Abstract
This paper presents the results of a study on the topics that Mexican television stations assign to web-distributed news content, after seeing a resounding drop in web traffic indicators. The inductive-hypothetico-deductive method was chosen for the work: in 2015, a six-month diagnosis was carried out at 14 Mexican television stations; three to identify the indexing points used by the television stations and three to carry out a hypothetical plan, aimed at comparing results. Among the findings, it is clear that television stations use Content Management Systems to automatically assign basic metadata such as newscast title, date of the news item, broadcasting schedule and anchors. Other metadata in which topics are described are, however, incorporated haphazardly or insufficiently. The evaluation showed that new methodologies for analysis and documentary treatment of videos must be applied so as to refine content description and representation. Furthermore, the hypothetical plan exercise showed viability of increasing search engine relevance and with it, visibility, web positioning and access.

1. Introduction
Thematic metadata used in Search Engine Optimization (SEO) strategies from television news broadcasts published on the web may facilitate retrieval and increase visibility, as required by companies leading these business-model-based endeavors.

The prevalence of reduced user traffic within these audiovisual contents, discovered through search engines, led to this study. A possible relationship was seen between thematic representation, increase in visibility and access. This paper presents a review of 14 Mexican television stations and the way they apply thematic metadata to common source code SEO tags on the web.

The objective is to make use of all the linguistic and semantic capital generated in audiovisual journalistic production for representation and thematic retrieval. This would establish the foundation for improving search engine positioning performance and user access. “However, without any terms to associate with a multimedia document as with images or music, there is an inherent problem indexing such objects. It is possible that the document has some metadata associated with it, but this is not always the case (e.g., on the web). With the web, multimedia documents are becoming increasingly more readily available, and mechanisms to access such information are sorely required” (MacFarlane 2016: 181).

Web access to television contents by a greater volume of users depends, to some extent, on indexing by search engines, essential tools in enabling people to retrieve both complete programs and sections of them. (Sobak and Pharo 2017:739). In this regard, the objectives of human-computer interaction call for developing systems that improve yield and user satisfaction. (Tsakonas and Papatheodorou 2008: 1.234-1.250).
Therefore, content representation links metadata management and use to information retrieval. Considered intellectual activities and information science approaches, this makes linguistic reasoning central to information science (Engerer 2017: 661).

2. Methodology
The inductive-hypothetico-deductive method was chosen for this research on representation and thematic retrieval metadata exploitation in the sphere of Mexican television newscasts on the web and their effect on phenomena such as visibility and access. In this sense, it begins with the idea that the inductive-hypothetico-deductive spiral contains two essential procedural steps.

- Heuristic or discovery phase: a phase involving observation, description, reflection and inductive generalization, intended to generate hypotheses (which could actually be a solution to the problem, answer to the question or explanation of the phenomenon).
- Justification-confirmation phase: the process of verifying the basis of a hypothesis by means of a procedure or device designed to do so (and susceptible to being reproduced) (Sarabia 1999: 55).

The author thus proposes observation and description of the problem to be studied, in this case by analyzing how television news stations manage contents. Specifically, they carry out a diagnosis and, after exploring the situation, define an explanatory hypothesis regarding the behavior, causes and effects of the phenomenon, to model a situation that enables them to confirm whether there is thus an improvement in the problematic situation and to guarantee a solution for it.

Additionally, the research conducted rests on the idea of attending to the needs of dynamic users who consume diverse content on the web, including news content. “According to findings of a 2011 survey (Purcell 2011), 92% of American adult Internet users use search engines to find information on the web, with 59% who do so on a typical day. This and other studies confirm our intuitions regarding the important role of web information. The web continues to provide extremely low cost means of publishing information, often coupled with high incentives for doing so, since web content can affect purchasing behaviors, opinions, and other important decisions of web users” (Kakol, Nielek, and Wierzbieki 2017: 1.043).

3. Findings
The first step in the methodological plan was diagnosis. In this case, it had to be done externally, because, as in several industrial sectors, secrecy characterizes television broadcasting companies and upper management, offices and departments related to “first screen” television journalism and news production and “second screen” distribution and publication on the web. The analysis period was February-April 2015.

The study was set up once the phenomenon had been detected in February 2015,
when web traffic indicators went down 45% (due to the appearance of Netflix and other types of fixed programming) with respect to the previous year, both on web platforms and mobile web (web on mobile devices).

To establish a strategy allowing for more in-depth analysis of the problem of representation and thematic retrieval, research was conducted on data that appear on the web, following this first step.

The instruments used were: a) selection of television stations for the study, b) user interface analysis and c) source code analysis. Among the strategies defined, first of all a group was formed of Mexican television stations that broadcast news on the web as well as on television screens.

The Mexican television stations selected for the diagnosis were: Aprende Televisión Educativa, CNNE (in Spanish), Canal Once, Efekto TV, Excélsior TV (*), Foro TV/Televisa (*), Fuerza Informativa Azteca, Milenio Noticias (*), MVS Noticias, Noticieros Televisa, Proyecto 40, Telefórmula, Televisión Metropolitana Canal 22 and TV UNAM. Television stations that produce and broadcast at least one news program a day were taken into account. The cases marked with an asterisk (*) are television stations with journalistic subject matter that broadcast 24 hours. They all broadcast their normal programming on pay television and a web site (desktop or mobile). Regarding open television, this only applies for Canal Once, Foro TV/Televisa, Fuerza Informativa Azteca, Noticieros Televisa, Proyecto 40 and Televisión Metropolitana Canal 22.

The second instrument was user interface analysis, which consisted of review and analysis of web pages that publish television news broadcasts or any of their substructures (news, reports, interviews, chronicles, sketches, etc.). The purpose is to determine presence or absence of those metadata that represent the topic of the content exhibited and which are considered relevant for search engines to improve positioning and visibility and, therefore, retrievability. An inventory was made of the metadata to be identified in order to establish definitive metadata from the research.

The third diagnostic instrument was source code analysis, consisting of review and analysis of web pages that publish television news broadcasts or any of their substructures (news, stories, interviews, chronicles, sketches, etc.), with Firefox® as the browser through use of the inspection tool and reading of title tags, meta, h1, h2, h”n”, body and content, among others commonly used in SEO strategies.

For each television station or production entity, the news broadcasts available through streaming or on demand (VOD) were inspected, as well as five video clips corresponding to substructures (sections: national, international, sports, entertainment and culture or their equivalents). For this aspect, an inventory of the use of semantic topics to identify, was made

1. Density of key words
2. Tag title
3. Meta tags
4. H1 and H2 tags
5. Analysis of key words in the competence
6. Selection of key words by seasonality
7. Idiomatic variations and concatenations of key words
8. Tagging (internal and social)
9. Optimization of images
10. Meta-descriptions
11. Microdata use
12. Key words in URL

From the findings identified upon concluding the analysis of television news broadcast user interface and the selection of video clips corresponding to substructures (news items, reports, interviews, etc.), the results were:

- All the television stations analyzed have content managers prepared to automatically assign the following metadata:
  - Title of the newscast/news item (the only one directly related to the topic)
  - Date of the newscast/news item
  - Broadcasting schedule/Date and posting time
  - Anchors
- 34% of the television stations do not include a summary or brief news item about the subject of the newscast/news video.
- 86% of the television stations do not include key words on the web pages published by the newscast/news video. The same proportion of proper names of people or places are not included as tags representing the topic.
- 79% of the television stations do not include tags on related topics that could be of interest to the audience and thereby keep it navigating within the web site.
- 14% of television stations do not include social share tools.
- 21% of television stations do not offer their news broadcasts live or on demand: Aprende Televisión Educativa, TV UNAM (platform under construction that broadcasts continuously with a video player) and CNNE. In the case of the first two, their main function is not journalistic, unlike CNNE which is.
- 100% of the television stations lack publication scripts, step outlines or transcriptions linked to their television news broadcasts.
- Only 7% do not offer news broadcast fragments: the TV UNAM platform, broadcasting continuously with a video player, is under construction.

Source code inspection produced the following results:
- Only 14% of the cases presented texts with over 15% key word density, placing them within the range. However, none of the digital properties inspected
presented over 30% density. In other words, even when web editors do publish “journalistic summaries”, they lack sufficient semantic elements for a search engine to establish aboutness.

- In 71% of the television stations, the source code for the digital properties does not include key words in the tag `<TITLE>`.
- 57% of the television stations do not include statements that describe the subject of the digital properties where their news broadcasts or video clips of substructures are hosted. The remaining 43% that do so suffer, however, from incorporating journalistic texts rather than texts with search-engine-appropriate key word density.
- 100% of the television stations use `<H1>` tags (H1 is the HTML element generally used to identify the most important headline on a web page. Examples of H1 would be a page title, post title, product name). In other words, they all place at least one title in the digital property that hosts their news broadcasts or substructure video clips. However, that title (which is published in the user interface) does not always include at least one key word representing the topic. Furthermore, none of the television stations use `<H2>` tags to create synopses, subtitles or subsections in the corpus of “journalistic summaries”, thereby wasting an opportunity to improve key word density.
- In 100% of the cases analyzed, it can be deduced that television station web editors pass up the opportunity to study the key words used by their competitors. There are even cases in which key words are not planted in digital properties. And reduced interest in key word use can be detected when comparing the web edition of the same subject or news item on different television stations.
- In 100% of the cases analyzed, use of seasonally selected key words was non-existent. Pending confirmation through a subsequent study, this may be because at certain times of the year consumption of this type of journalistic product heavily reduces user traffic, especially during vacation periods. Web editors could, however, add key words related to the trending topics of the day, yet, given the minor use of tagging, this resource seems wasted.
- In 100% of the cases analyzed, web editors were seen to have trouble defining a strategy to establish or select key words based on the application of idiomatic variations and concatenations. Instead, they tend to copy `<H1>` tag titles, which is neither wholly appropriate nor effective in terms of search engines.
- 57% of television stations do tagging. However, we note the lack of a clear strategy for doing so based on the creation of lists, indices, vocabularies or taxonomies. Efekto TV and Televisa stand out for being more structured, as they use key words in a way that is closer to thematic indexing. None of the television stations have a social tagging tool, limiting or reducing the possibility of learning
what words the audience uses to search and represent the topics addressed in news broadcasts and their substructures.

- In only 21% of the cases was it found that fixed images or the first image of the video player are not optimized, or rather are not utilized as a mechanism for reinforcing the topic, or that relevance or concordance between the image and content topic is lacking.
- In 79% of the cases, meta-description uses (tag use <META = DESCRIPTION>) was not identified as a tool to improve the thematic description of the content by means of a summary or abstract structured for search engines. It only appears with a greater degree of consistency in the digital properties of Televisa, Foro TV and Efecto TV news broadcasts. The latter case stands out because it uses a specific tag: <META = ABSTRACT>.
- 100% of the cases revealed an absence of microdata that search engines can promote as values added on to the relevance and potential attraction of digital properties, shown in snippets from the results pages of a particular consultation.
- Lastly, in 71% of the cases, key words are not included in URLs. In the cases analyzed, URLs are used with digital property numbers or with the brand of the production entity or the section, or with an operation indication such as "streaming", "VOD" and "live". In the case of Televisa and ForoTV news broadcasts, the URLs are especially long, as they incorporate the title. It would be more effective if URLs were to be used that included one or two key words, configuring the contents manager so that it carried this operation out automatically.

With the strategies and instruments used, it was possible to make a working hypothesis. If a model is applied or the labels filled out for title, topic, summary and social labeling, based on professional information criteria experienced in information indexing and retrieval, better positioning of the news is assured in search engines.

An intervention program was presented to the 14 television stations, which was accepted by five different newscasts that broadcast at different times. The intervention was applied to two newscasts, while the other three acted as a means of control. It was carried out from September through November 2015.

The intervention encompassed the following steps:
- Construction of a work group
- Discussion, design and implementation of a temporary policy for metadata use for thematic representation of journalistic content. The policy was based on the video analysis and documentary treatment model applied to the process of thematic representation of newscasts on the web.
- Training for web editors in managing and using metadata for thematic representation.
Implementation of the metadata policy when publishing newscasts or substructures of them, based on the extraction of descriptors and key words, as with the documentary analysis of newscasts, as well as the use of other elements such as scripts, transcriptions, translations and subtitles.

Automatized data collection (ComScore, Google Analytics, MyMetrix, Videolog) on user web traffic behavior by newscasts. Graphs are presented below showing the traffic curve (single browsers), where the greatest drop occurred in September and how, after implementing the intervention, it went back up.

![Figure 1](image)

Source: ComScore MyMetrix, Google Analytics.

4. Formulation of points learned

The field of information science clearly has opportunities to help enrich web content representation and thematic retrieval, in terms of research, formulating diagnostics and designing interventions for ongoing improvement and innovation. In the specific case of television newscasts on the web, an extra effort must be made. While there is a corpus of text that goes along with the multimedia resource, most of the enriched content is found within that element itself.

- Automatizing the use of metadata through television station content managers, predefining the topics by thematic sections or subsections.
• Applying methodologies for video analysis and documentary treatment, to refine content description and thematic representation (Soto-Hernández 2009: 173).
• Exploiting the value of semantic aspects to represent and retrieve content on the web (Soto and Naumis 2014: 85).
• Increasing content relevance for search engines, along with its visibility, positioning and access by applying SEO tools and better practices.
• Facilitating content representation and thematic retrieval, attempting to make it as ideal as possible and fit with users’ linguistic production style (documentary languages that approach natural language and social tagging). In this regard: “The main challenge in this context is to predict the most suitable retrieval model for a given user query and to cover the semantic gap between user information needs and retrieval models” (Ayadi et al., 2017: 1324).
• Improving the user experience for people entering the website.
• Contributing to the productivity expressed in the web traffic indicators and, hence, to the television station’s return on investment.

Text, audio, video and animation form a discourse whose very format calls for more specialized documentary analysis, where specific aspects are considered, including viewing, content analysis, summary elaboration and construction of thematic descriptions with semantic perspective for search engine optimization (SEO), as well as (controlled and social) tagging.

Documentary analysis of television news broadcasting on the web also demands that the personnel responsible for it stay updated and professionalized in competences that overlap with the skills of the web editor profile, in particular, and professionals in the creative industries in general (Mietzner and Kamprath 2013: 590).

In other words, for the human resource, it involves designing and structuring audiovisual products, keeping in mind linguistic and semantic relevance for improving representation, retrieval, visibility and user experience.

5. Final considerations

1. From the perspective of the worldwide television industry, this is a time for redefining plans aimed at audiences that currently play an ever-greater role as active audiovisual content users on the web than as conventional television viewers.
2. On the one hand, the previously passive and receptive audience now behaves actively and creatively, turning TV viewers into users. Therefore, contents must now be produced to be technically set up to distribute on conventional television and on web platforms and video on demand (apps and opt) that require thematic metadata for algorithms of representation, retrieval and recommendation in accordance with user habits.
3. Given the technological convergence that enabled the overwhelming evolution achieved by Netflix, among others, and the crisis of traditional business models, it is essential to have strategies for SEO and its enrichment through content analysis and video documentary analysis, in the case of newscasts on the web.

4. Thus, if “content is king”, as global television leaders point out, and that is supposedly what audiences with screen independence follow, then it appears that Google is the ‘Caesar’ which controls the web empire. Much more than a search engine, this global company sets policies and develops algorithms that rate the relevance of the thematic representation of content, also conditioning retrieval and, therefore, visibility and access. This has implications from the perspective of the information society and the wider economy, a discussion that, for practical reasons, will not be pursued here.

5. Despite the previous point, at least in the Mexican case, we observe an absence of organizational information policies regarding representation and thematic retrieval. This is not a strength of Mexican television stations, as reflected by their business indicators. A serious dilemma exists, because web publication demands investments and expenditures that should still be subsidized. Television reaches millions of people, while web publication in Mexico does not, thus compromising return on investment. It is a difficult situation, because television stations cannot sidestep the web, as they would no longer be in the market.

6. Finally, there is a demand for information professionals with competences to respond to the challenges of representation and thematic retrieval. Web editors, just like their supervisors, lack these specialized skills. Consequently, we can expect that in fewer years than you might think, there will be a great demand for specialized human resources.

References


