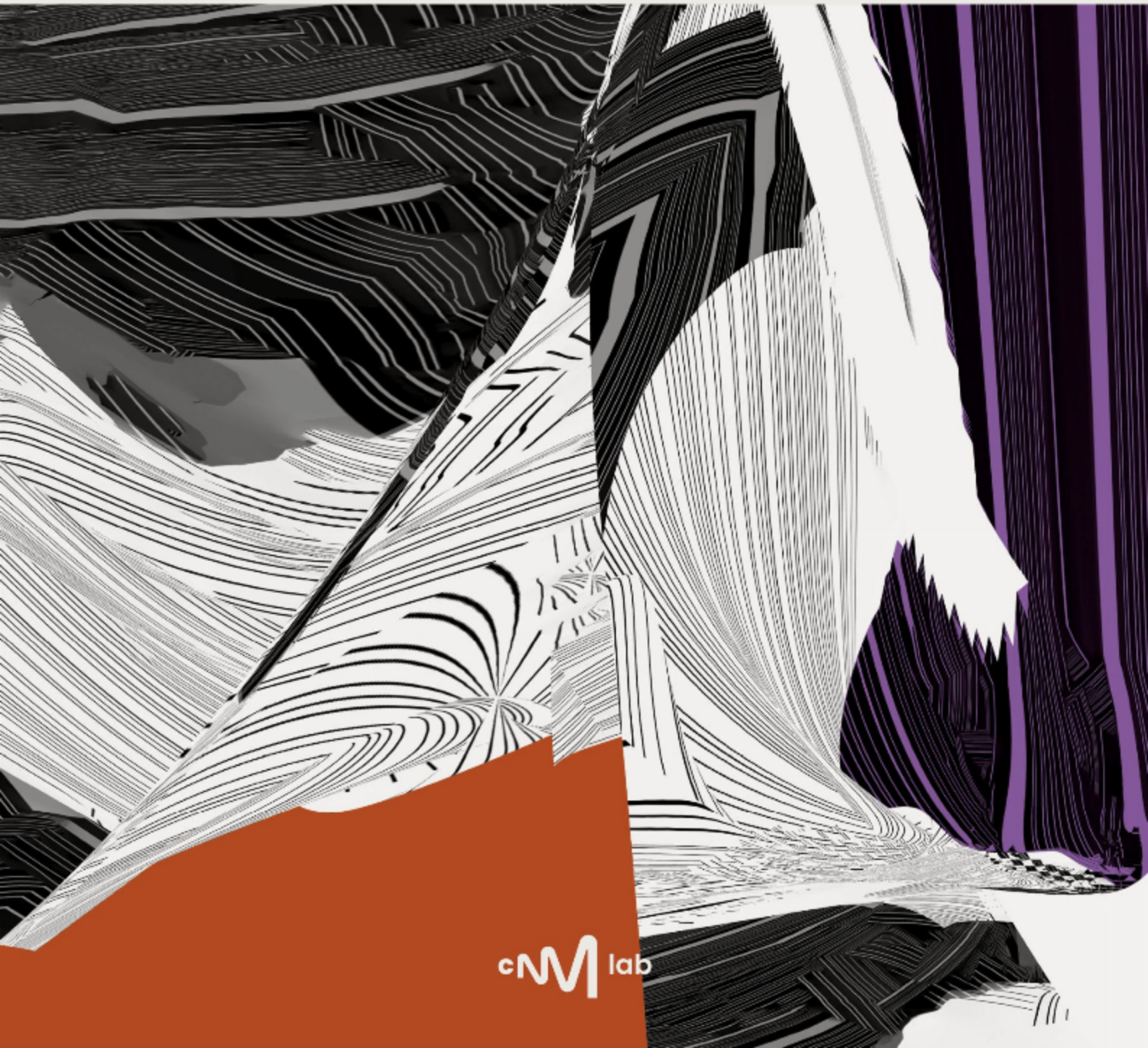
The Discoverability Index, Streaming and Content Diversity

By Jean-Robert Bisaillon



Bisaillon Jean-Robert

Jean-Robert Bisaillon is codirector of the Laboratoire de recherche sur la découvrabilité et les transformations des industries culturelles à l'ère du commerce électronique (Research laboratory on discoverability issues and the transformations of cultural industries in the digital era, LATICCE) at the Université du Québec à Montréal (UQAM). A musician with French B in the 1990s and a pioneer of sampling in Quebec, he founded MetaD, a company focused on research and training around the challenges of digital culture, in 2006.

Introduction

In a recording industry notable for an overabundance of options, concerns are arising not only about consumers' difficulty with decision-making but also about the concentration of these options on a small number of titles. These two phenomena can be explained by the new curation methods used by the online streaming platforms. Consequently, in order to maintain the cultural diversity of music available for streaming and to counterbalance certain effects of automated recommendations that have a knowing or unknowing bias, one approach consists in measuring and promoting the propensity for content to be discovered by users—a concept known as "discoverability."

However, this measurement is contingent upon precise documentation and adequate indexing of the content in question. [1] Successful completion of that work is dependent upon several critical factors, including the adoption of industry best practices, standards, and common rules in support of findability and traceability, as well as data governance to facilitate the use and protection of consumers' and artists' personal data, using transparent methods that are in compliance with regulations. The intent of this article, which deals with various types of institutional approaches and suggests potential tools and solutions, is to shine a light on the issues of discoverability and the functionality of algorithms. We are calling for a standardization of industry processes and encouraging a regulation of streaming, a broadcast mode that is still in development.

1. The Challenges of Discoverability in a Fast-Changing Environment

Discoverability

Our understanding of discoverability is in continual development. One widely accepted definition is derived from the Report of the Joint France–Quebec Mission on Online Discoverability of French-Language Cultural Content (RMFQD):

"The discoverability of a piece of content in the digital environment [2] refers to its availability online and its ability to be found amid a vast array of other content, especially by a person who was not specifically searching for it [3] ."

Energy is now being focused on best practices for stimulating and measuring discoverability.[4]

Among the initiatives that speak to the heightened interest in this issue was a colloquy held in the French Senate on October 3, 2022, on the diversity of cultural content in languages other than English on digital platforms. [4] That event presented an opportunity to call upon the institutions in charge of public policies to take a role in implementing open-access databases that inventory the cultural artifacts being marketed online.

At least two Canadian initiatives are in line with this proposition: Bill C-11 on online streaming [5] and the MetaMusic metadata capture project. [6] Let us emphasize the importance of mobilizing metadata in the documentation of content distributed on the platforms. Metadata are used by the online services' algorithms to contextualize and cross-reference works with the criteria used by the recommendation tools.

Meanwhile, LATICCE recently launched its "Echo Chamber Research Project: Enriched Metadata and Echo Chambers," in collaboration with the Mitacs program and the independent Quebec-based record label InTempo Musique. The project is aimed at more precisely defining the contours of content documentation through the use of metadata and ascertaining the increase in visibility stemming from the use of so-called "enriched" metadata.

All these initiatives are coinciding and making a mark at a time when speculation is rife regarding the near-future roles of recommendation systems, automated decision-making aids, algorithms, and artificial intelligence programs of all kinds in the choices we make about our event attendance and consumption of cultural products, as well as in the creation of future works.

However, our understanding of discoverability remains imprecise. Once we assert that there is a challenge to finding music in the context of a "hyper-offer," or saturated marketplace, the issue raises questions of a technical nature, as well as issues surrounding trade secrets, data governance, etc. LATICCE conducted a literature review on the topic [7] with a view to determining the main parameters, some of which are restated here. The objective is to refine our understanding of discoverability for the audio recording industry, to take stock of current knowledge, and above all to identify the areas of focus and the research questions that are most likely to guide our efforts in the most rewarding directions.

Discoverability Index

LATICCE has produced a prototype discoverability measurement index based on an equation in which P refers to presence, V refers to visibility, and R represents recommendation. This last variable can be further broken down into (c) concordance, (p) pertinence, and (n) novelty. Concordance corresponds precisely with the study subject's reference shopping cart—in other words, exact correspondence with certain variables in the person's listening history, especially the list of artists to whose work they have listened. Pertinence refers to a connection with various similar artists, as determined by collaborative filtering. Novelty refers to the presence of proposed items released within the

previous 30 days, as corroborated by the lists of new releases published by the Association québécoise de l'industrie du disque, du spectacle et de la vidéo (ADISQ) [8] .

$$ID\ (\%) = \left(\frac{P + V + R\left(\frac{c + p + n}{3}\right)}{3} \right)$$

These projects are attempting to measure not the effective consumption of digital feeds, as we would when drawing up a commercial chart, but to measure the actions taken in order to display certain content to certain audiences. Aggregated statistics regarding effective consumption (hit lists) or fraudulent practices (fake streams [9]) tell us little about the algorithmic approaches to content provision or about barriers to discoverability. The discoverability index constitutes a measurement paradigm of a different sort.

The index is a time-specific measurement that does not indicate the level of discoverability in an absolute sense, but rather indicates the longitudinal evolution (through taking periodic, regular, staggered measurements) of the observed variables, making it possible to estimate the progress and setbacks of the recommendation tools offered by the platforms. It is a technologically neutral tool that gauges the success of the recommendation without suggesting thresholds to be met or indicating consumption trends.

2. Institutional Approaches and Perspectives

National Inventory and Presence Among Offerings

The presence of works and repertoires within online catalogs is a key variable in evaluating compliance over time with potential regulations promoting diversity. Without presence, diversity and discoverability are obviously compromised. Presence must be measured in light of national inventories, on a country-by-country basis. In our view, this is embodied in databases—ideally, open and interconnected databases—containing enriched administrative and documentary metadata, listing the cultural artifacts made available online. Thus the formation of a national inventory of works and the associated general-interest (meta)data (GID) is a prerequisite for ensuring verification of presence and facilitating adjustments to that metric. [10]

Because it creates a public policy issue, this type of initiative represents a national undertaking. We must consider the role of legal copyright deposit, once it is enhanced to include sensitive patrimonial interest metadata in the context of a digital economy. In this respect, let us highlight the National Library of Finland's initiative to combine the efforts of all collective rights managers in the country on the topic of attributing ISNI (International Standard Name Identifier) to creators and arts organizations as a bridging identifier. [11]

Canada and Quebec form one of the three parties (the others being Germany and South Korea) that have signed on to date in support of implementing the Convention on the Protection and Promotion of the Diversity of Cultural Expressions. [12] As it appears that France will join them in the near future, the question of best practices and technical measures to be implemented in order to stimulate and evaluate the discoverability of online content will be an issue of public policy in the years ahead.

The Canadian Radio-television and Telecommunications Commission (CRTC), an independent public organization with a mandate to regulate and oversee broadcasting and telecommunications in Canada (equivalent to ARCOM in France), is mentioned in the so-called "Yale Report" (formally "Canada's Communications Future: Time to Act") [13] and is responsible for monitoring online content offerings in Canada.

In December 2022, the CRTC completed the process of reviewing its broadcasting policy, and some analyses suggest that this regulatory body has understood the importance of its role with regard to establishing a national inventory. There is an evident willingness to develop an open (meta)database to enable the identification of works.

The Commission is currently developing a digital monitoring system and an open database to simplify and automatize the process of identifying musical selections. This system relies on probative data [. . .] such as the International Standard Recording Code (ISRC), the International Standard Name Identifier (ISNI) and the International Standard Musical Work Code (ISWC), codes that can be used to confirm the accuracy of the information for any musical selection. [. . .] Once it is made public, this database will greatly facilitate identification of Canadian musical selections and mitigate the risks of non-compliance with regulatory requirements. [14]

Once the construction and use of a database of works is addressed, however, the quality, reliability, and exhaustiveness of that data remain an open question.

Commons, Unique Identifiers, and Interoperability

Questions have been raised for years regarding the official prescription, exhaustiveness, and degree of reliability of a database of cultural and patrimonial artifacts from the music sector.

A number of private initiatives, as well as crowdsourced digital commons initiatives, have addressed this matter. I have been maintaining a list of music metadatabases since 2010; it now has over 100 entries. [15] To date, no database has emerged as a common, reliable, and authoritative source of information. Moreover, few points of connection have been established among the various initiatives.

However, the Wikidata digital commons offers the option of entering and searching 7,709 types of unique standardized identifiers—machine-readable codes that make it possible to deduplicate and cross-reference works and/or artists. [16] This cross-referencing is helpful in distinguishing among works and attributing them to the correct artists and other

rightsholders. Such identifiers exist for the vast majority of works of intellectual property, such as ISBN for books and ISRC for recorded music. [17]

The Wikidata identifier for a work, [18] a person, or a legal entity thus becomes, by extension, a bridging identifier, facilitating interoperability among artistic contributors and repertoires and the unequivocal matching of works and their creators. In the Finnish example, ISNI functions as a bridging identifier to make such matches possible. A bridging identifier is a unique, standardized identifier that enables cross-referencing in order to remove any ambiguity and allow for the work of discernment and possible deduplication.

By the same token, the crowdsourced MusicBrainz database is an extremely rich source of musical metadata that must be incorporated into the dynamics of work and of the search for authoritativeness and exhaustiveness. Because it is created and maintained by amateurs (much like Wikipedia and Wikidata), the industry has tended to underestimate its impact. In the effort to construct unequivocal, high-quality data, however, no source ought to be overlooked. At a later stage, it will be necessary to establish suitable protocols for the cross-referencing, verification, and fusion of data.

Open Data

Aside from unique identifiers, the open and linked database Wikidata describes a multitude of properties associated with each listed element. Examples include the language or geographic origin of a work or person. It is possible to create a list of all items meeting specific criteria by conducting a search with the SPARQL query language. Property P407, for example, indicates the language associated with a work.

Let us take as an example a query for a musical composition—a song or piece of music with vocals—created by a French-speaking songwriter or performer born in Quebec or France. This query generates 3,373 results in 52 milliseconds. [19] The number of titles in open, public song catalogs that are adequately labeled as being of French or Quebec/Canadian origin and in the French language is well below the expected number. [20]

As we have just seen in the SPARQL query to Wikidata, an open database must, by definition, be automatically "harvestable" so that it can be reused by industry stakeholders or by the public institutions charged with monitoring broadcasters' compliance with content commitments. Over time, open data offers rightsholders the opportunity to ensure the quality of the data, to confirm or correct the data, and to have remedies available for that purpose. [21]

Indexing at the Source

Much like the National Library of Finland's scheme, the MetaMusic project is based on a joint initiative of all the rights management organizations and unions in the Quebec music industry, in the form of a nonprofit consortium. [22]

Practice now imposes preconditioning rules on theory, in order to minimize the barriers to discoverability. That is precisely what MetaMusic could make possible: encouraging new best practices with regard to preconditioning in order to ensure that the artifacts distributed through the value chain and via the platforms are adequately documented. Otherwise, algorithmic recommendation systems could overlook this content.

A good deal has been written about the digital online listening platforms' recommendation algorithms, in particular by Brian Whitman, co-founder of The Echo Nest, a recommendation project acquired by Spotify in March 2014. [23] Notably, Whitman addresses the limitations of collaborative filtering, which tends to produce filter bubbles and uninspired recommendations.

The consulting company Music Tomorrow describes industry practices aimed at a better understanding of recommendation dynamics using the term "recommender system optimization" (RSO), a reference to SEO (search engine optimization), which is aimed at improving organic results in web searches. RSO—particularly *technical* RSO—deals with administrative and documentary metadata, "optimizing the technical distribution pipeline to ensure that the artist catalog and metadata attached are complete and accurate." [24]

However, we do not yet know precisely which data are key to the promotion of optimal discoverability and, over time, an accurate rendering of accounts. While they may resemble general-interest data (GID), descriptive metadata appear to represent an area in which competition continues.

3. Tools and Solutions Developed by the LATICCE Team

Filter Bubbles and Enriched Metadata

This is where LATICCE's current work on enriched metadata comes into play. There exist Electronic Data Interchange (EDI) standards to fuel the music value chain. The Digital Data Exchange (DDEX) consortium was created to develop the electronic protocols needed for data transmission between each segment of the value chain for audio recordings and podcasts. The professional practices of non-Anglophone culture milieus absolutely must be based on this type of international standard.

The DDEX-MEAD choreography now makes it possible to share Media Enrichment and Description (MEAD) data among producers, digital distributors, and platforms. The MEAD format supports 30 categories of enriched data and an allowed value set (AVS), which can now be exchanged for each of those categories.

Our project is being developed in conjunction with the Quebec-based label InTempo Musique and several French research partners, including LabEx-ICCA, FÉLIN (a federation

of independent labels), and Musicovery. It is aimed at facilitating the analysis of the relationship among the existence of enriched metadata, recommendation quality, and the potential to break through echo chambers or filter bubbles.

When launched, the project established enriched metadata as follows:

- Song lyrics and keywords that help to accelerate content searches, with natural-language analysis, by concepts, places, moods, feelings, etc.
- The expressions used to describe a style, a period, a trend, or a one-hit wonder and that are used in processing voice requests.
- The digital watermarks and audio fingerprints that assist in track searches through listening to a signal via telephone. [25]
- The complete list of contributors and studios, facilitating connections among various recordings and projects.
- Entry in linked, open databases, permitting the reuse of biographical or phonographic data and information about periods and musical trends. This variable also makes it possible to feed certain services that provide matches to similar artists and genres.
- Technical, legal, and administrative metadata of use to platforms in subjecting the new proposals to catalog quality standards.
- Photographs of the main performers, to be changed out over the course of the artist's career. [26]

We note four areas of focus:

1) the typology and characterization of services, stakeholders, markets, and economic flows;

2) the measurement of content discoverability before and after documentary conditioning of the products with the aid of enriched metadata;

3) the defining of best business practices in the sector, in response to the new Canadian regulation of the platforms (documentation, recommendation, discoverability); [27]

4) the study of economic impacts on the sector, taking the various types of stakeholders and foreign markets into consideration.

Among the types of information now communicable via the DDEX-MEAD protocol, we highlight the following: key, time signature, tempo, theme, use of samples, mood, genre, use in advertisements or synchronization in television series or films, and awards won. It is also worth noting that it is possible to list these metadata in Wikidata, an open and linked environment, thus providing automatic access to the data. This could produce competition between those producers who manage to compile, communicate, and share enriched metadata and those who neglect to do so. The circulation of cultural objects is coming to be conditioned by these technical determinants.

The introduction of the new MEAD standard is a confirmation of the intuitions and best practices promoted by the MetaMusic project and firms such as Music Tomorrow, as well as the objectives of LATICCE's Echo Chamber Research Project, the goal of which is to break through the recommendation echo chamber to stimulate discoverability and, over time, music export.

The echo chamber (or filter bubble) phenomenon is defined as follows: automated decision systems [28] produce (based on their level of maturity) echo chamber or filter bubble effects that confine listeners to sound universes that are often sorely lacking in diversity, as they are conditioned by intended or unintended algorithmic biases. Depending on their size or hermetic nature, echo chambers will tend to hinder access to foreign markets, diverse audiences, and new domestic audiences by limiting audiences to sound profiles that are compatible with statistical norms. This phenomenon is referred to as "more of the same"; in other words, more tracks become available, but their characteristics remain static.

This echo chamber phenomenon is largely attributable to collaborative filtering, a foundational technology in early recommendation algorithms that remains widely used to this day. If this technology could make use of more data, however, it would likely produce more sophisticated results.

Testing in Progress

To this effect, our project led us to conduct some pretesting, in summer 2022, using Spotify's Get Recommendations API. [29] We found that the recommendations made for a "seed artist" and "seed track" were comparable regardless of the subscriber's listening history, location, or user token. We summarily conclude, on this basis, that collaborative filtering—in other words, how many times selection Y has been listened to by users who have listened to selection X, without any real consideration of the users' taste profiles—is always the essential factor in defining the distance between two tracks.

We also conclude that there is a pressing need to account for methods of access to the platform feeds in order to produce meaningful analyses of recommendation systems over time. The platforms use different algorithms with greater or lesser degrees of sophistication depending on these methods of access.

Typology of Methods of Access

Our preliminary work on defining the typology of methods of access has allowed us, for example, to list more than 30 access routines available on Spotify Mobile.

Table 1. Typology of Methods of Accessing Spotify Mobile

Spotify Mobile
Playlist (vertical widget)
Your favorite mixes

What's new from [. . .]
Hot album for you
What your friends are playing
For fans of [. . .] (link to programmed Artist Playlist)
Popular new releases
Popular album
More like [. . .] (link to programmed Artist Playlist)
More like [. . .] (link to programmed Genre Playlist)
Spotify Wrapped (a report on your year)
(Genre) Music picked just for you
Recently played
Play your old favorites
Daily mix
Daily Mix 1 (Made for [you])
Daily Mix 2 (Made for [you])
Daily Mix 3 (Made for [you])
Daily Mix 4 (Made for [you])
Daily Mix 5 (Made for [you])
Daily Mix 6 (Made for [you])
Daily Wellness
Discover Weekly
Radar releases
#SpotifyWrapped
Recommended stations
Recommended artists
Your playlists
Selected albums
Latest releases just for you
What your friends are playing
Popular stations
Search artist
Search title

Typology of Feeds and Searches

Our team is working to define a typology of user searches and feeds. This nomenclature has become necessary in order to qualify the listening modes for which potential discoverability measures could be enacted. It borrows from the concept of "on-demand"

listening used by the US-based data company Luminate, which produces statistics for on-demand audio and video. [30] LATICCE's current glossary is as follows:

1. On-demand audio: the user searches by the artist's name or by the title of a track or album
2. On-demand video: see audio
3. Curated audio programming: the playlist is created by professionals (for example, the platforms' curators)
4. Peer audio programming: the playlist is created by other users
5. Algorithmic audio programming: the playlist is automated
6. Curated video programming: see audio
7. Peer video programming: see audio
8. Algorithmic video programming: see audio
9. Extrapolated feeds: a list or search is followed by an automated feed
10. Hybrid programming: two or more methods of searching or programming are used in combination

Typology of Enriched General Interest Metadata

Finally, we addressed a typology of general interest enriched metadata that could, over time, serve to improve the documentation of artists or titles. We have taken the dataset prescribed by the DDEX-MEAD into account. By way of example, we have identified more than 70 actions to take and fields to fill in just for the MusicBrainz database, approximately 20 of which are to be considered as high priority; the list is below.

Table 2. Priority Enrichment Actions on MusicBrainz [31]

Add ISNI
Add Aliases
Link Wikidata
Link Apple ID
Link Spotify ID
Album Release Date
Link BandCamp
Link BandsIntown
Link Soundcloud
Songkick
Link Last FM
Link YouTube
Link VIAF
Link Amazon Music
Release Group – Enter Title

Album (Release) – Enter Title
Album (Release) – Enter Works
Album (Release) – Enter Musician Credits
Album (Release) – Enter Tracks

Using a "related artists" search tool from the independent label InTempo, it is possible to track the evolution of filter bubbles longitudinally, to the third degree of separation.

By way of example, our work demonstrates that French musician Clara Luciani's only path toward a North American audience (aside from the filter bubbles of French pop and chanson) is the Quebec-based artist Safia Nolin, who is located at the center of a bubble at the far left of a relationship graph produced with the support of our measurement tools, based on the Spotify platform's APIs for retrieval of data on similar artists.

In April 2023, LATICCE launched a research cycle aimed at monitoring the impact of data enrichment actions on these bubbles for a long list of independent artists from France and Quebec.

4. Small Data, Big Data: Sound Profiles and Artificial Intelligence

Each day, more than 100,000 pieces are added to the Spotify servers, [32] and 8.5 billion Google searches are performed. [33] The size of the web currently hovers around 35 billion pages. [34] Terms such as "hyper-offer," "infobesity," and "information overload" entered our vocabulary some time ago.

In his essay on Spotify data, journalist Philippe Astor stresses the massive quantity of data used in support of such a service.

When it migrated its data infrastructure to Google Cloud Platform, starting in 2016, Spotify had to transfer more than 100 petabytes of data from its data centers, according to Ramon van Alteren, the director of engineering who oversaw the operation internally. At the time, Spotify's pipeline was capable of carrying "more than 700,000 events per second worldwide," with "event" referring to any action undertaken by a user within Spotify's interface, such as adding a song to a playlist. [35]

Internet users browse these massive content "catalogs" by using natural, everyday language. Whether or not they realize it, however, they are guided in the process by robotic assistants in research and decision-making. This requires the organization of enormous quantities of information and content and curation efforts that, over time, are aided to some extent by algorithms (and in turn offer assistance to those algorithms, as this information feeds machine learning, also known as "deep learning"). These days, the online

commercial offering is built upon these massive deposits of digital data, which may be structured to varying degrees. These data repositories' state of organization can be termed "data lakes" or "data swamps." [36]

If we believe the motto generally attributed to French physician Claude Bernard, the experimenter who does not know what he is looking for does not understand what he finds. [37] That is a good summary of the dilemmas associated with the exploitation of mass data and one of the challenges involved in the use of artificial intelligence: the writing of "prompts," or questions in natural language, that we pose to AI chatbots and voice assistants.

Following a similar line of logic, what use is music recommendation if it is ill-suited and leads the listener to give up or lose interest?

In brief, the recommendation is based on the user's playlist in order to offer a so-called "personalized" radio feed; listeners are defined by their usage history, and their decision-making is handled by machines. Researchers Jean-Samuel Beuscart, Samuel Coavoux, and Sisley Maillard have attempted to define the role of recommendation systems in current listening habits, placing it between listener autonomy and imposed decisions (heteronomy). They found that algorithmic suggestions remain marginal and are based primarily on subscribers' libraries. [38]

As for LATICCE, a review of our work highlighted the current weakness of recommendation systems for a well-defined individual subject:

Despite the emphasis placed by the music services on the quality of the customized music experience they offer their subscribers, none of the 21 weeks of listening offered what our subject was expecting. The services offered reacted in highly varied ways to the issues encountered in a "cold start" situation (recommending relevant content with very little data history). [39]

In order to derive any real benefit from big data, small data—clear, granular, enriched data of verified quality—are needed, ensuring optimal usage of the recommendation processes. These data are open and shared so that they can be used, reused, updated, and continuously confirmed. The expression "garbage in, garbage out" is often used when describing data of insufficient quality to serve as the basis for an accurate rendering of accounts or an automated decision-making process. We therefore believe that small data must not be overlooked, nor should it be confined to a digital "black box" and treated as a trade secret.

5. Study of Socioeconomic Impacts

Meanwhile, Guy-Philippe Wells, a doctoral researcher at LATICCE, is pursuing a research project aimed at measuring the economic impact of the online listening platforms on the income of Quebec-based songwriters (lyricists and composers). It is necessary to study the

impacts of the digital transformation from a local perspective in order to verify whether the global dynamics are reproduced at the local level or if contradictory or divergent dynamics are observed at that level instead. The impact of the digital transformation on the music industry cannot be measured by means of a simple global aggregation. It must also be measured on the basis of the local industry networks that promote the creation of original music that distinguishes itself from the content produced by the "big three" global conglomerates (Universal, Sony, and Warner), thus ensuring artistic representation of the diversity of world cultures. Over the last few months, we have completed a first stage in our research: an online survey of Quebec-based artists who are self-produced or represented mainly by the independent labels. This survey consists of 18 questions aimed at describing the impact of the digital transformation on these artists' income. To date, more than 150 artists have participated in this survey, which is still ongoing. The preliminary results point toward a need for in-depth interviews in order to gain a better understanding of the state of play.

Data Governance and Public Policies

In order for streaming to become a lasting, viable mode of listening, the quality of the documentation of music tracks available for streaming, as well as the governance of the data and metadata that describe or are produced by that activity, must gain in maturity. Over time, it will also be necessary to ensure that artists are compensated and, consequently, that content and cultural diversity are continually renewed.

We wish to address the issues surrounding open general interest data in the sector, the governance of that data, feed monitoring, and respect for subscribers' data. These parameters are part of the new equation governing music listening, and they raise the question of the need for regulations requiring the efforts of the sector as a whole.

Naturally, governance issues arise when data are made open and reused and when tools are made available to update and confirm them. Undoubtedly, the best data governance approach is one guaranteeing that the subject maintains control over their data and authorizes the use thereof according to their own needs [40].

Nathalie Casemajor and Guillaume Sirois, researchers at the Institut national de la recherche scientifique (INRS) in Montreal, offer a clear definition of sensitive personal data and urge prudence with regard to such data:

Personal data confidentiality is a particularly significant issue when sensitive data are involved. Sensitive data are a subtype of personal data that, when divulged, may constrain the identified individual's exercise of their fundamental freedoms or place them in an undesirable situation. CNIL (the French National Commission on Informatics and Liberty) explains that these sensitive data may be connected with an individual's health, ethnic origin, religious or philosophical beliefs, or sexual orientation, among other factors. [41]

If origin and beliefs are included among sensitive personal data, do data regarding an individual's tastes and habits also qualify as sensitive? In fact, all of these data help feed "discriminatory models" and "model recognition," notably addressed by Clemens Apprich in the collective work *Pattern Discrimination*. [42]

6. The Right to Self-sovereign Identity

Individuals' capacity to decide for themselves, according to the principle of informational self-determination, when and to what extent information about their private lives can be communicated to others was first addressed in a German constitutional ruling ahead of the 1983 census. This capacity invariably rests on subjects' awareness of the transmission of information regarding their private life and on their ability to exercise their autonomy and power. This echoes the panopticon theory, originated by Bentham (1748–1832) and revisited by Michel Foucault, according to which efficient surveillance of a subject may be achieved through that subject's inability to know whether or not they are being watched. [43]

In order for a person to exercise their right to self-sovereign identity, a principle of database transparency and protected personal access to private information must first exist. The exercise of such a right undoubtedly rests on the governmentality of data, or a governance role undertaken by the state regarding issues of respect for and protection of data, as well as on their methods of effective governance—to wit, who holds the data and how the use and processing thereof is defined.

Regarding the ethical aspects of data governance, Stefaan G. Verhulst, Co-Founder and Chief Research and Development Officer of The GovLab at New York University, sets out the following framework.

Digital transformation has led to datafication, and that is where we need to focus our attention. [. . .] The real distinguishing feature in the current environment is that we can reuse data that was used for one purpose and use it for another purpose. [. . .] Data will be reused in a way that benefits society and improves people's lives [. . .] That will require what I call a new "social license" to leverage data for other purposes than initially intended.

How do we design the reuse of data for other purposes?

- The "why": Why do you need the data in the first place?
- The "what": What data are we actually talking about?
- The "who": Who has access to the data?
- The "how": How is the data going to be accessed?
- The "when": When is data going to be used, and when is it going to be deleted?

- The "where": Where will the data be stored, and in what jurisdiction will the data be accessed?
- The social contract: Is there a consensus on the use cases or purposes, and has a chief data steward been appointed? [44]

This last point raised by Verhulst is a particularly sensitive one. In order to carry out certain operations, the online user is dependent upon data that are sometimes used without the user's awareness; the question of governance then becomes essential.

The Solid project is currently being promoted by Tim Berners-Lee, who was the main inventor of the World Wide Web at the start of the 1990s as part of his work at CERN. The project proposes to use "pods" (personal online data stores) to refine control over personal data and institute data portability rationales and web users' right to informational self-determination upon creating a contributor or subscriber profile. This approach could be explored for use by rightsholders and their artifacts: "The idea of surveillance capitalism depends on your data going, by default, to somebody else, and Solid's default is that it goes to you." [45]

In the United Kingdom, the BBC is exploring Solid technology with a view to enabling profile-based content recommendations while making ethical use of personal data: "Technology like personal data stores could be transformative and support our ambition to create tailored and personalised content." [46]

7. Surveillance

At a time when we are being warned about the potential and ongoing slide into surveillance capitalism, [47] a question is being raised about the regulation of multinational companies and online platforms and supervision of their use of algorithms.

Shoshana Zuboff, professor emerita at Harvard Business School, cites Thomas Paine on the powers of the monarchy (and, by analogy, on the power of the major digital companies): "A body of men holding themselves accountable to nobody ought not to be trusted by anybody." [48]

Mathematician Cathy O'Neil emphasizes that web users are increasingly aware of algorithmic biases and that, over time, they will demand transparency, which the platforms will be unable to deny them altogether. [49]

In the same vein, the Yale Report, commissioned ahead of the ongoing review of Canadian broadcasting law, put forward the idea of subjecting digital services to rules promoting discoverability (recommendations 59, 61, 63, 65, and 73) and algorithm audit requirements (recommendation 63).

To ensure that Canadians are able to make informed choices and that Canadian content has sufficient visibility and is easy to find on the services that Canadians use, we recommend that the CRTC impose discoverability obligations on all audio or audiovisual entertainment media content undertakings, as it deems appropriate, including:

- catalogue or exhibition requirements;
- prominence obligations;
- The obligation to offer Canadian media content choices; and
- transparency requirements, notably that companies be transparent with the CRTC regarding how their algorithms operate, including audit requirements. [50]

Is it possible to speak of balance in surveillance? The CRTC is working on an open database of Canadian sound recordings; both MetaMusic and Echo Chamber Research Project are doing the same, to the extent they are able. These projects explore what economist Joëlle Toledanoasserts is a preferred path, in her book *GAFA. Reprenons le pouvoir*:

Interdisciplinary teams (IT and data processing specialists, as well as economists, legal experts, etc.) must be formed with a view to developing methods and analytical tools. [. . .] Public authorities must have access to them, but academic teams and NGOs should also have the means to successfully carry out research. We must learn to test algorithms' transparency and loyalty. [51]

In his book *Cyberstructure*, computer network engineer Stéphane Bortzmeyer points toward approaches to understanding certain technical determinisms introduced by the internet that shed light on our question:

The particularity of "internet governance" is that the internet does not have the well-defined structure of a nation-state, a company, or an association. Indeed, it has no clear structure at all. [. . .] But this lack of a center also has some advantages: it prevents the abuse of power by an authority. In many respects, then, the internet is a unique case in the world of political science. It is a common good in that it is a shared infrastructure—and one that does not function all on its own. How does it manage to function even though its stakeholders are competitors—or even, frankly, enemies? It probably boils down to this: it is in everyone's interest for the internet to work. [52]

Conclusion

The idea of a common good that is useful to everyone, mentioned above by Bortzmeyer, absolutely must be affirmed. The availability of general-interest cultural data in an open and linked format is, to this end, a clearly defined and fertile ground for exploration. It involves issues of respect for internet users' personal data, creators' right to informational self-determination, access to cultural works, and the framing of innovations (namely, artificial intelligence and algorithms) to develop and secure popular trust in them. In the absence of a basis for such trust, the promises of digital technology could be irreversibly

shattered.

To avoid a scenario in which trust is based only on subjective criteria, it will be wise to base it on a framework of standards, upon which public policies in turn can draw. For the culture industries, it is necessary to force the adoption of best practices for content preconditioning by producers and artists, making that a condition of access to production subsidies. We use the term "force" advisedly, to emphasize the necessity of regulating the use of metadata (including enriched metadata) in order to give us greater assurance that audiences are able to discover what interests them. This is the area we are currently exploring.

If we impose discoverability rubrics on the platforms and require them to contribute to financing new productions, as in the case of Canada's Bill C-11, it makes sense for the industry to assume its share of responsibility and deliver music catalogs in compliance with the resulting best practices so that they can access that support. This is also the spirit of the UNESCO Canada-Quebec digital roadmap and operational directives.

Acknowledgments

Antoine Beaubien, Denis Bouchard, Philippe Bouquillion, Vincent Castaignet, Michelle Chanonat, Véronique Desjardins, Pierre B. Gourde, Jean-Baptiste Le Friant, Céline Lepage, Mirjana Milovanovic, Jacinthe Plamondon, Michèle Rioux, Alain Saulnier, Keani Schuller, Brice-Armel Simeu-Tagno, and Guy-Philippe Wells.

1. This process, also called "preconditioning," is the foundation for the decision-making processes in algorithmic recommendation systems.
2. We consider the term "digital environment," used in the definition given by the joint France–Quebec mission, to include all forms of online public provision of digitized content.
3. "Rapport de la mission franco-québécoise sur la découvrabilité en ligne des contenus culturels francophones," Quebec City: Ministry of Culture and Communications of Quebec and Ministry of Culture of France, 2020, p. 8. https://cdn-contenu.quebec.ca/cdn-contenu/adm/min/culture-communications/publications-adm/rapport/Decouvrabilite-Rapport.pdf.
4. The symposium in the French Senate is available for viewing online: http://videos.senat.fr/video.3007738_633abd7fd0f89.decouvrabilite-des-uvres-dexpression-francophone-sur-plateformes-numeriques-apres-midi .
5. See the Online Streaming Act, Bill C-11 441. House of Commons of Canada, "Bill C-11: An Act to amend the Broadcasting Act and to make related and consequential amendments to other Acts," 2022: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-11/first-reading.
6. See the official website of the MétaMusique project: https://metamusique.ca/a-propos
7. Jean-Robert Bisaillon, "Être ou ne pas être découvrable? Une revue de littérature," Montreal, LATICCE (2022). https://ceim.uqam.ca/db/spip.php?page=article-ceim&id_article=13145.
8. Michèle Rioux (ed.), "Être ou ne pas être découvrable? Présence, visibilité et recommandation des propositions culturelles en ligne. La musique et l'audiovisuel," public scientific report, UQAM-CEIM-LATICCE (2021). https://ceim.uqam.ca/db/spip.php?page=article-ceim&id_article=13145.
9. "Manipulation des écoutes en ligne," Centre national de la musique (2023). https://cnm.fr/faux-streams-vrai-phenomene-le-cnm-avec-les-professionnels-pour-lutter-contre-la-fraude.fr/faux-streams-vrai-phenomene-le-cnm-avec-les-professionnels-pour-lutter-contre-la-fraude.
10. See Joëlle Farchy and Juliette Denis, La culture des données. Intelligence artificielle et algorithmes dans les industries culturelles (Paris: Presses des Mines, 2020), and Claudine Duchesne, Laurent Cytermann, Laurent Vachey, Mathieu Morel, and Tristan Aureau, "Rapport relatif aux données d'intérêt général," CGE/IGF (2015). https://www.economie.gouv.fr/files/files/PDF/DIG-Rapport-final2015-09.pdf.
11. National Library of Finland, "ISNI Project Launched in Copyright Management Organizations" (2022). www.kansalliskirjasto.fi/en/news/isni-project-launched-copyright-management-organizations.
12. See the roadmaps or the full set of statements addressed to the French Senate on October 3, 2022, as well as Véronique Guèvremont's statement at Conference No. 4, "Vers un nouvel instrument juridique de l'UNESCO sur la diversité

linguistique des contenus culturels en ligne. La formation d'une coalition internationale." https://praxis.encommun.io/n/Wb5v6Zf1xBuniCjMGY27oINmrhQ/. See also Véronique Guèvremont et al., "Les mesures de découvrabilité des contenus culturels francophones dans l'environnement numérique : compte rendu des tendances et recommandations," Chaire UNESCO sur la diversité des expressions culturelles (2019). http://www.unescodec.chaire.ulaval.ca/sites/unescodec.chaire.ulaval.ca/files/rapport-decouvrabilite-10_decembre_2019_-_final.pdf.

13. Innovation Canada, "Canada's Communications Future: Time to Act (Yale Report)," Ottawa, 2020: https://ised-isde.canada.ca/site/broadcasting-telecommunications-legislative-review/en/canadas-communications-future-time-act.
14. "Revised Commercial Radio Policy" (Ottawa: Canadian Radio-television and Telecommunications Commission, 2022). https://crtc.gc.ca/eng/archive/2022/2022-332.htm.
15. List of 100 music databases and unique identifiers in this sector: bit.ly/musicalmetadata
16. "List of Properties," Wikidata, accessed January 2, 2023, https://www.wikidata.org/w/index.php?limit=7000&title=Special:ListProperties/external-id.
17. Unique identifiers are machine-readable codes that meet standards established by specialized organizations such as the International Organization for Standardization (ISO). See, for example, the International Standard Recording Code (ISRC) for audio recordings: https://www.iso.org/standard/9515.html.
18. See, for example, WKID Q925657: https://www.wikidata.org/wiki/Q925657.
19. According to the following query made on January 2, 2023: w.wiki/5Dg8.
20. LATICCE is currently considering whether to initiate or promote a "datathon" focused on enrichment of Property P407 in order to study the potential for collective engagement around the issues of contribution to an open database of recorded French-language music. The CRTC affirms that it, too, is working to develop an open database that would make it possible to attribute a language and Canadian origin to a musical selection.
21. By way of example, see the landing page for ISNI database correction requests: isni.oclc.org/xslt/DB=1.2/SET=1/TTL=1/WEBCAT?CI_FORMINDEX_COMMAND=update&cmd=update&PPN=466242395&formcode=IS COMMENT&CLT=ISNI.
22. This project was inspired in part by a 2013 paper emphasizing the need for artists and rightsholder to contribute directly to the establishment of a reliable open database: Jean-Robert Bisaillon, "Métadonnées et politique numérique du répertoire musical québécois. Un essai de mobilisation des connaissances dans le nouvel environnement numérique" (master's thesis, Université du Québec à Montréal, 2013).
23. Brian Whitman, "Comment fonctionne la recommandation musicale?," Medium saignant (blog), 2019: http://mediumsaignant.media/comment-fonctionne-la-recommandation-musicale/.
24. Dmitry Pastukhov, "Towards Recommender System Optimization: How Can Artists Influence the Spotify Algorithm?," Music Tomorrow (blog), 2022: https://www.music-tomorrow.com/blog/towards-recommender-system-optimization-how-can-artists-influence-recommendation-algorithms.
25. These technologies are called "watermarking" and "fingerprinting." The former consists of embedding inaudible information into an audio file; the latter involves establishing an audio reference database.
26. Jean-Robert Bisaillon, "10 nouvelles raisons d'état d'indexer nos œuvres avec des métadonnées," LATICCE - Wiki UQAM (blog), 2019: https://wiki.uqam.ca/pages/viewpage.action?pageId=54433018.
27. House of Commons of Canada, "Bill C-11 441: An Act to amend the Broadcasting Act and to make related and consequential amendments to other Acts," 2022: https://www.parl.ca/DocumentViewer/en/44-1/bill/C-11/first-reading.
28. Term used by the Canadian government. See "Bill C-11 432: An Act to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act and to make consequential and related amendments to other Acts," 2020: https://parl.ca/DocumentViewer/en/43-2/bill/C-11/first-reading.
29. Spotify for Developers, "Get Recommendations": https://developer.spotify.com/documentation/web-api/reference/get-recommendations.
30. See, by way of example, Luminate Data, « U.S. Midyear Report for 2022," p. 3: https://luminatedata.com/reports/luminate-2022-u-s-mid-year-report/.
31. Protocol currently under development by the label InTempo Musique and LATICCE, with support from the Mitacs program and the Joint France–Quebec Mission on Online Discoverability of French-Language Cultural Content.
32. Tim Ingham, "It's Happened: 100,000 Tracks Are Now Being Uploaded to Streaming Service Like Spotify Each Day," Music Business Worldwide, October 6, 2022, http://www.musicbusinessworldwide.com/its-happened-100000-tracks-are-now-being-uploaded.
33. See https://www.worldometers.info.
34. See http://www.worldwidewebsize.com/.
35. Philippe Astor, "Les big datas musicales, une question de souveraineté culturelle qui n'est pas posée," Music Zone, November 30, 2022: https://musiczone.substack.com/p/les-big-datas-musicales-une-question.
36. Bisaillon, "Être ou ne pas être découvrable?" op. cit., p. 2.
37. "L'expérimentateur qui ne sait pas ce qu'il cherche, ne comprend pas ce qu'il trouve," motto attributed to physician Claude Bernard and cited on the website of the French Senate: https://www.senat.fr/connaitre-le-senat/lhistoire-du-senat/dossiers-dhistoire/le-senat-sous-le-second-empire-et-napoleon-iii/claude-bernard.html.
38. Beuscart et al., op cit.
39. Rioux (ed.), "Être ou ne pas être découvrable? op. cit., p. 23.
40. Bisaillon, "Être ou ne pas être découvrable?" op. cit., p. 2.
41. Nathalie Casemajor and Guillaume Sirois, "La gouvernance des données d'usage. Enjeux éthiques et perceptions des publics dans les bibliothèques et archives," INRS, 2021, p. 4.

42. Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl, Pattern Discrimination (Minneapolis: University of Minnesota Press, 2018).
43. Michel Foucault, Surveiller et punir. Naissance de la prison (Paris: Gallimard, 1975).
44. Centre for International Governance Innovation, "Digital Technologies: Building Global Trust," YouTube, June 15, 2021: https://www.youtube.com/watch?v=IOc7Lo4ACEs.
45. John Harris, "Tim Berners-Lee: 'We Need Social Networks Where Bad Things Happen Less,'" The Guardian, March 15, 2021: https://www.theguardian.com/lifeandstyle/2021/mar/15/tim-berners-lee-we-need-social-networks-where-bad-things-happen-less.
46. Eleni Sharp, "Personal Data Stores: Building and Trialling Trusted Data Services," BBC, September 29, 2021: https://www.bbc.co.uk/rd/blog/2021-09-personal-data-store-research.
47. See Apprich et al., op. cit.; Alain Saulnier, Les barbares numériques. Résister à l'invasion des GAFAM (Montréal: Écosociété, 2022); and Harris, op. cit.
48. Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (New York: PublicAffairs, 2019), p. 513.
49. Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (New York: Broadway Books, 2017), p. 210.
50. Innovation Canada, "Canada's Communications Future: Time to Act (Yale Report)," Ottawa, 2020: https://ised-isde.canada.ca/site/broadcasting-telecommunications-legislative-review/en/canadas-communications-future-time-act.
51. Joëlle Toledano, GAFA. Reprenons le pouvoir ! (Paris: Odile Jacob, 2020), p. 130.
52. Stéphane Bortzmeyer, Cyberstructure. L'Internet, un espace politique (Caen: C & F Éditions, 2018), pp. 90–92.