Metadata in Mexican Television News Broadcasts on the Web

Silvano Soto-Hernández[1][0002-8299-1519] and Catalina Naumis-Peña[2][0003-3152-3958]

[1] IDMS, Ciudad de México, México
silvano_soto@idms.com.mx
[2] Instituto de Investigaciones Bibliotecológicas y de la Información Universidad Nacional Autónoma de México, Ciudad de México, México
naumis@unam.mx

Abstract. To increase visibility and maintain return on investment of television news broadcasts ("digital property", as termed in the television industry), web distribution is needed in addition to television broadcasts. This paper presents the results of a study on the use of metadata within Mexican television stations, analyzing those that fulfill a relevant function for representation and thematic retrieval on the web. Action Research was selected for the study, which assessed 14 Mexican television stations for a six-month period in 2015: three to identify metadata use and three to carry out an intervention. One of the findings shows that television stations use Content Management Systems to automatically assign basic metadata such as newscast title, news item date, broadcasting schedule and anchors. Other metadata that describe subject matter are, however, incorporated ambiguously or insufficiently. The evaluation showed the need to apply new methodologies for analysis and video documentary treatment so as to refine content description and representation. The intervention made it possible to demonstrate the viability of increasing search engine relevance, along with visibility, web positioning and access.

Keywords: Content Metadata, Thematic Metadata, Indexing, Thematic Representation, Thematic Retrieve, Web visibility, Mexican Television News Broadcast, Search Engine Optimization.

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” [1].

Access to television contents by a greater volume of users, through the web, depends, to some extent, on indexing by search engines, essential tools in enabling people to retrieve both complete programs and sections of them [2]. In this regard, the objectives of human-computer interaction call for developing systems that improve yield and user satisfaction [3]. Therefore, content representation links metadata management and use to information retrieval. Considered intellectual activities and approaches of information science, this makes linguistic reasoning central to information science [4].

2 Methodology

Action Research was the method chosen for this research on the exploitation of representation and thematic retrieval metadata in the sphere of Mexican television newscasts on the web and their effect on phenomena such as visibility and access. Kemmis, McTaggart and Nixon [5] have thoroughly studied, analyzed and systematized knowledge from this methodological perspective starting with an extensive specialized literature review, and the planner focus was taken from them. "[...] participatory action research expresses a commitment to bring together broad social analysis, the self-reflective collective self-study of practice, and transformational action to improve things" [6].

Other aspects of the focus mentioned were considered when defining methodology, including components such as the necessary diagnostics, definition of intervention strategies and formulation of points learned, using Technical Action Research as the basis. Comfort highlights the influence of the method when pointing out that: “Action research, in contrast, emphasizes the importance of fitting the research process to the action context. Simulated emergencies designed to train personnel in public” [7]. Furthermore, Argyris used this form of social research, alternating the substantial distinction concept between basic and applied research. The assumption is that if the social research is worthy, it reports on action; in other words, it must be within an action context [8].

Additionally, the search that was done rests on the idea of attending to the needs of dynamic users who consume diverse content on the web, among which is the news. “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" [9].

3 Findings

The first point 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 journalism, television news production and distribution on the "first screen" and web edition and publication, or "second screen". The analysis period was February - April 2015 [10].

Table 1. Mexican Television News Broadcasts on the Web Television stations according to Distribution Platform.

Television station	Open TV	Pay TV	Web TV
Aprende Televisión Educativa (educational TV)	Not applicable	X	X
CNN en Español (in Spanish)	Not applicable	X	X
Canal Once (Channel 11)	X	X	X
Efekto TV	Not applicable	X	X
Excélsior TV (*)	Not applicable	X	X
Foro TV/Televisa (*)	X	X	X
Fuerza Informativa Azteca	X	X	X
Milenio Noticias (*)	Not applicable	X	X
MVS Noticias	Not applicable	X	X
Noticieros Televisa (Televisa newscasts)	X	X	X
Proyecto 40	X	X	X
Telefórmula	Not applicable	X	X
Televisión Metropolitana Canal 22 (Channel 22)	X	X	X
TV UNAM	Not applicable	X	X

Note 1: Television stations that produce and broadcast at least one news program a day were taken into consideration. The cases marked with (*) are television stations with journalistic topics that broadcast 24 hours.
Note 2: Data from February-April 2015

To establish a strategy allowing for more in-depth analysis of the problem of representation and thematic retrieval, research was done with data that appear on the web, as of 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 of Mexican television stations that broadcast news on the web as well as television screens was formed. 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 that search engines consider relevant for improving positioning and visibility and, therefore, retrievability. An inventory was made of the metadata to be identified in order to establish the definitive ones from the research.

1. Title of the news program/news item
2. Date of the news program/news item
3. Broadcasting schedule/Date and posting time
4. Anchors
5. Summary of the news program/news item
6. Key words or descriptors
7. Proper names of people or places
8. Related topics
9. Extension of content on social media
10. Newscast offered in complete version, live or on demand
11. Offers script, step outline, transcription or translation/subtitles of the entire news program
12. Offers newscast fragments on demand

The third diagnostic instrument was source code analysis, consisting of review and analysis of web pages which 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, too, an inventory was made of the use of semantic topics to identify.

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. KeywordsinURL

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). o Date of the newscast/news item
  - Broadcasting schedule/Date and posting time
  - Anchors
- 34% of the television stations *do not include a synthesis 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. In the same proportion, *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 attractive 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 CNN Español. In the case of the first two, their main function is not journalistic, whereas for CNN Español it is.
- 100% of the television stations *lack publication scripts or step outlines or transcriptions* linked to their television news broadcasts.
- *Only 7% do not offer news broadcast fragments*: The TV UNAM platform that broadcasts continuously with a video player is under construction.

The source code inspection found that:

- *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, however, suffer from incorporating journalistic texts rather than texts with search-engine-appropriate key word density.

- *100% of the television stations use <H1> tags*. 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 nonexistent.* Pending confirmation through a subsequent study, this may be because in certain seasons 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 not totally appropriate and effective in terms of search engines.
- 57% of television stations do tagging. However, the lack of a clear strategy for doing it based on the creation of lists, indices, vocabularies or taxonomies has been noted. EfektoTV and Televisa stand out for more structured work, as they use key words in a way that is closer to thematic indexing. *None of the television stations have a social tagging tool.* This limits or reduces the possibility of learning what words the audience uses to search and represent the topics addressed on 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 relevance or concordance between the image and content topic is lacking.
- In 79% of the cases, *meta-description use* (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 **EfectoTV** news broadcasts. The latter case stands out because it uses a specific tag: <META = ABSTRACT>.
- 100% of the cases reflected *absence of microdata* that search engines can promote as values added on to the relevance and attraction of digital properties, shown in the snippets of the pages of results from 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 "stream-

ing", "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 were URLs 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 proposal for management and use of metadata for representation and thematic retrieval, in order to suggest ways of solving detected weaknesses. 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 journalism contents. The policy was based on the video analysis and documentary treatment model applied to the process of thematic representation of newscasts on the web.

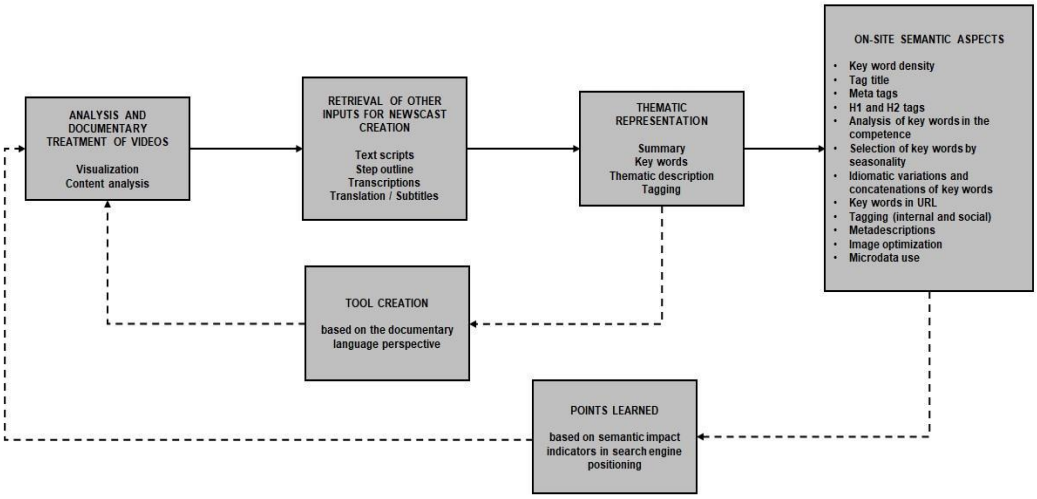


Fig. 1. Video Analysis and Documentary Treatment Model Applied to the Process of Thematic Representation of Newscasts on the Web.

- Training for web editors in metadata management and use 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 of 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, as of the intervention, it went back up. Indicators are also presented for five newscasts, where 1, 2 and 3 were means of control and 4 and 5 the intervened contents.

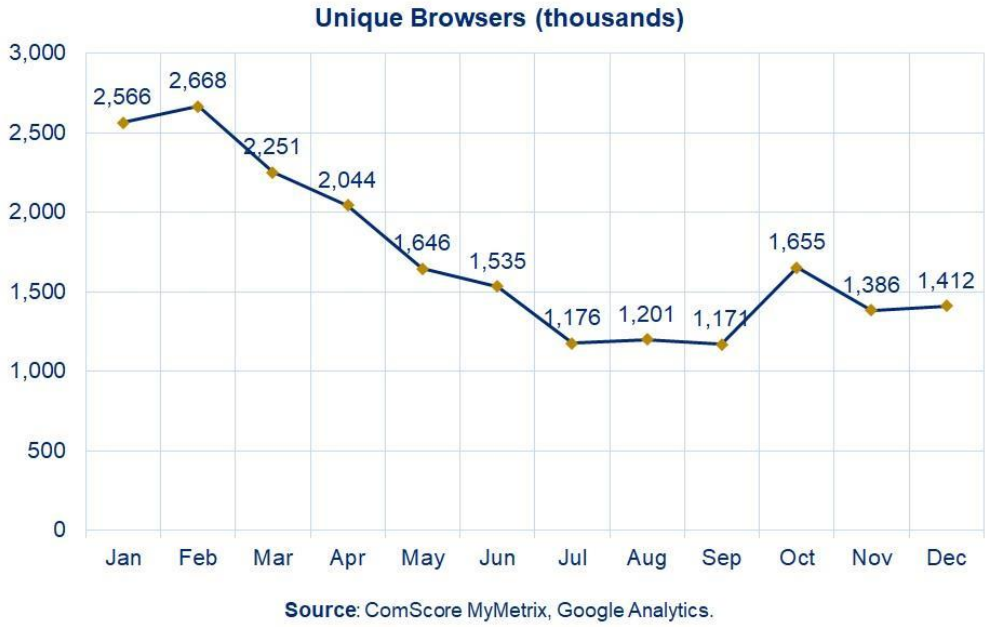


Fig. 2. Unique Browsers (thousands)

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 actually 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, in order to refine content description and thematic representation [11]

- Exploiting the value of semantic aspects to represent and retrieve content on the web [12]
- Increasing content relevance for search engines, along with their 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" [13]
- 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

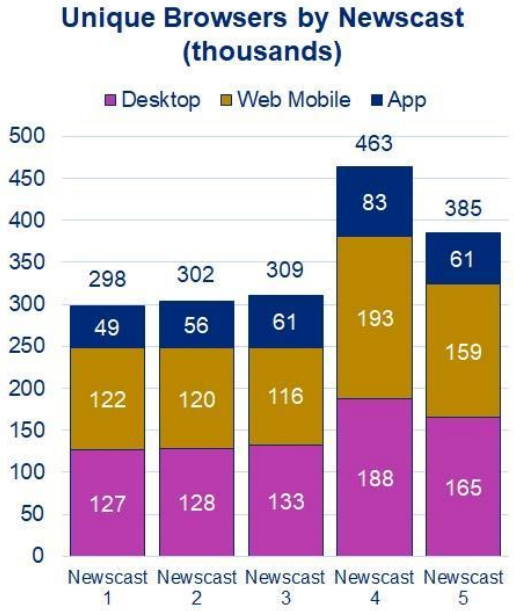
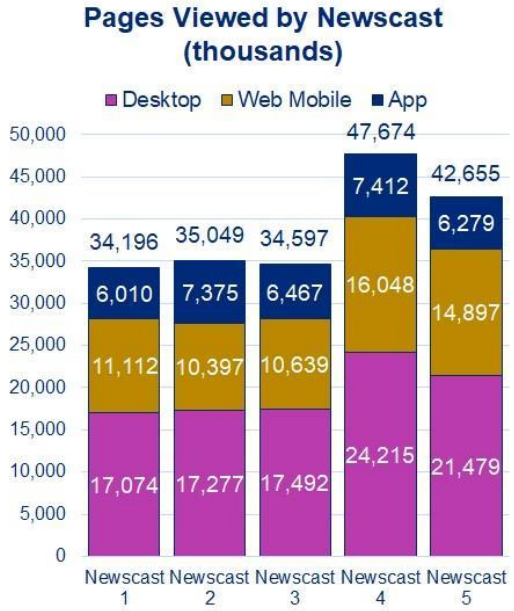


Fig. 3. Unique Browsers



Source: ComScore MyMetrix, Google Analytics.

Fig. 4. Pages Viewed

Text, audio, video and animation form a discourse whose very format calls for more specialized documentary analysis, where specific aspects are taken into account, 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 broad-

casting 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 [14]. 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.

Fig. 5. Videos Viewed

Source: ComScore MyMetrix, Google Analytics.

Fig. 6. Browsing Time

5 Final considerations

From the perspective of the worldwide television industry, this is a time for redefining plans aimed at audiences that currently take an ever-greater role as active audiovisual content users on the web than as conventional television viewers.

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 so as to be technically set up to distribute on conventional television and on web platforms and video on demand (apps and ott) that require thematic metadata for algorithms of representation, retrieval and recommendation in accordance with user habits.

Given the technological convergence that enabled the overwhelming evolution 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.

So, if "content is king", as world television leaders point out, and that is what supposedly audiences with screen independence follow, apparently Google is the 'Caesar' that 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 a content, also conditioning the retrieval and, therefore, visibility and access. This has implications from the perspective of the economy and information society, a discussion that, for practical reasons, will not be pursued here.

Despite the previous point, at least in the Mexican case, absence of *organizational information policies* regarding representation and thematic retrieval is observed. 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.

Finally, there is a demand for information professionals with competences to respond to the challenges of representation and thematic retrieval. Web editors, just as their supervisors, lack those 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

1. MacFarlane, A. Knowledge Organization and its Role in Multimedia Information Research. Knowledge Organization, 43 (3), 180-183 (2016).
2. Sobak, V., Pharo, N. Decentralized Subject Indexing of Television Programs: The Effects of Using a Semi-Controlled Indexing Language. J. of the Association for Information Science and Technology, 68 (3), 739 - 749 (2017).
3. Tsakonas, G., Papatheodorou, C. Exploring Usefulness and Usability in the Evaluation of Open Access Digital Libraries. Information Processing and Management 44, 1234–1250 (2008).
4. Engerer, V. Exploring Interdisciplinary Relationships between Linguistics and Information Retrieval from the 1960s to Today. J. of the Association for Information Science and Technology, 68 (3), 660-680 (2017).
5. Kemmis, S., McTaggart, R., Nixon, R.: The action research planner: doing critical participatory action research (p.26-28). Springer (2014).
6. Kemmis, S., McTaggart, R., Nixon, R.: The action research planner: doing critical participatory action research (p. 12). Springer (2014).
7. Comfort, L. Action Research: a Model for Organizational Learning. Journal of Policy Analysis and Management, 5 (1), 100-118 (1985).
8. Argyris, C. Inner Contradictions of Rigorous Research. Academic Press, New York (1980).

9. Kakol, M., Nielek, R., Wierzbieki, A. Understanding and Predicting Web Content Credibility Using the Content Credibility Corpus. Information Processing and Management 53, 1043 - 1061 (2017).
10. Soto-Hernández, S. Tratamiento documental: representación y recuperación de los noticieros en la Web. Doctoral thesis, UNAM, Mexico (2017).
11. Soto-Hernández, S. Tratamiento documental del video: Propuesta metodológica. Master's thesis, UNAM, Mexico (2009).
12. Soto-Hernández, S., Naumis-Peña, C. C. Análisis bibliotecológico de los noticieros televisivos mexicanos en la Web. El Profesional de la Información, 23(1), 80-86 (2014)
13. Ayadi, H., Torjmen-Khamakhem, M., Daoud, M., Xiangji Huang, J., Ben Jemaa, M. Mining Correlations between Medically Dependent Features and Image Retrieval Models for Query Classification. J. of the Association for Information Science and Technology 68 (5), 1323 - 1334 (2017).
14. Mietzner, D., Kamprath, M. A Competence Portfolio for Professionals in the Creative Industries. Creativity and Information Management, 22(3), 280-294 (2013).