VOL.40 Nº1 2023
https://doi.org/10.22491/rmct.v40i1.9267.en

Recommendations for filling geospatial data abstracts

Carolina Coutinho Salustiano Silva[1], **Ivanildo Barbosa**[1]
[1]**Military Institute of Engenharia, Praça General Tibúrcio, 80, 22290-270, Praia Vermelha, Rio de Janeiro, Brazil**
carolsalustiano@gmail.com
ivanildo@ime.eb.br

ABSTRACT: As spatial data infrastructures evolves, geospatial data producers became able to provide them to a wide scope of potential users. This user, either human or a search engine, decides to adopt that dataset based on the analysis of the correspondent metadata by comparing data characteristics and their expectations. The Brazilian Geospatial Metadata Profile (in Portuguese, PMGB) offers guidelines to support agents in charge to fill geospatial metadata. However, they often meet the expectations of web search engines. This study aims to propose guidelines to fill the metadata element Abstract to make data more attractive to both human and machine users. This study created alternative versions of Abstracts of geospatial data available at the web based on search engines optimization techniques and the PMGB guidelines. A group of expert users assessed the alternatives by considering their preferences regarding their perception of gain of information between the proposed alternatives. In total, 84.6% of respondents approved the proposed guidelines for filling the Abstract metadata element.

***KEYWORDS**: Geospatial Metadata; Search Engine Optimization; Free-Text; Abstracts*

RESUMO: Com o avanço das Infraestruturas de Dados Espaciais, os produtores de dados geoespaciais podem disponibilizá-los para um amplo número de potenciais usuários. A decisão do usuário, humano ou motor de busca na Web, por acessar esses dados se baseia, prioritariamente, na análise dos seus metadados, ponderando as características do dado disponibilizado e as suas expectativas. As recomendações de preenchimento de metadados do Perfil de Metadados Geospaciais Brasileiro (PMGB) servem como referência aos agentes responsáveis pelo preenchimento de metadados, porém nem sempre atendem aos critérios adotados pelos motores de busca na Web. O objetivo deste trabalho é apresentar propostas de diretrizes de preenchimento do elemento de metadados Resumo, com o intuito de tornar a descrição do dado geoespacial mais atrativa para usuários humanos e máquinas. Neste trabalho, foram criadas versões alternativas de resumos de dados já disponibilizados na Web, aplicando técnicas de otimização para motores de busca e as recomendações de preenchimento indicadas no PMGB. Em seguida, um grupo de usuários avaliaram a sua percepção de ganho de informação dentre as opções apresentadas. Observou-se que 84,6% dos respondentes aprovaram a sistemática de preenchimento proposta.

***PALAVRAS-CHAVE**: Metadados Geoespaciais; Otimização de Motores de Busca; Texto-Livre; Resumos*

1. Introduction

The implementation of Spatial Data Infrastructures (SDI), regardless of their hierarchical level, allowed for the optimization of processes of dissemination and user access to Geospatial Data Sets (CDG) [1]. In this context, *Catalog Services for Web* (CSW) stand out for facilitating the search for CDG based on metadata elements such as title, abstract, and keywords.

Broadening the scope to data search through web search engines, it becomes imperative that one selects words that increase the chances a search engine will present the CDG as an answer to a related query.

Logical and semantic inconsistencies may arise during metadata filling, due to a mistaken understanding of the meaning of metadata elements, concomitantly with a lack of knowledge of the dataset to be documented [1].

An example of this misunderstanding is the description of data attributes rather than the information pertaining to the data itself. In addition, the author has the possibility of perceiving the relevance of abstract contents as subjective, whereas this is an important parameter for search engine indexing and classification.

Several CDGs arouse interest for academic research, governments, companies or related activities, and the importance of an adequate filling of metadata is evident, since inadequate filling prevents the search engine from correctly locating and indexing spatial data on the Web.

A ranking was elaborated through a study by Benjelloun *et al*. [2] to list the notability of metadata elements in scientific texts published on the internet.

According to the authors, the abstract influences 100% in the discovery of a data set, occupying the first position in the list of most relevant elements.

Consequently, it makes clearer the need for proper completion of the abstract. To mitigate the problem of metadata filling, the following steps are advisable: proper documentation, including the characteristics and information of geographical data and those made available on the Web by the producers; associated with knowledge on document and data indexing strategies, in addition to understanding of search engine operation.

This work aims to propose filling guidelines for the Abstract metadata element, in the context of an IDE, so to make the CDG description more attractive to human and machine users.

After this contextualization, Section 2 presents the concepts that support the guidelines herein. Section 3 describes the methodology used to verify the gains obtained with the existing recommendations and Section 4 presents and discusses the results obtained. The fifth section presents the final considerations of the study.

2. Conceptual review

2.1 Search Engine Optimization

Metadata search engines are typically implemented on Web catalog service platforms or other specific local repositories for spatial data. In these cases, textual search occurs only between repository records, including fields such as keywords, title, and abstract, considering that querying in web search engines has become almost instinctive, since not all geospatial data of interest are part of some IDE and its respective metadata catalog.

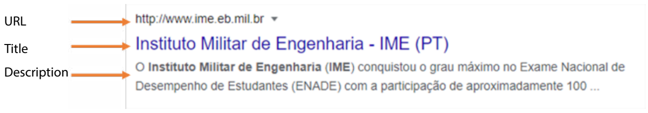
The operation of a search engine is divided into four main aspects and algorithms. First are the crawlers, which look for new content on the Web. Then comes indexing, which registers important information in the engine's search index, such as keywords. Third, a search engine that ranks and organizes hundreds of billions of web pages, analyzing factors such as query words and usability and knowledge of sources and settings. The weight applied to each factor varies according to the type of search. Fourth is the display of results, presenting the user with pages related to the search. Some preponderant factors help to determine the result, such as: understanding of natural language, which involves interpreting typos; finding commonalities frequently appearing in titles, headings, or text body; interpreting contexts, as the user's location, language preference, and search history, for example. [3] [4].

The search results that appear to the user are listed according to what is considered most relevant from the words used in the search for titles and abstracts. Inadequate filling of metadata can generate a non-coherent indexing of keywords in search engines, causing a difficulty of access.

In this context, techniques or tricks were identified to improve a system or Web page, to optimize its indexing by search engines. This process is known as "search engine optimization" (*SEO*) [5, 6, 7, 8, 9, 10].

Among the good optimization practices, known as *white hats*, we can mention the use of meta descriptors, used by search engines to display the text results of the second and third lines of search results, just below the title of the sites (Figure 1). Meta descriptors have a character limit, around 120 to 153, depending on the search engine used. They themselves do not increase the reputation of a page, but if users find what they are looking for in the text of the meta description, the chance of accessing the site increases [11].

Figure 1 - Example of a Google search using meta description *tags*



http://www.ime.eb.mil.br ▾

Instituto Militar de Engenharia - IME (PT)

O **Instituto Militar de Engenharia** (**IME**) conquistou o grau máximo no Exame Nacional de Desempenho de Estudantes (ENADE) com a participação de aproximadamente 100 ...

Source: prepared by the authors and adapted from Google.

The second best practice to highlight is to avoid identical or similar descriptions in all metadata when individual pages appear in web results. This goes against the practice of creating long *templates* for dynamic content generation, reinforcing the need to emphasize the uniqueness of each CDG in the abstract.

The development of a fluid text is also recommended, to the detriment of long sequences of keywords. First, it makes the abstract human-readable; then, it allows natural language processing algorithms to process the context and eliminate any ambiguities.

Datasets are easier to find when they provide supporting information, such as name, description, creator, and distribution formats like structured data. Based on the approach that Google applies to the discovery of datasets, the use of 'schema.org' is recommended, as well as other metadata patterns that can be added to the pages describing the datasets.

The purpose of this information is to improve discovery of datasets from fields such as geosciences, life sciences, social sciences, machine learning, civic and government data, and more. Thus, abstracts are more attractive not only to human users, but also to search algorithms that refine search results based on the interpretation of the available text [12].

Finally, the first two sentences are often displayed in search engine results. Therefore, making them attractive, with relevant keywords, encourages people to click through to the page to read the content in its entirety. Ideally, one should try to repeat these keywords three to six times in the abstract, maintaining the readability of the text [13]. The naturalness of the textual production must be maintained, in addition to being composed of clear and concise central points, respecting the limit between 50,000 and 5,000 characters [12].

This function is intuitive for queries by text documents, image captions and video descriptions. In the case of CDG, the initiatives for indexing CSW services and map servers aim to make such content visible to search engine indexers. An example is the implementation of the *Geosearch* module on *Geoserver* map servers [14]. Another solution is the generation of web pages with content based on the title, abstract, and keyword elements, for publication and indexing by search engines.

2.2 Geospatial metadata

The most simplistic concept of metadata is that it consists of the description of a given data [15]. The purpose of metadata is to document and organize, in a systematic and structured way, the data of organizations to facilitate their sharing and maintenance, discipline the production and storage of data, and guide the use of data in different applications.

The concept extends to bibliographic and objects cataloging in digital format. Different metadata schemes have been proposed to suit the characteristics of the described objects, such as the *International Standard Bibliographic Description* (ISBD) [16] and the *Dublin Core Metadata Initiative* (DCMI) [17].

In the context of geospatial data, different metadata profiles have been proposed: *Content Standard for Digital Geospatial Metadata* (CSDGM) [18], ISO 19115 [19], and several national profiles drawn from the latter. The Brazilian Geospatial Metadata Profile (MGB Profile) [20] is the standard adopted in Brazil, based on [19], and recently adapted to suit the updates made by ISO 19115 in 2014 [21].

Each profile presents dozens of metadata elements that aim to describe technical, legal, and identification aspects, so that users can query the repository content to discover the CDGs related to their expectations. Some of these elements have values that can be automatically filled — the spatial extension and the reference coordinate system of a vector file, for example. A second set of elements have a controlled domain, i.e., they can only be filled with values defined in a pre-established list. Finally, another set of elements are filled in free text format, which may give rise to subjectivity on the part of the agent who fills in the metadata.

In section 2.1, two metadata elements with great influence on the search for resources on the Web were mentioned: keywords and abstract.

The keyword element is designated to describe a feature, its aspect, or the source. The selection of terms can be facilitated by using the controlled list, MD_TopicCategoryCode [19, 20], which has a category code, as they contain defined themes and taxonomy. Another way to favor the choice of keywords is the use of a lexicon or *thesaurus* [20].

The Abstract element is defined as a brief summary of the resource [19, 20, 21, 22] and "must synthesize the fundamental aspects of the resource in terms of content, geographical extent, date, scale, series name, producer or responsible entity, sources used, etc." [20]. The subjectivity inherent in the term "brief" can induce overly simplified filling, failing to register content that can be found in keyword searches. On the other hand, long-winded summaries may omit relevant information to the detriment of aspects that do not add value to the search, in addition to unnecessarily taking up more storage space. Below, respectively, are examples of overly simplified and wordy summaries:

• The Digital Elevation Model, which is part of the RJ-25 project, represents the numerical model of the surface altimetric characteristics, articulated by sheets according to the framework of the Brazilian systematic mapping. It covers a geographical square of 07'30' 'latitude by 07'30'' longitude [23];

• "URBAN WATER SERVICE INDEX – IN023. Indicator of the National Sanitation Information System SNIS. Calculation formula. IN023 = AG026_R / G06a * 100. Percentage unit. Reference Year 2011". [24].

"Several states today monitor the quality of surface water in their territory and pass on the data to ANA. However, from a national perspective, it is not always possible to compare the data generated, since the states adopt different criteria, methodologies and parameters, and there is no standardization on a Country scale. The National Water Quality Monitoring Network (RNQA) is the main component of the National Water Quality Assessment Program (PNQA), and its main objective is to standardize and expand monitoring in the country, eliminating existing temporal and geographical gaps. The points of RNQA were determined based on a point allocation methodology developed by ANA and were later analyzed together with all states and the Federal District to seek to take advantage of the monitoring points of existing networks. In addition, the ANA is responsible for the operation of the National Hydrometereological Network, which contains fluviometric stations and generates river flow information throughout the country. In part of these stations, approximately 1600, there is also the monitoring of four water quality parameters measured with multi-parametric probes (Dissolved Oxygen, Turbidity, Temperature and pH)" [25].

2.3 Related works

Studies were developed describing the SEO line.

Cahill and Chalut [5] examined techniques used by marketing for optimization: the different optimization tactics between the terms "white hat" and "black hat" were observed, and why it was important for librarians to understand these techniques and the impact on search engine results pages. They also looked at ways library staff could help their users develop awareness of the factors influencing search results and better assess quality and relevance on the results page.

Shih, Chen, and Chen [6] developed a search engine optimization that could be used by a company. Social networking sites were included in the Internet marketing strategy. The proposed technique was applied in the operations of an online e-bookstore. Website rankings were monitored in two search engines: Google and Yahoo. The results revealed that a well-designed SEO, with the incorporation of social networks, can effectively increase website visibility and exposure.

Zilincan [7] looked at the most important factors that can help improve placement in search results. He points out that no technique can guarantee high ranking, because search engines have sophisticated algorithms that measure the quality of web pages and derive their position in search results. Zilincan also developed a website for the purpose of implementing and testing key SEO techniques. Then, the relevant optimization factors that influenced the search engine increased the ranking of your site, in addition to subsequently verifying higher traffic.

Katumba and Coetzee [10] identified and categorized the search terms typically employed by users when searching for geospatial resources on the Web. Guided by these terms, metadata on geospatial sources was published "directly" on the Web and empirical tests were performed with search engine optimization (SEO) techniques. Two sets of HTML pages were prepared and registered in Google and Bing, respectively. The metadata in one set was tagged with Dublin Core, the other with Schema.org.

3. Methodology

The methodology used was divided into four parts: extraction of metadata from the INDE repository, analysis of the filling structure, analysis of the filling structure used in the INDE, and compilation of recommendations. The validation of results obtained consisted of the evaluation performed by users, choosing the option with the highest semantic representativeness.

3.1 Extraction of metadata

The process of extracting metadata from the INDE catalog was based on the script developed and documented in [26]. The identifiers, titles, and abstracts of 5,808 metadata stored in the INDE repository on 11/05/2020 were extracted. CSW services allow the preparation of HTTP requests, receiving responses in XML format, which can be interpreted and stored in a structured way. In this work, data were stored in comma-separated value format.

3.2 Analysis of the filling structure

The MGB Profile is the Brazilian normative reference for filling in geospatial metadata. As mentioned in Section 2.3, it specifies the information on some aspects of the data that should be included in the summary. In addition, this Profile, as well as dozens of other initiatives around the world, is based on the specifications of ISO 19115 [19]. The same occurs in international standards such as INSPIRE [27, 28] and IDE Espanha [29], which were chosen by this study, due to greater maturity and, consequently, availability of documentation with a higher level of detail and greater adherence by member countries.

3.3 Analysis of the filling structure used in INDE

This step aims to understand what content the agents responsible for filling in geospatial metadata use in the preparation of summaries.

One hundred and four (104) metadata were selected from the population of 5,808 metadata extracted from the catalog in Section 3.1. The selection took place based on the diversity of types and themes of the data; varied producing institutions and, mainly, an attempt to avoid duplication of abstracts.

This sample size implies a 95% confidence level with a margin of error of 9.4%. However, it was observed that some producers followed filling templates for dozens of products, which could influence the statistics of model identification.

The standardization of metadata filling in the abstracts was analyzed based on the requirements suggested by the PMGB. The results of this stage were obtained from the presence of the following factors in abstracts: geographical area, date, scale, series name, producer, and sources used. Each aspect of this filling structure was analyzed and classified with the following criteria: a) fully meets the requirement, or b) partially meets the requirement, or c) does not meet or not found.

3.4 Compilation of recommendations

The recommendations were divided into two groups: in terms of form and content.

As for the form, the concatenation of the items for the preparation of the abstract text followed the SEO recommendations whenever possible, by using the most relevant keywords with a repetition of 3 to 6 times naturally in the abstract, including the main keywords in the first two sentences of the abstract, avoiding oblique and wordy texts and avoiding repetition of abstract templates for different geospatial data in an IDE.

As for the content, Table 1 compares the filling contents in the abstracts. Items presented in it are assumed to be the basis of the recommendations.

Table 1 - Incidence of requirements on published recommendations

Item	PMGB	INSPIRE	IDE Espanha
Geographical area	X	X	X
Date	X	X	
Scale	X	X	
Sources Used:	X	X	X
Work Importance			X
Key Attributes			X
Producer	X	X	
Legal references			X
Grade	X	X	

Source: Prepared by the authors.

Some adaptations were necessary to meet the recommendations. In short, the content of the prepared abstract should bring together nine elements:

- Theme – subject to which it refers, or seeks to develop, or the proposition to be addressed. Main theme of the CDG;
- Product – result of the survey, i.e., what was produced;
- Spatial framework – result of the survey, i.e., what was produced;
- Time frame – main locations according to the scale: less than 1:10,000,000 (country); less than 1:5,000,000 (states and capitals); less than 1,000,000 (cities with more than 1,000,000 inhabitants);
- Scale – denominator compatible with the dimension of the smallest detail representable in the data. Use the scale for vector CDGs. For matrix data, use the scale compatible with the smallest detail representable on the ground or the spatial resolution, expressed in meters;
- Series name (applicable to a series or data collection) – name of a cartographic series, sheet-by-sheet documents or fieldwork documents, for the constitution of a given resource. The name is applicable to a series or collection of data;
- Producer and/or entity responsible for the institution and/or responsible for the geospatial data;
- Sources used – origin of spatial data, for example aerophotogrammetric survey, charts, maps, images, mosaics, cartographic bases, etc.;

Contextualization (which helps to understand the data produced) – describes the purpose of the data, presents a context for the creation of the data.

3.5 Validatio

Validation aims to evaluate the gain obtained with the use of the proposed completion guidelines, i.e., the perception of completeness of the information contained in the abstract when compared to the original texts.

To this end, a questionnaire was developed according to the self-explanation model, in which the form is delivered to the respondents to be completed without the intervention of the researcher [19, 30].

The questions within each section were designed following complexity and respondents' reflection levels [30]: the initial questions in each section asked for quick answers, while the final ones were more reflective, complex, and abstract. The form was released during a scientific event and published through emails and publications in groups of geotechnologies users on *Facebook* and *Whatsapp*. Responses were received for approximately two months (between November 2020 and January 2021).

The form was developed in *Google Forms* and divided into two fundamental sections: characterization of respondents and validation of the abstract filling systematization.

Respondent characterization aimed to quantify the expertise of the respondents regarding academic training and time of experience in the use of geospatial data. The following are the main questions of the form for the characterization:

1. To which of the categories below do you belong?

- Undergraduate Student;
- Graduate Student;
- Technical Course Student;
- Professor/Researcher;
- Professional/Producer;
- Professional/User;
- Others.

2. How long have you been in contact with geospatial data, either in your training or professional experience (years)?

- 0 – 1;
- 2- 5;
- 6 – 10;
- More than 10.

In the second section, respondents compared the abstracts of three products originally made available in the INDE metadata catalog with other versions, written in accordance with existing guidelines in the PMGB (see section 2.2) and the ones compiled in this study. The aim of the analysis of responses obtained in the form is to verify whether the abstracts prepared based on the proposed structure are clearer and more representative, compared to the abstracts based on the PMGB guidelines or those originally filed in the INDE.

If the original abstract did not contain all the items necessary for the preparation of the alternative abstracts, it would be necessary to manually locate the complete metadata in the INDE repository to consult the content of other elements.

The abstracts were rewritten, keeping the original texts of the INDE and adding the missing information from the PMGB guidelines in one version, and adding information according to the recommendations suggested in this study in the other.

This measure was taken to minimize the influence of the form of writing from one author to another, so that the respondent could exclusively evaluate the filling structure of content presented in the abstract. The following are the main form questions for the abstract comparison section.

3. Which Abstract option do you consider the most representative and clear?

- “The Digital Elevation Model, which is part of the RJ-25 project, represents the numerical model of the surface altimetric characteristics, articulated by sheets according to the framework of the Brazilian systematic mapping. It covers a geographical square of 07’30’ ’latitude by 07’30’’ longitude [23];
- “The Digital Elevation Model, which is part of the RJ-25 project, represents the numerical model of the surface altimetric characteristics, articulated by sheets according to the framework of the Brazilian systematic mapping. It covers a geographical square of 07’30’ ’latitude by 07’30’’ longitude in the locality of Folha de Guaçuí. This feature was created in 2008/08/08 with a scale of 1:25,000 and the name of the series is 1:25,000 Digital Elevation Model, developed at IBGE by the Cartography Coordination. Aerial photographs obtained from an aerial survey carried out by the company Base Aerofotogrametria e Projetos S.A. were used as a data source;
- The Digital Elevation Model, which is part of the RJ-25 project, covers a geographical square of 07’30’ ’latitude by 07’30’’ longitude, covering the area of Folha Guaçuí. This feature was created in 2008/08/08 with a scale of 1:25,000 and the name of the series is 1:25,000 Digital Elevation Model, developed at IBGE by the Cartography Coordination. Aerial photographs obtained from an aerial survey carried out by the Base Aerofotogrametria e Projetos S.A. company were used as a data source. This survey aims to represent, through a numerical model, the altimetric characteristics of the surface.

4. Which Abstract option do you consider the most representative and clear?

- Several states today monitor the quality of surface water in their territory and pass on the data to ANA. However, from a national perspective, it is not always possible to compare the data generated, since the states adopt different criteria, methodologies and parameters, and there is no standardization on a Country scale. The National Water Quality Monitoring Network (RNQA) is the main component of the National Water Quality Assessment Program (PNQA), and its main objective is to standardize and expand monitoring in the country, eliminating existing temporal and geographical gaps. The points of RNQA were determined based on a point allocation methodology developed by ANA and were later analyzed together with all states and the Federal District to seek to take advantage of the monitoring points of existing networks. In addition, the ANA is responsible for the operation of the National Hydrometereological Network, which contains fluviometric stations and generates river flow information throughout the country. In part of these stations, approximately 1600, there is also the monitoring of four water quality parameters measured with multi-parametric probes (Dissolved Oxygen, Turbidity, Temperature, and pH)

- Several states today monitor the quality of surface water in their territory and pass on the data to ANA. However, from a national perspective, it is not always possible to compare the data generated, since the states adopt different criteria, methodologies and parameters, and there is no standardization on a Country scale. The National Water Quality Monitoring Network (RNQA) is the main component of the National Water Quality Assessment Program (PNQA), and its main objective is to standardize and expand monitoring in the country, eliminating existing temporal and geographical gaps. The points of RNQA were determined based on a point allocation methodology developed by ANA and were later analyzed together with all states and the Federal District to seek to take advantage of the monitoring points of existing networks. In addition, the ANA is responsible for the operation of the National Hydrometereological Network, which contains fluviometric stations and generates river flow information throughout the country. In part of these stations, approximately 1600, there is also the monitoring of four water quality parameters measured with multi-parametric probes (Dissolved Oxygen, Turbidity, Temperature and pH). These data comprise the Brazilian territory. The date identifies when the appeal was issued on 2016-03-22, with the scale 1:1000000, developed by the National Water Agency;
- The points of the National Water Quality Monitoring Network (RNQA) were determined based on a point allocation methodology developed by ANA and were later analyzed together with all states and the Federal District to seek to take advantage of the monitoring points of existing networks. The appeal was issued on 2016-03-22, with the scale 1:1000000, developed by the National Water Agency. This survey aims to provide geospatial information on the panorama of water resources in Brazil.

5. Which Abstract option do you consider the most representative and clear?

- URBAN WATER SERVICE INDEX – IN023. Indicator of the National Sanitation Information System SNIS. Calculation formula. IN023 = AG026_R / G06a * 100. Percentage unit. Reference Year 2011. The research covered 4864 municipalities, making up 87% of the possible sample universe. The scale of these data is 1:2500000, developed by the Ministry of Cities. This survey aimed to generate information necessary for actions related to future land use;
- URBAN WATER SERVICE INDEX – IN023. Indicator of the National Sanitation Information System SNIS. Calculation formula. IN023 = AG026_R / G06a * 100. Percentage unit. Reference Year 2011. The research covered 4864 municipalities, making up 87% of the possible sample universe. The scale of these data is 1:2500000, developed by the Ministry of Cities.
- URBAN WATER SERVICE INDEX – IN023. Indicator of the National Sanitation Information System SNIS. Calculation formula. IN023 = AG026_R / G06a * 100. Percentage unit. Reference year 2011.

6. What items do you think are important in the abstract? (There is no limit of options for the answer)

- Theme;
- Spatial framework;
- Time frame;
- Date;
- Series name;
- Scale;
- Product;
- Producer and responsible entity;
- Sources used;
- Context;
- Status;
- Reference system;

- Distribution format;
- Type of representation;
- Language.

The respondent was encouraged to contribute to the construction of abstract recommendations, opening the opportunity for interaction based on a subjective question, which allowed the inclusion or removal of information different from those mentioned in previous questions.

7. Would you like to add any additional information regarding the completion of the Abstract?

To quantify the users' acceptance of the proposal to systematize the filling structures, a scoring metric was established: each question or alternative accepted by the user was assigned a value of 1 (one) point; in the answers that generated doubts about the user's opinion, the value of 0.5 (half) point was assigned; finally, 0 (zero) for those that were not accepted by the respondents. The validated abstracts were considered and the systematization of the abstract filling by the users was approved, when the score was equal to or greater than 70%, this percentage represents how much the proposals made in this study were accepted by the respondents.

Complementing the validation of the proposed recommendations, the respondent was also consulted on the relevant information to include in an abstract. No option limits were established to be chosen. The nine recommended items were presented as options, as well as five items unrelated to the recommendation: status, reference system, distribution format, type of representation, and language. These items were selected among elements of the summarized MGB profile [13].

The metadata element Data is present in Table 1 and, despite its temporal character, has a more limited meaning than the time frame, designed to characterize periods represented in the geographical data instead of a single reference date. The question was presented so that respondents did not know how to distinguish the recommended items from the additional items.

4. RESULTS OBTAINED

From the main recommendations for the Abstract (Table 1) in the PMGB, the following requirements were extracted: geographical area, date, scale, series name, producer, and sources used. This information was evaluated and it was verified if these elements appeared in the abstracts, presented in Table 2.

Table 2 - Incidence in percentage of the requirements for summary in the PMGB in the INDE metadata analyzed

Element	Requirement	INDE
Abstract	Geographical area	49%
	Date	22%
	Scale	33%
	Series name	17%
	Producer	23%
	Sources Used	31%

Source: Prepared by the authors.

One can see that geographical area, scale, and sources used are the most found, however with percentages below 50% of incidence. Based on this result, it is worth mentioning the importance of a greater dissemination of good metadata filling practices among geospatial data producers.

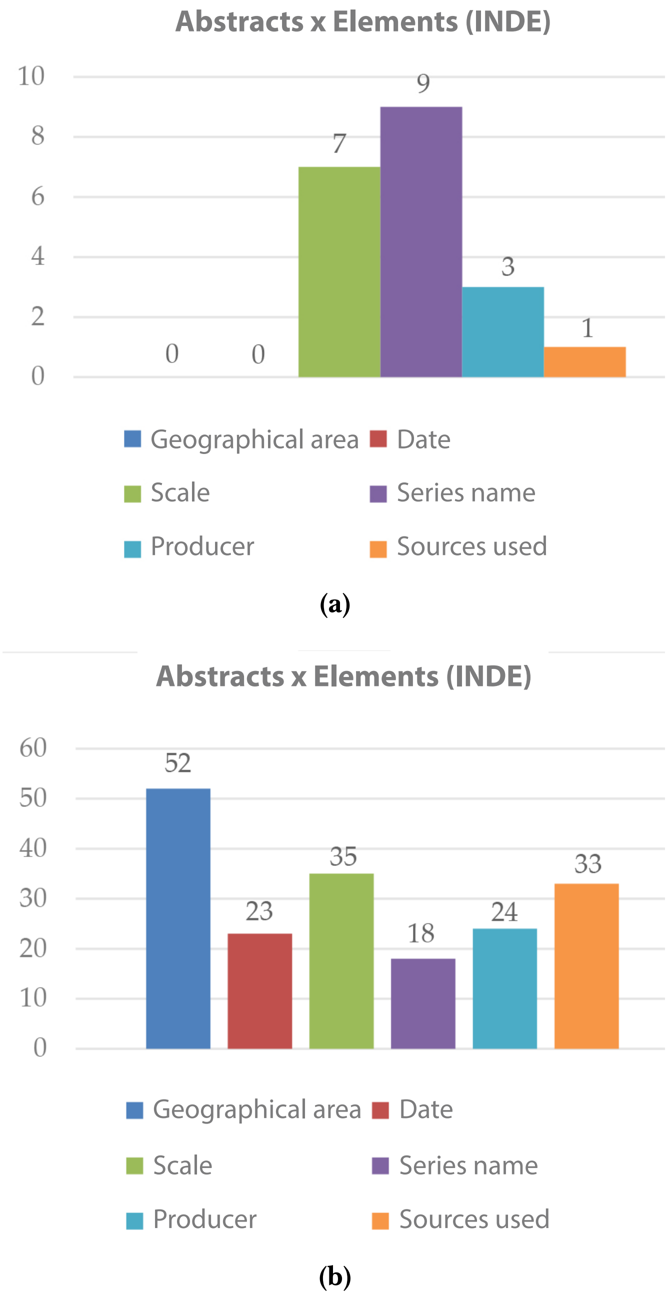
The analysis of abstracts of the metadata sample analyzed indicates that not all recommended items were met (Figure 2), but there were also unsuggested items such as the periodicity of data updating, the calculations and descriptions of the methodology used in the data attributes, the names of projects mentioned instead of the series name, in addition to explaining the operation of these projects, among others.

Seventy-five responses to the questionnaire were obtained. Despite being a quantity that makes the representativeness of the results obtained questionable, the profile of the respondents is composed of 66% of teachers, researchers and professionals, and about 60% of people with 10 or more years of experience with geospatial data.

The respondents were 29 teachers/researchers, 16 graduate students, 13 professionals/producers, 9 undergraduate students, 7 professionals/users, 1 high school teacher, and no technical course students.

Regarding the time of contact with geospatial data by training or professional experience, 44 reported having more than 10 years, 18 reported having between 6 and 10 years, 11 reported having between 2 and 5 years and 2 reported having between 0 and 1 year.

Figure 2 - Comparison between the items suggested by the PMGB in the abstracts referring to (a) systematic and (b) thematic data



(a)

(b)

Source: Prepared by the authors.

The characterization of respondents included a group with different levels of training and experience. However, most were research professors, with a level of experience of more than 10 years. The qualification of respondents and the time of interaction with geospatial data demonstrate reliability in the responses received.

Results of the comparison between the evaluated abstracts is shown in Table 1. Each row corresponds to the respective abstract, and the columns indicate the number of times each alternative was selected. One can observe that the abstract prepared following the recommendations was selected more frequently, in all cases. However, the abstract prepared only with the recommendations of the MGB profile already shows a significant gain in two of the three abstracts evaluated. The third summary stands out from the other two by originally describing the context in detail, being the only one to have alternatives of reduced size.

Chart 1 - Summary of responses regarding users' preference for the abstracts presented

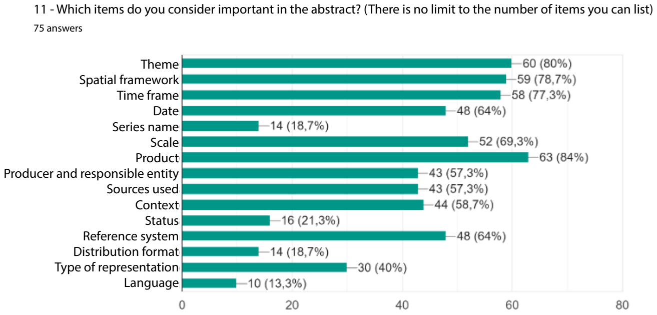
Responses / Abstract	INDE (responses)	PMGB (responses)	Alternative (responses)
1	15	29	31
2	11	30	34
3	19	17	39

Source: Prepared by the authors.

According to the score established to quantify the acceptance of the proposals, abstracts 1 and 2 received 0.5 points, as the difference between the values of the second and third columns are almost identical. However, in abstract 3, the alternative abstract based on the recommendations was selected as preferred by more than half of the respondents. This means assigning 1 point to this abstract.

The second stage of validation consisted of consulting the opinion of users regarding the items deemed important in the elaboration of a clear and comprehensive abstract. The graph illustrated in Figure 3 contains the compilation of the responses.

Figure 3 - Representation of the choice of the most important requirements for the abstract



Source: Prepared by the authors.

In the fourth quartile of responses, with the items most selected by the respondents, there are the items product, theme, spatial framework, and time frame (more than 77% of selections), all included in the recommendations. In the third quartile of responses, there are the items scale, date, and reference system (more than 59% of selections). With the exception of the reference system, these items are among the recommendations specified in the MGB Profile. It should be noted that the users treated the time frame item (proposed as the period represented in the geographic data) differently from the date (proposed as a metadata element indicated in Table 1, referring to a single date, which may be the date of publication of data or the oldest input used). The second quartile of responses (more than 31% of selections) includes the elements of producer, sources and contextualization (recommended in the PMGB specifications) and the representation format (not mentioned in those recommendations). Items such as series name, distribution format, status, and language were the least selected.

In analyzing this result, one may notice that not all recommended elements were the most voted by the respondents, so that the items Date and Reference System are among the most indicated items, to the detriment of the items Sources and Contextualization.

The validation of the abstracts accumulated 10 out of 13 points and an acceptance with the respondent of 84.6%.

5. Conclusions

The movement of geospatial open data has increasingly motivated data sharing. Therefore, the way metadata is described has become paramount, as it is directly related to the search engine's ability to locate the material made available on the internet.

Currently, the metadata filling instructions presented in the MGB Profile are not sufficient to efficiently describe the interesting characteristics of the product to its users. In addition, some factors are considered harmful in this process:−producers who use the PMGB recommendations based on their own criteria, i.e., subjectively; the recommendations need to be reviewed periodically to monitor technological developments, which are increasingly updated; the diverse environment of INDE, in which the multiplicity of consumers and data producers was observed, with different training and expertise; among other factors.

This study aims to propose guidelines for filling in the Abstract metadata element, in the context of an IDE, to make the description of the CDG more attractive to human and machine users.

To this end, the recommendations for filling in the abstracts in the PMGB were raised, criteria for analyzing the abstracts made available in the INDE catalog were established, and guidelines for filling in the abstracts based on SEO techniques in the PMGB and in the abstracts analyzed by INDE were developed. The analysis indicated the need for producers to insert in the abstract content that mentions theme, product, spatial framework, time frame, scale, series name, producer, sources used, and contextualization.

Considering that search engines are in constant change, in which they evoke a continuous refinement for SEO techniques, the main focus was directed at the quality of the content useful to the audience to be reached. In concrete terms, sharing information relevant to one who generated the data and one who will use it. In the responses obtained from the consumers and producers consulted, there was a gain in representativeness in the proposed alternative abstracts, i.e., the respondents validated the proposed recommendations with the preference of the product and theme requirements.

Based on this knowledge, metadata filling guidelines were developed for geospatial data summaries with a simple and natural language for users and aimed at meeting the demands of search engines. Such recommendations can be employed in developing solutions for suggesting abstracts for geospatial resources published in an IDE or in a catalog of geospatial products on the Web.

Finally, it should be noted that this topic requires constant updating, since the volume of data made available increases daily and the profile of producing and consuming agents changes continuously. Some improvements in the method presented in this work include the use of complementary analysis techniques, as well as the experimentation of other databases, in the deepening and addition of geospatial metadata elements, in the development of natural language processing techniques, machine readability compared to natural language, in the automation of abstracts, among others.

References

[1] IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. *Acesso e uso de dados geoespaciais, Manuais técnicos em Geociências*. n. 14. Rio de Janeiro: IBGE, 2019. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101675.pdf Acesso em: 20 out. 2021.

[2] BENJELLOUN, O.; CHEN, S.; NOY, N. Google Dataset Search by the Numbers. *arXiv, Cornell* University, Nova York, 2020. DOI: https://doi.org/10.48550/arXiv.2006.06894

[3] CENDÓN, B. V. Ferramentas de busca na Web. *Ciência da Informação*, [*s. l.*], v. 30, n. 1, p. 39-49, 2001.

[4] GOOGLE. Home Search Work. *Google*, [*s. l.*], 2019. Disponível em: https://www.google.com/search/howsearchworks/crawling-indexing/ Acesso em: 18 dez. 2019

[5] CAHILL, K.; CHALUT, R. Optimal results: what libraries need to know about google and search engine optimization. *The Reference Librarian*, [*s. l.*], v. 50, n. 3, p. 234–247, 2009. DOI: https://doi.org/10.1080/02763870902961969

[6] SHIH, B.-Y.; CHEN, C.-Y.; CHEN, Z.-S. Retracted: An empirical study of an Internet Marketing Strategy for Search Engine Optimization. *Human Factors and Ergonomics in Manufacturing & Service Industries*, [*s. l.*], v. 23, n. 6, p. 528–540, 2012. DOI: https://doi.org/10.1002/hfm.20348

[7] ZILINCAN, J. Search engine optimization. *CBU International Conference Proceedings*, [*s. l.*], v. 3, p. 506–510, 2015. DOI: https://doi.org/10.12955/cbup.v3.645

[8] TAYLOR & FRANCIS. *Writing your paper*. Milton Park: Taylor & Francis, 2021. Disponível em: https://authorservices.taylorandfrancis.com/resources/writing-paper-ebook/. Acesso em: 2 mar. 2021.

[9] GABRIEL, M. *Marketing na era digital*: conceitos, plataformas e estratégias. São Paulo: Novatec, 2010.

[10] Katumba, S.; Coetzee, S. Empregando técnicas de otimização de mecanismos de busca (SEO) para melhorar a descoberta de recursos geoespaciais na Web. *ISPRS International Journal of Geo-Information*, [*s. l.*],v 6, n. 9, p. 284, 2017. DOI: https://doi.org/10.3390/ijgi6090284

[11] ROCK CONTENT. SEO 2.0: o guia definitivo e atualizado para conquistar a primeira página do Google. Ebook. Disponível em: https://rockcontent.com/materiais-educativos/seo-o-guia-definitivo-da-rock-content/. Acesso em: 10 ago. 2019.

[12] CENTRAL DA PESQUISA GOOGLE. Conjunto de dados. *Google*, [*s. l.*], 2021 Disponível em: https://developers.google.com/search/docs/advanced/structured-data/dataset. Acesso em: 27 abr. 2021.

[13] TAYLOR & FRANCIS. *Writing your paper*. Abingdon: Taylor & Francis, 2021. Disponível em: https://authorservices.taylorandfrancis.com/resources/writing-paper-ebook/. Acesso em: 27 abr. 2021.

[14] OSGeo – OPEN SOURCE GEOSPATIAL FOUNDATION. Layer. 2024. Disponível em: https://docs.geoserver.org/stable/en/user/data/webadmin/layers.html. Acesso em: 30 ago. 2024.

[15] PRADO, B. R.; HAYAKAWA, E. H.; BERTANI, T. C.; SILVA, G. B. S.; PEREIRA, G.; SHIMABUKURO, Y. E. Padrões para metadados geográficos digitais: modelo ISO 19115:2003 e modelo FGDC. *Revista Brasileira de Cartografia*, Uberlândia, v. 62, n. 1, p. 33-41, 2010. DOI: https://doi.org/10.14393/rbcv62n1-43665

[16] IFLA – International Federation of Library Associations and Institutions. ISBD International Standard Bibliographic Description. *IFLA*, [*s. l.*], 2011. Disponível em: https://repository.ifla.org/bitstream/123456789/786/1/ifla-isbd-international-standard-bibliographic-description-2011.pdf. Acesso em: 14 out. 2019.

[17] DCMI – Dublin Core Metadata Initiative. 2022. Disponível em: https://dublincore.org/specifications/dublin--core/. Acesso em: 16 mar. 2022.

[18] FGCD – Federal Geographic Data Committee. Content Standard for Digital Geospatial Metadata. *FGCD*, Washington DC, 1998. Disponível em: https://www.fgdc.gov/standards/projects/metadata/base-metadata/v2_0698.pdf. Acesso em: 30 jun. 2019.

[19] ISO 19115. Geographic Information – Metadata. 2003.

[20] CONCAR – Comissão Nacional de Cartografia. *Perfil de Metadados Geoespaciais do Brasil*. Perfil MGB. Brasília, DF: Ministério do Planejamento, 2009.

[21] ISO 19115. Geographic Information -Metadata. 2014. Disponível em: https://www.iso.org/standard/53798.html. Acesso em: 23 mai. 2019.

[22] DCMI – Dublin Core Metadata Initiative. 2022. Disponível em: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/abstract/https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/abstract/. Acesso em: 12 jul. 2022.

[23] IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA *Modelo Digital de Elevação 1:25.000 - GUAÇUÍ SF-24-V-A-IV-4-NO 2613-4-NO*. 2008. Disponível em: https://metadados.inde.gov.br/geonetwork/srv/por/catalog.search#/metadata/ea5778a6-e4db-495a-9b1c-f7a22976921f. Acesso em: 30 ago. 2024.

[24] MINISTÉRIO DO PLANEJAMENTO, ORÇAMENTO E GESTÃO. *Índice de atendimento urbano de água com rede de abastecimento - Valor realizado*. 2011. Disponível em: https://metadados.inde.gov.br/geonetwork/srv/por/catalog.search#/metadata/e466791c-0a49-4a8f-b8fe-e31f8a16bf15. Acesso em: 30 ago. 2024.

[25] AGÊNCIA NACIONAL DE ÁGUAS. *Evolução da Rede de Monitoramento de Qualidade da Água*. 2015. Disponível em: https://metadados.inde.gov.br/geonetwork/srv/por/catalog.search#/metadata/5f5da94d-f61b-4706-857f-63f7152618eb. Acesso em: 30 ago. 2024.

[26] GOTTARDO, T. V. *Proposta de metodologia para avaliação de ide através de indicadores sobre a disseminação de dados especiais*. Rio de Janeiro: Instituto Militar de Engenharia, 2018.

[27] SILVA, H.; SERRONHA, A. *Metadados Inspire*. Portugal: Direção-Geral do Território, CCDR-LVT, CCDR Algarve, CCDR Norte, CCDR Centro, CCDR Alentejo, 2015.

[28] INSPIRE – Infrastructure for Spatial Information in Europe. *Technical Guidance for the implementation of INSPIRE dataset and service metadata based on ISO/TS 19139:2007*. [*S. l.*]: INSPIRE, 2017.

[29] GOBIERNO DE ESPAÑA, Catálogo de Datos y Servicios IDEE. Disponível em: https://www.idee.es/csw-inspire-idee/srv/por/catalog.search;jsessionid=9F3F01DDBFAD57D7FDB569FCD810EA47#/home. Acesso em: 30 ago. 2024.

[30] VIEIRA, S. *Como elaborar questionários*. São Paulo: Atlas, 2009.