Information Services & Use 43 (2023) 243–247
DOI 10.3233/ISU-230204
IOS Press

Unanticipating metadata: Metadata in the ages of the internet and AI

David Weinberger*
Berkman Klein Center for Internet & Society, Harvard University, Cambridge, MA, USA

Abstract. This article is based upon the open plenary talk given at the NISO Plus Conference on February 14, 2023. David Weinberger, an American author, technologist, and speaker, explores the effect of the Internet and AI on metadata. He discusses how traditional metadata has had to anticipate the uses to which it will be put, who will use it, how users will navigate it, how it will be encoded, and how much metadata is enough. He asserts that the Internet has changed that because it encourages the adoption of the strategy of *unanticipation*.

Keywords: Metadata, internet, artificial intelligence, future, unanticipation

The Age of Artificial Intelligence (AI) is changing the role and nature of metadata even before the Internet is done messing with it. The Age of AI is so young that it is hard to predict its effects on anything, but perhaps the Internet's transformation of metadata can give us some clues about how metadata may be evolving or, perhaps, is being disrupted.

1. Metadata and anticipation

Three major and completely obvious characteristics of the Internet have resulted in important changes in the nature of metadata: The Internet is huge, it's open to everyone, and it's digital.

These characteristics have had an effect on what we can take as the traditional purpose and value of metadata: to make things findable; to make things interoperable; and to help us make sense of things.

For metadata to achieve these purposes, we have had to anticipate what we are going to use it for. There is nothing unusual about this. Our relatives hundreds of thousands of years ago (and perhaps three million years ago [1]) were creating tools, anticipating their use and preparing for the occasion. Anticipate-and-prepare has been our strategy of strategies since we became recognizable as people. We are never going to give up on that strategy, if only because we do not want to be run over by a bus. But much of what is distinctive about the Internet comes from a very different strategy: unanticipation. And this has had important effects on metadata.

One way to see this new strategy at work is to compare product design before and after the Internet became a force. For example, in 1908, Henry Ford and a handful of engineers designed the Model T,

*E-mail: david@weinberger.org.

0167-5265 © 2023 – The authors. Published by IOS Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (CC BY-NC 4.0).

anticipating market needs so well that fifteen million of the cars were sold over the course of nineteen years with only the most modest of changes [2]. Henry Ford anticipated what the market wanted so successfully that he's still admired despite being a supporter and enabler of Hitler, and a proud antisemite [3].

Dropbox took a different approach when it launched its first product in 2008 [4]. Rather than trying to anticipate what people would want to do with its cloud backup technology, it launched a product with the most minimal feature set it could while still launching successfully, and then waited to see how people used it, what they wanted from it, what they were saying to one another on the Internet, and so forth. This validated the market desire for the product and helped Dropbox figure out what features to add to its next releases. This is the Minimum Viable Product (MVP) strategy that has become common on the Internet. Why anticipate when you can learn along the way?

A similar strategy explains the success of the iPhone. There were basically no features of that phone that had not been in prior phones except for one, added a year after launch: the App Store. The App Store was Apple acknowledging that it couldn't know everything that people around the world might want to do with a palm-sized, wireless, connected computer. Instead, Apple decided to let everyone with an idea create apps that Apple could not have anticipated and probably would not have had the resources to create.

Computer games have long taken this approach, allowing users to create modifications — "mods" — that add new characters, weapons, and maps, as well as altering the rules. For example, Minecraft, the most popular game in the world — one hundred and seventy-six million people played it in January 2023 [5] — has achieved this position in part by enabling mods. There are currently over one hundred thousand of them [6]. Each extends the game in ways that Minecraft could not have anticipated and would not have had the resources to create if they had.

Unanticipation is also the strategy behind open source, open access, open APIs, and the Internet itself which was designed to support as many applications and new protocols as the world cares to make. That is why Net Neutrality is so essential, for it is the policy that stops (or should stop) commercial Information Service Providers (ISPs) from deciding that the Net is "really" for streaming commercial video, for consuming rather than for creating [7], for listening to the top-selling musicians, or anything else. The Internet is really for inventing applications and ways to communicate that the architects of the Internet could not have anticipated.

Traditional metadata has had to anticipate the uses to which it will be put, who and how users will navigate it, how to encode it, and how much metadata is enough. The Internet has changed that.

2. Three orders of metadata

Traditionally, we have had three "orders of order", each with their own type of metadata [8]. The first order of order is the physical arrangement of objects, such as books in a library. Everything has to be in a place. The metadata in this order is physically attached to the objects that they label. The space is typically small so the metadata is limited; consider the size of a library label on the spine of a book.

With the second order, the metadata is separated from the physical object to which it refers. For example, it might be put on index cards that are then filed alphabetically. The index cards can contain more metadata than the label on the physical object, and can present multiple ways of locating the items: title, author, subject, and perhaps a few more. But still the number of sorting possibilities and the amount of metadata that can be presented is limited.

With the third order, the metadata and the content are digital. Now the amount of metadata and the number of sort orders can be indefinitely large. This opens the possibility of users adding their own metadata to digital objects and this changes the fundamental relationship of data and metadata.

Take a simple example: You can't remember the name of the whaling book written by Herman Melville. So you go to your favorite search engine and type in "Melville whale book" and it returns "Moby-Dick". Minutes later you've forgotten the author's name, so you type in "Moby-Dick author" and the search engine helpfully responds, "Herman Melville". Then you forget both these facts and type in "Book with 'Call me Ishmael'" and the search engine cheerily responds, "Moby-Dick."

This is all perfectly ordinary, of course, except for your terrible memory. But it is also quite remarkable, for in each case, metadata and data switch: At one moment, "Herman Melville" is metadata and in the next it's the data that "Moby-Dick author" finds.

Our experience on the Internet has decisively taught us that the only difference between data and metadata is a functional one: Metadata is what you know and data is what you're looking for. In fact, "The author who had a picnic with Hawthorne on top of Monument Mountain" is metadata that will return the datum "Melville".

In our most common information retrieval tasks, we are no longer stuck with the metadata we can anticipate. We have entered the Age of Metadata Unbound…unbound from always needing to be anticipated.

3. Metadata in the age of AI

Now AI has come along to make metadata yet more complicated. There are perhaps four ways we can use AI — by which I mean machine learning — as a metadata tool.

First, AI can do at least some of the work of classifying items into established, structured metadata categories, although the difficulty Google Books had in algorithmically deriving bibliographic data from scanned texts [9] is a reminder that this can be harder than over-enthusiastic technologists may think.

Nevertheless, it is entirely plausible to think that AI is going to be helpful in assigning data to structured classifications, especially if there is a "person in the loop" who reviews the AI's suggestions. Not only would human discretion be helpful, but AI systems that are trained on language are subject to "hallucinations," which is the technical way of saying that they sometimes just make things up. This problem is obviously the subject of a great deal of concern and research [10]. In fact, auto-categorization is already happening. Here's a small-scale example: Matt Web trained a machine learning system to assign a Dewey Decimal System number to one thousand episodes of the BBC Radio's *In Our Time* [11].

Second, AI can discover its own categories, and not just sort items into existing ones [12]. Such a system clusters works according to correlations it finds in the works themselves. This can enable it to assign tags without reference to a pre-existing taxonomy.

Third, machine learning determines its outputs based on a degree of confidence typically expressed as a number from 0 to 1. So, if it were asked to come up with subjects to file *Moby-Dick* under, the system could state its confidence about its assignations: 0.96 confident it's fiction, 0.81 confident it's about obsession, 0.43 it's an adventure, 0.01 that it's a romantic comedy. Exposing metadata's weight or degree of likelihood opens up possibilities for interfaces that give more control to users. It can also remind users that metadata need not be binary, and that machine learning systems speak not as gods on high.

Fourth, machine learning enables a rich form of discovery via what we might call latent metadata, although it stretches the metadata-data model itself. For example, in March 2023 (when I'm writing this),

I prompted chatGPT with: "Dante postulated an afterlife with three levels. What are ten artworks that also show something in three parts?" It came back with an excellent list, with the tripartite nature of each work cogently explained.

Likewise, one can imagine works being clustered by similarities in style, by algorithmically-created summarizations, by similar ideas in very different disciplines, by works that are maximally opposite, and so forth. Some of these might be spurred by humans asking chatty AI questions such as "What are the least romantic British novels written in the 19th century?", or "What cookbook is closest to *The Origin of Species*?" In the right circumstances, anything can be metadata, so why not?

4. Metadata tomorrow

Many of these new roles for metadata are possible because machine learning is able to deal with complexities that computers cannot explain. This enables them to take in more than we humans can of the ultimate black box: the world.

The black box nature of the world is also the ground that made anticipation a first-recourse strategy. It was the best way we could control our environment, despite the enormous hidden cost of the over-preparation, under-preparation, and mis-preparation that that strategy inevitably entails. But now that our new technology enables us to wring benefit from patterns in data that are unintelligible to us, we are better able to acknowledge and appreciate the overwhelming and chaotic nature of our situation in the world.

To gain advantage from such a universe - and perhaps even to survive it - we will demand more of our metadata:

- Once we have a taste for metadata without bounds, silo-ed data may start to feel like an offense to our humanity. Given all that we can find and know now, why can't we learn from what is in this particular silo? Security concerns will of course necessitate the continued existence of some silos, but the rise in the power of metadata is going to continue to fuel the cultural drive towards open access data. And vice versa.
- Even when two data sources are open, if the metadata isn't interoperable — either natively or through programmatic transformations — then those data sources are functionally silos. Interoperable metadata is likely to become ever more important, even if the interoperability is accomplished computationally on the spot rather than through global agreements on standards.
- Now that our machines enable it, we will expect metadata to become more and more contextually aware, both of the subject matter and of the needs and capabilities of the person (or system) using the metadata.
- Increased contextual awareness will enable systems to deliver responses more relevant to the individual. The individual's own work and play patterns are important data for this. We can only hope against hope that we will simultaneously be given more control over what our machines know about us.
- We can also hope that it becomes routine for systems to accompany their outputs with a statement of their confidence in that output. Humility is a virtue for machines as well as for humans.
- We need machine learning systems to be as transparent as they can be about how they came up with their responses. This is a knotty problem in multiple dimensions, but being explicit about the features that led to a particular output is one approach; features are in some odd sense a type of metadata. But knowing the features involved in a model's output can be pointless if those features are not intelligible by humans, and when there may be an overwhelming number of them. Transparency about necessarily obscure factors does not clarify anything.

All this does lead one to wonder how important the division between metadata and data will be in the future. It certainly will be key in traditional data solutions, and when the data set is small or well structured enough.

But the Internet has already trained us to rely on a fluid and functional division between data and metadata. Our encounters with machine learning that works by finding hidden currents in a swirling ocean of data stretches the concept of metadata, perhaps to the breaking point. Are features, vectors, and complex patterns metadata? Is that a useful framing?

It's too early to tell. And, as the idea of unanticipation expresses, it is actually always too early to tell.

About the Author

Dr. David Weinberger over the past twenty years has been a fellow, senior researcher, and member of the Fellows Advisory Board at the Berkman Klein Center . Trained as a philosopher, with a doctorate from the University of Toronto, he was co-director of the Harvard Library Innovation Lab, and a journalism fellow at Harvard's Shorenstein Center. Dr. Weinberger has also been a marketing VP at pioneering Web companies, an adviser to high tech companies and to presidential campaigns, and a Franklin Fellow at the U.S. State Department. For two years he was a writer-in-residence at Google AI's People and AI Research (PAIR) group, and was recently an independent editor-in-residence in Google's Moral Imagination group. He edits the *Strong Ideas* open access book series for MIT Press. E-mail: david@weinberger.org.

References

[1] H. Docter-Loeb, A 3 million-year-old discovery may rewrite the history of intelligent life on earth, *Vice*, 2023. https://www.vice.com/en/article/88x4vv/oldest-oldowan-tools-intelligent-life, accessed September 23, 2023.
[2] History.com editors, "Model-T". The History Channel, Apr. 26, 2010. https://www.history.com/topics/inventions/model-t, accessed September 23, 2023.
[3] J.R. Logsdon, Power, Ignorance, and Anti-Semitism: Henry Ford and His War on Jews, https://history.hanover.edu/hhr/99/hhr99_2.html, accessed September 23, 2023.
[4] A. Baus, Examples of successful apps that were MVPs first, *Decode*, 2022. https://decode.agency/article/app-mvp-examples, accessed September 23, 2023.
[5] A. Sharma, How many people play Minecraft? 2023 player count, Charlie Inttel, Feb. 15, 2023. https://www.charlieintel.com/minecraft/how-many-people-play-minecraft-2022-player-count-195437/, accessed September 23, 2023.
[6] *Minecraft* modding, Wikipedia, https://en.wikipedia.org/wiki/Minecraft_moddin, accesses September 23, 2023.
[7] We have lost that battle in the U.S. where the big commercial carriers offer downloads in the gigabit range, but still 10–30 mbps for uploading.
[8] I discuss this in my 2007 book, *Everything Is Miscellaneous,* https://en.wikipedia.org/wiki/Everything_Is_Miscellaneous, accessed September 23, 2023.
[9] L. Miller, The trouble with Google Books, *Salon*, Sept. 9, 2010, https://www.salon.com/2010/09/09/google_books/, accessed September 23, 2023.
[10] For example, Tim Simonite, AI Has a Hallucination Problem Thath's Proving Tough to Fix, *Wired*, Mar. 9, 2018, https://www.wired.com/story/ai-has-a-hallucination-problem-thats-proving-tough-to-fix/, accessed September 23, 2023.
[11] M. Webb, New thing! Browse the BBC In Our Time archive by Dewey decimal code, BBC Interconnected, Feb. 7, 2023, https://interconnected.org/home/2023/02/07/braggoscope, accessed September4 23, 2023.
[12] For an introduction, see Onesmus Mbaabu, Clustering in Unsupervised Machine Learning, *Section*, Nov. 18, 2020. https://www.section.io/engineering-education/clustering-in-unsupervised-ml/, accessed September 23, 2023.