Report on the metadata needs of STT's clients

Author: Martin Kjellman

Date: 22 December 2020

Executive summary

Metadata, the information that describes news content, are important in many internet-era news practices. This report explores the metadata needs of Suomen Tietotoimisto's (STT) clients through in-depth interviews with managers representing five Finnish news organisations. By focusing on why and how the organisations use both metadata and content provided by STT, the report also considers STT's role in the metadata practices of its clients.

Key findings:

1. Newsroom staff have enough steps in their work practices. This means extra steps risk being poorly executed, and that the number of actions carried out by individual reporters should not be added to.

2. Keyword tagging is difficult. This means individual reporters adding keywords is not always the best practice. If tagging is done manually, the organisation needs a) clear rules for tagging and b) a CMS that allows flexible tagging.

Conclusions:

1. STT content should not create unnecessary work steps.

- Metadata should allow smooth handling of STT content.

2. STT content should contain metadata that support automation.

- Basic metadata: section, location, relevance etc.
- Also metadata for connecting pictures and text...
- … and for updating STT articles automatically.

Contents

1. Introduction

In the context of the news industry, the term metadata typically refers to “the semantic description and the automated exchange of data” (Pellegrini, 2017, p. 10). This type of metadata usage was popularised by the library and information sciences, was adopted by newspapers as early as the 1960s and has evolved significantly due to the rise of the internet as an increasingly important publishing platform (Pellegrini, 2017). Going by the most common definition, according to which metadata is “data about data” (Pomerantz, 2015, p. 19), the data being described in the news industry is the content: articles, pictures etc.. Essentially, metadata is all information about the content, such as the name of the author or photographer, length of an article or size of a picture. It can also be topical information: Is it domestic or international news? Sports or culture?

Metadata are particularly interesting for organising and automating the distribution of news content, for example different kinds of personalisation (Diakopoulos, 2019, pp. 195–196). After examining the current developments in the Finnish news industry in general and STT’s past experience with metadata in particular, topical keyword tagging and location data were singled out as the most relevant types of metadata to explore further. Metadata are immensely important in the case of visual news content, but for the sake of narrowing the scope, this report is limited to news texts.

The purpose of this report is to gain an understanding of STT clients’ metadata needs. In-depth interviews with news managers representing multiple types of STT clients (N=5) were used in order to provide a nuanced account of how metadata functions in the organisations. The scope of this report can be described as socio-technical (cf. Geels, 2004) as opposed to purely technological. This means that the report is mainly an attempt to look at metadata in a larger organisational context. By keeping in focus the clients’ relationship to STT, specifically the news content provided by STT, the report also attempts to shed some light on how STT can approach metadata.

2. Method & participants

Semi-structured interviews were deemed the most effective way to a) get detailed descriptions of how the clients use metadata, and b) explore their priorities and reasoning with regards to metadata a bit more in-depth.

In the interest of providing as broad a perspective as possible, interviews were carried out with representatives from a wide variety of clients. The interview participants therefore represent small local newspapers as well as regional newspapers and large consortia. When choosing participants, people in management positions were prioritised, since these were assumed most qualified to speak to their organisation's overall needs and practices. In one of the interviews, the participant representing a managerial role was accompanied by the head of the IT-department who provided technological viewpoints. Thus, this interview took on elements of a focus group discussion, which also proved fruitful.

The interviews followed an interview guide which was prepared by STT's media services unit, thus utilising in-house knowledge of both metadata usage in general and the specific situations of STT's clients. The interview guide consisted of two separate parts, the first one focusing on what kind of metadata the clients use and the second part focusing on how it is used (i.e. for what purposes). Each part was structured according to a few themes which were derived from how metadata is currently used in the news media industry in general.

The interviews took place between September 15th and October 27th 2020, lasting between 38 and 48 minutes. Due to Covid-19 restrictions, all of them were done over video call. Subsequent to the first interview, which effectively functioned as a pilot interview, the interview guide was revised (cf. Appendix A for the revised interview guide). One key addition was to ask the participants to talk about their content management system (CMS). This allowed them to get more particular in their accounts of their work practices, which both provided relevant context and helped them be more specific about the metadata they use.

In the transcription process, the interviews were translated into English (all interviews were conducted in Finnish). The interviews were then analysed using a phenomenological approach that included structuring the interview content into themes. Some of them, such as "Keyword tagging" and "Personalisation" were directly derived from the main themes of the interview guide. Others, such as “Types of metadata” and “Work practices”, were constructed based on what topics emerged during the interviews.

3. Results

In this section, the results of the analysis are presented. The first part summarises what metadata are being used in the organisations, the second part why they use metadata and the third part how they work with metadata. An overview of how the participants’ organisations work with metadata is provided in Appendix B.

3.1 What metadata?

Keyword tagging

In most of the organisations examined, keyword tagging is considered increasingly important. Keywords function as information that describes the topical content, i.e. what is an article about? They are mainly used to package articles that are related (for example all articles about a specific election). In a similar fashion, keywords are used to recommend more related articles. Combined with user information (either mined or provided) keywords function as the basic tools for personalisation. They also allow newsrooms to analyse more in-depth how their content performs.

The participants describe a wide variety of approaches to keyword tagging. Some of the organisations utilise a system best described as “keyword trees”, meaning that an article is described by classes of keywords. In this case, persons, companies and location are all considered keywords, but they might be classified as different types of keywords. One of the participants gave the keyword Nokia as an example: news about the city Nokia are not necessarily relevant for someone reading an article about Nokia the company. Other organisations classify to a much smaller degree or not at all, in which case keywords are basically flat text.

Although both IPTC and IAB tagging standards are either being used or considered among the organisations, no standard was essential to any of the organisations.

Keyword tagging is done either manually or semi-automatically. One of the most important interview findings is that all the participants agreed that manual keyword tagging is problematic. It requires clear and well-understood rules and adds an extra step to the work process. Both these factors cause the tagging to be poorly executed. Depending on their resources, the organisations handle this issue in a few different ways: Two organisations have introduced automation software that suggests keywords for the reporter to choose from. Both of them claim this to be very effective.

In my experience, it is incredibly difficult for humans to grasp a large collection of keywords and think about what keyword is right for this particular topic. So in a way the machine does a better job of making the basic suggestions and finding the connections than an untrained human. - Participant E

One client has doubled down on their guidelines, made them as clear as possible and made the reporters learn how to follow them. Finally, two clients have simply chosen not to make much use of the keywords since the quality of tagging is considered unreliable.

Location

In general, participants consider location data to be important. Especially the participants representing regional newspapers emphasise the importance of being able to use good location metadata.

We want to do things in the future with the data that we currently don’t even know about. The most obvious example is expressly some kind of location data for which there might be some applications in the future that we are not yet aware of. - Participant B

Among the examined organisations, location mainly functioned as a type of keyword. Some also use specific location metadata, i.e. data about location that is not necessarily used for recommendations and analytics but instead function as hidden info used for organising and handling content. For example, the regional newspapers might have a specific web page section for collecting articles about a specific municipality.

For STT, a clear use case for location metadata is comprehensive nationwide reporting, for example data-driven reporting. Headline-generation based on location was mentioned by one participant as an interesting prospect. Other types of geographical personalisation could also be viable. This is of course already being done (e.g. Lukiovertailu), but location metadata could potentially raise the degree of automation in the distribution process.

Other metadata

With the exception of the specific questions about location and keyword tagging, the semi-structured nature of the interviews allowed the participants to talk about the types of metadata they deemed to be important. They ended up recounting a wide variety of metadata relevant to their work practices. Among other things, the type of metadata that is relevant depends on the type of organisation. The larger organisations need to distribute content to multiple publications, while the smaller ones are able to cater to more specific needs of their audience, such as providing hyperlocal news. But since the larger organisations are either planning to or have already started to increase personalisation (cf. section 3.2), these distinctions are getting blurred at least in terms of distribution.

Information about section (international, politics, economy, sports etc.) and payment level (content open or behind a paywall) is metadata that is relevant to most organisations. One organisation includes metadata about article importance. This is manually determined by the newsroom in all cases except for STT articles, for which the metadata already included in the XML-file is converted automatically. Another example is metadata about the style or mood of an article. Finally, information about the planned production and publication time as well as data about content performance used for analytics can be considered planning metadata. These kinds of metadata were not subject to much in-depth discussion in the interviews. However, it is clear that such planning metadata that allows STT’s clients to handle STT content seamlessly is highly relevant. Most importantly, this includes metadata that would allow clients to automatically update STT articles on their webpage, as well as metadata that pair articles and pictures.

3.2 Why metadata?

In general, the participants' organisations use metadata either for better analytics or for automating certain processes. Keywords in particular allow the organisations deeper insights into the performance of their content. They also allow the organisations to automatically link together content. This can happen either during the production process in which case the reporter gets recommendations of older content on the same topic, or it can happen after publishing in which case it is the reader who gets recommendations. These reader recommendations are typically referred to as personalisation, but the term might refer to a variety of practices.

Personalisation

All but one of the participants stated that their organisation is looking to increase their degree of personalisation. The organisations are approaching personalisation in a number of different ways. One option is to allow the readers (or users, depending on the amount of agency they are ascribed) to choose which topics they are interested in. This content can be provided either as a unique view on the website or for example via newsletters. Another option is to recommend articles automatically based on what content readers have accessed previously. A third variation, brought up by a representative for a regional newspaper, is to personalise on content level (this can be limited to headlines or be done for the actual content of an article). In such a use case, the degree of control given to the reader might vary: personalisation can be done either using information that the reader provides actively (such as age) or using other metadata such as location. But regardless of how it is done, personalisation is a way to keep readers engaged.

The idea here is perhaps that all metadata should help us digitally do Netflix better, that is saying that if you were interested in this, you might also be interested in this. - Participant E

Not all of the participants are unreservedly positive about personalisation though. One of the participants simply considered their audience to be too small for it to be worthwhile. Another participant raised the question of editorial responsibility:

Then there is another side to personalisation, which is that we still want to make journalistic choices … We have this belief that for a newspaper it is valuable if a person can trust that if something important happens in the world, they can read about it in the newspaper.
- Participant B

The participants also brought up a few processes which, depending on how we choose to define it, might also be described as types of personalisation. One such process is targeted advertising, in which information, i.e. metadata, about the content a reader "consumes" is used to tailor the ads they receive. Another variation is to distribute ads based on the topic of the articles. Advertising is also a use case for metadata about style or mood – for example, articles with a tragic subject matter such as accidents might be exempt from ads. As far as personalisation, the organisations are generally farther along in the personalisation of advertising than they are in the personalisation of editorial content. So even though the organisations are actively pursuing personalisation, most of them are still in the planning stages.

Analytics

Analytical data are used to some extent by all of the participants' organisations. All or at least most of the aforementioned metadata can be used as analytical units. These data are combined with information about how the content is being consumed, which is also a type of metadata although one not discussed in-depth in the interviews conducted for this report. Metadata such as section info or keywords are used to draw more nuanced conclusions about the content than what could be done by looking merely at the performance of individual articles.

One usage for metadata is that we are able to combine data about how content is consumed, how our articles are being read and what has gained us subscriptions, with other metadata. This can create tools for management, so basically using data to steer the newsroom.
- Participant C

As is the case with most metadata, analytical data can be used to automate certain work practices. One such use case, which is being actively pursued by one of the larger organisations, is a dynamic paywall. Most importantly, such a solution involves varying when an article is put behind a paywall. This can be done automatically based on the performance of an article, but it can also be done semi-automatically with an algorithm doing the analysis and an editor making the final decision. The latter is a noteworthy example of a frequently recurring theme, which is automation assisting editorial staff in their decision-making (semi-automated keyword tagging being the other prominent example of such hybrid practices).

Archive and print

The print paper is still important for most of the organisations but the participants do not consider metadata to be as important for print as it is for online publishing. Still, some metadata such as information about section (domestic, sports etc.) and size are very important. Some organisations are also looking to raise the degree of automation in their print production, a process in which metadata is key.

Assembling print is time consuming as hell. The newsroom should be able to focus more on creating the best possible content. - Participant A

The participants agree that good metadata, keywords in particular, is important for efficiently using their archive. Participant C also noted that there is a retroactive aspect to archive tagging: their organisation is looking to organise their existing archival content, although this has proven quite a challenge. Currently, they do not necessarily even have very good insights into their archive, meaning that they are not making the most of it.

We have a need to classify large masses of content … If a salesman came to me and said here's a device that I can use to classify and in some automagical way tag our existing archive, I would be terribly interested. - Participant C

But the participants indicate that the way they approach both keyword tagging and archiving, hierarchical structures, such as the ones supported by for instance the IPTC standard, are losing relevance. Instead, free text search is believed to be sufficient provided that the metadata is good.

If the article has the keyword "ice hockey", you don't necessarily need to know that it belongs to the category "ball games", which in turn belongs to the category "team sports", which in turn belongs to the category "free time". In a way the structure over there in the background is not important. It's enough that you find articles about ice hockey.
- Participant E

3.3 How metadata?

Work practices

The work practices around metadata described in the interviews point towards two sides of the same coin. On the one hand, metadata demands work. Managing metadata – adding keywords, setting parameters such as whether an article is behind a paywall – might involve adding actions to the work process. On the other hand, metadata reshapes work. Since metadata are used for automation (e.g. organising content online), it has the potential of making certain processes smoother. This duality means that metadata is both desired and approached with caution. One of the participants recalled their organisation wasting time adding metadata that ended up not having clear business advantages. Another perceived danger is poorly executed metadata management, which the participants generally believe causes more harm than it does good. In the participants' experience, keyword tagging is difficult to begin with. The fact that work practices are at a point of saturation in many organisations exacerbate these issues.

Every year there is something new added to the work process ... This sits well with some but others struggle against it and experience it as very hard and time-consuming ... Perhaps for this reason we are quite careful about what we add to the process. - Participant D

The rapid influx of new work operations has to do with the long-established fact that the main focus of publishing is transitioning to the web. Publishing has become something more fluid, with real-time analytics and personalised recommendations merely the two most obvious examples. Thus, metadata are becoming integral parts of the work practices. Recycling older content is standard practice. This is done either by manually bundling together packages of articles or automatically recommending related content. But the participants' accounts of their work practices also involve reporters accessing the archives in the production phase, a process largely dependent on quality metadata.

The organisations that have transitioned to hybrid work practices, with automation assisting in for example tagging, express a rather pragmatic view. One of the main perceived benefits of automatic keyword tagging it simplifies multiple aspects of the work process. Both the participants whose organisations are using automated tagging concede that the automation does cause mishaps, some for example having to do with confusing words that might have different meanings. Still, they emphasise that the positive aspects outweigh the negative ones.

The effort-reward ratio is awfully good … It's effective. It's good enough and very effective.
- Participant E

Their experiences also indicate that hybridity is essential: any automation of the workflow needs to be flexible enough to allow manual corrections. This has to do with the aforementioned desire to keep making editorial decisions, and with the fact that automation has its limits.

Thinking about the future, we strongly desire to automate the print side as well. The more automated the better … but those skills aren't actually that easy to automate. -Participant B

The need to automate extends to STT content as well. In general, STT content is easily handled in the organisations. Some specific disconnects in how STT metadata is integrated were mentioned, but the participants did not indicate any structural issues. Online, the participants' organisations either publish STT content automatically or wish to be able to update STT articles automatically when new versions of an article are received. Being able to automatically pair articles and pictures was another desired feature. Finally, an issue related to work practices was raised by a participant representing one of the smaller organisations: in their online publishing, the organisation is reluctant towards overstuffing the web page. The reasoning behind this is that all content is competing among itself for the audience's attention, meaning that weaker (i.e. less interesting) content undermines stronger content. While this is an issue perhaps more related to the actual content, it is worth thinking about in terms of metadata: since metadata can be used to better organise and package content, metadata management potentially does affect these issues.

CMS and other limitations

All of the organisations examined in this report use multiple content management systems (CMS). Some of them differentiate between a production system and a publishing system. A rough categorisation can in most cases be done between the system used to produce the print newspaper and the web publishing system, but in some cases the web publishing systems used can function as production systems as well.

The content management systems used play a big role in how the organisations approach metadata. One of the participants directly stated that their technology does not support the use of metadata, which of course makes any metadata strategy redundant. Similar experiences were described by other participants as well: innovations that affect work practices, including metadata practices, can not be implemented until software-related updates are completed. In fact, the majority of the organisations were in the process of switching CMS at the time of the interviews.

In a similar vein, rigid or complicated technology was brought up as a limiting factor: any technology introduced needs to be easy enough to use. This is just another example of how flexibility emerged as a key concept of the interviews. The standards for keyword tagging, such as IPTC, were themselves seen as problematic in this respect.

I remember for example when we still had IPTC-classifications and people had to choose some IPTC-class. Those were quite difficult for the reporters to grasp and they just chose whatever because when you have to choose something you always just go with something like human interest. Then there is no use if the keyword tagging is done badly. - Participant B

As hinted at in the previous section, work practices are perhaps the most limiting factor of metadata usage. Some metadata can be added or adjusted automatically (e.g. automatic word or sign count as a type of size metadata), but metadata nonetheless requires its own set of practices. From the participants' experiences, it is clear that reporters often lack either time or motivation or both to add new steps to their work practices.

Somewhat paradoxically, the flexibility desired by the participants ties together with a factor that can itself be limiting: editorial control. As already implicitly stated, metadata usage is affected by the conflict between journalistic values and business values. All-out personalisation, perhaps the strongest metadata use case, was in part perceived to be at odds with journalistic decision-making. Although by no means an inherently bad phenomenon, the practice of claiming editorial independence in order to slow down innovation in news media is well-documented (Deuze, 2005; Ess, 2014).

4. Conclusions

The organisations’ relationship to their audience is changing, as demonstrated by the emphasis on analytics and personalisation. The same developments have brought on a lot of changes in their work practices. This means that metadata practices have to take into account something resembling “innovation fatigue” (Posetti, 2018, p. 7). Extra steps in the work process risk being poorly executed, and the number of actions carried out by individual reporters should not be added to. In order for metadata usage to reach high enough standards, the staff responsible for it must be a) onboard and b) educated. In other words, metadata policy has to be clear and well-integrated in the workflow.

Anyone providing these organisations with content and services needs to be sensitive to these factors. In STT’s case, any measures that can be taken to prevent adding extra elements to the work practices of the client organisations are worth considering. Metadata that enable smooth handling of STT content are therefore recommendable. In particular, clients are interested in metadata (and practices) that allow them to connect pictures and articles, and to automatically update STT articles online.

Unsurprisingly since they represent a wide variety of organisations, the participants described a range of different metadata needs. Location metadata and topical keywords were the types of metadata most interesting to all of them. Location metadata are perceived to have a lot of potential. Besides establishing that municipality is one relevant unit for location metadata, the interviews did not bring forth any other clearly relevant type of location (such as address or coordinates) metadata.

Keywords are perhaps the type of metadata with the clearest use-cases since they can be used for steering production (most notably through analytics) and distribution (e.g.

personalisation). But since the quality of tagging has to be high, it is not easy for the organisations to do it manually themselves. The ones that have experience with automation speak to its benefits. Automation is also changing the way keywords and archives are understood: detailed hierarchies of keywords are being replaced by looser but allegedly more efficient practices. Partly for these reasons, the international standards for keyword tagging appear to be losing importance among STT's customers. This does not mean that tagging standards are not relevant for STT, but it calls into question how they are applied. The IPTC keyword tagging is perhaps not indispensable, but since many clients do not have automation in place, STT still needs to have a clear strategy for keyword tagging.

One way for STT to use its position as a centralised institution is to alleviate the workload of its clients. Good metadata practices on STT's part has obvious ripple effects, since all clients benefit from this instead of all of them having to handle all aspects of metadata themselves. However, this path needs to be carefully tread. Any metadata management done by STT certainly needs to have clear benefits both for STT as an organisation (for instance with regards to its archive) and for STT's clients. Given the extremely saturated work practices of most if not all newsrooms, raising the degree of automation is always going to be one such benefit. The interviews show that STT's clients have very varying needs. They have specific ways of distributing their content, different sections, different approaches to personalisation and so on. One implication of this is that even though keywords are clearly important, they are not the be-all and end-all organising unit.

For STT, enabling personalisation could be worthwhile, since it would potentially allow organisations to make better use of STT content. As one of the participants pointed out: overstuffing one's website with online content might lead to one's content competing for attention. In a worst-case scenario, STT content starts smothering local news content. Personalisation offers a solution to this problem. An imagined generic reader of a local newspaper's website will certainly be more interested in most local content than in most STT content, but with personalisation, this dichotomy becomes less relevant. It is actually quite likely that every newspaper has a type of user profile for whom certain parts of STT's content is extremely interesting.

References

Deuze, M. 2005, "What is journalism? Professional identity and ideology of journalists reconsidered", *Journalism,* vol. 6, no. 4, pp. 442-464.

Diakopoulos, N. 2019, *Automating the news: how algorithms are rewriting the media,* Harvard University Press, Cambridge, Massachusetts.

Ess, C.M. 2014, "Editor's Introduction: Innovations in the newsroom–and beyond", *The Journal of Media Innovations,* vol. 1, no. 2, pp. 1-9.

Geels, F.W. 2004, "From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory", *Research policy,* vol. 33, no. 6-7, pp. 897-920.

Pellegrini, T. 2017, "Semantic metadata in the publishing industry–technological achievements and economic implications", *Electronic Markets,* vol. 27, no. 1, pp. 9-20.

Pomerantz, J. 2015, *Metadata,* The MIT Press, Cambridge, Massachusetts ;.

Posetti, J. 2018, *Time to step away from the 'bright, shiny things'? Towards a sustainable model of journalism innovation in an era of perpetual change*, Reuters Institute for the Study of Journalism.

Appendix A

Revised interview guide:

What does the term metadata mean to your organisation?

- What CMS do you use?
- What works well?
- What could be improved?

What kind of metadata do you use for news articles?

- What kind of classes (names, organisations etc.)?
- Keyword tagging?
- Specific standards for tagging?
- What about location info? How specific?
- Any tools for handling metadata (automation or manual)?

For what do you use metadata?

- Personalisation?
- Steering content through different sections/channels on web page?
- Analytics?
- Advertising?
- Print news?
- Archival purposes?

What kind of strategies do you have for metadata?

What do you need from STT in terms of metadata?

Appendix B

Participant	**Keyword tagging**	**Automation**	**Tagging standard**	**Location**
A	Manual - transitioning to automated.	Print production	None/IPTC	Not in use
B	Automated	Print production + paywall settings	None/IPTC	Municipality
C	Manual	-	None/IAB	Keyword class
D	Manual	-	None	As in-text lead-word in print.
E	Automated	Personalisation	None	Normal keyword

Table 1. Overview of how the participants' organisations handle keyword tagging, what is or will likely become automated, what tagging standards are in use, and how they handle location data.

Participant	Other metadata	Personalisation	STT content	Future plans
A	-	Not in use	Interested in content personalised for regions or municipalities.	Personalisation
B	Priority, style, sections.	Manual choices made by readers and editors.	Automatically tagging of STT content works but needs to be improved further.	Dynamic paywall
C	Payment level, section, topics, persons.	Based on user profile/type	Metadata provided by STT is underused.	Management through data (analytics).
D	Analytical, paywall, section.	Not in use	STT content auto-published (need to differentiate between articles with or without pictures).	Personalisation (not urgent)
E	Planning data (where/when something is to be published).	Front page, recommendations etc.	Need to update STT articles automatically.	Auto-connect content in production-phase.

Table 2. Overview of what other metadata (except keywords and location) were brought up during the interviews, how the organisations approach personalisation, (some of) their specific desires with regards to STT content, and future plans.