

Why untidiness is good for us

Books and formal papers make knowledge look finite, knowable. By embracing the unfinished, unfinishable forms of the web we are truer to the spirit of enquiry – and to the world we live in, says **David Weinberger**

IN RECENT years, controversies over issues ranging from the possibility of faster-than-light neutrinos to the wisdom of routine screening for prostate cancer have increasingly raged outside the boundaries of peer-reviewed journals, and involved experts, know-nothings and everyone in between. The resulting messiness is not the opposite of knowledge. In the internet age it is what knowledge looks like, and it is something to regret for a moment, but then embrace and celebrate. Knowledge is fast reshaping itself around its new, networked medium – thereby becoming closer to what it truly was all along.

Our old idea of knowledge shaped itself around the strengths and limitations of its old medium, paper. We all understand those strengths: paper is cheap, displays text and graphics, lasts a lot longer than hard drives, and no technology is needed to make it work. But there's a price: paper doesn't scale or link, which has made knowledge and science both what they are and less than they could be.

Paper fails to scale in two directions. First, there are limits on what can be published: if it is not considered important enough, even good science is rejected. That is a reasonable response to the cost of journal printing and shelf space but it is far from the ideal of science, where all data and all hypotheses are welcome. With such limitations, regimes

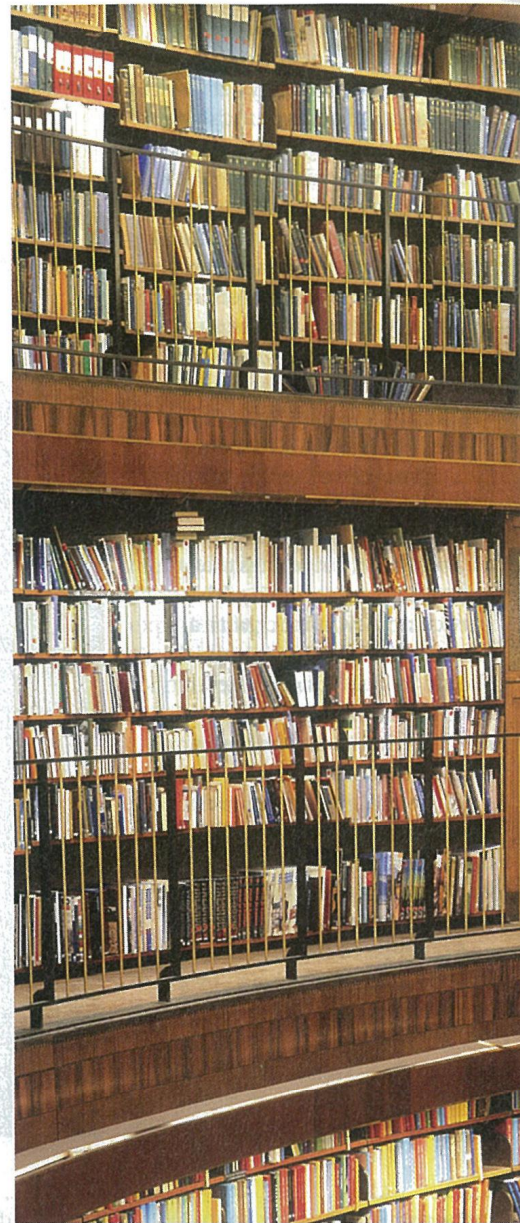
emerge to dole out the scarce resource. Second, printed articles rarely contain all the data on which their conclusions are based, and literature reviews are kept to a reasonable length – reasonable being dictated by the economics of atoms.

The other limitation has an arguably greater effect: printed matter does not link. Each book is its own thing. The references to other books don't work, no matter how hard you click them. That means the author has to cram everything the reader needs into one volume, summarising references to other books in a single paragraph, and pulling sentences out of their rich context.

Printed books are also disconnected from the discussions that appropriate them into the culture – and that correct them. Authors must anticipate objections because once published, books cannot be altered. In the Age of Paper, knowledge looks like that which is settled, or settled enough to be committed to paper.

These limitations led to the typical rhythm of scientific discourse. Do your research. When you're as sure of it as you're going to be, make it public. Only then is it officially yours. If someone publishes before you, you lose. And once it is public, it becomes hard – and often embarrassing – to change even a word.

Science's new medium overcomes both limitations. We have scarcely plumbed its capacity, and it is so hyperlinked that if digital content is not linked, it is essentially unpublished. This fundamentally changes the understanding of the nature of knowledge and science prevailing in the west for the past 2500 years: to know X was to know its essence, its place in the rational order. That order



consisted of a set of coordinated definitions based on essential differences and similarities. That's one reason Charles Darwin spent seven years discovering whether barnacles were molluscs, as Linnaeus said, or crustaceans. The result was a two-volume work to establish a single fact: they are crustaceans.

These days we don't care nearly as much, in part because we recognise that how we classify things depends on our interests. Studying the evolution of marine creatures? Classify them based on their genetic history. Studying how to keep hulls smooth? Then lump barnacles with rust. The internet has decisively moved us from belief in a knowledge of universal essences because it has made plain two facts: we don't agree, and we can't let that stop us.

PROFILE

David Weinberger is a senior researcher at Harvard University's Berkman Center for the Internet and Society. This essay is based on his new book, *Too Big to Know* (Basic Books)



For example, at the online Encyclopedia of Life (EOL), you can look up an organism by any name you want, and obtain information about it within any of the taxonomies it supports. So two scientists who disagree can collaborate because they know they are both talking about the creature on the same page of the EOL. This is an example of “namespaces”; domains that bestow a set of unique identifiers on their objects. These names can be mapped so we can work together without having to agree. The result is a sloppy mess with names and categorisations overlapping unevenly. But the different schemes add information and meaning: so long as we can map them, we are better off not waiting for resolution.

We see the same approach with another

The limitations of paper were crucial in shaping our western notions of knowledge

promising development: the rise of “big data”. Organisations are releasing gigantic clouds of data for public access. Since these are too voluminous for cranial processing, the data is increasingly released in the linked data format recommended by Tim Berners-Lee. In this format, data consists of “triples”: subject, object and a relationship connecting them.

This might look like a further atomisation of facts and information, but it is the opposite since each element of a triple should ideally consist of a link pointing to some spot on the web. For example, in the triple “barnacles have shells”, the word “barnacle” might link to an

EOL entry, “have” to a site that explains how creatures can have components, while “shell” might point to the relevant Wikipedia entry. This technique helps computers see that the triple “Cirripedia are crustaceans” refers to the same thing as triples calling them “barnacles”. Linked data are facts literally consisting of links. The resulting tangle of pointers is quite unlike our old view of facts as well-defined building blocks.

And these clouds of data are being released without being thoroughly vetted. For example, the US government website, Data.gov, says that the data it has gathered from federal agencies is raw. We might prefer

“Knowledge will live in a web of discussion, debate, elucidation, disagreement”

tidy, vetted data, but that doesn’t scale; we do better to have lots of data, even if it’s not perfectly structured or completely reliable. Messiness is the price of scaling.

Further, rather than working in private and publishing to a select group, we are finding tremendous value in posting early on non-peer-reviewed sites, and letting everyone chime in. We saw this when the scientists who discovered what might be faster-than-light neutrinos posted their work at arxiv.org, the pre-print site. The discussion sprawled across the internet, with amateurs and professionals weighing in, with kind-hearted experts explaining it to lay people, with insightful and pointless ideas stirred together – and all without prior peer review and outside the standard journals.

The result was that if you wanted to see where the knowledge about neutrinos “lived”, you wouldn’t go to the library or online versions of the standard journals. The knowledge lived in the loose web of discussion and debate. All this happened faster, wider and deeper than if science had stayed in its paper comfort zone. Even after the question is settled, the knowledge will live not in the final article but in that web of discussion, debate, elucidation and disagreement. It’s messy, but messiness is how you scale knowledge.

Knowledge has inherited many other of the web’s properties. It is now linked across all boundaries, it is unsettled, it never comes fully to rest or agreement, and we can see that it is bigger than any of us could ever traverse. But doesn’t that make internet-based knowledge and science more like the very human world into which we have been thrown? ■