VISION PAPER

Open Science and the Hype Cycle

George Strawn†

US National Academy of Sciences, Washington DC 20418, USA

Keywords: Open Science; Hype cycle; FAIR data; Digital Objects

Citation: Strawn, G.: Open Science and the hype cycle. Data Intelligence 3(1), 88-94 (2021). doi: 10.1162/dint_a_00081

ABSTRACT

The introduction of a new technology or innovation is often accompanied by "ups and downs" in its fortunes. Gartner Inc. defined a so-called hype cycle to describe a general pattern that many innovations experience: technology trigger, peak of inflated expectations, trough of disillusionment, slope of enlightenment, and plateau of productivity. This article will compare the ongoing introduction of Open Science (OS) with the hype cycle model and speculate on the relevance of that model to OS. Lest the title of this article mislead the reader, be assured that the author believes that OS *should* happen and that it *will* happen. However, I also believe that the path to OS will be longer than many of us had hoped. I will give a brief history of the today's "semi-open" science, define what I mean by OS, define the hype cycle and where OS is now on that cycle, and finally speculate what it will take to traverse the cycle and rise to its plateau of productivity (as described by Gartner).

1. TODAY'S "SEMI-OPEN" SCIENCE

Today's science is semi-open in that only scientific articles are available, and often at a significant price. Semi-open articles are available because the western printing press was invented in Germany in the 15th century. And as Elizabeth Eisenstein has argued [1], it revolutionized western society, including science. "Only" two centuries later the Royal Society in England encouraged scientists (called natural philosophers in those days) to publish their results rather than keeping them private [2]. The number of journals supporting this mission grew slowly at first (consistent with the slow growth in the number of scientists), but in the 19th and 20th centuries both numbers exploded, especially after Germany created the research university in the early 19th century and other western nations followed suit (e.g., Johns Hopkins in the US in 1876).

† Corresponding author: George Strawn (Email: gostrawn@gmail.com; ORCID: 0000-0003-4098-0464).

2. TOMORROW'S OPEN SCIENCE

The printing press was the *innovation trigger* that enabled semi-Open Science (OS). High performance networked computers are today's equivalent of the printing press. They are a thousand times more powerful and a thousand times cheaper than computers were a generation ago: A multi-terabyte disk costs a hundred dollars, a thousand dollar computer can execute several billion instructions per second and the network that connects millions of such computers has a bandwidth approaching a terra-bit per second. Hardware with these capabilities is the innovation trigger that will enable *Open Science*, which is *the "publication" of all relevant science results: data, software, workflows, etc., in addition to summary articles describing those science results, and accessibility of all these elements (at least all those produced with public money) for the cost of an Internet connection.* Publication takes on a new meaning here, since the computers must "understand" all of these artifacts well enough to provide us novel assistance.

Science articles, of course, have been digital for some time, but neither open in the sense of OS nor available for semantic analysis. Software is digital by nature and the open source movement has already made much important software part of OS. Scientific workflows are currently undergoing extensive automation that will more easily make them part of OS. This article focuses on scientific data, which is in process of becoming part of OS.

Of course, hardware is only half of the IT proposition. And whereas the hardware is OS-ready, the software is not. I whimsically describe three generations of computing, only a third of which can support OS. Generation One was of isolated (also low performance and expensive) computers—1950 to 1995. The emergence of the commercial Internet in 1995, which has connected an exponentially increasing number of computers, ushered in Generation Two. This generation is still in existence—1995 to 2025 or 2030? An early workstation manufacturer correctly described this generation as one in which "the network is the computer" (but with many non-interoperable data sets). The Third Generation of computing will emerge when advanced software *enables the interoperability of heterogeneous data*. So not only will be only one computer (the network), but that computer will effectively have only one data set.

Interoperability of data requires a semantic layer. Semantics is one of the long-studied areas of artificial intelligence and it entered a new phase in the new century. Twenty years ago when Web designers created the *Semantic Web* [3], an observer noted that "the only new thing about the Semantic Web is the Web." Nevertheless, just as the Web revolutionized access to information, the Semantic Web *may* revolutionize access to knowledge. Another effort making progress in the interoperability of heterogeneous data is the *Digital Object (DO) Architecture* [4]. It was also designed several decades ago, but it, too, has only become practical with today's high performance connected computers. When efforts such as these or others achieve their goals, we will have the third generation of computing, where not only is the network the computer, but, as indicated above, "all" of its data are interoperable. As the word "data" is used here, it includes articles, data, software, workflows, etc. In such a technical environment, OS will be enabled.

3. THE HYPE CYCLE

Of course, being enabled and being widely adopted are two different things. The path between enablement and adoption, also known as the path from innovation to production, has been characterized by Gartner as a *hype cycle* [5]. The following graph clearly represents the possible ups and downs along this path (Figure 1).

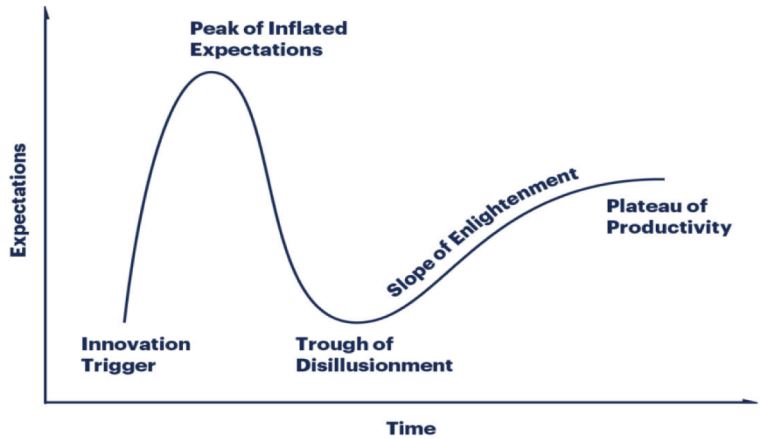

Figure 1. The possible ups and downs of a hype cycle. Note: Source Ref. [5].

4. INNOVATION TRIGGER

Gartner asserts that this curve represents a common pattern that arises with many new technologies or other innovations. As stated above, the hardware innovation trigger has been in evidence for ten years. At that time, the US Federal agencies established an interagency senior sterling group to research *big data* [6]. The thesis of this group was that we could now store more data than we could effectively process. Several years later, the US Office of Science and Technology Policy tasked the research-oriented agencies to make all scientific articles and data created with Federal support available to the public (Open Data (OD)) for the cost of an Internet connection [7]. Several years after that, a conference at the Lorentz Center in Leiden University asserted that merely "opening" that data was not nearly enough to assure interoperability. The conference created the acronym FAIR data, identifying some of the characteristics that OD need to be maximally useful: findable, accessible, interoperable, and reusable [8]. As stated above, "open" scientific data are part of OS along with articles, software, workflows, etc. But in a broader use of the word, articles, software, workflows, etc., are all data or *DOs*.

So for the last decade, many people could see that OS was no longer a pipe dream. As we advanced in our terminology from big data to OD to FAIR data our expectations have continued to rise, even though FAIR data were a policy prescription, not working software.

5. THE PEAK OF INFLATED EXPECTATIONS

I suspect that we may be at or near a peak of OS expectations now. The European Open Science Cloud (EOSC) is being funded, the US Federal agencies are beginning to implement their OD requirements, and the global Research Data Alliance (RDA) [9] has been functioning for almost a decade. FAIR data *could* have helped the world in its fight against the coronavirus, and FAIR data proponents say this has created a spur to action and that we will be better prepared for the next pandemic.

6. THE TROUGH OF DISILLUSIONMENT?

After a decade of pursuing OS, I fear that our expectations may be about to be tested. We are beginning to acknowledge that the software is not ready and that we have not solved critical problems in the business and social dimensions, which are of course more difficult than software problems. Although I wish the EOSC and the US agencies the best of luck, I fear that their initial efforts will produce data that are far from FAIR. And the RDA has been working on components, not data systems. These potential disappointments could cause many persons to assume that OS is not all it was cracked up to be. I want to stress that I see a *potential* for disillusionment, not a certainty. On the other hand, I am certain that OS will, with or without a trough, rise to a plane of productivity.

Regarding the business dimensions of OS, the publishers are concerned about their viability if they loose their subscription model of pricing. And since the publishers include scientific societies as well as commercial firms, an important part of the science community is concerned. As an example of the active resistance, the US government intended earlier this year (2020) to put more teeth into the agency open data requirement, but a major lobbying effort [10] has, at least for the time being, sidetracked that plan. OS advocates also fear that commercial publishers want to take over the scientific data market as they have for years held much of the scientific article market. If they were to succeed in locking away data under copyright protection as they have with articles, the disillusionment would be palpable.

As difficult as the business problem may be, the "social problem" may be the greatest (temporary) impediment to OS. The social problem as I perceive is that many (but not all) scientists believe that *OS may be good for science, but it is bad for scientists*. So far (at least in the US) scientists have been given sticks but no carrots to adopt open practices. Until scientists are properly rewarded for sharing all their outputs (as they are currently only for their articles), such sharing is a personal imposition even if it is good for the science enterprise as a whole.

7. THE SLOPE OF ENLIGHTENMENT

After all this apparent naysaying, it is time for optimism to emerge. Semantic software *will* emerge, intellectual property laws *will* recognize that information is different, and scientists *will* be rewarded for and will embrace OS. I suspect that software will be first and intellectual property laws second on the slope

of enlightenment. Then when (a new generation of?) scientists embrace OS, we will have reached the plateau of productivity.

As mentioned above, semantic software in the context of the Web and elsewhere is twenty years old. It has had a slower uptake in part because there has been no "major force" backing it. In the case of the Internet, there was first the major effort of the US Defense Advanced Research Projects Agency (DARPA), which invested twenty years of continuous development (first with NCP then TCP/IP). Then the US NSF (National Science Foundation) invested ten years, which resulted in the Internet becoming the network infrastructure for US higher education. Note that it was only after these thirty years of development that the Internet "burst upon a surprised public," as if it had just been created. DARPA also supported research in the late 1990s that led to the semantic Web, but neither it nor NSF followed through with the same level of support that they had given to the Internet. For these and other reasons, data software has been left to the private sector, which of course, prefers vendor lock-in to a vendor-neutral solution. Thus we see the continuing search for *de facto* data standards. The Internet was also a *de facto* standard, but after the thirty years of federal support, it became a *de jure* standard. I suggest and hope that such standards and software to implement them will be developed in the 2020s.

Intellectual property laws were developed long before information technologies made copying an information product a free good. And as the historian Harari has observed [11], the West has only produced one new slogan since "Liberty, Equality, Fraternity" in 1800, and that new slogan is *"Information Wants to be Free"* [12]. In the provocatively titled book, *Postcapitalism* [13], Paul Mason asserts the increasing array of "free" goods will undermine the market pricing mechanism of capitalism and the only options for our economic future are either monopoly and locked-up data or free goods and new mechanisms. In *Understanding Media* [14], Brian Winston observed that the status quo always seeks to suppress the revolutionary potential of a new technology. So there will be continuous and vigorous lobbying in support of the status quo. But societies that make full use of information via the free copy path may be the successful societies of the future. Jared Lanier (and others) has raised a related question of *Who Owns the Future?* [15] as the public gives away its information in return for free services plus advertisements. I would venture that by the late 2020s or early 2030s at least some societies will have decided in favor of free coping and against monopolies. One of the beneficiaries of such a decision will be OS.

8. THE PLATEAU OF PRODUCTIVITY

As mentioned above, when scientists embrace OS, the plateau of productivity will be upon us, and we may find that it creates a scientific revolution as profound as that of the 17th century by accelerating scientific discovery. But the question remains: *which* scientists will do the embracing? As mentioned above, I suspect that it will be a *new generation* of scientists who will understand the value of the broader sharing of results *and* who will also be rewarded for their OS participation.

One anecdote will illustrate my point. Over the last two decades, researchers at the US National Library of Medicine developed an experimental knowledge overlay for their popular database *Medline*, which