computers

MDPI

Article

Rating the Dominance of Concepts in Semantic Taxonomies

Gerasimos Razis [1,2,*], Ioannis Anagnostopoulos [1] and Hong Zhou [2]

[1] Computer Science and Biomedical Informatics Department, University of Thessaly, 35131 Lamia, Greece; janag@uth.gr
[2] Atypon Systems LLC, Santa Clara, CA 95054, USA; hzhou@atypon.com
* Correspondence: razis@uth.gr or grazis@atypon.com

Abstract: The descriptive concepts of "semantic" taxonomies are assigned to content items of the publishing domain for supporting a plethora of operations, mostly regarding the organization and discoverability of the content, as well as for recommendation tasks. However, either not all publishers rely on such structures, or in many cases employ their own proprietary taxonomies, thus the content is either difficult to be retrieved by the end users or stored in publisher-specific fragmented "data-silos", respectively. To address these issues, the modular and scalable "Dominance Metric" methodology is proposed for rating the dominance and importance of concepts in semantic taxonomies. Our proposed metric is applied both on the vast multidisciplinary Microsoft Academic Graph Fields of Study taxonomy and the MeSH controlled vocabulary in order for their enhanced and refined versions to be produced. Moreover, we describe the cleansing process of the resulting taxonomy from Microsoft's structure by deduplicating concepts and refining the hierarchical relations towards the increase of its representation quality. Our evaluation procedure provided valuable insights by showcasing that high volume, namely the number of publications a concept is assigned to, does not necessarily imply high influence, but the latter is also affected by the structural and topological properties of the individual entities.

Keywords: taxonomies; tags; importance; dominance; Microsoft Academic Graph; MeSH

Citation: Razis, G.; Anagnostopoulos, I.; Zhou, H. Rating the Dominance of Concepts in Semantic Taxonomies. *Computers* **2022**, *11*, 35. https://doi.org/10.3390/computers11030035

Academic Editors: Katia Lida Kermanidis, Manolis Maragoudakis and Phivos Mylonas

Received: 7 January 2022
Accepted: 1 March 2022
Published: 2 March 2022

1. Introduction

A plethora of operations in the publishing domain rely on the utilization of "semantic" taxonomies, which are hierarchical and directed structures of descriptive concepts. Those concepts are assigned to content items for facilitating their organization and discoverability [1], as well as for further improving recommendation tasks [2]. These semantic taxonomies are utilized by hosting platform providers of scientific content for satisfying the academic publishers' needs in terms of semantic classification of content items (e.g., publications, videos, and digital objects).

The semantic taxonomies employed by the academic publishers are either publicly available ones, such as the MeSH (https://www.nlm.nih.gov/mesh/meshhome.html, accessed on 6 January 2022) controlled vocabulary, which may not always sufficiently cover all content aspects, or proprietary ones tailored to the publishers' needs and content, as happens in the majority of the cases (more details are provided in Section 5). These proprietary taxonomies are created by third-party vendors with the aid of taxonomists, subject matter, and domain experts [2], which is an intensive, costly, and time-consuming process [1,2]. Eventually, once the taxonomy is available then the automatic or manual classification takes place and concepts are assigned to the content.

As already mentioned, in the majority of the cases proprietary taxonomies are used for tagging the scientific content, which ends up being stored in publisher-specific "data-silos" in a fragmental fashion. Therefore, the content providers, hosting platforms' publishers, or end users (e.g., researchers, and clinicians) are not able to effectively tag, retrieve, recom-

mend, or even query the content by adopting a consistent approach [3,4]. This conventional tagging approach leads to the following drawbacks from the publishers' viewpoint:

- Subject matter, domain experts, and taxonomists are employed for generating a taxonomy [3,4];
- The taxonomy creation process is intensive, costly, and time-consuming [1,3,4];
- Expansion of existing taxonomies is also a laborious task [3,4];
- Publicly available taxonomies may not sufficiently cover all content aspects;
- Tailor-made taxonomies are not easily extended to be applied to other context or scientific domains [2];
- Tagging the content and maintaining those tags requires a lot of effort; and
- Client-specific intelligence is not efficiently associated with the global research context (thus solutions such as TrendMD [4] emerge in order to partially solve this issue).

With the purpose of our study to provide an efficient solution for publishers to overcome the aforementioned restraints, the "Dominant FoS" (DoFoS) taxonomy is proposed and created by leveraging the vast and multidisciplinary Microsoft Academic Graph (MAG) [5] Fields of Study [6,7] (FoS) semantic taxonomy (https://academic.microsoft.com/topics, accessed on 30 December 2021). This taxonomy consists of approximately 720 K concepts, structured in a six-level hierarchy, covering the 19 disciplines presented in Table 1.

Table 1. The 19 disciplines of the MAG FoS taxonomy.

Art	Biology	Business	Chemistry	Computer Science
Economics	Engineering	Environmental science	Geography	Geology
History	Materials science	Mathematics	Medicine	Philosophy
Physics	Political science	Psychology	Sociology	

Practically, the whole MAG FoS taxonomy (99.9%) has been automatically generated, based, among others, on the UMLS and Wikipedia data resources, whereas only the concepts of the first two levels (313 in number) are manually defined [6,7]. These remain constant in subsequent versions of the structure and are based on the ScienceMetrix classification (https://science-metrix.com/?q=en/classification, accessed on 6 January 2022). Therefore, the DoFoS covers the publishers' need for a refined, cleansed, and more compact structure of the MAG FoS, consisting of the most characteristic and dominant tags in terms of their importance. Consequently, the establishment of the DoFoS introduces an enhanced, modern, and disruptive element to the scholarly ecosystem taxonomic and tagging workflow, where the stakeholders of the industry (i.e., content providers, hosting platforms, publishers, and researchers) are benefitted as follows:

- The requirement for tailor-made taxonomies is drastically reduced, as a result of the multidisciplinary nature and concept coverage of the DoFoS;
- In comparison to the current workflow, the DoFoS can be immediately adopted, thus saving time and resources, as opposed to the creation of a tailor-made taxonomy;
- The adoption of DoFoS can lead to improved content organization and enhanced content discoverability;
- The fragmented scientific knowledge is consolidated as the content is classified based on a single taxonomy;
- Publisher-specific taxonomic silos are decreased both in volume and number; and
- Large-scale recommendations and analytics across all publishers and disciplines are feasible.

On top of that, a major percentage of scientific content publishers are focused on the bio-medical domain and are intending, or already, relying on the MeSH vocabulary for content classification purposes. Our aim is to facilitate publishers overcoming the aforementioned taxonomic and tagging restraints of the conventional approach, by also

providing a compact version of the MeSH structure, since, in many cases, only a subset of this structure is, or intended, to be utilized.

There are two aims for this study. Firstly, a modular and scalable methodology is proposed for measuring the dominance and importance of concepts in semantic taxonomies of any given structure and size. Secondly, as case studies, the proposed methodology is applied on the MAG FoS and MeSH taxonomies in order to produce their more compact iterations, consisting of the most important tags across all disciplines and root concepts, respectively.

The remainder of the study is organized as follows. In the next section, an overview of the related studies is presented for discovering impactful entities in graph-based structures of different types, such as Online Social Networks (OSNs) and bibliographic knowledge bases. Then, in Section 3, the proposed methodology for measuring the dominance of concepts in taxonomic structures is analytically described. In Section 4, two real-case scenarios are presented where our proposed methodology is applied on the MAG FoS and MeSH taxonomies. Moreover, the cleansing process of the derived DoFoS is also described along with the Inter-Annotator Agreement evaluation procedure. In Section 5, we discuss the establishment of the taxonomies in the scholarly publishing domain, along with benefits deriving from the establishment of the DoFoS. Finally, in Section 6, the conclusions of our study are provided, by summarizing the outcomes, in addition to our considerations for future directions.

2. Related Work

Identifying and discovering entities exerting influence, as well as quantifying their impact, is a research topic of high interest spanning all domains, since it can be applied to a wide range of structures and sciences (e.g., computer, social, publication, and even conceptual networks). In spite of the fact that in the literature the term "influence" has not been defined uniformly, recent studies ([8–10]) conclude that the number of direct or indirect peers, (i.e., popularity an entity has) does not guarantee high impact. This observation has also been confirmed by our "Dominance Metric" methodology [11], since the popularity of an entity is only one of the factors affecting its impact. Moreover, typically these studies consider factors involving the individual characteristics and attributes of the entities under investigation, their topological and structural properties, or a combination of these. Our proposed methodology falls under the latter category.

The following studies rely on the individual characteristics and attributes of the individual entities under investigation for measuring their impact. The work in [8] proposes the "Influence Metric" methodology for quantifying the impact of Twitter accounts. Specifically, the metric considers three social aspects, namely, the posting frequency of a Twitter account (i.e., activity), its popularity in terms of followers and following accounts (i.e., social degree), and its content acknowledgment via network diffusion (i.e., qualitative impact). The authors conclude that high popularity or activity do not necessarily imply high influence in OSNs.

A similar conclusion is also discussed in [10], where its authors introduce three types of influence for Twitter accounts. Specifically, the "In-degree" metric depends on the number of followers, the "Retweet" one on the number of retweeted user's posts, and the "Mention" one on the number mentions a user receives. The experiments revealed a correlation between the "Mention" and "Retweet" metrics, but not with "In-degree" one.

The "InfluenceRank" framework is proposed in [12] for also measuring the influence in Twitter. For this purpose, the most recent tweets of the accounts are collected along with their profile characteristics, including the total number of tweets, the number of followers and following, as well as the items in the accounts' lists. Then, a machine learning technique based on regression analysis is used for measuring the accounts' impact and the effect the selected features have. Regardless of the promising results, due to the small size of the training set the model was not very accurate.

Apart from OSNs, impactful entities can also be identified in mobile networks. To this end, the authors of [13] introduce the concept of entropy for the "friend" and "interaction frequency" parameters, for describing the complexity and uncertainty of social influence. Based on a real-world mobile communication dataset, transformed into a weighted graph containing the users and their relationships the "direct", "indirect", and "global" influence scores are calculated.

A quantification of the impact of research publications is examined in [14] where the "Relative Citation Ratio" is described, while the authors are questioning the performance of the established bibliometric metrics. The proposed methodology relies on the publications' co-citation networks, while the number of citation scores is normalized depending on the scientific discipline, based on the other publications' obtained citations in that field.

The following works, apart from the individual characteristics and attributes of the nodes, also utilize their topological and structural properties in the networks, in order for impactful entities to be identified. In comparison to these works, where only a section of the entire network is utilized, usually because of complexity or hardware limitations, our methodology is not bound by such restrictions. Community detection algorithms are employed in [15] towards the identification of influential users in OSNs. To this end, the structural characteristics of the nodes are identified, such as centralities, degree counters, and shortest paths, whereas their weights depend on the global network characteristics. Finally, principal component analysis is applied, and the most influential nodes are identified.

Another work relying on centrality metrics, and specifically on the eigenvector one, by also examining the topological and structural properties of the network is the "Influence Rank" metric presented in [16]. The algorithm is based on the PageRank and relies on the Twitter accounts' followers and following relationships, retweets, favorites, and mentions, for evaluating the ability of "opinion leaders" to influence other accounts. The authors conclude that the OSN users are influential, only if their direct peers are also influential. This has also been reported by our previous work [11] as important nodes in the hierarchical structure denote major taxonomic branches (Section 3).

A hybrid framework is presented in [17] integrating both the structure of the Flickr network and the individual characteristics of its nodes in order to measure their influence. The application of graph-based and graph analysis algorithms, such as PageRank, eigenvector, and in-degree, reveal the topological details of the nodes, whereas their properties (e.g., activity) derive by adapting a similar metric of the Flickr platform.

The authors in [18] present the "Influence Ranking for Testers" methodology for not only measuring the impact of software testers in bug-tracking systems, but also for assessing the contributions of individuals. The actions of filing new bug cases and commenting on existing bug reports are treated as equivalent to posting and replying in the OSNs. Furthermore, it is assumed that influence is not only affected by the number of a node's direct peers, but also by the influence exerted from each peer. By relying on the aforementioned comments and relationships, the influence is quantified based on random walk traversals.

Finally, the authors in [19] propose the "CITEX" citation metric for calculating the impact of authors and publications in the scholarly domain. To that end, a graph is created consisting of nodes, representing the aforementioned entities, and edges, representing the citation and authorship relationships. The methodology is applied on the graph, which structure is analyzed, and influence values are assigned to the nodes, until converging to a threshold. Higher values indicate greater impact.

3. Defining Dominance Metrics

As already mentioned, taxonomic structures are directed and hierarchical of a specific depth consisting of concepts, often called tags. If represented in a graph, they would appear as nodes, while the edges connecting them are usually called "Parent–Child" relations. Formally these depict the "Supertype–Subtype" relations. In the scholarly domain, the "semantic" tags, representing descriptive concepts, are associated with the scientific content for improving content discoverability and organization [1]. Clearly, in any taxon-

omy there are tags with greater importance than others, as is the case in any graph-based structure [8,18]. In this section, we analytically describe the methodology towards the calculation of that importance and dominance in taxonomic structures. As described in our previous work [11], our proposed methodology is based on the principals of "Influence-Metric" discussed in [8] and considers a combination of factors, involving the individual characteristics and attributes of the entities under investigation, along with their topological and structural properties.

However, specific pieces of information need to be identified prior to the quantification of this importance, covering both aspects of the whole semantic taxonomy under investigation and the characteristics of the individual concepts. Specifically, the following values need to be calculated:

- "Depth": The depth of the taxonomy, as an integer of one or greater positive value;
- "Level": The level of a concept in the taxonomy, defined as Depth plus one, as an integer of zero or greater positive value, where the lowest value is assigned to the root node;
- "Descendants (direct and inferred)": The number of direct descendant concepts of a concept along with its inferred ones (all the descendants of its descendants), as an integer of zero or greater positive value; and
- "Tagged Objects (direct and inferred)": The number of tagged objects directly associated with a concept along with the tagged objects of its inferred descendants, as an integer of zero or greater positive value.

Obviously, the nature of these values is generic and can be applied to any type of taxonomies associated with content. Once these prerequisite values are calculated, the dominance score of each concept in the taxonomy under investigation can be measured.

As the works in [8–10,18] conclude, a dominance metric should not depend only on the social popularity. In the analogy of the publishing domain, we suggest that the number of "Tagged Objects" a concept has is not a representative measurement, even if that number is high enough and numerous objects are associated with that concept. Due to the characteristics deriving from the structural position in the hierarchy, the root nodes always have the greatest values of this factor, similarly to the "friends" or "followers" of well-known accounts in OSNs. Thus, the "Tagged Objects" factor is not sufficient by itself. Therefore, the "Descendants (direct and inferred)" factor should also be considered, as the structural position is also affected by the in and out degree of a node, a direct analogy of the "following" and "followers" attributes of Twitter [8,10]. However, this factor is also insufficient, as due to the structural position of the root and leaf nodes, these will always be assigned to very high or low values respectively. In order to handle these edge cases, the "Tagged Objects to Descendants" (TOtD) ratio is introduced, a ratio similar to the one discussed in [8], combining the topological properties with the individual characteristics. This ratio is used in two ways; on the one hand, to provide to some extent an "equal" distribution over the number of tagged objects per concept, while on the other hand to indicate the dominance of a taxonomic "branch". Higher values of that ratio indicate greater importance of a concept. Since the "TOtD" ratio can generate extremely high or low values, the value of the "Level" of each concept is considered as another key factor. The factor is not only used to balance the aforementioned "TOtD" values, but also to identify strong or weak taxonomic leaf nodes. Finally, the generic information about the "Depth" of the taxonomy is also utilized, in order for the value of the "Level" to be more accurately represented as a structural characteristic (e.g., in the case of "Level" 3 and "Depth" 6, as opposed to the case of 3 and 14). Finally, our proposed "Dominance Metric" relies on all of the aforementioned characteristics and properties of the examined taxonomy and its concepts and is defined in Equation (1). The denominator of the "TOtD" ratio is increased by one so as to prevent it from being equal to zero in cases of leaf nodes, namely nodes without descendants.

$$DominanceMetric = \left(\frac{TaggedObjects}{\frac{Descendants + 1}{\left(1 - \frac{Level}{Depth}\right)}} \right), \quad (1)$$
$$where\ 0 \leq Level, Level \in Z, 1 \leq Depth, Depth \in Z,$$
$$0 \leq TaggedObjects, TaggedObjects \in Z,\ 0 \leq Descendants, Descendants \in Z.$$

The aforementioned factors, namely the "TOtD" ratio, the "Level", and the "Depth" are generic characteristics of the taxonomies and can be measured for any such structure having been used for classification purposes. Our aim is to propose an extensible and modular framework, therefore additional factors deriving from the characteristics of the examined taxonomy can be easily introduced.

4. Use Cases

In this section, we present two real case scenarios where our proposed methodology is applied; firstly, to the vast MAG FoS semantic taxonomy (approximately 720 K tags) and, secondly, to the MeSH controlled vocabulary (approximately 29 K topical descriptor tags), in order for their compact versions to be derived, consisting of the most important and representative concepts.

4.1. MAG FoS Taxonomy

In addition to applying the "Dominance Metric" methodology to the MAG FoS taxonomy, we present the cleansing process of the derived structure for increasing its representation quality. Specifically, we deduplicated and merged the noisy concepts existing in the original taxonomy, as well as refined the derived hierarchical structure by deleting and correcting erroneous parent–child relations.

4.1.1. Adapting the "Dominance Metric" Methodology

As already mentioned, our aim is to facilitate publishers overcoming the taxonomic and tagging restraints of the conventional approach discussed in Section 1. Towards this end, the "Dominant FoS" (DoFoS) was created by leveraging the MAG FoS taxonomy in order for its compact and cleansed version to be derived, consisting of the most important and representative concepts of the 19 disciplines (Table 1), so it can be horizontally applied across all domains and publishers.

Towards this, the methodology presented in Section 3 will be employed. Apart from the four generic factors of the core methodology applicable to all taxonomies, two new factors tailored to the MAG FoS structure will be incorporated. Specifically, for each concept the following values need to be calculated:

- "UMLS relation": Indicates whether a concept is related with a UMLS term, as a Boolean value.
- "Source relation": Indicates whether a concept is related with external knowledge sources (e.g., Wikipedia), as a Boolean value.

Especially for the case of the MAG FoS taxonomy, the "UMLS relation" and "Source relation" factors are expanding the core methodology and aim at increasing the dominance values of the individual concepts in case these properties exist. We consider the relation of a concept with the UMLS (Unified Medical Language System) resource as more significant compared to the relation with any other knowledge source, therefore it is associated with a higher weight. By considering the aforementioned, the tailored "Dominance Metric" values for the MAG FoS taxonomy derive from Equation (2). Since the maximum taxonomic depth of this structure is equal to six, the "Depth" factor of Equation (1) is replaced by that number. Moreover, the generic "Tagged Objects" term of Equation (1) has been replaced by the "Publications" term.

$$DominanceMetric_{FoS} = \left(\frac{Publications}{\frac{Descendants+1}{\left(1-\frac{Level}{6}\right)}} \right) \times UMLS_{Factor} \times Source_{Factor}, \quad (2)$$
$$where\ 0 \leq Level \leq 5, Level \in Z,$$
$$0 \leq Publications, Publications \in Z, 0 \leq Descendants, Descendants \in Z,$$
$$UMLS_{Factor} = \begin{cases} 1.2, UMLS = 1 \\ 1, UMLS = 0 \end{cases}, \quad Source_{Factor} = \begin{cases} 1.1, Source = 1 \\ 1, Source = 0 \end{cases}.$$

4.1.2. Generating the DoFoS

In order for the DoFoS to be generated, a framework consisting of five phases was developed, presented in the flowchart of Figure 1.

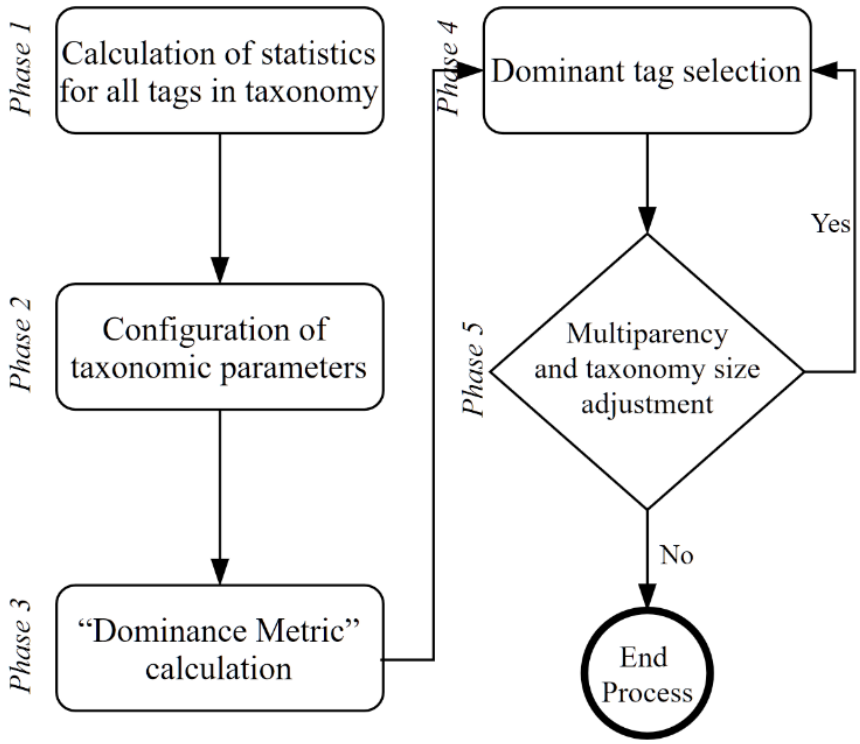

Figure 1. An overview of the five phases of the proposed framework.

The first phase involves the calculation of the prerequisite pieces of information presented in Section 3. To that end, we rely on the most recent MAG dataset available via Azure subscription. These metrics are calculated on a 32 GB RAM machine almost instantly, the actual processing time is less than a minute, apart from the "Publications (direct and inferred)" metric. Since its calculation relies on the processing of a 40 GB file containing the associations among the publications and the MAG FoS tags, a small SPARK cluster is utilized producing the results in approximately 15 min.

During the second phase, the taxonomic parameters are configured, namely, (a) the desired size of the DoFoS, i.e., the number of top-*k* dominant tags from the initial structure, and (b) the disciplines (i.e., root tags) to be considered in the process. For the DoFoS case, *k* is set equal to 200 K, and all 19 disciplines are included.

Since the configuration parameters have been set and the prerequisite information has been calculated, the third phase starts, and the adjusted proposed "Dominance Metric" methodology (Section 4.1.1) is applied to the MAG FoS taxonomy. Eventually, all tags are rated while their dominance is indicated by a numeric value.

The fourth phase involves the selection of the dominant tags for the taxonomy. A simple selection of the top-*k* tags across the selected roots is not adequate, despite the

fact that the data analysis revealed similar patterns. Specifically, the distributions of the "number of publications over tags" as well as the "number of tags per discipline" are both typical power-law ones. The distribution of the latter is depicted in Figure 2. The vertical axis represents the number of tags being part of a discipline, while the horizontal axis presents the IDs of these 19 disciplines. We measured that the first three disciplines of Figure 2 consist of approximately 80% of the MAG FoS tags.

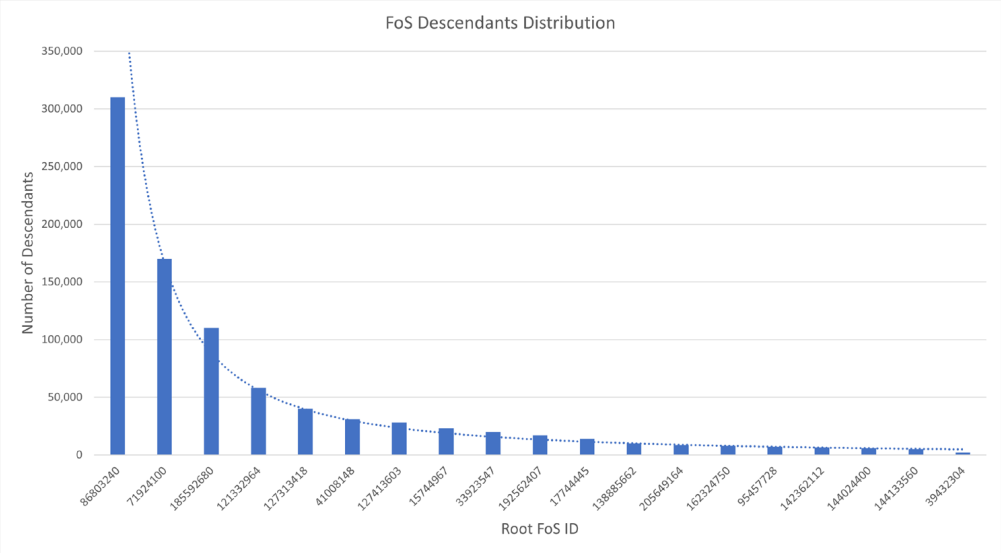

Figure 2. The distribution of tags per discipline in MAG FoS.

This is an essential characteristic of the structure of the MAG FoS taxonomy, and also indicative of the knowledge representation and coverage of each discipline. This characteristic is incorporated into the derived DoFoS by selecting proportionally the number of top-k tags per discipline according to their distribution in the MAG FoS taxonomy. Furthermore, since the root tags (Level 0) and their direct children (Level 1) have been manually defined, based on the ScienceMetrix classification, they are considered as very important and are always present in the derived DoFoS despite that they could be placed below the threshold of the kth position.

The taxonomies are multi-parent structures by nature, thus the final fifth phase is responsible for measuring the number of distinct tags in the derived taxonomy via an iterative process and thus identifying any duplicates which may have been included due to the multiple parents of these tags. In cases that the number of distinct tags is less than the specified threshold k (defined in the second phase), the fourth phase is performed again until this threshold is reached, by adding proportionally new tags to all disciplines.

Indicative examples of the "Dominant Metric" values are presented in Table 2, according to the characteristics of the MAG FoS taxonomy and its tags. This table provides quantifiable evidence of what it also reported in the recent literature ([8–10]) and is also validated by our research, namely that an entity's characteristic of belonging to high structural positions or being related with millions of peers (in our case publications) do not necessarily guarantee high impact both on an individual and on a graph level.

Table 2. Calculating the "Dominance Metric" in MAG FoS.

Tag ID	Publications	Descendants	Level	UMLS	Source	Dominance Metric
ID_01	14,709	0	3	0	1	32,359.8
ID_02	4318	0	5	0	1	28,498.8
ID_03	14,656	7	4	0	1	6045.6
ID_04	970,638	375	1	0	1	3407.56
ID_05	4,052,723	3630	0	0	1	1227.76
ID_06	12,989	56	2	0	1	376
ID_07	383	12	3	0	1	64.82
ID_08	6	0	5	1	0	43.2
ID_09	9	1	4	1	0	16.2
ID_10	6	0	3	0	1	13.2
ID_11	11	2	3	1	0	8.8
ID_12	1	0	5	1	0	7.2
ID_13	1	0	2	0	1	1.65

4.1.3. Cleansing Process

As discussed before, the concepts of the MAG FoS taxonomy [6] (99.9%) have been automatically created by primarily utilizing the UMLS and Wikipedia data resources, contrary to the manually defined concepts of the first two levels (313 in number), while the hierarchical relations of the MAG FoS concepts rely on a more sophisticated term co-occurrence measurement [6].

Due to the absence of any manual curation on the MAG FoS taxonomy, neither on a concept or on the hierarchy level, data cleansing processes in these areas of the derived DoFoS need to be performed in order for the representation quality of the structure to be improved.

DoFoS Deduplication

Our first data cleansing stage of the derived DoFoS focused on the elimination of the duplicate concepts existing in the original MAG FoS taxonomy. Characteristic cases of "noisy" terms representing the exact same concept involve British and American spelling, misspellings, singular and plural form, existence of additional characters or special symbols (e.g., spaces, dashes, apostrophes, and parentheses), and so on. Due to this wide spectrum of cases generating noise, a human intervention is required for them to be identified and for the DoFoS concepts to be cleansed, so that the quality of the structure and consequently of the content classification process to be improved.

To address this issue, the Levenshtein distance was employed for identifying candidate pairs of conceptually identical taxonomic nodes. The Levenshtein distance measures the character difference of two terms, thus providing the minimum number of single-character changes (i.e., substitutions, insertions, or deletions) required to transform one string into the other. To materialize our aims, two dataset types of candidate tag pairs were created: the first with distance equal to one (D1) and the other equal to two (D2), containing 15 K and 105 K records, respectively. Indicative examples of such pairs are the {work organi**s**ation-work organi**z**ation} of the D1-type, and {breast tumor-breast tumo**u**r**s**} of the D2-type.

Five curators from the University of Thessaly, who attended post-graduate courses with related project experience, performed the manual cleansing of the two aforementioned datasets. The D1/D2-type datasets were divided into five segments each, consisting of approximately 3 K and 21 K pairs, respectively. Each curator was assigned two groups of the D1/D2-type segments, whereas each group was allocated to exactly two curators. This enabled us to not only enhance the quality of the derived annotations, but also to measure their confidence, whereas, as expected, the same two groups could not be assigned to more than one curator. Eventually, approximately 48 K concept pairs, consisting of two groups of two segments of the D1/D2-type datasets, were allocated to each curator for manual inspection.

The curators were asked to firstly examine whether the two concepts of each pair are the same, and in case of a positive assessment, which of them should be considered as the preferred concept. The latter is of high importance, because when two or more concepts have been marked as the same conceptual entity and must be merged, only one of them appears in the cleansed version of the DoFoS. All pieces of information of the "duplicated" concepts are used for expanding the semantic coverage of the preferred concept. For example, the labels of the duplicate tags are used as alternative labels of the primary concept, namely, as synonyms.

Each concept of the shared datasets was associated with its ID, normalized name, URL redirecting to its unique landing page in the MAG FoS portal, as well as its adapted "Dominance Metric" value (presented in Section 4.1.1). In the event of further information or investigation was required due to the ambiguity of a concept, additional resources or search engines should be used by the curators.

In order to assess the overall curation agreement and quality of the annotation process, we relied on the Inter-Annotator Agreement (IAA) based on the Cohen's Kappa [20] coefficient. Kappa values ratings [20] are as follows:

- [0] indicate no agreement;
- (0, 0.2) indicate slight agreement;
- [0.2, 0.4) indicate fair agreement;
- [0.4, 0.6) indicate moderate agreement;
- [0.6, 0.8) indicate substantial agreement; and
- [0.8, 1] indicate nearly perfect agreement.

In Table 3 the curators' Cohen's Kappa scores are presented regarding whether their assigned concept pairs refer to the same conceptual entity or not. In order to provide more detailed insights, the aforementioned ratings have been expanded with the symbols of minus and plus, indicating whether a value lies in the first or second half of the rating range, respectively. Overall, as presented in the bottom row of the table, the annotations are qualitative and in substantial agreement.

Table 3. Agreement on Identical Concepts.

Ranking	Curator ID	Cohen's Kappa	Rating
1	C05	0.83	Perfect−
2	C04	0.79	Substantial+
3	C01	0.76	Substantial+
4	C03	0.62	Substantial−
5	C02	0.58	Moderate+
Overall: 0.72 (Substantial+)			

As already mentioned, in the cases of conceptually identical tags, the curators were also tasked with identifying the primary concept to be kept into the cleansed version of the DoFoS. Towards this end, we employed a secondary metric for quantifying this agreement, presented in Table 4, by calculating the overlap of these annotations. As before, we have expanded the ratings with the symbols of minus and plus. These annotations are also in high agreement, as presented in the bottom row of the table.

In order to resolve any cases of disagreement, we relied on the aforementioned scores, based on the IAA, to keep the curator's annotations with the higher value. Since the quality of the derived DoFoS is essential for its establishment in the scholarly domain, an additional evaluator from University of Thessaly also manually resolved specific cases of disagreement, including only 440 (2.9%) for the D1-type dataset and 2425 (2.3%) for the D2-type dataset. As a consequence of the deduplication process, approximately 4.3 K concepts (2.1% of the top-200 K) were marked as duplicates and merged (~2.8 K from the D1-type and ~1.5 K from the D2-type). Eventually, the DoFoS was updated by eliminating the duplicated concepts and preserving the preferred ones. Thus, the representational

quality of the DoFoS was increased. On top of that, the primary concepts were expanded with synonym names of the duplicated concepts, further increasing the semantic coverage of the taxonomy.

Table 4. Agreement on Primary Concept.

Ranking	Curator ID	Agreement	Rating
1	C01	0.90	Perfect+
2	C02	0.89	Perfect−
3	C03	0.85	Perfect−
4	C04	0.84	Perfect−
5	C05	0.84	Perfect−
Overall: 0.87 (Perfect−)			

DoFoS Hierarchy Refinement

Our second data cleansing stage of the derived DoFoS focused on the refinement of its hierarchical relations by deleting and correcting erroneous parent–child relations existing in the original MAG FoS taxonomy. Characteristic cases of erroneous hierarchical relations are the "Diabetes Mellitus" and "Jazz (music)" or the "Telecommunications" and "Words per Minute" concepts. Due to the multi-parent nature of the MAG FoS taxonomy, the concepts are usually related to more than one parents, thus in the majority of the cases simply deleting the erroneous relations is sufficient. Therefore, a human intervention is required for such cases to be identified in order for the DoFoS concept relations to be cleansed, so that the quality of the structure and, consequently, of the content classification process to be improved. This is also a very important step, as usually the publishers rely on the "aggregation" tagging function, which associates all direct and indirect ancestors of an assigned tag to a content item.

To address this issue, the Cosine Similarity metric was employed for identifying candidate erroneous hierarchical pairs. The Cosine Similarity is a semantic similarity metric of two terms measuring how similar their meanings are, independent of any syntactic differences (e.g., "hound", and "dog"). Specifically, relying on word embeddings each concept name was transformed into a real-valued vector using a pre-trained (https://code.google.com/p/word2vec/, accessed on 6 January 2022) Word2Vec model [21,22], trained on a Google News dataset. This model contains the vectors of approximately three million English words and phrases. The adopted metric compares the similarity of the examined vectors as the cosine of the angle between them. Since it does not consider the magnitudes of the vectors but rather their directions, it can identify terms of opposing meaning. Assuming that the vectors V1 and V2 represent two concept names, then a value of Cosine Similarity equal to 0 indicates no correlation between them, equal to 1 perfect match, and equal to −1 opposing meanings.

Using the Cosine Similarity metric, the similarity of the names of all concepts participating in a parent–child relationship was measured. Approximately 350 K pairs were examined and their similarity values were then assigned to 20 groups ranging between (–1, 1), with an incremental step of 0.1. These groups are presented in Table 5. It should be noted that 13% of the hierarchical pairs were not rated, since their names were not part of the pre-trained Word2Vec model. The Cosine Similarity value of the aforementioned "Diabetes Mellitus" and "Jazz (music)" concepts is 0.13.

To materialize our refinement aim, a dataset of candidate parent–child concept pairs was created consisting of 21 K records from the tags of the range (−0.2, 0.1). Twelve curators from the University of Peiraeus, who attended post-graduate courses with related project experience, performed the manual cleansing of the dataset.

Thus, the dataset was divided into 12 segments, consisting of 1.75 K pairs, respectively. Each curator was assigned two segments, whereas each segment was allocated to exactly two curators. This enabled us to not only enhance the quality of the derived annotations, but also to measure their confidence. Obviously, the same two segments could not be

assigned to more than one curator. Eventually, 3.5 K concept pairs, consisting of two segments of the dataset, were allocated to each curator for manual inspection.

Table 5. Distribution of Cosine Similarity Values of Hierarchical Concepts.

Similarity Range	Percentage
[−1.0, −0.2)	0%
[−0.2, −0.1)	0.002%
[−0.1, 0.0)	0.44%
[0.0, 0.1)	5.28%
[0.1, 0.2)	12.03%
[0.2, 0.3)	14.09%
[0.3, 0.4)	12.36%
[0.4, 0.5)	9.15%
[0.5, 0.6)	7.31%
[0.6, 0.7)	6.55%
[0.7, 0.8)	7.07%
[0.8, 0.9)	8.56%
[0.9, 1]	4.07%

The curators were tasked with identifying whether the two concepts of each pair form a valid or an erroneous parent–child relation. Each concept of the shared dataset was associated with its ID, normalized name, and URL redirecting to its unique landing page in the MAG FoS portal. In the event of further information or investigation was required, additional resources or search engines should be used by the curators.

In order to assess the overall curation agreement and quality of the annotation process, as before, we relied on the Inter-Annotator Agreement (IAA) based on the Cohen's Kappa [20] coefficient.

In Table 6 the curators' Cohen's Kappa scores are presented regarding their agreement on the hierarchical relation, including both the cases of valid and erroneous relations. As before, the aforementioned ratings have also been expanded with the symbols of minus and plus. Overall, as presented in the bottom row of the table, the annotations are in moderate agreement. These scores are lower compared to the ones of the deduplication-cleansing phase, mainly due to the complexity and often the subjective nature of the task.

Table 6. Agreement on Hierarchical Relation.

Ranking	Curator ID	Cohen's Kappa	Rating
1	C08	0.81	Perfect−
2	C07	0.79	Substantial+
3	C09	0.78	Substantial+
4	C03	0.76	Substantial−
5	C04	0.58	Moderate+
6	C05	0.55	Moderate+
7	C06	0.53	Moderate+
8	C10	0.38	Fair+
9	C11	0.36	Fair+
10	C12	0.35	Fair+
11	C01	0.31	Fair+
12	C02	0.29	Fair−
Overall: 0.54 (Moderate+)			

As a consequence of the hierarchical refinement process, approximately 13.8 K concept relations (66% of the selected 21 K) were annotated similarly by both curators. Specifically, ~11.5 K (83%) relations were marked as valid, whereas ~2.3 K (17%) as erroneous. Contrary to the deduplication–cleansing phase, we did not resolve the disagreement cases (34% of the 21 K) based on the curators' IAA scores, in order to have a more direct control over

this process. Eventually, the DoFoS was updated accordingly, by eliminating the erroneous hierarchical relations, and preserving the ones marked as valid. As a result, the quality of the DoFoS was further increased.

4.2. MeSH Taxonomy

A major percentage of scientific content publishers are focused on the bio-medical domain, relying on the MeSH controlled vocabulary. In many cases, only a subset of this large structure (approximately 29 K topical descriptor tags) is or intended to be utilized. To this end, we apply the "Dominance Metric" methodology to the MeSH taxonomy in order for its compact version to be derived consisting of the most important concepts.

4.2.1. Adapting the "Dominance Metric" Methodology

As already discussed, our aim is to facilitate publishers overcoming the taxonomic and tagging restraints of the conventional approach discussed in Section 1, by also following the requirements of providing a compact version of the MeSH structure. Towards this end, the "Dominant MeSH" (DoMeSH) taxonomy was derived consisting of the most important and representative topical descriptor concepts of the 16 root concepts of the initial structure. Since one root does not contain any topical descriptors, but tags of other utility categories, such as "publication type", the 15 root concepts presented in Table 7 are eventually considered, by excluding any other concepts not being of the "topical descriptor" type.

Table 7. The 15 Root Concepts of the MeSH taxonomy.

Anatomy	Organisms	Diseases	Chemicals and Drugs	Analytical, Diagnostic and Therapeutic Techniques, and Equipment
Psychiatry and Psychology	Phenomena and Processes	Disciplines and Occupations	Anthropology, Education, Sociology, and Social Phenomena	Technology, Industry, and Agriculture
Humanities	Information Science	Named Groups	Health Care	Geographicals

As in the case of the DoFoS, the methodology presented in Section 3 will be employed. Apart from the four generic factors of the core methodology applicable to all taxonomies, one new factor tailored to the MeSH structure will be incorporated. Specifically, for each concept the following value needs to be calculated:

- "Registry relation": Indicates whether a concept is related with a term from an external registry (i.e., CAS, EC, FDA, and NCBI), as a Boolean value.

Especially for the case of the MeSH controlled vocabulary, the "Registry relation" factor is expanding the core methodology and aims at increasing the dominance values of the individual concepts in case this property exists. By considering the aforementioned, the tailored "Dominance Metric" values for the MeSH taxonomy derive from Equation (3). Since the maximum taxonomic depth of this structure is equal to 14, that number replaces the "Depth" factor of Equation (1). Moreover, the generic "Tagged Objects" term of Equation (1) has been replaced by the "Publications" term.

$$\begin{gathered} DominanceMetric_{MeSH} = \left(\frac{\frac{Publications}{Descendants+1}}{\left(1-\frac{Level}{14}\right)} \right) \times Registry_{Factor}, \\ where\ 0 \leq Level \leq 13, Level \in Z, \\ 0 \leq Publications, Publications \in Z, 0 \leq Descendants, Descendants \in Z, \\ Registry_{Factor} = \begin{cases} 1.2, Registry = 1 \\ 1, Registry = 0 \end{cases}. \end{gathered} \tag{3}$$

4.2.2. Generating the DoMeSH Taxonomy

In order for the DoMeSH to be generated, the same five-phased framework was employed, presented in the flowchart of Figure 1.

As previously mentioned, the first phase involves the calculation of the prerequisite pieces of information presented in Section 3. To that end, we rely on the most recent dataset (http://participants-area.bioasq.org/datasets/, accessed on 6 January 2022) of the BioASQ competition, containing annotated articles from PubMed. These metrics are calculated on the same machine mentioned in Section 4.1.2 in less than a minute, apart from the "Publications (direct and inferred)" metric. Since its calculation relies on the processing of a 25 GB file containing the associations among the publications and the MeSH tags, a small SPARK cluster is utilized producing the results in approximately 5 min.

During the second phase, the taxonomic parameters are configured, namely, (a) the desired size of the DoMeSH, i.e., the number of top-*k* dominant topical descriptor tags from the initial structure, and (b) the root tags to be considered in the process. For the DoMeSH case, *k* is set equal to 10 K, and the 15 roots of Table 7 are included.

Since the configuration parameters have been set and the prerequisite information has been calculated, the third phase starts, and the adjusted proposed "Dominance Metric" methodology (Section 4.2.1) is applied to the topical descriptor tags of the MeSH vocabulary. Eventually, all appropriate tags are rated, while their dominance is indicated by a numeric value.

The fourth phase involves the selection of the dominant tags for the derived taxonomy. However, a simple selection of the top-*k* tags across the selected roots is not adequate. By analyzing the BioASQ dataset and the MeSH taxonomy, similar patterns were revealed. Specifically, the distributions of the "number of publications over tags" as well as the "number of tags per discipline" are typical power-law ones. The distribution of the latter is depicted in Figure 3, although it is not as smooth as the one of Figure 2. The vertical axis represents the number of tags belonging under a root tag, while the horizontal axis presents the IDs of these 15 root tags. We measured that the first three disciplines of Figure 3 consist of approximately 72% of the MeSH tags.

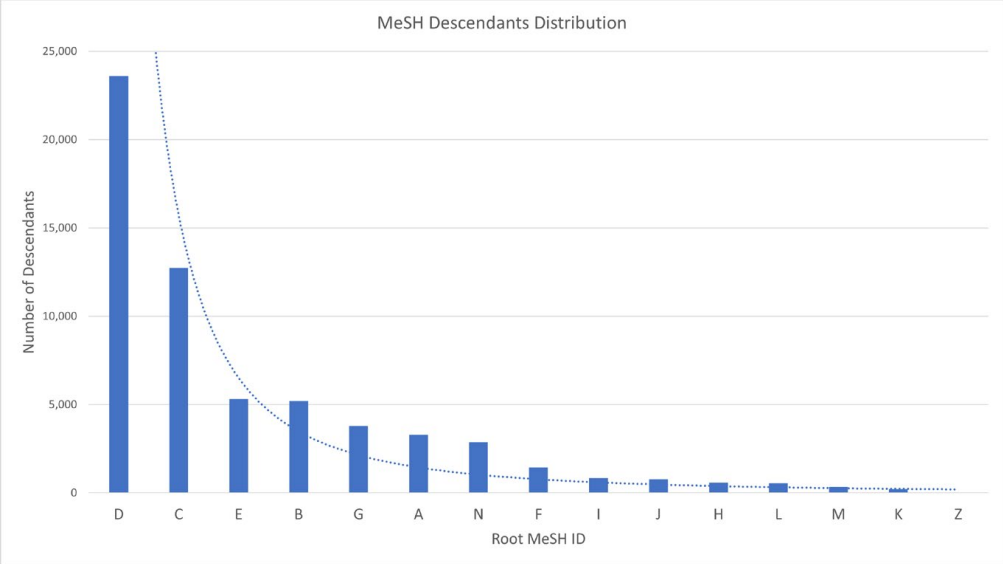

Figure 3. The distribution of tags per root in MeSH.

This is an essential characteristic of the structure of the MeSH taxonomy, and also indicative of the knowledge representation and coverage of each root. This characteristic is incorporated into the derived DoMeSH by selecting, proportionally, the number of top-*k* tags per root according to their distribution in the MeSH taxonomy.

Since the MeSH taxonomy is also a multi-parent one, the final fifth phase of the workflow takes place, similarly to the DoFoS case, until the specified threshold k (defined in the second phase) of distinct tags is reached.

Indicative examples of the "Dominant Metric" values are presented in Table 8, according to the characteristics of the MeSH taxonomy and its tags. As before, it is showcased that an entity's characteristic of belonging to high structural positions or being related with a huge number of publications does not necessarily guarantee high impact.

Table 8. Calculating the "Dominance Metric" in MeSH vocabulary.

Tag ID	Publications	Descendants	Level	Registry	Dominance Metric
ID_01	1,623,587	1635	3	1	1515.68
ID_02	135,895	568	5	1	445.82
ID_03	13,588	12	7	1	2508.55
ID_04	2,983,654	18,233	1	1	211.46
ID_05	1,988,774	3710	1	1	692.56
ID_06	15	0	12	0	105
ID_07	15	1	11	0	35

5. Discussion: Taxonomies in the Scholarly Publishing Domain

As already mentioned, a wide range of vital operations in the scholarly publishing domain relies on the utilization of the taxonomic hierarchical structures. These taxonomies can be split into two types, the "utility" and the "semantic" ones. The former type is used for facilitating fundamental operations, such as marking the usage rights of a content item (e.g., "download", and "print"), or its current version in the publishing workflow (e.g., "accepted manuscript", and "version of record"). Contrary, the taxonomies of the "semantic" type, which consist of descriptive concepts, are employed for supporting advanced operations and services, such as for enhancing content organization and discoverability, and for further improving recommendation or search engine optimization tasks. These semantic taxonomies are utilized by Atypon (https://www.atypon.com/, accessed on 6 January 2022), which is a platform provider that hosts approximately the 45% of the scientific content written in English and satisfies the academic publishers' needs as far as the semantic classification of their content items (e.g., publications, videos, digital objects, etc.) is concerned.

An analysis on that platform provider's clients reveals valuable insights in terms of the usage of the taxonomic structures in general, as well as of the semantic ones in particular. The analysis was performed on 97 publishers and revealed that 49.5% of the publishers utilize semantic taxonomies, while 50.5% do not. Moreover, 95% of these semantic taxonomies are proprietary, while the rest of them publicly available ones. There are only a few cases where publishers employed both proprietary and publicly available semantic taxonomies. Obviously, all of the publishers rely on the "utility" type taxonomies for facilitating the fundamental operations of the publishing process.

However, the publishers' semantic taxonomies vary considerably in terms of size and hierarchical depth, and consequently in the semantic coverage of the described domain. These taxonomies have been categorized in four groups depending on their size, as presented below:

- "Small" ranging from one to 100 concepts;
- "Medium" ranging from 100 to 500 concepts;
- "Large" ranging from 500 to 1000 concepts; and
- "Huge" containing more than 1000 concepts.

These four groups can be viewed in Figure 4, where additional information is also visible. More specifically, for each group, we present the absolute number of publishers participating in it, their percentage, as well as the average number of concepts (tags) per taxonomy. As we can see, 44% of the taxonomies belong to the "Small" group represented by the blue color, since 21 publishers rely on taxonomies of an average size of 37 concepts.

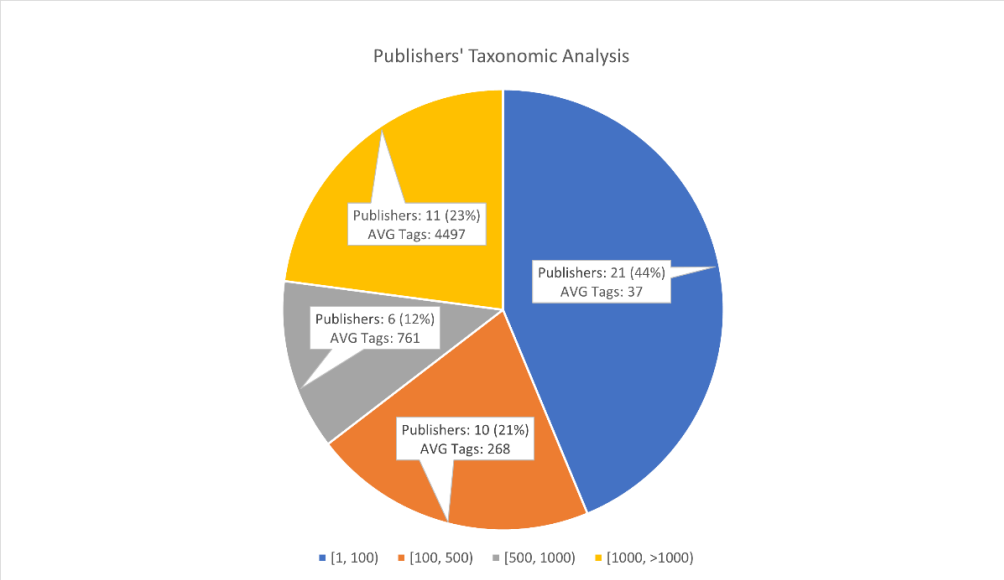

Figure 4. The analysis of the publishers' semantic taxonomies.

Our analysis led to two outcomes. Firstly, half of the publishers (50.5%) do not semantically annotate their content, thus making it difficult for both end users (e.g., researchers) and systems (e.g., recommenders) to accurately discover and retrieve that content. Secondly, the size of the semantic taxonomies for the majority of the publishers (64.6%) relying on these structures is relatively small, especially when considering the volume of the content items, which may be up to 500 K for these publishers. As it can be inferred, these taxonomies are not very detailed or deep enough, and, consequently, the content is classified only on a relatively high level. An indicative example is the classification of content items with the broad term "Artificial Neural Network" existing in a semantic taxonomy, as opposed to the more specific terms "Convolutional Neural Network", "Recurrent Neural Network", and "Binary Neural Network", which are absent from that taxonomy.

To address these issues, along with the drawbacks of the conventional tagging, and facilitate both the publishers and the end users, the "Dominant FoS" is created by leveraging the vast and multidisciplinary Microsoft Academic Graph Fields of Study semantic taxonomy. The establishment of this structure introduces an enhanced, modern, and disruptive to the scholarly ecosystem taxonomic and tagging workflow, where all stakeholders of the scholarly publishing domain are benefitted, as presented in Section 1.

In order for the "Dominant FoS" to be generated, the five-phased framework of Figure 1 was employed, where our proposed "Dominance Metric" methodology (Sections 3 and 4.1.1) is a core component, as it was applied on the vast MAG FoS taxonomy for rating the dominance and importance of its concepts. At the end, the procedure of merging duplicate concepts and removing erroneous hierarchical relations cleansed the derived structure.

Due to the characteristics of the "Dominant FoS", namely its size and discipline coverage, it can be employed by all publishers, independent of already relying or not on a semantic taxonomy. The new structure can be used for semantically annotating their content in case such a taxonomy is not present, or is of medium size, or it can act as a complementary to their existing large semantic taxonomy for consolidating the fragmented scientific knowledge.

6. Conclusions and Future Work

The contributions of this study are two-fold. Firstly, we described a modular and scalable methodology towards the quantification of the dominance and importance of concepts in semantic taxonomies. Our "Dominance Metric" is an expandable, multi-factor,

and parameterizable methodology, applicable to any hierarchy of arbitrary structure and size. Secondly, two use cases were presented where our proposed methodology was extended and applied on well-known structures, the vast MAG FoS [6] taxonomy and the MeSH controlled vocabulary, in order for their compact versions to be derived, consisting of the most dominant and important concepts.

The MAG FoS taxonomy was leveraged for deriving the "Dominant FoS", a structure consisting of 200 K concepts spanning over 19 disciplines, to be horizontally applied across all scholarly publishers and domains. Moreover, this semantic taxonomy was manually cleansed by removing duplicated concepts and erroneous hierarchical relations, in order to increase its semantic coverage. For further facilitating the publishers of the biomedical domain to overcome the taxonomic and tagging restraints of the conventional approach discussed in Section 1, the "Dominant MeSH" taxonomy was also derived consisting of the most important and representative topical descriptor concepts of the 15 root nodes of the MeSH vocabulary.

Furthermore, our experiments revealed the existence of patterns in terms of the examined entities' dominance characteristics and the factors affecting their impact dynamics. Specifically, we showcased that the popularity (i.e., number of direct peers an entity has), or, in our case, the number of publications a taxonomic concept is related to, is not a precise assessment of its importance and dominance in any given structure. The reason for this is that the numbers of direct or indirect descendants, as well as the topological and structural properties of a concept have a significant effect on its impact dynamics, which is calculated by our "Dominance Metric". This conclusion is aligned with the correlation of an OSN entity's number of direct peers and the derived influence score, as presented in the works of [8,10].

In the future, we plan to continue our research. Firstly, we aim at further expanding the proposed core "Dominance Metric" methodology by incorporating additional factors in Equation (1), such as the semantic similarity and the "trendiness" of the tags, evaluating in parallel their impact. Secondly, we intend to utilize the results of the manual data cleansing processes in order to evaluate methodologies towards the automatic identification of duplicate concepts, as well as of erroneous hierarchical relations. Given the lack of studies in the literature for quantifying the importance of concepts in taxonomic structures, we will compare our methodology against works identifying influential entities relying on graph centrality metrics.

Author Contributions: Conceptualization, G.R. and H.Z.; methodology, G.R.; software, G.R.; validation, G.R., I.A. and H.Z.; formal analysis, I.A.; investigation, G.R.; resources, H.Z.; data curation, G.R. and I.A.; writing—original draft preparation, G.R.; writing—review and editing, I.A.; visualization, G.R.; supervision, I.A.; project administration, H.Z.; funding acquisition, G.R. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Atypon Systems, Inc.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carelli, B. TrendMD: Using AI to enhance discovery and achieve publisher goals. *Inf. Serv. Use* **2020**, *39*, 335–346. [CrossRef]
2. Sujatha, R.; Rao, B.R. Taxonomy Construction Techniques-Issues and Challenges. *Indian J. Comput. Sci. Eng.* **2011**, *2*, 661–671.
3. Shen, J.; Wu, Z.; Lei, D.; Zhang, C.; Ren, X.; Vanni, M.T.; Sadler, B.M.; Han, J. HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18), London, UK, 19–23 August 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 2180–2189. [CrossRef]

4. Tuan, L.A.; Kim, J.; Kiong, N.S. Taxonomy Construction Using Syntactic Contextual Evidence. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 810–819. [CrossRef]
5. Sinha, A.; Shen, Z.; Song, Y.; Ma, H.; Eide, D.; Hsu, B.-J.; Wang, K. An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion), Florence, Italy, 18–22 May 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 243–246. [CrossRef]
6. Shen, Z.; Ma, H.; Wang, K. A Web-scale system for scientific knowledge exploration. In Proceedings of the ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 87–92. [CrossRef]
7. Shen, Z.; Wu, C.-H.; Ma, L.; Chen, C.-P.; Wang, K. SciConceptMiner: A system for large-scale scientific concept discovery. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, online, 1–6 August 2021; pp. 48–54. [CrossRef]
8. Razis, G.; Anagnostopoulos, I. Semantifying Twitter: The Influence Tracker Ontology. In Proceedings of the 9th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Corfu, Greece, 6–7 November 2014; pp. 98–103. [CrossRef]
9. Romero, D.M.; Galuba, W.; Asur, S.; Huberman, B.A. Influence and passivity in social media. In Proceedings of the 20th international conference companion on World Wide Web (WWW '11), Hyderabad, India, 28 March–1 April 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 113–114. [CrossRef]
10. Cha, M.; Haddadi, H.; Benevenuto, F.; Gummadi, P.K. Measuring user influence in Twitter: The million follower fallacy. In Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM '10), Dublin, Ireland, 4–8 June 2010; The AAAI Press: Palo Alto, CA, USA, 2010.
11. Razis, G.; Anagnostopoulos, I.; Zhou, H. Identifying Dominant Nodes in Semantic Taxonomies. In Proceedings of the 16th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP), Corfu, Greece, 4–5 November 2021; pp. 1–6. [CrossRef]
12. Nargundkar, A.; Rao, Y.S. InfluenceRank: A machine learning approach to measure influence of Twitter users. In Proceedings of the International Conference on Recent Trends in Information Technology (ICRTIT '16), Chennai, India, 8–9 April 2016; pp. 1–6. [CrossRef]
13. Peng, S.; Yang, A.; Cao, L.; Yu, S.; Xie, D. Social influence modeling using information theory in mobile social networks. *Inf. Sci.* **2017**, *379*, 146–159. [CrossRef]
14. Hutchins, B.I.; Yuan, X.; Anderson, J.M.; Santangelo, G.M. Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLoS Biol.* **2016**, *14*, e1002541. [CrossRef] [PubMed]
15. Jaitly, V.; Chowriappa, P.; Dua, S. A framework to identify influencers in signed social networks. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21–24 September 2016; pp. 2335–2340. [CrossRef]
16. Hajian, B.; White, T. Modelling Influence in a Social Network: Metrics and Evaluation. In Proceedings of the IEEE 3rd International Conference on Privacy, Security, Risk and Trust and IEEE 3rd International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011; pp. 497–500. [CrossRef]
17. Almgren, K.; Lee, J. Applying an influence measurement framework to large social network. *J. Netw. Technol.* **2016**, *7*, 6–15.
18. Li, H.; Gao, G.; Chen, R.; Ge, X.; Guo, S.; Hao, L.-Y. The Influence Ranking for Testers in Bug Tracking Systems. *Int. J. Softw. Eng. Knowl. Eng.* **2019**, *29*, 93–113. [CrossRef]
19. Pal, A.; Ruj, S. CITEX: A new citation index to measure the relative importance of authors and papers in scientific publications. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 1256–1261. [CrossRef]
20. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
21. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781002E.
22. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2 (NIPS '13), Lake Tahoe, NV, USA, 5–10 December 2013; Curran Associates Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119.