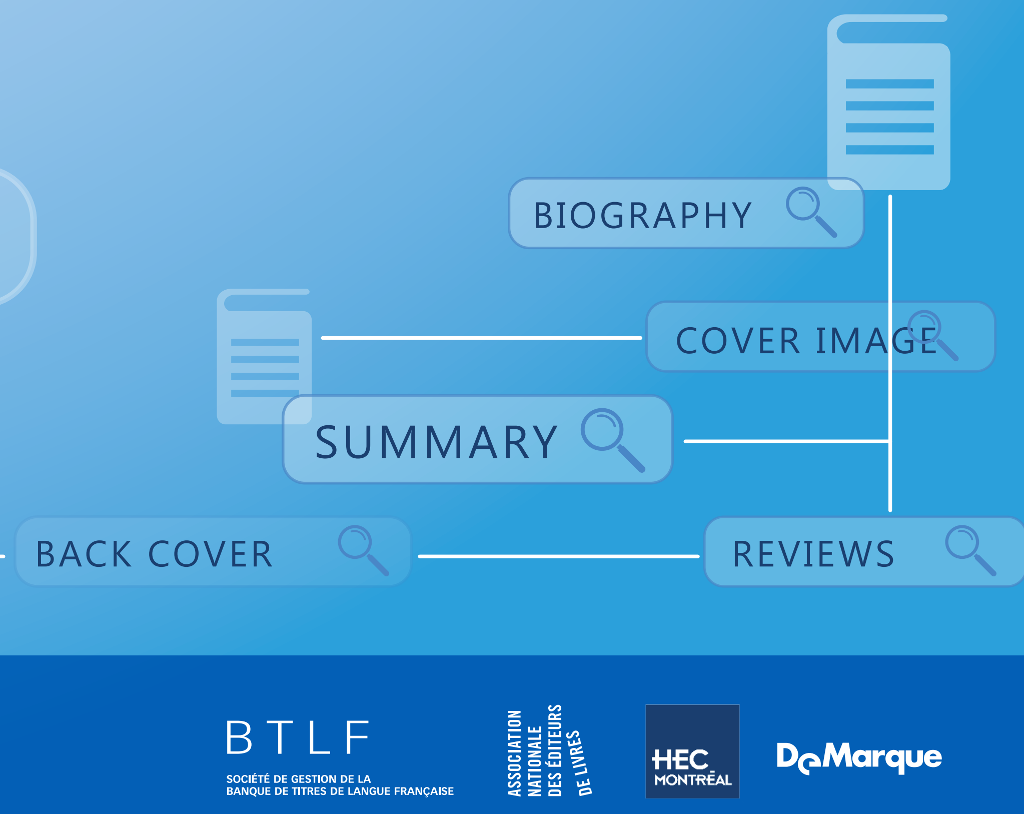
The Effect of METADATA on Book Sales

Analysis of the Relationship Between Enriched Metadata and Book Sales

A study conducted by **La Société de gestion de la BTLF** in partnership with **L'Association nationale des éditeurs de livres** and **HEC Montréal**'s Professorship en données massives pour les arts et la culture

September 2020



Credits

Analyst

Yves Leblond, Lecturer and Senior Analyst, Département de sciences de la décision, HEC Montréal

Edition

Yves Leblond, Lecturer and Senior Analyst, Département de sciences de la décision, HEC Montréal

Renaud Legoux, Professor, Professorship en données massives pour les arts et la culture, HEC Montréal

Data collection and preparation

Michel Cervellin, IT Director, BTLF

Marco Omiccioli, Operations Manager for Gaspard, BTLF

Mélissa Haquenne, Product Expert, Cantook Hub, De Marque

Coordination

Isabelle Gaudet-Labine, Project Manager, BTLF

Supervision

Patrick Joly, Executive Director, BTLF

Administration

Eveline Favretti, Project Manager, ANEL

Serge Grenier, Accountant, BTLF

Revision

Isabelle Gaudet-Labine, Project Manager, BTLF

Coralie Piotr, Executive Assistant, BTLF

Nicolas Montagne, Translator

Design and layout

Nathalie Duperré, La Boîte de Pandore

Acknowledgments

La Société de gestion de la BTLF would like to thank L'Association nationale des éditeurs de livres for its collaboration in the project's conception and administration.

La Société de gestion de la BTLF and its collaborators would also like to acknowledge the support of The Canada Council for the Arts.

Canada Council for the Arts | Conseil des arts du Canada

Table of Contents

Context

Bibliographic, enriched and commercial metadata linked to a book are essential to its visibility and discoverability. Several book industry participants must cooperate to produce and disseminate that metadata. In order to minimize the gap between the best practices and metadata as it is issued and transmitted, it is important for La Société de gestion de la Banque de titres de langue française (BTLF) and L'Association nationale des éditeurs de livres (ANEL) to collaborate in the development of tools allowing industry participants to better work together.

Therefore, thanks to a grant from The Canada Council for the Arts' Digital Strategy Fund, BTLF, in partnership with ANEL and HEC Montréal's Professorship en données massives pour les arts et la culture, initiated the development of a toolkit to analyze metadata that is aggregated by BTLF, intended for publishers, distributors, sales representatives, booksellers, librarians and authors. The study *The Effect of Metadata on Book Sales – Analysis of the Relationship Between Enriched Metadata and Book Sales* is the first component of this kit.

This analysis was made possible thanks to the two information systems managed by BTLF: Memento and Gaspard. The Memento database contains a set of metadata about French-language books published in Canada while Gaspard allows a book's economic performance to be measured. De Marque's involvement made it possible to apply the analysis to digital books.

Until now, the sector did not have a reference that could demonstrate in a consistent manner the impact of the book industry participants' work regarding metadata. This study concretely measures the effects enriched metadata production, transmitted prior to the publication date, has on book sales. Publishers, but also sales representatives, will notice the importance of integrating metadata production into their business strategies. We hope that these new insights encourage the emergence of better practices for the benefit of publishers, creators and all the stakeholders of the book industry.

Summary

This document presents a statistical analysis of the quantity of books sold at BTLF-registered points of sale in the presence or absence of specific types of metadata. The objective is to measure the effect of metadata on book sales. A first analysis focuses on print titles published by Quebec publishers[1] between January 1st, 2016 and May 19, 2018. The chosen method is the use of statistical tests of difference on sales with and without metadata by controlling[2] the impact of the number of titles published annually per publisher and the number of copies sold in the past per author.

The general conclusion of this analysis is that the sales of titles with enriched metadata are significantly higher than the sales of similar titles without it. Specifically, results show that the effect of metadata is particularly strong on publishers with medium and high annual publishing volumes. The effect is also more noticeable for authors considered medium and top-sellers (authors whose recorded sales are considered medium and high according to the categories presented at pages 8 and 9 of the study).

The metadata that seems to have the greatest effect is the cover image, particularly for medium and high-volume publishers and for medium and top-selling authors. The cover image has a stronger effect when it is added and released prior to the publication date rather than after it. However, this *prior to the date* effect is mostly noticeable for authors that usually sell fewer copies.

For the other types of metadata, the addition of a summary alone, without a cover image, has no significant effect on sales but its addition combined with the presence of a cover image significantly increases sales for top-selling authors and medium and high-volume publishers.

For medium and high-volume publishers, the average copies sold is significantly higher when four of the five types of metadata are available compared to the average observed when only the cover image is present.

Lastly, a similar analysis has been conducted on a smaller scale for digital books. As is the case for print books, it is observed that the average copies sold per title is higher when metadata is provided prior to the publication date rather than after it, for medium and high-volume publishers.

1 In this study, ''Quebec publisher" refers to a publishing house located in Quebec that publishes and sells in the Quebec market French-language titles under one or several brands. French-language book publishers from outside Quebec (Canada) are also included in this study. In Gaspard, each brand is counted as a separate publisher.

2 Statistical Control: ''Statistical techniques for excluding the influence of specified variables in an analysis." Oxford University Press, https://www.encyclopedia.com/social-sciences/dictionaries-thesauruses-pictures-and-press-releases/statistical-control (Retrieved on September 12, 2020)

Introduction

This report demonstrates the effect of the presence of metadata on French-language book sales in bookstores and other BTLF-registered points of sale[3]. For this purpose, BTLF provided a database on titles from Quebec publishers with at least one recorded sale between September 1st, 2015 and May 15, 2019, a total of 19 558 titles.

We have adopted the definition of metadata from the Grand dictionnaire terminologique de l'Office québécois de la langue française : « *Ensemble structuré de données accompagnant un ouvrage et servant notamment à en décrire le contenu et le format, à assurer son indexation dans les moteurs de recherche et les bases de données, et à faciliter la gestion des droits d'auteur qui y sont liés.* » (Structured set of data, linked to a book, notably used to describe its content and format, to ensure indexing in search engines and databases and to ease copyright management). This definition was also used in OCCQ's 2017 report *État des lieux sur les métadonnées relatives aux contenus culturels*.

The types of metadata considered in this study are the following: the presence or absence of a summary, a back cover, a cover image, a biography and reviews. The date of appearance of metadata is provided for back covers, cover images and reviews, in order to measure the effect on sales of the availability of this data prior to the publication date. A similar analysis on a smaller scale has been conducted for digital books.

3 Through Gaspard, its sales information tool, BTLF collects sales data for approximately 60% of the Canadian French-language book market. Registered retailers include independent bookstores, school cooperatives, bookstore chains and big-box stores.

1. Methodology

1.1 Print Books

Data related to bibliographical metadata is sourced from BTLF's Memento online catalogue of French-language publishing products, while sales data comes from Gaspard. Titles included in the analysis are those published by a Quebec publisher[4] between January 1st, 2016 and May 19, 2018 with at least one recorded sale. A total of 12 852 titles were retained for analysis. For each year studied, the number corresponds, for publishers, to the number of titles published during the preceding year and, for authors, to the number of copies sold from 2010 up to the year studied.

For each title, the variables being considered are the presence or absence of each of the five types of meta-data (summary, back cover, cover image, biography and reviews), the date of appearance of the cover image and the number of copies sold (retail and collectivity) during the first year following the publication date.

In this study report, ''title without metadata" means that none of the five types of metadata are present. Also, in order to isolate, as much as possible, the effect of metadata on sales, the number of copies sold in the past per author and the number of titles published annually per publisher were considered. Because these two variables have a clear impact on the quantity of copies sold and because practices relative to metadata are influenced by the type of author and publisher, it was imperative to control their effect in the analysis.

For this purpose, authors and publishers were categorized. For publishers, it is based on the distribution of the number of titles published during the year prior to the year studied. Thus, for titles published in 2017, publishers with three titles or less published in 2016 belong to the first quartile of this distribution and are labeled as ''small-volume publishers". Those with four to twenty titles published are between the first and third quartile[5] and are labeled as ''medium-volume publishers". The others, with over twenty titles published in 2016, are considered ''high-volume publishers".

A similar classification has been done with authors, based on the number of copies sold since 2010 up to the year studied. Thus, for the year 2017, authors considered ''low-sellers" are those with 0 to 302 copies sold between 2010 and 2016; ''medium-sellers" are those with 303 to 6526 copies sold and ''top-sellers" being those with more than 6526 copies sold. Note that ''low-seller" and ''small-volume publisher" labels were also assigned to titles for which the author or publisher had no titles published in the past.

It is important to keep in mind that the labels ''low-selling", ''medium-selling" and ''top-selling" for auth-ors and ''small-volume", ''medium-volume" and ''high-volume" for publishers are solely used for statistical grouping and do not represent value judgements of any kind.

4 French-language publishers located outside Quebec (Canada) with a minimum of one sale during the studied period at a BTLF-registered point of sale are also included in our data.

5 A quartile is one of the three values that divide the sorted data in four equal parts.

For the purpose of this analysis, each title has been assigned a type of publisher and a type of author.

For instance, for titles published in 2017, the categorization for publishers and authors is presented in Table 1 while Table 2 and Table 3 show different statistics related to each type of publishers and authors. Note that, in both cases, the percentage of titles with at least one type of metadata increases according to the type of publisher or author.

Table 1

Definition of the Types of Publishers and Authors for Titles Published in 2017

	Low/Small	Medium	Top/High
Author (nb of copies sold 2010-2016)	0 to 302	303 to 6 526	more than 6 526
Publisher (nb of titles published in 2016)	3 or fewer	4 to 20	more than 20

Table 2

Statistics Related to the Types of Publishers for Titles Published Between January 1st, 2016 and May 19, 2018

Type of publisher	Nb of publishers	Avg quantity sold per title	Nb of titles published	% of titles with at least one of 5 types of metadata
Small-volume	389	274	1 423	76%
Medium-volume	178	461	3 349	82%
High-volume	80	720	8 084	90%

Table 3

Statistics Related to the Types of Authors for Titles Published Between January 1st, 2016 and May 19, 2018

Type of author	Nb of authors	Avg quantity sold per title	Nb of titles published	% of titles with at least one of 5 types of metadata
Low-seller	3 710	409	6 545	84%
Medium-seller	2 285	415	3 357	85%
Top-seller	809	1 246	2 954	94%

Instead of commenting on averages in a strictly descriptive manner, statistical tests have been carried out to determine if the difference between the average number of copies sold per title with and without metadata (or depending on the number of types of metadata present) is significant. The statistical significance was evaluated with the help of a statistical test, the Satterthwaite t-test[6]. The significance threshold was set at 5%, which is the standard in this type of studies. However, the fact that a difference is not considered to be statistically significant does not necessarily mean that there is no difference, only that it cannot be demonstrated with the available data. This lack of significance may be due to too much variability in sales and/or the sample size being too small to detect a difference.

1.2 Digital Books

For digital books, the data used was provided by De Marque, an aggregator, distributor and publisher of digital content. Titles included are those of digital format published by a Canadian French-language publisher between January 1st, 2016 and May 19, 2018. Types of metadata taken into account are the same that were used for print books. Sales for these titles were compiled by format (mainly PDF and ePub) for the twelve months following their publication date. However, no analysis of the effect of the presence or absence of the five types of metadata has been carried out considering the fact that metadata is available for the vast majority of digital titles. On the other hand, information related to the date of availability of metadata made possible to analyse the effect of the presence of metadata prior to the publication date (rather than after it) on sales.

This analysis was carried out per type of publisher. These types were defined by the same methodology used for print books.

6 Two-sample statistical test used to test the hypothesis that two populations of different sizes have equal means.

2. Results for Print Books

2.1 Presence of Metadata

Table 4 outlines the presence of metadata for the 12 852 titles analysed. The results show that 13.6% of these titles have none of the five types of metadata considered. This percentage decreases according to the publisher's volume: the greater the number of titles published annually, the lesser the number of titles without any of the five types of metadata. While almost one out of four titles had none of the five types considered for small-volume publishers, the ratio is one out of ten for titles published by high-volume publishers.

Table 4

Percentage of Titles with Metadata per Type of Publisher

Type of publisher	No metadata	Cover image and summary only	Cover image w/o summary only	Cover image, summary and others	Other types of metadata w/o cover image	Total
Small-volume	24.0%	12.4%	9.6%	45.8%	8.2%	100%
Medium-volume	18.0%	9.7%	6.3%	61.9%	4.1%	100%
High-volume	9.9%	18.8%	6.4%	58.2%	6.6%	100%
Total	13.6%	15.8%	6.7%	57.8%	6.1%	100%

We can also observe the strong representation of the cover image in metadata. Indeed, the cover image is present in slightly over 80% of titles overall. Lastly, almost 58% of titles have a cover image, a summary AND at least one of the other three types of metadata considered (reviews, biography or back cover).

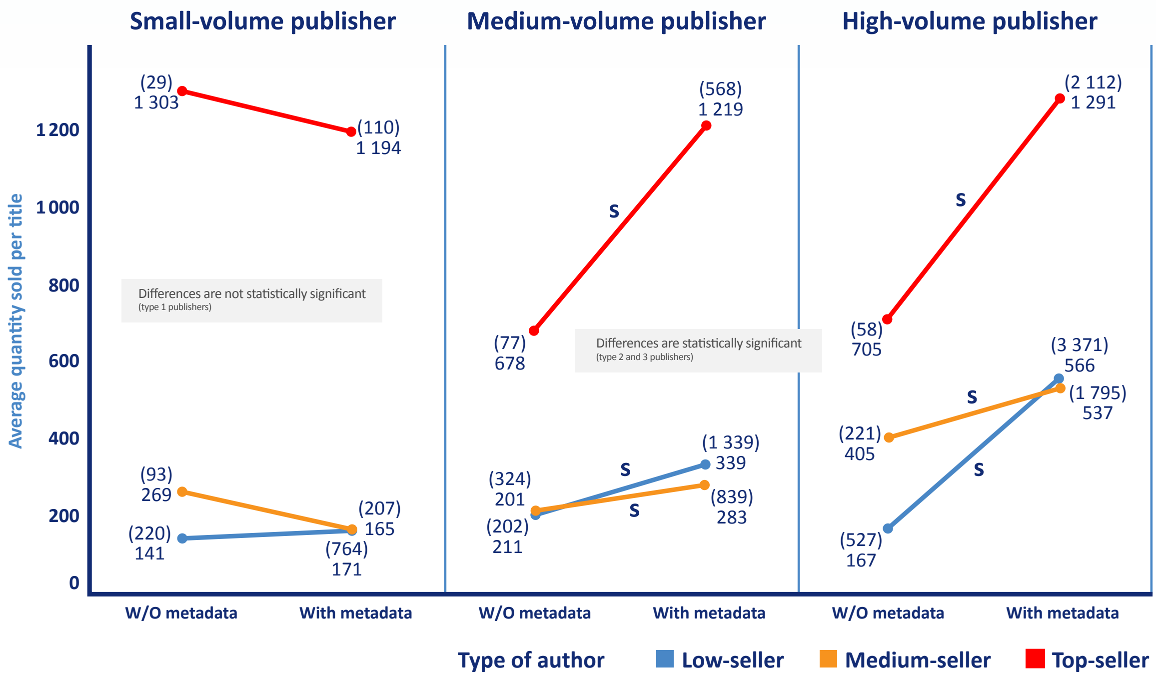
2.2 The Effect of the Presence of at Least One Type of Metadata

A first analysis consists in the comparison, according to the type of publisher and the type of author, between the number of copies sold for titles with at least one type of metadata and the number of copies sold for titles without any metadata. Results in Graph 1 show that there is, for medium and high-volume publishers, a significant positive difference (t-test under the 5% threshold) in the number of titles sold with at least one type of metadata, regardless of the type of author. For instance, for titles published by high-volume publishers coming from low-selling authors, the 3 371 titles with at least one type of metadata sold an average of 566 copies during the first year following their publication compared to an average of 167 copies for the 527 titles without metadata. For small-volume publishers, the differences are not statistically significant.

In the graphs, the straight lines annotated with the letter "S" indicate a statistically significant difference. Straight lines without annotation indicate that differences are not considered statistically significant[7].

Graph 1

Average Quantity of Copies Sold per Title With at Least One Type of Metadata and Without Metadata, per Type of Author and Type of Publisher*



* *Bracketed figures relate to the number of titles processed overall while figures not between brackets relate to the average quantity sold per title.*

7 The fact that a difference is not considered to be statistically significant does not necessarily mean that there is no difference, only that it cannot be demonstrated with the available data. This lack of significance may be due to too much variability in sales and/or the sample size being too small to detect a difference.

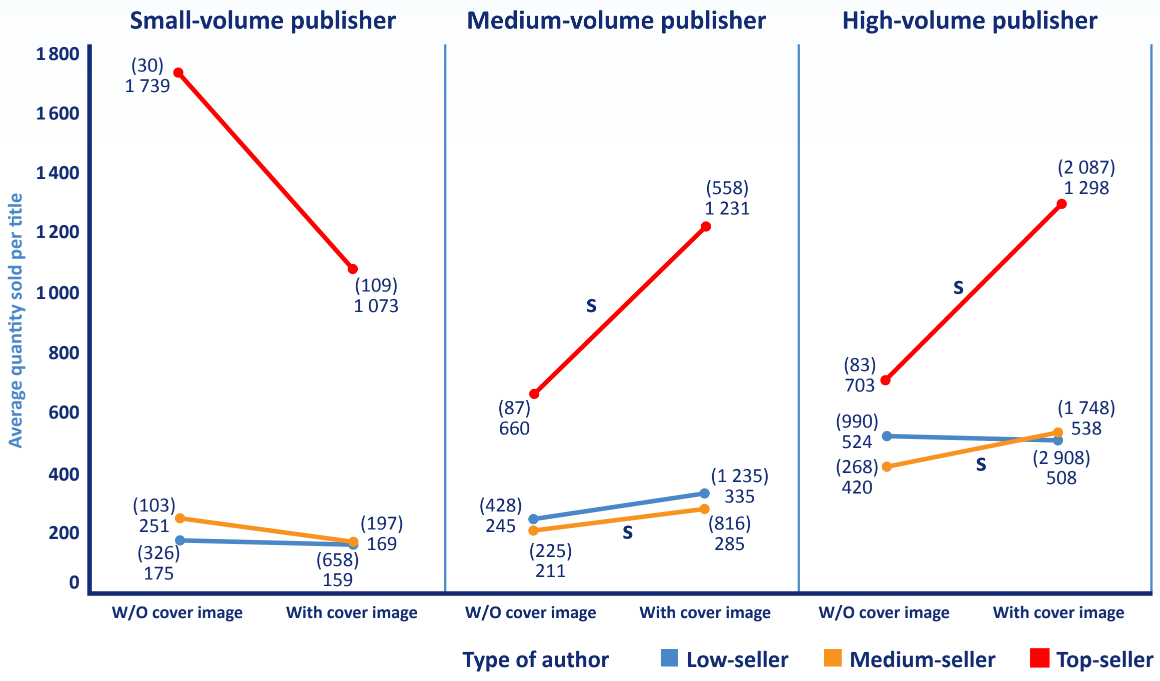
2.3 The Effect of the Cover Image

Given the important presence of the cover image as a type of metadata (over 80% of titles), the average quantity of copies sold with a cover image was compared to the average quantity sold without it. Results in Graph 2 indicate that, for medium and high-volume publishers and medium and top-selling authors, the average quantity of copies sold during the first year following the publication is significantly higher for titles with a cover image.

Note that the percentage difference between the average quantity of copies sold with and without a cover image is higher for top-selling authors, with an increase of almost 100% in sales. The effect is of lesser importance, while still being significant, for titles from medium-selling authors (average of 285 copies sold with a cover image vs 211 without a cover image for medium-volume publishers, compared to an average of 538 copies sold with a cover image vs 420 without a cover image for high-volume publishers).

Graph 2

Average Quantity of Copies Sold per Title With and Without a Cover Image, per Type of Author and Type of Publisher



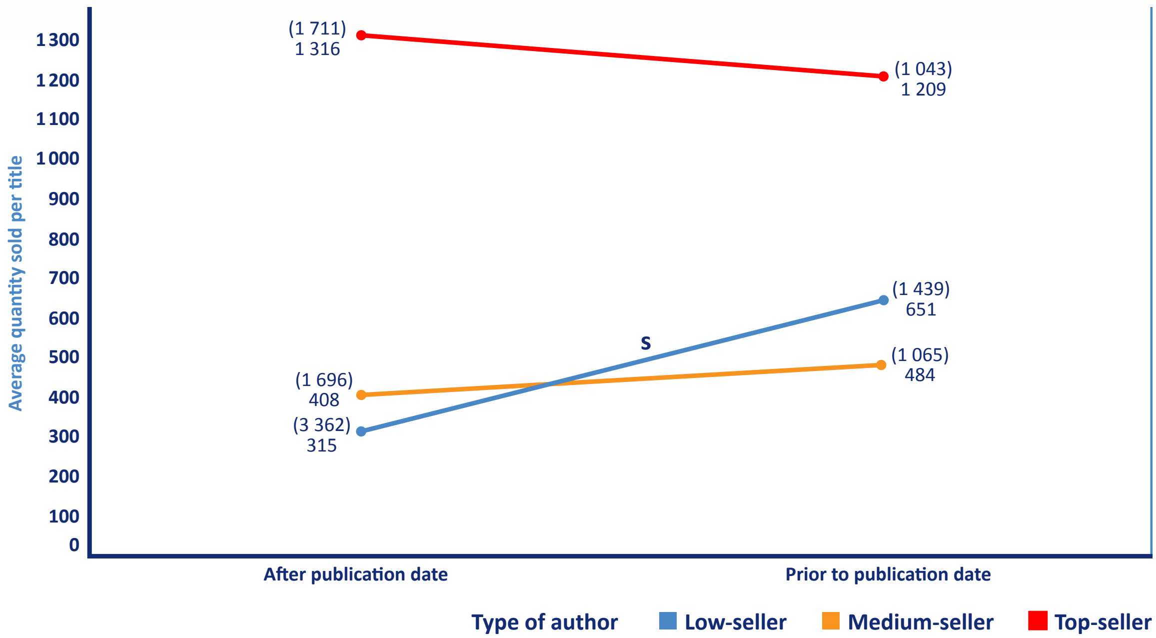
Bracketed figures relate to the number of titles processed overall while figures not between brackets relate to the average quantity sold per title. The letter S indicates a statistically significant difference between the two figures at the right.

2.4 The Effect of the Presence of Metadata Prior to the Publication Date

Also of interest is the measured effect on sales of the time when metadata becomes available. An analysis was conducted for the cover image type of metadata. The total number of titles with a cover image was divided in two groups: titles for which the cover image as metadata was added **prior** to the publication date (34% of titles) and titles for which the cover image was only available **after** it (66% of titles). For titles from low-selling authors (category with the smallest quantity of titles sold), the number of copies sold for titles with a cover image added **prior** to the publication date is significantly higher than the number for titles with a cover image added **after** it, regardless of the type of publisher (overall: 651 copies sold on average vs 315). For titles from medium and top-selling authors, differences are not generally significant, regardless of the type of publisher. Graph 3 shows the results per type of author.

Graph 3

Average Quantity of Copies Sold per Title According to the Availability of a Cover Image Prior to the Publication Date or After it, per Type of Author



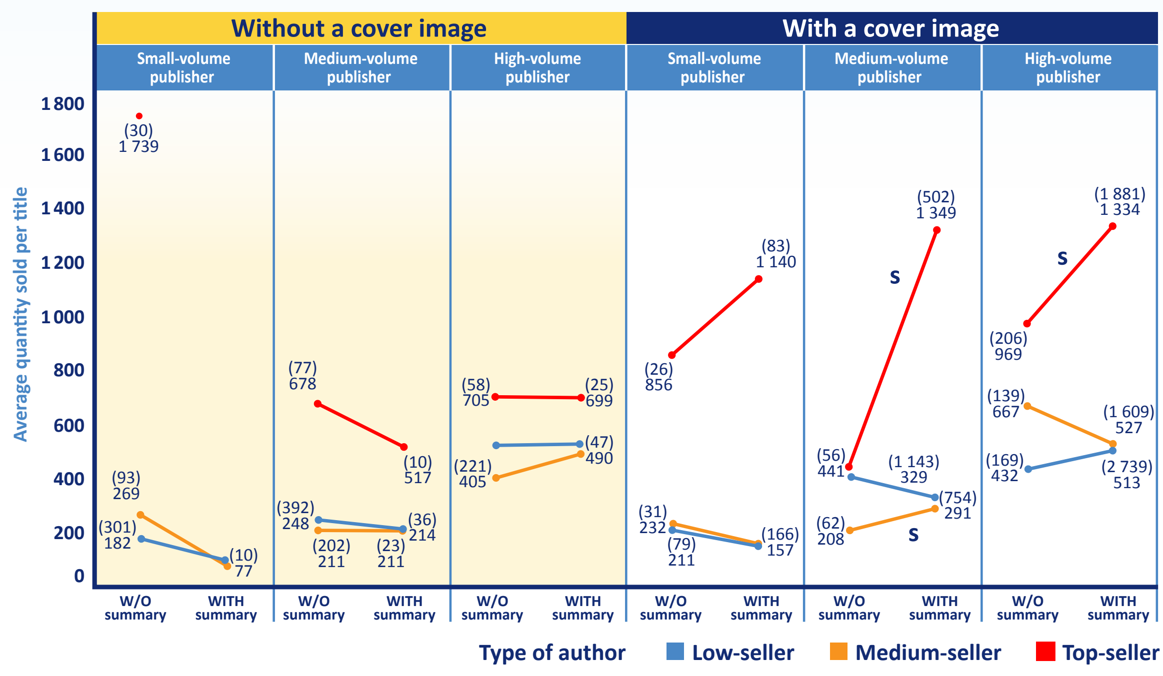
Bracketed figures relate to the number of titles processed overall while figures not between brackets relate to the average quantity sold per title. The letter S indicates a statistically significant difference between the two figures at the right.

2.5 The Effect of the Summary, With and Without a Cover Image

Another analysis consisted in verifying the effect of the summary as a type of metadata in the presence or absence of a cover image. Results show that the addition of a summary when the cover image is absent does not significantly increase the number of copies sold, regardless of the type of publisher or author. However, when a cover image is present, the average quantity of copies sold for titles to which a summary was added is significantly higher than the average for titles without a summary for top-selling authors published by medium and high-volume publishers (see Graph 4). The highest increase of the number of copies sold is found in top-selling authors published by medium-volume publishers (almost 300%).

Graph 4

Average Quantity of Copies Sold per Title With or Without a Summary in the Presence or Absence of a Cover Image, per Type of Author and Type of Publisher



Bracketed figures relate to the number of titles processed overall while figures not between brackets relate to the average quantity sold per title. The letter S indicates a statistically significant difference between the two figures at the right.

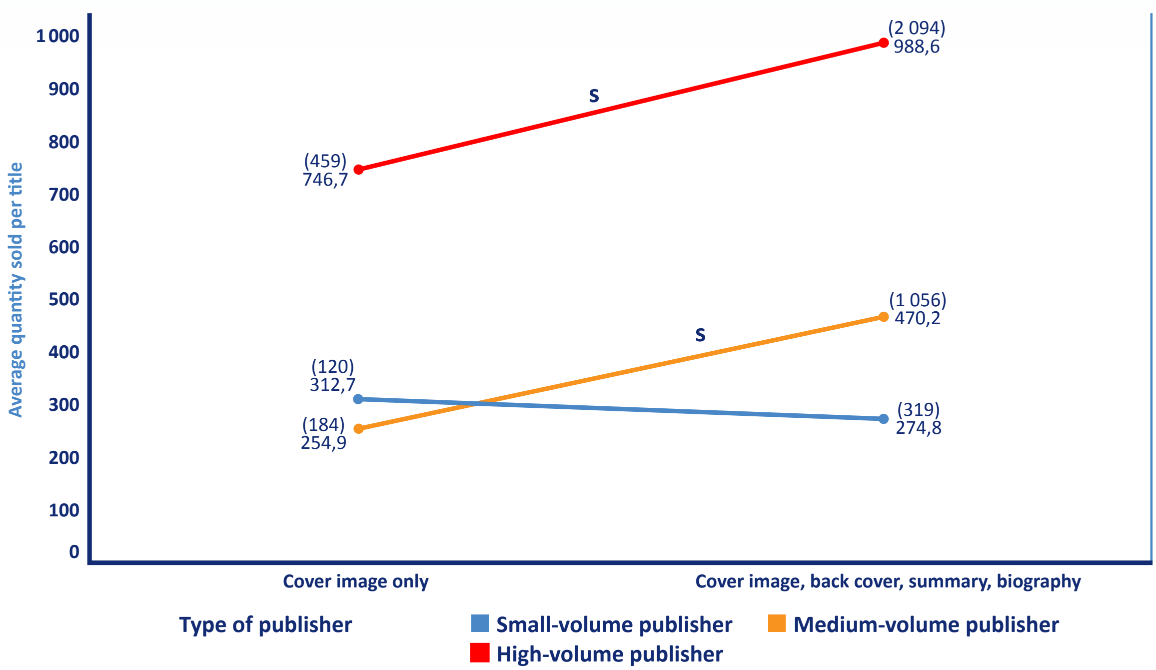
2.6 The Effect of the Amount of Metadata

The last analysis consisted in verifying the effect of the number of different types of metadata present (one to four types) on the quantity of copies sold per title. To do so, we compared the average sales for titles with only the cover image present as metadata to the average for titles with a cover image, a back cover, a summary and a biography. Results are shown in Graph 5 per type of publisher (the conclusions regarding the significance of differences per author did not distinguish themselves from those obtained per type of publisher).

The main conclusion is that a significant difference can be observed between the average quantity of copies sold per title according to the number of types of metadata present for publishers with high and medium volumes (respectively, an increase of 84% and 32% when four types of metadata are present).

Graph 5

Average Quantity of Copies Sold per Title with only a Cover Image or with Four Types of Metadata (Cover Image, Back Cover, Summary and Biography), per Type of Publisher



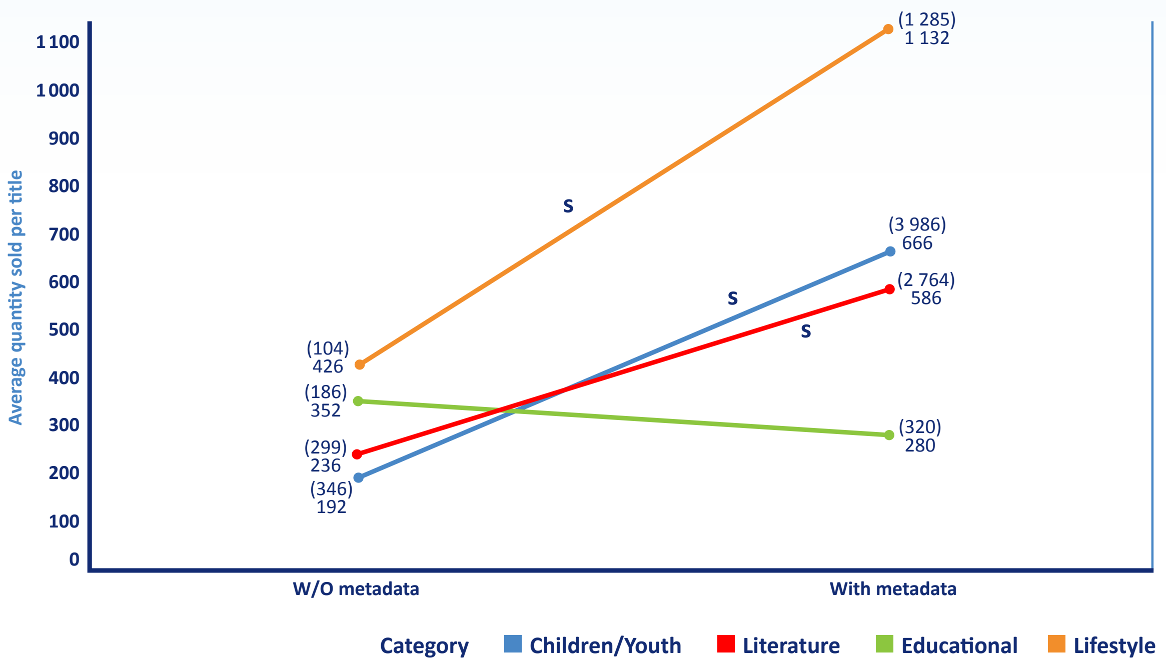
Bracketed figures relate to the number of titles processed overall while figures not between brackets relate to the average quantity sold per title. The letter S indicates a statistically significant difference between the two figures at the right.

2.7 Analysis per Category of Books

In accordance to BTLF's nomenclature, each title is associated to one of 22 book categories defined in Gaspard. The four main categories in terms of number of titles published during the reference period are Children/Youth, Literature, Lifestyle and Educational. For each of these categories, the analysis of the effect of the presence of metadata was conducted by taking into account all publishers. Only the Educational category does not show a significantly higher number of copies sold in presence of metadata. The highest ratio is found in the Children/Youth category, where the number of copies sold with at least one type of metadata is 3.47 times higher than the number for titles without metadata (676 vs 192).

Graph 6

Average Quantity of Copies Sold per Title Without Metadata and With at Least One Type of Metadata, per Category



Bracketed figures relate to the number of titles processed overall while figures not between brackets relate to the average quantity sold per title. The letter S indicates a statistically significant difference between the two figures at the right.

3. Analysis for Digital Books

For digital books, a database provided by De Marque that includes every title published between January 1st, 2016 and May 19, 2018 was used to analyse the effect of metadata on sales.

The main goal of this analysis was initially to replicate the analysis done for print books to each type of digital format (audio, PDF, ePub and mobile). The comparison of sales in the presence or absence of metadata could not be conducted due to the fact that the vast majority of titles contain metadata, resulting in an insufficient number of digital books without metadata as a benchmark (see Table 5. Note that the cover image is not included in this table since all titles included it).

Table 5

Distribution of Digital Titles According to the Type of Metadata and the Type of Format

Types of metadata	Audio	ePub	Mobile	PDF	Total	Total %
None		20	1	48	69	0.40%
Biography	1	113	1	186	301	1.60%
Summary	1	78	15	95	189	1.00%
Biography and summary	311	7 503	51	9 008	16 873	90.40%
Reviews and biography				1	1	0.00%
Reviews and summary				1	1	0.00%
Reviews, summary and biography	7	60		83	150	0.80%
Back cover and biography		1		7	8	0.00%
Back cover and summary				7	7	0.00%
Back cover, summary and biography		382		610	992	5.30%
Back cover, reviews and biography				1	1	0.00%
Back cover, reviews, summary and biography	1	25		38	64	0.30%
Total	321	8 182	68	10 085	18 656	100%

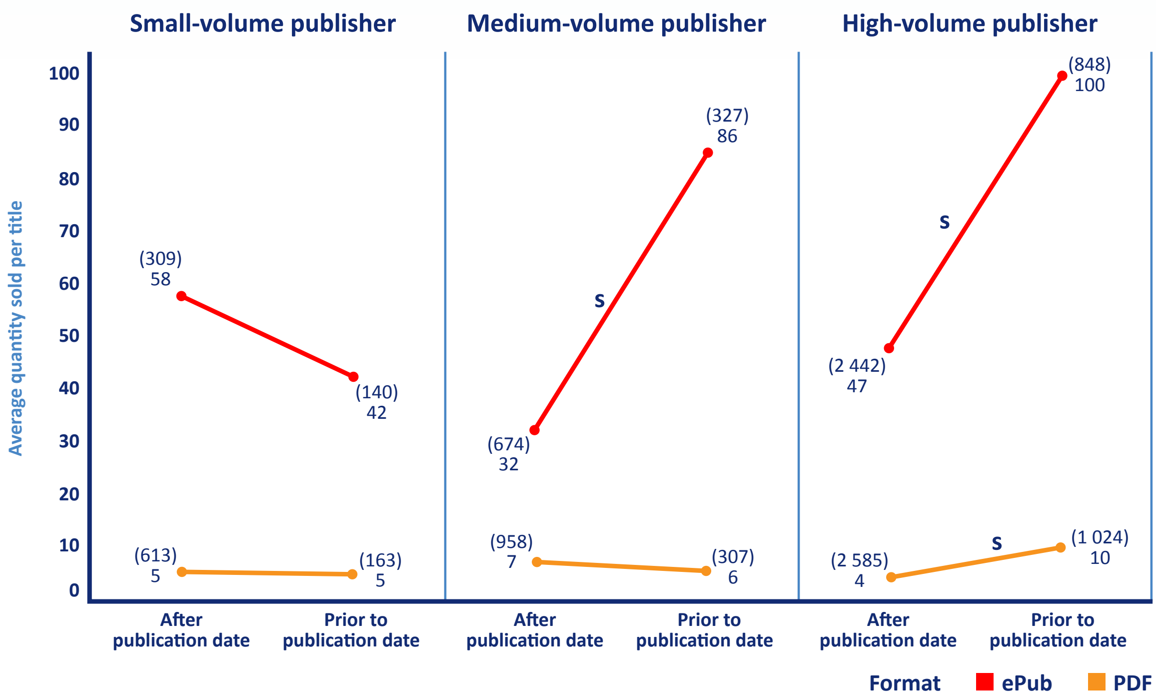
90% of digital titles systematically contain a cover image, a biography and a summary as available types of metadata.

3.1 The Effect of the Presence of Metadata Prior to the Publication Date

De Marque's database indicates if metadata is present or not prior to the title's publication date. This information allowed the comparison of the quantity of copies sold when the availability of metadata precedes the publication date to the quantity of copies sold when metadata is only available after it. This analysis was conducted per type of publisher (defined in a way similar to print titles) and type of digital format (PDF and ePub[8]). Results for medium and high-volume publishers show a significant difference in the average number of copies sold for titles in the ePub format when metadata is available **prior** to the publication date: indeed, the average number of copies sold per title is more than doubled when compared to titles for which metadata is only available **after** the publication date (see Graph 6). As for the PDF format, the only significant difference is seen regarding high-volume publishers.

Graph 7

Average Quantity of Copies Sold per Digital Title According to the Date of Availability of Metadata, per Type of Publisher and Type of Digital Format



Bracketed figures relate to the number of titles processed overall while figures not between brackets relate to the average quantity sold per title. The letter S indicates a statistically significant difference between the two figures at the right.

8 PDF and ePub formats make up 98% of total titles.

Conclusion

This study takes part in the ongoing efforts made by the book industry to encourage the discoverability of Quebec and Canadian French-language books. Results show that quality metadata transmitted on time is generally linked to higher sales. The study also provides indications on the context in which metadata is the most efficient. While we cannot identify the precise mechanism that leads to these positive results, we can affirm with relative certainty that metadata is a key factor in book discoverability and that it has a considerable effect on sales. In the future, a study to identify the metadata users throughout the book creation chain could provide a better understanding of the effect we have measured.

The present study is a complement to prior research, notably Walter[9] and Breedt and Walter[10], on the impact of metadata on book sales in the United Kingdom. In addition to validating the results on Quebec's book industry, our study, like its predecessors, confirms a general effect of metadata on sales as well as a specific impact of the cover image and of the availability of data prior to the publication date. Our analysis also examines the impact of metadata depending on the annual volume of published titles per publisher and on the number of copies sold in the past per author. We also present an analysis for digital books and for prints. As is the case for earlier reports, we can only identify a correlation relationship between metadata practices and sales and not a causal effect.

A key component of this analysis is the fact that the visual aspect of metadata (cover image) seems to be an important vector of discoverability. The fact that making metadata available prior to the publication date has such an important impact also demonstrates the necessity for actors of the book chain to actively and strategically manage its process of production and distribution.

It is worth noticing that some external factors, that could not have been taken into account in this study, might amplify the effect of metadata. For instance, media coverage surrounding a book release, promotional expenses and the time of the year a book is published could moderate its effect.

In conclusion, it is worth repeating that the present study focused on the business performance of books linked to good practices of production and transmission of metadata. It would be interesting to compare the effect of metadata on retail and collectivity sales. Further efforts could also aim at having a better understanding of the effect of other types of metadata on sales, such as Thema categories, for instance.

BTLF, ANEL and De Marque wish to work together with the other participants of the book chain in order to improve the discoverability of books published in French and to support their marketing with the help of metadata. This study demonstrates the importance of including metadata management in any business strategies related to books. BTLF also intends to refine data in Gaspard and Memento in order to conduct always more thorough analyses and meet the needs of the stakeholders of the industry.

9 David Walter, Nielsen BookScan, Nielsen Book UK Study: The Importance of Metadata for Discoverability and Sales, UK, 2016.
10 Andre Breedt and David Walter, Nielsen BookScan, The Link Between Metadata and Sales, UK, 2012.