From the president

Boris Veytsman

> Something there is that doesn't love a wall
>
> Robert Frost

Attentive readers of *TUGboat* might have noticed a change over the last few issues: articles are now accompanied by unique strings of letters and numbers, their Digital Object Identifiers, or DOIs for short. This would be impossible without hard work done by the *TUGboat* team, and especially Karl Berry, who made many updates to our styles and scripts. Some of these updates may help other publishers. They are released to CTAN and CPAN.

This is an important change. Science and technology are collective endeavors rather than the heroic product of inspiration of singular geniuses. We all build the edifice started by our predecessors, and our papers or programs contain acknowledgments of them. Historically formatting of citations is defined by the "house style" of the publisher, and it took much manual effort to find the canonical citation of the referenced work (my colleagues and I published some research about this on these pages). Exact citation data are useful for the funders (is my money doing good work?), for the practitioners (is my work visible?) and for those studying the science of science (how is knowledge generated and spread?). However, it is very difficult to work with the data without associating a unique identifier with each entity — and this is what DOIs are for. Each paper gets a unique DOI, and it makes it much easier to study the citation relationships between the different papers — especially when authors' own unique identifiers (called ORCIDs) are used.

Assigning DOIs requires some important decisions. The first is, what should get a DOI, and what should not? Every technical paper, sure. Editorials and columns like this one, yes. What about tables of contents? Indices? Advertisements? We made the decision not to assign DOIs to these ephemera, which have no assigned author. This decision might be controversial, but any choice can be criticized. Note that Crossref, the authority controlling the registration of scientific DOIs, elegantly sidestepped this choice: member organizations like TUG must make it themselves for their publications. The entities that get DOIs are not just papers: any "digital object" can obtain one, including programs, datasets, etc. This is a smart decision: if DOIs were assigned to papers only, then we would have a hard question: what is a paper? In a recent work my colleagues and I catalogued the use of scientific software, and found that the question, what is software?, is rather difficult to answer: the differences between software packages, algorithms, Web services, etc., are not evident. Are Google spreadsheets software or a Web service? What about LAPACK with its multiple implementations: is it an algorithm or software?

Another hard question is whether the entity to be given a DOI is one, or many. We do not assign DOIs to separate sections of the same paper. What if the paper is split into parts and published in several consecutive issues? Yet another question is raised when a paper is updated. *TUGboat*, being a physical journal (even if many readers get it in the digital form), has a reference version of each paper. Not so for some other publishers like preprint servers, which allow posting new version of the same paper. Do these versions deserve a separate DOI each? What if between updates the title or the abstract changed? Or if the team of authors changed (strangely, this happens too)?

These questions remind me of the famous problem of the ship of Theseus in philosophy. If we change a rotten plank on this ship, it is still the same ship. What if we change two planks? Three? What if in several centuries we have changed all planks — is it still the ship of Theseus?

There are several ways to treat this paradox. My favorite one is to recognize that the world is infinitely complex, and our labels, like "scientific paper", "ship", etc., are just approximations of it. These labels are useful, but they are not the reality. Thus they are bound to break in some cases.

A similar problem occurs in a field better known to TeXnicians. As many of us know, TeX itself is a program that puts letters on paper. Formats like LaTeX or ConTeXt consider the structure of the text, so `\section{Introduction}` in LaTeX means that (a) we start a new unit of text, called "Introduction", and (b) tell TeX to typeset it in a certain way (bold, or in small caps, or maybe in larger size, etc.), possibly starting the next paragraph without indentation, update the table of contents, etc. Ideally a document in these systems should not have any visual TeX commands, as they relate to how the text is shown rather than what is it.

However, as anybody using these systems knows, more often than not this is not true. Even if the authors do not abuse visual commands, adding unnecessary spaces or decorations, they often need some presentation effects that are simply not expressible by the high-level commands. It looks like our ideas of what constitutes text structure are as approximate as any other ideas — like the ideas of "scientific paper"

doi.org/10.47397/tb/43-3/tb135pres

or "ship". It is not a coincidence that our colleagues who represent text in XML format use "extensibility" in the name of their approach (XML stands for eXtensible Markup Language). They implicitly admit that any specification is not complete and may need to be extended.

Perhaps nowhere is this problem so evident as in the language of mathematics. As editors of mathematics journals know too well, mathematicians are fond of inventing new notation, stretching the limits of typesetters and TEXnicians. I think it is not a coincidence or another manifestation of the peculiar nature of mathematicians themselves. Rather, the existing notation reflects our existing knowledge, and if we actively work on extending the knowledge, we necessarily need to extend our notation.

As an aside, this makes Knuth's structure of mathematical formulae, with eight classes of mathematical objects, even more surprising since it has endured several decades of heavy use. In general, TEX mathematical typesetting is still very close to that of the original version. A notable exception is the `\middle` primitive added by ε-TEX to `\left` and `\right`. LuaTEX adds a number of other primitives (and development continues there). It will be interesting to see whether they will be adopted by other engines in the future.

At the end of the day these questions, from the ship of Theseus to DOI assignment to mathematical notation, are questions of classification. Jorge Luis Borges touches this problem in his famous essay *The analytical language of John Wilkins*, where his protagonist invents a meaningful language, which is also "a secret encyclopedia":

> He divided the universe in forty categories or classes, these being further subdivided into differences, which was then subdivided into species. He assigned to each class a monosyllable of two letters; to each difference, a consonant; to each species, a vowel. For example: *de*, which means an element; *deb*, the first of the elements, fire; *deba*, a part of the element fire, a flame. In a similar language invented by Letellier (1850) *a* means animal; *ab*, mammal; *abo*, carnivore; *aboj*, feline; *aboje*, cat; *abi*, herbivore; *abiv*, horse; etc. In the language of Bonifacio Sotos Ochando (1845) *imaba* means building; *imaca*, harem; *imafe*, hospital; *imafo*, pesthouse; *imari*, house; *imaru*, country house; *imedo*, column; *imede*, pillar; *imego*, floor; *imela*, ceiling; *imogo*, window; *bire*, bookbinder; *birer*, bookbinding. (This last list belongs to a book printed in Buenos Aires in 1886, the *Curso de Lengua Universal*, by Dr. Pedro Mata.) *[Translated from Spanish by Lilia Graciela Vázquez; edited by Jan Frederik Solem with assistance from Bjørn Are Davidsen and Rolf Andersen.]*

This long quotation reveals the major problem with any classification: it is always "a secret encyclopedia". However, an encyclopedia requires absolute knowledge. If our understanding of nature is not complete, the classification cannot be right. John Wilkins in the 17th century considered fire a primary element, and his language reflected it. We now consider fire a complex physico-chemical process. Thus Wilkins' language is inadequate for us. A similar problem is with Linnaeus' nomenclature in botany and zoology. It reflects our knowledge of the relationship between the species. If the knowledge is updated, for example, due to molecular genetics methods, the nomenclature becomes "wrong". Since our knowledge is never absolute, our classification is never right.

Yet another way to express this is that stable mathematical notation is possible only for *dead* mathematics, a dry school subject rather than living and developing field. The latter requires constant changing of the notation.

One can compare our efforts in classification to wall building. We constantly build walls around the things we know, delineating our knowledge. However, as Robert Frost noted (*Mending Wall*, 1914),

> Something there is that doesn't love a wall,
> That sends the frozen-ground-swell under it,
> And spills the upper boulders in the sun,
> And makes gaps even two can pass abreast.

Does this mean that our efforts are futile? No. This wall building is our effort of understanding the infinitely complex world, creating better and better models for it — while always remaining imperfect.

Some prefer to think of this work as building roads rather than building walls. In this case another metaphor is apt, this one by Ukrainian poet Ivan Franko (*The Stonecutters*, 1878): *we are but stone cutters for the road of progress.* It is interesting that both Frost and Franko think about stone as the hard material we are working on. For a long time stone was used to make inscriptions for the future — something intimately close to us TEXers.

Like Frost's protagonist and his neighbor, "we wear our fingers rough" creating imperfect inscriptions with our imperfect tools. We are bound to do this till the end times.

◇ Boris Veytsman
`president (at) tug dot org`