It’s often contrasted with Big Data and the problems associated with that. Long-tail data might not suffer from the problems of Big Data such as large volumes, or rapid velocity of production, but it does have its own challenges.

Big Data, as a term, does suffer a little bit from scope creep, basically coming down to the subjective question of “what do you mean by big”? A big dataset 5 years ago won’t be considered Big Data now, and size, though an important factor, doesn’t really capture all of the nuances of the challenges of dealing with Big Data.

I find it helpful to think about data using the terms “artisanal” and “industrial.” These are by no means a comment about the quality of the data. From our own real-world experience, I’m sure we’ve all seen situations of high-quality, effectively produced goods from both industrial and artisanal processes (and, unfortunately, the opposite). These adjectives encourage us to think about the “why” and the “how” of the data production process. Who collected the data, and what purposes did they have in mind for it? All these questions need to be considered, not only when thinking about data curation and preservation, but also when we start thinking about how best to publish data or share it with other users.

I like to think of artisanal data as the output from single researchers or small teams, working in (relative) isolation—whether that’s working away at the blackboard or carefully taking measurements at a field site or a lab bench. Artisanal data can be small volume and is generally organized in the way that suits the creator best, using the most convenient and effective tools and software for them, which may or may not be standardized. Oftentimes, the creator will not even be considering the lifespan of the data after they’ve been analyzed, and the results published in a journal article—the data are simply the means to an end. These sorts of data are the types that wind up stored on CDs or memory sticks in a desk drawer, or on the hard disk of a laptop, which then crashes—not good for preservation or for ensuring the reproducibility and verifiability of the scientific record.

It’s important to remember that these artisanal producers of data often simply don’t have the resources to spend the time to learn (for example) ISO standard data preservation techniques. We can get around this by encouraging the use of data repositories, both disciplinary and institutional, but for many researchers, even these repositories don’t exist. It’s a continuing effort to encourage and educate the producers of artisanal data to deposit their data in repositories and use what standards exist. I feel that preserving the scientific record when it comes to data is a group responsibility, requiring collaboration between data producers, data repository managers, journal editors, and funders. Regardless, change is coming, but making it easy for people to do the right things with their data does help a great deal.

By contrast, the industrial production of data is done by large groups and consortia and is often international in scope. Because they’re dealing with massive quantities of data, the only way to be able to manage the data are by building infrastructures and having dedicated support staff, along with standardized processes and conventions for everything from file-naming conventions, to archival policies, to sharing agreements and licensing.

Big Data tends to be described by a varying number of “V”s—for example: volume, velocity, variety, variability, veracity, visualization, and value. These are not all unique to Big Data; in fact, when you look at long-tail data as an aggregate of many parts, variety, variability, veracity, and value are just as important problems as they are for Big Data.

The key distinction is that industrial data are generally standardized, within the domain of production at least (one can’t expect genomics data to comply with the same conventions and file formats as climate modeling data). Industrial data can take advantages of economies of scale and generally has data managers whose full-time jobs are to think about and care for the data. Artisanal data are those bespoke, one-of-a-kind creations that may need special care and attention because they might not fit into the standardized “boxes” produced for industrial data.

This is why part of the standard procedure for submitting articles to Patterns involves asking the authors the questions: “does your manuscript report new largescale datasets” and “does your manuscript report custom computer code or introduce a new algorithm”? If the answer is yes, then further information is requested, as access to these is important for the reviewers (and later on, the readers) to be able to evaluate the article appropriately. This is also why, as standard, we have a data and code accessibility statement as part of all our research
articles and encourage data and code citation as a way of perma-
nently identifying and linking to the underlying code and data (a
discussion of the relationship between code and data is left for
another time).

There is, and will always be, space for both industrial and arti-
sanal data in science. Every large-scale data-producing experi-
ment had as a precursor a smaller-stage experiment where new,
artisanal datasets were produced. Researchers start small and
build up, and the same is true for data. The difference is now
that a “small” dataset is a lot larger than has ever been consid-
ered small in the past, and these changing perceptions of data-
set size will continue as our data creation and management tech-
nologies improve.

Patterns is a home for all data, and all data science techniques
and technologies. Whether your data are lovingly hand collected
by an individual or produced in great volume by a clever, mod-
ern, large-scale process, we don’t mind. What we care about
is sharing the story behind your data and your data science tech-
niques, so other people can learn from your experience and
appreciate all the effort that you’ve put into creating datasets,
tools, services, techniques, and systems that work for you.

Sarah Callaghan
Editor-in-Chief, Patterns
https://doi.org/10.1016/j.patter.2020.100100