JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12777

JLIS.it

New Challenges in Metadata Management between Publishers and Libraries

Pietro Attanasio[(a)]

a) Associazione Italiana Editori, http://orcid.org/0000-0001-7410-6682

Contact: Piero Attanasio, piero.attanasio@aie.it
Received: 25 August 2021; **Accepted:** 12 September 2021; **First Published:** 15 January 2022

ABSTRACT

Identifiers, bibliographic metadata, thematic category schemes are at the heart of the functioning of the book supply chain. There are international standards for all these elements, which allowed e-commerce to develop in the book trade before any other sector.

The dialogue on metadata management between the book industry and the library community is not always as intensive as desirable. The challenges that the whole book world must cope with today and in the near future pressure us into change. Building on lessons learned from the past, the article focuses on some upcoming challenges, such as big data and artificial intelligence applications, with the aim of identifying fields for a future collaboration.

KEYWORDS

Metadata; Publishing; Big data; Artificial intelligence.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12777

Introduction

The article focuses on metadata management from the publishing industry point of view, which is slightly different from that of the library community. In the first part I introduce the work made in this field by the Italian publishers association (AIE) and describe our approach in order to identify the reasons why the library approach is different. This is a prerequisite to setting a strategy to bridge the gap between the two.
In the second part I focus on the factors that today are disrupting the traditional context, which are related to the entrance of new players in the book sector and to the impact of big data (vs. metadata) and artificial intelligence.
I conclude that the changes that are occurring call both the book industry and the library community to build a new alliance for a fair and open book data management, starting from some core principles that we share, notwithstanding the differences between commercial purposes and the public sector mission, which will remain.

Publishers approach to book data

The Associazione Italiana Editori (AIE, the Italian publishers association), besides being a trade association representing the Italian publishers' interests at a national and international level, is characterized by a peculiarity which is probably unique: we have a research and development team within the association that is primarily engaged in the fields of book standards and metadata. We develop technologies in these areas, with particular attention – in the last 10 years – to the management of rights metadata, in line with the principle of the Copyright infrastructure launched by the European Council in 2019 and then indicated by the European Commission in relation to the European recovery and resilience plan. The AIE R&D team has been coordinating important European initiatives in the field, such as ARROW – dedicated to the management of rights metadata in digital library initiatives – and the more trade oriented ARDITO.
Linked to this experience, AIE representatives have been and still are in the governance bodies of standard setting organisations such as EDItEUR, ISBN International Agency, IDF (International DOI Foundation), W3C Digital Publishing Business Group, and EDR-Lab (European Digital Reading Lab).
According to our approach, metadata originate from events. Therefore, we place the "event" – rather than the "document" – at the core of our metadata analysis[1]. In this view, metadata start existing before a book is published (or, in general, before any document is produced). The first event to be considered is: "author *A* creates the work *W*", which is relevant even before publishing that work in the form of a document. Such an event originates the need for:

[1] This is the ontological difference between the <in*d*ecs> data model and the FRBR. See Rust and Bide (2000), in particular chapter 4.3. "The Commerce View", where the role of the events is described in the terms used in this article. In the FRBR model the *event* is instead one of the "entities that serve as the subjects of intellectual or artistic endeavour". Cf. IFLA (1997).

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12777

a) Uniquely identifying A and W, e.g. with an ISTC[2] and an ISNI;
b) Metadata for describing A and W;
c) A qualifier to identify the relation between A and W: in this case: "A is the author of W".

The second event in the typical life of a literary work is "*A* assigns publication rights *PR* in *W* to publisher *P*", which generates similar metadata needs, i.e. identification and description for the assigned rights and the publisher. Every following event generates needs for new metadata for further editions, translations, transposition for cinema or theatre etc.

More in general, these events can be described as "People make stuff" in the first case, and "People do deals about stuff"[3], in the second case.

Saying that metadata *originate* from events does not mean that metadata are directly *generated* by the events. A common definition of metadata is "An item of metadata is a relationship that someone claims to exist between two entities", which emphasises that there is a level of discretion in making that claim, and thus "the identification of the person making the claim is as significant as the identification of any other entity" (Rust and Bide 2000).

Since metadata are "claims", the objective of the claimer is as important as the nature of the relationships that are described. To understand differences and similarities between the approach to metadata of publishers and that of librarians, it is useful to look at the purposes of the "claimers" in the two cases.

The first purpose in metadata management in the industry is to increase the efficiency in the supply chain. Typically, an important metadata item in our world is the weight of the book, a crucial piece of information to maximize the efficiency of logistics. But the main data items that make a difference between a books-in-print database in a specific country and – for example – the national bibliography in that same country are the book price and its availability (P&A). This little difference (it is a matter of few metadata items) creates a big distance in the management of the two catalogues. P&A data are subject to change over time, which does not happen for other metadata[4], and this implies that a books-in-print database must manage changes in the existing records on a daily basis, whilst the national bibliography is enriched with new titles but the existing records change rarely.

If the need to serve the supply chain determines a big difference, improving the discoverability of books is the main objective that the two communities have in common. Both the industry and libraries need to assist their clients (book buyers or library users) by facilitating as much as possible how they look for and find books. Books-in-print databases and library OPACs shared this pur-

[2] The International Standard Text-work Code (ISTC) was the ISO standard to uniquely identify text-based work. Because of very limited use by the industry the standard has been recently withdrawn, though the need for identifying text-works remain. The International Standard Name Identifier (ISNI) is the ISO standard for identifying contributors to creative works and those active in their distribution. See https://isni.org.

[3] Rust and Bide (2000), p. 4. See also Paskin (2006).

[4] Since metadata are "claims" about a relationship, all metadata are not written in stone: a claim may change if there was a mistake or if there is a change in the way claims are expressed in a standard metadata language. In the case of P&A, however, there are continuously new events that originate new relationships and thus the need for new metadata. Prices may change from time to time, and availability changes continuously, both at manifestation and at work level. When dealing with digital library programmes, the metadata element "the work W is out of commerce" is very important and in the EU carries important juridical consequences, after the approval of Directive 790/2019.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12777

pose since the origins, back in the Seventies. With the advent of the Web this aspect became even more crucial in any service provided to readers. In both communities the awareness on the importance of quality and richness of descriptive metadata grew in last 25 years. The Internet made the role of metadata in search engines crystal clear: to improve discoverability and to provide data to readers to allow them to make informed decisions. In spite of this, there are still differences in one crucial aspect related to discoverability: the subject classification scheme. In particular, in my opinion, the library world did not pay a desirable level of attention to the big effort of the industry to build Thema[5].

The third purpose for metadata is to elaborate statistics about the use of books. In the language I am using, metadata serve to build data about the third kind of events cited in the <in*d*ecs> model, when "People use Works", i.e. when a person buys, or borrows or makes any use or re-use of a book. Statistics are useful to make decisions both for publishers and librarians. The difference, here, is in the perception of the value of a standard vocabulary. Since sales data are produced further down in the supply chain, publishers need standard ways to collect them. Conversely, any library produces data from its users directly, and standardisation is needed only for comparisons with other libraries. This has created more standardisation needs in the trade than in the library world.

The disruption: from metadata to big-data

Metadata, in the traditional meaning understood by publishers and librarians, played an important role in the first phase of the Internet. In mid-Nineties, the book sector was the only sector that had databases containing standard identification and rich description of millions of items, ready to be posted on the Internet, and standard messaging for tele-ordering. This was the reason why e-commerce was developed for selling books before any other good or service. Similarly, library OPACs were the first public service transferred online, in the same years.

The context was disrupted by the (so-called) Web 2.0, i.e. when the Internet started to be characterised by the meta-intermediation of web platforms on one side and user-generated content on the other side[6]. Tracking events of the kind "people-use-stuff" opened a completely different scenario.

Let me start from one specific event:

> (A) Reader *R buys* books *B1*, *B2* and *B3* in bookshop *BS*

Such a simple event generates a number of data:

- The relation "buy" between R and each of the 3 books;
- The relation among the 3 books due to the circumstance that they were bought during the same event;

[5] See <https://ns.editeur.org/thema/en>. A short illustration of the origin, purpose and main characteristics of Thema is in Bell and Saynor, 2018.

[6] The evolution between the two phases is well narrated by Foer, 2018. A brilliant - though not rigorous, from a scientific point of view - description of the same evolution is in Lanier, 2011 and 2019 and in many posts of the same author here: www.jaronlanier.com.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12777

- The relation between the 3 books and BS;
- The relation between R and BS.

The two people (the natural person R and the legal person BS) and the 3 manifestations are (or could be) described by metadata, which per se multiply the relationships between the entities. E.g.: if R is 28, a graduate, an Italian citizen, living in Rome, etc.; this creates a relation between all the metadata items of R and the 3 books, and all the metadata associated with each of the 3 books (e.g. all the 3 books are crime novels).

- These metadata may be registered in different sources:
- R may have a BS fidelity card where that information is registered;
- The books' metadata are in a books-in-print database;
- BS metadata are in the database of the Italian bookshops.

Later on, R borrows a book from the public library L, where he/she is registered with another data-set. Then R posts a comment on social media SM about one of those books…

Collecting data of this sort is not new. It is the basis of any statistic on reading, to estimate, for example, how much young, well-educated Italians like reading crime novels. Which was usually done by interviewing a sample of readers.

The disruption lies in the fact that machines are now able to track millions of similar events and the current computing power and memory allow to elaborate all the generated data through powerful algorithms. In principle this allows to collect data about events involving millions of readers that buy or borrow books and post comments etc. All in all, we have billions of data generated by events that machines are able to track.

Combining human intelligence and professional skills with good algorithms, such big data would enable publishers to design outstanding editorial plans and marketing strategies, and librarians to have the perfect collection and reading promotion strategies for their patrons.

Are we still speaking about metadata? If we consider the <in*d*ecs> definition above ("An item of metadata is a relationship that someone claims to exist between two entities") we can easy appreciate the difference: in registering the events here described we have not "someone claiming": it is a matter of machines registering events and extracting data from the events, usually according a pre-defined model[7].

Opportunity or threat?

Machines are able to track any event in our life. Tracking what we read is a very delicate issue, since it involves our thoughts, our lifestyle, our opinions and thus our fundamental rights of freedom of thought and expression. The issue should be treated with all possible care.

In the examples above, R participated in events that produced data which were then controlled by a bookshop, a library and a social media platform (BS, L and SM), each independent from each other. Only R has all the information about the whole picture, and legislation limits the possibility of BS, L and SM to exchange (personal) data about R.

[7] Machines may also produce metadata as defined in the <in*d*ecs> model. There is extensive literature about the automatic extraction of metadata (keywords, subject, etc.) from texts. See, for example, the recent Li 2021, useful also for the reference list. In this case there is "someone claiming": it is the machine, with the algorithm or, better, the person who runs the machine for that purpose.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12777

At the same time, data have an economic value, and determine more and more market power. When R buys all books from one Internet shop, together with many other goods, and posts reviews of the books in the same shop, and uses the cloud services and the platform of the same company for audiobooks, e-books and videos, etc., that single company acquires information and know-how that other competitors can never reach. Data control is a key driver to market power in the digital economy, as is also recognised by the proposal for a Regulation on Contestable and fair markets in the digital sector (known as DMA – Digital Markets Act), which emphasises the presence of "data driven advantages" (Recital 2), the existence of barriers-to-entry generated by data control (Rec. 3), stressing the "potential advantages in terms of accumulation of data, thereby raising barriers to entry" (Rec. 36)[8].

The reasons why data are so relevant in digital markets are well explained by the literature. "The quintessential task of many digital platforms is that of making predictions of various sorts (…) Data is the oil that powers these predictions" (Calvano and Polo, 2020). The more data they accumulate the better their knowledge of the market and the distance with competitors becomes. "Platforms can use this information asymmetry to facilitate interaction and increase welfare for users. These data externalities attract users to the platform" (Martens, 2020) triggering a circle: "The collection and use of big user data enables [platforms] to continuously improve the quality of their offerings" (Fast et al., 2021) creating network effects that "may result in monopolistic market power of platforms which they can use for their own benefit, at the expense of users" (Martens 2020).

This market evolution calls for new regulations, to better protect personal data and to ensure a level-playing field in digital markets, but this is out of the scope of this article. Here I would like to call for more collaboration within the book value chain, involving publishers, booksellers and librarians.

We, in the book community, share some key objectives. We all aim at better understanding readers' needs to offer them the best content and services. We also share some fundamental values: the respect for personal data and – above all – freedom of expression and pluralism, which, in market terms, also means fair competition and absence of monopolistic positions.

Because we share goals and values, we need to design a context where cooperation will enable all citizens and SMEs to access relevant information and intelligence derived from book-related data sets (i) at fair conditions, (ii) while respecting personal data (iii) and commercial confidentiality.

Technologies offer opportunities besides threats. The potential offered by artificial intelligence and data analysis can be exploited by the cultural sector too. In a market where network effects give immense advantage to few, cooperation among many can be the answer.

[8] Issue related to data exclusivity by the market "gatekeepers" are also enlightened in Recitals 43-45, 54-56 and 61. See European Commission 2020-b.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12777

References

Bell G. and Saynor G. (2018), *Thema: the Subject Category Scheme for a Global Book Trade*, Editeur: https://www.editeur.org/files/Thema/20180426%20Thema%20briefing.pdf.

Calvano E. and Polo M. (2020) Market Power, Competition and Innovation in Digital Markets: A Survey, *Information Economics and Policy*, Vol. 54, 100853, https://doi.org/10.1016/j.infoecopol.2020.100853.

European Commission (2020-a) *Making the Most of the EU's Innovative Potential. An Intellectual Property Action Plan to Support the EU's Recovery and Resilience*, COM(2020) 760 Final, 25 Nov 2020.

European Commission (2020-b), Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act), Brussels, COM (2020) 842 final, 15 Dec 2020.

European Council (2019) *Developing the Copyright Infrastructure - Stocktaking of work and progress under the Finnish Presidency*, https://data.consilium.europa.eu/doc/document/ST-15016-2019-INIT/en/pdf.

Fast V., Schnurr D. and Wohlfarth M. (2021), Regulation of Data-driven Market Power in the Digital Economy: Business Value Creation and Competitive Advantages from Big Data (January 31, 2021). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3759664.

Foer F. (2018), *World Without Mind: The Existential Threat of Big Tech*, Penguin Putnam.

IFLA International Federation of Library Associations (1997), *Functional Requirements for Bibliographic Records*. https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

Krämer J. and Wohlfarth M. (2018), Market Power, Regulatory Convergence, and the Role of Data in Digital Market, *Telecommunications Policy* Vol. 42, pp. 154–171. https://doi.org/10.1016/j.telpol.2017.10.004.

Lanier J. (2011), *You Are Not a Gadget: A Manifesto*, Penguin Books.

Lanier J. (2019), *Ten Arguments for Deleting Your Social Media Accounts Right Now*, Vintage Publishing.

Li J. (2021), A Comparative Study of Keyword Extraction Algorithms for English Texts, *Journal of Intelligent Systems*,9 July 2021. https://doi.org/10.1515/jisys-2021-0040.

B. Martens (2020), *An Economic Perspective on Data and Platform Market Power*, JRC Digital Economy Working Paper 2020-09.

Mazzucchi P. (2021), Copyright Infrastructure: l'innovazione che fa bene alla cultura del Paese, Agenda Digitale, 9 Jun 2021.

Paskin N. (2006), Interoperability. A Report on Two Recent ISO Activities, *D-Lib Magazine*, Vol. 12, No. 4.

Rust G. and Bide M. (2000), The indecs Framework - Principles, Model and Data Dictionary. https://www.doi.org/factsheets/indecs_factsheet.html.

Vuopala A. (2021), "Copyright Infrastructure. A Recipe for Recovery and Resilience of the Creative Sectors", IPR Info, n. 2 / 2021.