Amazônia and Amazon
Domain Analysis with Iramuteq in Scopus and LISA Databases

Abstract:
The study reports the comparative analysis between the results of the search queries for the terms Amazônia and Amazon in Scopus and LISA databases, in the period from 2008 to 2018. Concept Theory and Domain Analysis were used in conjunction with IRaMuTeQ software in order to identify, quantify and analyse semantic distances in a sample consisting of 80 abstracts from retrieved articles.

1.0 Introduction
Inspired by the decision of ICANN - Internet Corporation for Assigned Names and Numbers, which manages the Internet address system, it was decided on 21st May 2019 that the technology company Amazon has the right to use the internet domain “amazon.com” with its variations, despite protests from South American countries that shelter the Amazon rainforest, including Brazil.

This empirical research was developed based on Hjørland's Domain Analysis (2002) and Dahlberg’s Theory of Concept (1993) to investigate if the dispute between the terms Amazon and Amazônia would have any influence on the results of search expressions in the retrieval of documents in large databases as Scopus and LISA.

According to Dahlberg, “any organization of knowledge must be based on knowledge units”, that is, on concepts. Concepts are knowledge units that form the elements of knowledge systems. The concept cannot be “represented unless it is presented by knowledge units and their many possible combinations in words/terms or statements”. (Dahlberg 1993, 211).

In this sense, Hjørland argues that "it should be possible to develop the concept of the semantic distances as investigated by Brooks (1995, 1998), and to consider distances between groups and between queries and document representations". (Hjørland 2002, 445).

The present study conjugates the concept of semantic distances with the similitude analysis (graph theory) performed by IRaMuTeQ (Interface "R" for Multidimensional Analysis of Texts and Questionnaires) software applied to the terms Amazônia and Amazon as search queries.

2.0 The method
This research was based on a sample of 20 most cited articles retrieved from the same query search for Amazônia and Amazon terms in Scopus and LISA, from 2008 to 2018. The IRaMuTeQ textual corpus comprised a total of 80 analysed abstracts.

The similitude analysis results were compared based on the Domain Analysis and Theory of Concept approach. Subsequently, we also considered the applications of FAIR principles - Findable, Accessible, Interoperable, Reusable (Wilkinson et al. 2016).
Despite the existence of differences between the Scopus and the LISA databases in relation to their retrieval structure, vocabulary control, and the number of indexed journals, it was possible to analyse semantic distances assigned to the terms Amazônia and Amazon as knowledge units. The four questions indicated by Hjørland were important to conduct a qualitative-quantitative analysis in order to evaluate domains behavior and the possible transfer of meanings between them. (Hjørland 2002, 448).

To analyse terms, concepts, and keywords in the results, we used principles of the Knowledge Organization Systems (KOS), which are formed by classification systems comprising different types of relationships. (Dalhberg 1993, 212).

3.0 Results from databases

3.1 Results from Scopus

The similitude analysis of the term Amazônia unveiled that the term forest has a nuclear position, and it is formed by subclusters represented by the terms: tree, climate, amazonia, scale, pattern, datum, and specie. The central cluster is linked to the cluster represented by the term carbon. It was noticed a convergence of the subclusters to the central cluster of the term forest.

In the similitude analysis of the term Amazon, it was not possible to identify a common core term as in the Amazônia graph. It was observed that the clusters are apparently dispersed and organized in the following syntaxes: 1) the right position detaches the following terms: mechanical, turk, amazon, and show; 2) the lower position the terms: online, research, review, product, and datum; 3) the upper position the terms: mturk, datum, and system.

3.2 Results from LISA

The similitude analysis of the term Amazônia revealed two central clusters: a cluster identified by the term research which is composed of subclusters: brazil, brazilian, document, and digital. In addition, a cluster represented by the term information where it was noticed the presence of the term ‘amazon’ but not in a relevant position.
The similitude analysis of the term Amazon exposed a central cluster represented by the term amazon connecting the clusters identified by the terms: book, online, and mechanical, in an apparent dispersion.

![Figure 2: Similitude analysis between Amazônia and Amazon from LISA](image)

### 4.0 Discussion

In Figure 1 (Scopus), the Amazônia graph showed a closer semantic relation to the forest universe. Nevertheless, in Figure 2 (LISA), the analysis of the Amazônia graph did not apparently correspond to the syntax of the terms in the forest universe or Amazônia forest. Perhaps, because LISA is an Information Science domain database.

In Figure 2 (LISA), despite the fact the term amazon is in the center, the analysis of the Amazon graph did not disclose a connection with the term forest, neither with the term Amazônia.

It was perceived in both databases that the Amazon graphs are related to the technology of the company Amazon. In this perspective, we may consider that by tagging procedure sequences in both databases, practically the same syntagmatic and lexical construction patterns are obtained.

It is worth mentioning that Scopus displayed subject areas based on its special language control in accordance with the search queries results. Also, Scopus's subject areas reflect apparently specific knowledge domains. However, those procedures are not available at LISA. Table 1 exposes the comparison of the number of documents retrieved from the search queries for the terms Amazonia and Amazon in Scopus.

<table>
<thead>
<tr>
<th>Subject Areas</th>
<th>Amazônia</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>12,014</td>
<td>39,031</td>
</tr>
<tr>
<td>Agriculture and Life Sciences</td>
<td>4,097</td>
<td>7,613</td>
</tr>
<tr>
<td>Environmental Science</td>
<td>2,033</td>
<td>3,757</td>
</tr>
<tr>
<td>Earth and Planetary Sciences</td>
<td>1,483</td>
<td>2,826</td>
</tr>
<tr>
<td>Computer science</td>
<td>76</td>
<td>5,759</td>
</tr>
</tbody>
</table>
The total of Amazon documents retrieved represent three times more than Amazônia ones. The distribution of documents retrieved in both terms is proportional in each subject area, except for Computer Science, which stood out in the volume of documents only in Amazon.

5.0 Considerations

The similitude analysis demonstrated that partitive, complementary or opposite syntax relations are semantically restructured to assure database retrieval procedures. In this regard, Posner's theory is relevant to discuss the term Amazon as a “comsymbol”, because the term is “independent of the recipient's context of action” (Posner 1982, 4).

According to Hjørland’s fourth assumption regarding language for special purposes, the results revealed that: “when documents are merged in databases information about implicit meanings from the prior contexts are lost” (Hjørland 2002, 446). In this sense, the results may indicate that the Amazon domain becomes dominant and merge into the Amazônia one; likewise, prior information and implicit meanings of the Amazônia cultural, social and native ethnic contexts are similarly dispersed. Furthermore, for some users, the terms Amazônia or Amazon would be the same because it is just a question of translation into the English version. For other users, the adoption of those terms may represent a strong difference regarding their value system and cultural domains (Smith 2013).

Regarding FAIR principles, the semantic distances depicted by similitudes graphs 1 and 2 pointed out that the terms Amazônia and Amazon are not interoperable, because their clustering terms distribution presented different complementary, opposition and partitive relations. Therefore, the results confirmed that “knowledge organisation systems and IR should be developed to cope with this loss of implicit information by making it explicit (database semantics)” (Hjørland 2002, 446).

This study demonstrated that the similitude analysis applied to FAIR principles is an important technical device for evaluating automatic indexing procedures in databases. The results suggested that, although we may deal with different epistemic communities with their organization of concepts and diversified genres of documents, FAIR principles are attainable goals for information retrieval systems. In addition, the findings indicate that Domain Analysis and Theory of Concept are useful analytical tools to develop databases indexing languages.

In the future, we intend to expand this method to other databases in order to improve semantic analysis in the retrieval process of terms and concepts.

Acknowledgement

This research dataset is available at the Laboratório em Rede de Humanidades Digitais (Larhud) community in the Zenodo digital repository, following Open Science practices.1

References


1 Dataset accessible at http://doi.org/10.5281/zenodo.3751623

