Linking Metadata, Validation, and Democracy

© by Wolfgang Senges



graphics design © by Wolfgang Senges

[Free download available[1].]

Looking back, it seems I am kind of obsessed with concepts and taxonomies since I was about 13 years old. That's when I fell in love with building taxonomies. Later, as a teenager, I wanted to get hold of and control any information buried in my fast-growing music collection. On the next level, my master thesis in Computational Linguistics focused on a conceptual clustering algorithm to find patterns and relations between one-syllable German nouns and the attribution of their respective gender. Then, a query sent out to broadcasters to understand their progress in building media databases landed my first job, during which I learned the fundamental lessons of the value of metadata. Finally, I discovered licensing to be the most crucial pillar in the music business, deeply intertwined with the challenge of metadata.

But while constantly working on the topic ever since, metadata has evolved—a lot.

The narrative of the metadata chaos: not solved but outshined.

Whenever the topic of metadata comes up in the music business, we all agree that the industry is suffering from a massive problem for decades. Despite all efforts, it still is a clusterf*** of sorts. However, some progress has been made in the fact that we are aware of the problem. Also—surprise!—licensing requires a strict management by *universally unique identifiers* (UUIDs). Still, the mantra is: "It's a metadata chaos. Data is incorrect; data is missing; money is lost."

[1] https://bit.ly/data-democracy-wsenges

As the nature of business is, talk is cheap, and time flies. Meanwhile, the issues in the music business by far have been outpaced by the "real" world outside of our small niche. A related challenge, yet of so much more impact on democracy and devastatingly more complex, has already punched our collective stomach multiple times: denial of facts, fake news, and deep fakes.

After all, it is about time to focus on one specific parameter in working with metadata: validity. During the process of developing an approach for validation, creative industries might act as a sandbox, or they might at least benefit from general research. Looking at real-world threats by missing validation, you might even want to forget about headaches in the music business or any other industry altogether.

With this article, I would like to shine a light on how the metadata space evolved and draft a basic model helping to understand where to start.

The linear use of metadata

Placing signs on trees and stones to mark good locations for food or shelter in the early days of humankind might have been where it started. In 280 BC, tags for title and author were attached to script roles at the Great Library of Alexandria[2]. All of this represents a simple kind of metadata to find and remember objects.

Let's not forget one category of lists that was very popular until the middle of the 20th century: By taking the name of an object as its referencing identifier, alphabetically sorted lexicons and almanacs have been a valuable tool to retrieve more information and data connected to the object in question. Even today, the format has evolved with Wikipedia and all kinds of Wikis.

The world of potentially accessible items grew, and the data vaults did the same.

The metadata space

Digital data, along with globalisation, led to an explosion of information. There was more administrative data, as well as there was more descriptive data. This first culminated with statistical data, including data describing the relation between objects like Fuzzy Semantics.

Data referring to objects evolved from one-dimensional lists and a two-dimensional canvas of taxonomies to displaying relations in a three-dimensional space. Finally, in the late 1960s, the term "metadata" was coined.[3]

Naturally, digital processing and machine readability and storage on punched tape, magnetic tape and modern digital storage supported and promoted the storage of increasingly vast corpora of data.

Enter: media.

The age of media, and indeed most of all the internet, changed the role of metadata. Before, metadata was helpful as a tool in administration. As a result, metadata defined the time-to-

[2] https://www.dataversity.net/a-brief-history-of-metadata/
[3] https://www.dataversity.net/a-brief-history-of-metadata/

market for all kinds of media: print, broadcasters, and the internet. Content became king, and content was accessible only by the availability of metadata.

The development of *Media Asset Management* (MAM) systems as the digital backbone of media corporations marked the final turn towards the commercialisation of content: content was considered an asset. However, the core of any MAM system is the linking of content stored within a file and metadata by a universally unique identifier (UUID).

During the 1990s, a broadcaster's archive might have reached the size of several petabytes, considered an unfathomably huge size at that time. In addition, archives depended on a wealth of metadata to make use of their content vault, and providing metadata and annotation imposed an enormous challenge to handle. Consequently, early tools for automated annotation emerged.

Metadata became the key to information; it became the tool to retrieve invisible patterns; and it helped to identify objects—and people. Business models based on metadata evolved.

Legal aspects

Without the ability to retrieve content by metadata, content is dead. But even a rich set of metadata is not sufficient.

The accessibility of content and its commercial use inevitably led to another layer of required metadata: legal information. Without the availability of data regarding the copyright holder, the value of the content is equal to zero, despite being accessible. Moreover, any person appearing in a media file's content can stop the content owner from public release for privacy reasons. While representing an additional obstacle in content use, the reinforcement of rightsholders and the right to a person's privacy cannot be lauded enough by creatives and society alike.

So, the content container is accessible by descriptive, referential, administrative, and statistical metadata. Yet, the legal metadata wrapper provides a shield. It is the point of handshake between the creator, content owner, and user. Without legal metadata, signing a valid licensing contract is not legally possible. At least, it should not.

Meanwhile, as mentioned above, the struggle for clean and validated data in the music business, with its widely scattered and varying data in countless data silos, is still continuing. While the management of exchanging metadata might be solved in the upcoming years, we face the following challenge: validation.

Vulnerability of data and truth

In the 21st century, metadata is ubiquitous in availability and utilisation. As a result, data is taken for granted, and it is no surprise that metadata is the Achilles' heel for the economy, society, and democracy.

The act of manipulating the perception of truth by metadata and media is a rather ancient one. *Damnatio memoriae* (condemnation of memory)[4] describes acts of memory sanctions like deleting or erasing names, pictures, and statues of emperors, for example in ancient

[4] https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095658990

Rome or Egypt. Later, in the 20th century, Stalin's photographic retouchings[5] intendedly generated a fake representation of reality.

Today, common technologies can provide false representations by misuse of their power. As mentioned above, Media Asset Management systems provide the backbone technology of media corporations. Apart from annotation and documentation, they offer tools for the ingest of media, administration, deriving and producing new content, play-out, and archiving. Working at one of these providers, I remember that at least as early as back in the 90s, customers started to ask for options to add a slight delay of some minutes to the broadcast of live events. This helped unwanted incidents to be stopped from being aired. What might have been a naked person then is today's symbolic gesture of Iran's national soccer team at the world championships in Qatar.[6]

Manipulated content, omitting context or content, and finally, Deep Fakes lead to [7]misinformation. It destabilises society and democracies because it is propaganda's most beloved tool. Sometimes the metadata is corrupt, and sometimes the content is.

Validated metadata, though, can help check the context and provide proof of the content's validity.

… and here goes that same old insight yet again: Whatever technology one implements is worthless if the data is wrong or inconsistent. Relying on false data can be useless at best and devastating in the worst case if someone has wilfully manipulated the data.

The key to solving the problem is two questions:

- How to keep metadata from being manipulated?
- Even more important: How to tell if metadata is true?

Unfortunately, we are becoming increasingly indifferent toward tweaking truth and shaping reality. What appears to be an innocent and light-hearted edit of a picture is, as a matter of fact, the start of manipulating the truth.

Source checking

It is the core challenge of any (meta)data model to provide a check for validity. Of course, one helpful approach is cross-checking data from comparable data sources or approving the validity based on consistency and context. Also, probability proved beneficial when there is sufficient statistical data, for example in controlling or tax payments.

Any combination of these options is a good starting point. Yet, it is required to optimise the sourcing workflow and the adding of data for one and evaluate the reliability of the sources and contributors as well.

The workflow design of annotation must support comfortably contributing correct data. For example, a creator is supposed to be the best source of data on the composition and performance of their work. But suppose creators are urged to keep notes on which analogue

[5] https://www.openculture.com/2017/08/long-before-photoshop-the-soviets-mastered-the-art-of-erasing-people-from-photographs-and-history-too.html

[6] https://ga.de/news/wirtschaft/regional/qvest-group-in-koeln-infrastruktur-fuer-medienunternehmen-weltweit_aid-84159083

This article, retrieved from the German newspaper "General-Anzeiger", is paid content in German language that describes the work of Qvest, a company that sells media production technology.

instruments are used by whom with whatever parameters. In that case, they often must disrupt the creative flow, which definitely is not intended.

Regarding reliability, it is a huge difference if someone is paid for contributing data or whether it is voluntary or altruistic work, and the latter should be preferred. What is the respective person's knowledge about the importance of the data? What about the level of education and experience? Is the provider or source of data a reliable institution?

The relevance of validity

Interestingly, there has been a notable shift of relevance for data validity. So far, whether data is valid or not has been mainly of economic interest. Now, it is the focus of society and politics.

Of course, it still is vital if a company can generate business and revenues based on given data. The importance of validity, though, is extraordinarily more relevant when it comes to manipulating elections and, ultimately, the perception of truth.

Considering the fast rise of artificial intelligence just in the last two years, more than just a change of perception has to be feared. The basis of AI is vast amounts of data that we refer to as knowledge. Whenever this data source is – wilfully or not – biased or manipulated in any way, the algorithm's output will be too.

Maybe even worse, the foundation to corrupt knowledge silently is in the making. Studies have shown that “research that is less likely to be true is cited more”[8].

Sure, the AI algorithm might have rules implemented which maybe can filter some inconsistency. However, humans have coded these same rules. Also, a knowledge that apparently has been proved in research makes fertile soil.

You may have noticed: history and insights keep repeating. It's not about the technology but the data's validity, which depends on the people providing it.

Everyone's hot for AI. That is where future money is. It can do this, and hey, even that. But if an AI is fed and tweaked by flat-earthers, good old Mother Earth might be hammered flat.

Shouldn't we intensify research for the validation of data first?

The role of a metadata score

Metadata bears considerable value. That is trivial. It is not as trivial to realise there is no standardised way to measure the value of data. Its value equals its validity.

If the draft of a scheme for validating (meta)data is successful, it is possible to define a metadata score. It is a value measuring the likeliness of the data's correctness. In linguistic terms, the value measures the probability of correctness for the proposition of a statement.

Let's assume that metadata does come with information regarding its reliability. That will instantly lead to an increase in the metadata's value. The proof of how reliable a single piece of information or an entire database is, represents a qualitative property. The proven reliability of data—and truth—awards economic value and social and ethical advantage.

[8] https://phys.org/news/2021-05-replication-crisis-true-cited.html

Threats and disadvantages

Nevertheless, the existence and availability of metadata referring to persons is a threat. People can be identified by metadata; metadata means surveillance.[9] It's a primary challenge to address. Recent approaches in anonymising metadata might help here.[10]

If a reliability score of data sources, including persons, can be calculated, it tackles the vulnerable matter of privacy additionally. Any trust-based score can only be sanctioned if it is strictly anonymised.

Also, how to tell if omitting facts is "good" or not? It is absolutely fine not to disclose your skin colour, sexual orientation or other facts when applying for a job. However, for political positions, the public should know a bit about your mindset. It is tough to draw a line here.

Furthermore, there is another essential disadvantage which—like in the field of artificial intelligence—often is ignored. It is the massive consumption of power. Any common approach to calculating a reliability score for metadata will threaten our climate.

So, if any form of reliability score for metadata is required, disadvantages and threats should be identified, addressed, solved, and eliminated first. No matter how tempting a straight approach is, the result is useless if fundamental obstacles are not addressed before working on an algorithm.

Nevertheless… How to trust the architecture of trust? How to be sure about the algorithm being non-biased and independent? Openness and transparency are essential, but how to realise their implementation and maintenance?

The conclusion

Whether it is blockchain, AI, or any other technology we should come across next, it is all about the validation of data and metadata. If we don't get this straight, the future may sport a black dress.

Data defines the matrix we live in, and data validation shows the way out. Therefore, it is no less than the prerequisite of truth.

***

I wrote this article after noticing some correlations and possible implications described here. It is a humble step towards addressing specific issues in metadata research where there is a lack of awareness. For the most part, the article relies on personal experience. Nevertheless, I hope to investigate the matter according to higher standards and in more detail in the future. If you would like to share additional insight and join the exchange of thoughts on metadata, please do so—I look forward to your input.

Contact:

Wolfgang Senges
wolfgang.senges@contentsphere.de
https://www.linkedin.com/in/senges/

[9] https://www.schneier.com/essays/archives/2014/03/metadata_surveillanc.html
[10] https://psoido.com