This is a preprint of an article published in International Journal on Digital Libraries.
The final authenticated version is available online at: https://doi.org/10.1007/s00799-021-00304-z

What is all this new MeSH about?

Exploring the semantic provenance of new descriptors in the MeSH thesaurus

Anastasios Nentidis[1,2] · **Anastasia Krithara**[1] · **Grigorios Tsoumakas**[2] · **Georgios Paliouras**[1]

Received: date / Accepted: date

Abstract The Medical Subject Headings (MeSH) thesaurus is a controlled vocabulary widely used in biomedical knowledge systems, particularly for semantic indexing of scientific literature. As the MeSH hierarchy evolves through annual version updates, some new descriptors are introduced that were not previously available. This paper explores the conceptual provenance of these new descriptors. In particular, we investigate whether such new descriptors have been previously covered by older descriptors and what is their current relation to them. To this end, we propose a framework to categorize new descriptors based on their current relation to older descriptors. Based on the proposed classification scheme, we quantify, analyse and present the different types of new descriptors introduced in MeSH during the last fifteen years. The results show that only about 25% of new MeSH descriptors correspond to new emerging concepts, whereas the rest were previously covered by one or more existing descriptors, either implicitly or explicitly. Most of them were covered by a single existing descriptor and they usually end up as descendants of it in the current hierarchy, gradually leading towards a more fine-grained MeSH vocabulary. These insights about the dynamics of the thesaurus are useful for the retrospective study of scientific articles annotated with MeSH, but could also be used to inform the policy of updating the thesaurus in the future.

Keywords MeSH · terminology extension · semantic indexing · biomedical literature

[1]National Center for Scientific Research "Demokritos", Athens, Greece
E-mail: {tasosnent, akrithara, paliourg}@iit.demokritos.gr
[2]Aristotle University of Thessaloniki, Thessaloniki, Greece
E-mail: {nentidis, greg}@csd.auth.gr

1 Introduction

The *Medical Subject Headings* (MeSH) thesaurus[1] is a collection of hierarchically organized entities for annotating biomedical knowledge, primarily literature in PubMed/MEDLINE[2], with topic labels. The basic conceptual entity is the MeSH *concept* which is a collection of synonymous terms for a particular domain meaning. Each concept has a *preferred term*, which is also used as the name of the concept. MeSH concepts are not directly used for annotating the literature. Instead, closely related concepts are grouped into MeSH *descriptors* that constitute the main MeSH elements for annotating biomedical literature with topic labels. Although a descriptor can consist of several concepts, each MeSH concept belongs to exactly one MeSH descriptor. All the concepts and terms of a descriptor are equivalent for the purposes of indexing and searching MEDLINE. Beyond MeSH concepts and descriptors, MeSH also provides some *Supplementary Concept Records* (SCRs) that are directly used for annotating articles with labels for substances, rare diseases, and organisms.[3]

As the MeSH hierarchy evolves through annual updates, new descriptors are introduced that were previously unavailable. This evolution of MeSH is essential, in order to follow the development of knowledge in the field. For example, new descriptors can be more fine-grained than old ones, providing a level of detail previously unavailable in the vocabulary. On the other hand, new high-level descriptors can also be added, providing new groupings of topics, under the light of the current understanding of the domain. In some cases, the topics

[1] `https://meshb.nlm.nih.gov/`
[2] `https://www.nlm.nih.gov/bsd/pmresources.html`
[3] `https://www.nlm.nih.gov/mesh/intro_record_types.html`

covered by the new descriptors may have been present in MeSH previously, covered by older descriptors. However, some new descriptors may cover topics that are totally new to the vocabulary, representing emerging concepts in the domain.

Despite their necessity for keeping MeSH up-to-date, the introduction of new descriptors raises practical challenges. Several applications, such as (semi-)automated semantic indexing of biomedical literature with MeSH labels, are based on supervised learning techniques that exploit accumulated data from previous use of the vocabulary. However, for new descriptors, no such annotated literature is available at the time of their introduction. Therefore, it becomes important to devise a mapping of existing literature to the new descriptors. Towards this direction, for each new descriptor, we are interested in whether the corresponding topic was already covered by old descriptors in MeSH or not.

In this context, the basic questions motivating this study on the provenance of new descriptors are the following:

- To what extent do the new MeSH descriptors cover emerging domain concepts that are really new for the MeSH thesaurus?
- For those new descriptors that do not cover emerging domain concepts, can we identify older descriptors that were used to cover these concepts?
- What is the current relation of the new descriptors with the old ones that they are related to?
- Is there any pattern over time concerning the introduction of new descriptors in MeSH and how the new descriptors relate to the old ones?

The main contribution of this work consists in developing a conceptual framework for exploring the provenance of new MeSH descriptors considering the hierarchical structure of the thesaurus. In particular, we describe an approach for identifying predecessor descriptors, that used to cover the topic of a new descriptor previously. Namely, a coding system is introduced for organizing the new descriptors based on two key dimensions: a) whether and how they have been covered in MeSH prior to their introduction, and b) their current position in the hierarchy in relation to their predecessors. In addition, a method is developed for the computational identification of predecessors and conceptual provenance codes for new MeSH descriptors. Finally, based on the proposed framework we perform an analysis that sheds light on the conceptual provenance of descriptors introduced in MeSH during the last fifteen years.

The rest of this paper is structured as follows. In Section 2 some background knowledge is summarized, regarding the structure of basic elements of the MeSH thesaurus and their relationships. In section 3 we provide a brief overview of work related to the extension of biomedical thesauri with new concepts, focusing on the MeSH thesaurus. In section 4 we propose a framework for identifying the predecessors of new descriptors and introduce new types of conceptual provenance to characterize the current relation of new MeSH descriptors with their predecessors. In section 5 we propose a method to automatically analyze the versions of the MeSH hierarchy, in order to identify the various types of provenance. In section 6 we present and discuss the results of this analysis, which lead to useful insights about the evolution of MeSH. Finally, in section 7 we conclude and indicate potential uses of our results in future research.

2 MeSH structure

The MeSH hierarchy is elaborately structured to efficiently represent and organize terms, concepts, and topics from the complex domain of biomedicine. This elaborate structure of MeSH is manually maintained by the US National Library of Medicine (NLM) through annual version releases. Each new version may incorporate new vocabulary terms. For example, since 2018 a project is running in NLM to incorporate the vocabulary of the NCBI taxonomy in MeSH, starting with terminology Viruses and extending to Archaea, Bacteria, and Fungi.[4] In addition, potential issues about inconsistencies, errors, or outdated information are also addressed during the annual maintenance of MeSH. For example, two closely related descriptors can be merged into one, as done with "Cascara" and "Rhamnus" in 2020.[5] This continuous maintenance and evolution of MeSH ensures that this unique resource remains as up-to-date and free of errors as possible.

In the MeSH hierarchy, each descriptor has exactly one *preferred concept* and may also have some subordinate (narrower, broader, or related) concepts that attach additional terms to the descriptor. For example, the descriptor for Dementia (Fig. 1) consists of three concepts. The preferred concept, which has two synonymous terms ("Dementia" and "Amentia") and two narrower concepts with a single term each. The preferred concept is the reference point for defining the subordinate concepts as narrower, broader, or related. Therefore, we consider the preferred concept as the dominant entity representing the main topic of a descriptor.

[4] `https://www.nlm.nih.gov/pubs/techbull/nd20/nd20_mesh_ncbi_taxonomy.html`

[5] `https://www.nlm.nih.gov/pubs/techbull/nd19/nd19_medline_data_changes_2020.html#concept_merge`

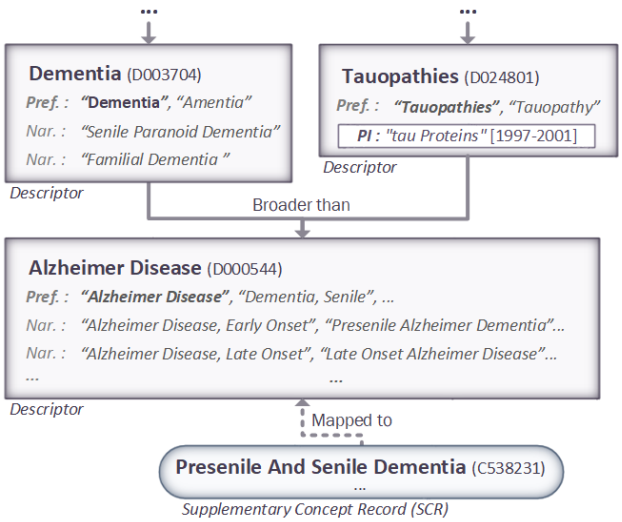

Fig. 1 MeSH concepts are grouped into descriptors, which are hierarchically organized and can also have a "Previous Indexing" note (PI). Each Supplementary Concept Record (SCR) is mapped to at least one descriptor.

The MeSH descriptors are hierarchically organized so that most descriptors have at least one broader descriptor as a parent. For example, the "Alzheimer Disease" (AD) descriptor has two parents, namely the descriptors "Dementia" and "Tauopathies", as shown in Fig. 1. Additionally, there are some top-level descriptors that have no parents and are the roots of the trees in the MeSH hierarchy, which are called MeSH *trees*.[6] The MeSH trees are grouped into sixteen MeSH *categories*[7], and each descriptor belongs to one or more MeSH trees and corresponding MeSH categories. For example, the AD descriptor belongs to the "Nervous System Diseases (C10)" tree in the "Diseases" (C) MeSH category and to the "Mental Disorders (F03)" tree in the "Psychiatry and Psychology" (F).

The exact position of a descriptor in a tree is determined by one or more *tree numbers* or *tree paths*. Each MeSH tree number of a descriptor is extending a tree number of some parent, and recursively includes the tree numbers for a series of ancestors reaching up to the corresponding root. For example, the AD descriptor has two tree numbers extending the tree numbers of "Dementia" (`F03.615.400.100`, `C10.228.140.380.100`) and one tree number extending the tree number of "Tauopathies" (`C10.574.945.249`).

The SCRs, that are also known as *Supplementary Chemical Records*, have a similar conceptual structure to descriptors, with one preferred MeSH *concept* and potentially some subordinate ones, but they are not part of any descriptor and are not directly included in the MeSH hierarchy. However, they are mapped to at least one descriptor. In Fig. 1 for example, the SCR "Presenile And Senile Dementia" is mapped to the AD descriptor as indicated by a dotted arrow towards the latter. In practice, this means that when indexers in PubMed/MEDLINE use this SCR to annotate an article, the article will also get automatically annotated with the mapped descriptors [12]. This mapping is important because it defines which descriptors cover the meaning of each SCR at the level of main MeSH topic annotations, that are primarily used for indexing and searching the literature.

3 Related work

The MeSH evolution is often studied in the broader context of the evolution of dynamic biomedical terminologies [8]. In this area, the effort has often been on defining and studying elementary and composite changes, that require one or more basic operations, e.g. adding, removing, merging, splitting, editing, and moving elements of a biomedical terminology. For instance, the CONCORDIA framework, which stands for "CONcept and Change-Operation Representation for any DIAlect" was proposed for representing, reporting, and documenting different types of change in medical terminologies [14], and MeSH was explicitly considered in this study. Studying different types of change in MeSH is also the focus of the work presented in this paper. The basic premise is that by studying the basic operations that lead to a change, one can identify the conceptual source of new elements.

Though MeSH is not an ontology, it is often mentioned or even treated as such in the relevant literature, and it has also been transformed into a meta-ontology, in an effort to formally express all knowledge about semantically indexing MEDLINE [1]. McCray and Lee [11] studied the evolution of MeSH in an ontological context, as a conceptualization of the biomedical domain. They focused on the evolution of the MeSH category "Psychiatry and Psychology" (F), capturing and quantifying change at the level of descriptors, as well as at more fine-grained terminological and lexical levels. In particular, they investigated whether change reflects the evolution of corresponding knowledge in the biomedical domain. Their results reveal that change in MeSH reflects both the evolution of biomedical knowledge, as well as some internal ontological restructuring efforts, such as the separation of behaviors from disorders.

[6] `https://www.nlm.nih.gov/mesh/intro_trees.html`

[7] `https://www.nlm.nih.gov/bsd/disted/meshtutorial/meshtreestructures/index.html`

Recently, Balogh *et al.* [3] studied the evolution of MeSH as a network focusing on the addition and removal of links between MeSH descriptors, which they call "attachment" and "detachment" of links respectively. Interestingly, they investigated whether these re-wiring events are associated with certain descriptor properties, such as the number of parents or descendants in the hierarchy. Their results suggest that old MeSH descriptors with many descendants appear to receive and lose children descriptors more than expected by chance. On the other hand, descriptors with many ancestors appear to receive and lose children descriptors less than expected by chance.

More recently, Cardoso *et al.* [5] suggested the interlinking of distinct versions of MeSH developing a historical knowledge graph, to extend queries for biomedical literature retrieval and for supporting the maintenance of semantic annotations. In particular, they introduce "evolution connections" between descriptor elements (e.g. terms) in different versions. In some cases, these evolutionary relationships can indicate the conceptual provenance for new descriptors, whereas in other cases they express more fine-grained internal restructuring, such as relocating a term from one MeSH concept to another. Identifying the conceptual provenance of descriptors is also central in the work presented in this paper. However, the focus here is at the level of topics and the goal is to capture additional provenance connections, beyond MeSH concepts, namely through SCRs and Previous Indexing information.

Some related studies also focus on the identification of the elements that need to change in MeSH and try to automate this process. For example, Sari [15] proposed an approach for propagating changes already incorporated in the Gene Ontology into appropriate changes in MeSH. However, other studies attempt to predict the extension of MeSH based on different approaches. For instance, Fabian *et al.* [10] proposed a method for finding siblings to a set of MeSH terms, analyzing the structure and the content of HTML pages on the Web. Guo *et al.* [17], on the other hand, proposed a "structure-based" method for recommending new siblings for MeSH descriptors, that was exclusively based on the positions of existing terms in the MeSH hierarchy. Eljasik-Swoboda *et al.* [9] proposed an embedding-based method for suggesting new sub-topics for existing topics of MeSH, combining both knowledge about the hierarchy and the analysis of documents already annotated with specific MeSH labels.

Other studies proposed approaches that beyond the analysis of annotated corpora and the structure of the hierarchy, they incorporate temporal information for the history of MeSH as well. Tsatsaronis *et al.* [16] proposed a method to predict which MeSH descriptors should be expanded with new children. This method combined information about the number of articles annotated with each descriptor in PubMed with information about the hierarchical position of each descriptor in MeSH and temporal features that capture changes. Cardoso *et. al* [6] also proposed a method for identifying concepts that require revision, based on structural and temporal information, as well as information from other resources including the UMLS and article annotations in PubMed. Beyond the expansion of concepts with more children, their method also suggested other types of revision, such as removal and relocation.

Finally, the MeSH thesaurus has also been considered in some studies in the context of topic modeling and evolution in the biomedical domain. In these works, MeSH labels are treated as keywords to develop a network of keyword co-occurrence from a document corpus, where latent (meta-)topics can be detected as clusters or communities. In this context, Castillo *et al.* [7] aligned such (meta-)topics of MeSH terms extracted from different time intervals based on the similarity of the corresponding sets. Then, they present an overview of the evolution of these matched (meta-)topics as a phylogeny-inspired network, where evolutionary events like merge and split can be identified. Balili *et al.* [2] propose the TermBall approach for both tracking and forecasting the evolution of such (meta-)topics of MeSH terms, treating them as evolving communities in the term co-occurrence network. The work presented here investigates evolutionary relations between biomedical topics as well, however, we focus on the level of MeSH descriptors, as used for indexing in PubMed, rather than the broader level of latent (meta-)topics reflecting the evolution of a domain.

In this study, we focus on newly added descriptors during the extension of MeSH. In particular, we study whether the meaning of each new descriptor has been covered by old descriptors previously, and if so, how its new position in the hierarchy relates to those of its predecessors. In contrast to most related work in the biomedical terminology evolution context, which focuses either on the operations to implement a change (addition, merge, etc) or on general features of the descriptors, such as depth in the hierarchy, we aim at characterizing the new descriptors according to their conceptual provenance. In other words, how they are related to their predecessors in previous versions of MeSH. In order to achieve this characterization, we investigate how we can identify the predecessors and introduce provenance types that provide new insight into the study of MeSH evolution.

4 A conceptual model for descriptor provenance

In this section, we introduce a conceptual model to characterize and group new MeSH descriptors based on their conceptual provenance. That is, we investigate the cases of previous coverage of new descriptors during the extension of MeSH. We define the notion of Previous Host (PH), as a predecessor of a new descriptor, and describe categories of descriptors based on how these predecessors can be identified. Subsequently, we introduce types of conceptual provenance, to characterise interesting cases of new descriptors, based on their current relation with each of their PHs in the hierarchy of MeSH.

4.1 MeSH extension and provenance

As the MeSH hierarchy evolves, the new descriptors introduced may cover domain concepts that are not totally new to the vocabulary. Some concepts may have been explicitly present in the previous version of MeSH. In particular, a concept of a new descriptor may have been available as a subordinate concept of an old descriptor or as an SCR concept. The latter case, of turning an SCR concept into a descriptor, is usually reported in a textual note in the new descriptor, called *Public MeSH Note* (PMN). For example, the "Adenocarcinoma of Lung" descriptor that was introduced in 2019, shown in Fig. 2, has a PMN field indicating its previous state as an SCR mapped to the "Adenocarcinoma" and "Lung Neoplasms" descriptors. Therefore, literature for "Adenocarcinoma of Lung" annotated in 2018, can be found with "Adenocarcinoma" and "Lung Neoplasms" topic labels.

In addition, even in cases where the concepts of the new descriptor have not been explicitly available as such, their meaning may have been implicitly covered by old descriptors. Such information is usually available as a Previous-Indexing note (PI) in the new descriptors, as in the case of the "Tauopathies" descriptor in Fig. 1. The PI note indicates that some old descriptors were used to annotate literature for the topic of the new descriptor, during a specific period prior to its introduction. For example, the "tau Proteins" descriptor was used to annotate articles about "Tauopathies" since 1997. This changed in 2002, with the introduction of a descriptor for "Tauopathies".

In this work, we refer to the MeSH version of the introductory year of a new descriptor as *version 1*, and the last year before *version 1* as *version 0*. In addition, we refer to such old descriptors that were used to annotate literature for the topic of the new descriptor

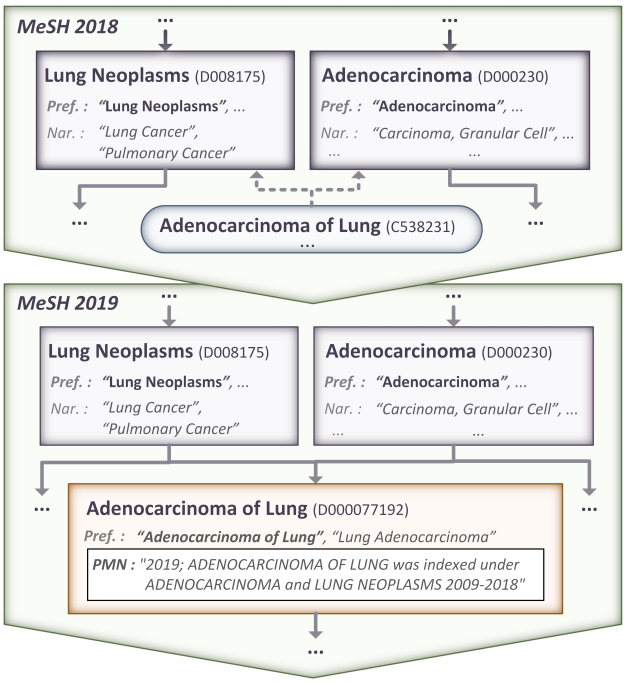

Fig. 2 The promotion of the SCR "Adenocarcinoma of Lung" into a descriptor in 2019.

in the version prior to its introduction (*version 0*), as its *Previous Hosts* (PHs). Apart from identifying the PHs of a new descriptor, the current relation of the new descriptor with its PHs is also important. For example, the new descriptor "Adenocarcinoma of Lung" was positioned in the hierarchy as a child of its two PHs (Fig. 2). Therefore, literature for "Adenocarcinoma of Lung" is still covered by "Adenocarcinoma" and "Lung Neoplasms" as done prior to the introduction of the new descriptor. On the other hand, the new descriptor for "Tauopathies" is not hierarchically related to its PH "tau Proteins". As a result, literature for "Tauopathies" is not covered by the "tau Proteins" descriptor after the introduction of the new descriptor in 2002.

Although the above cases are common, they are not the only types of relation one encounters between a new descriptor and its PH(s). Furthermore, as the MeSH hierarchy keeps evolving, the relation of a descriptor with one or more of its PHs can change in subsequent years, complicating the situation further. Therefore, this relationship depends on the version of MeSH considered, which we call *reference version*. In this work, aiming at a profound understanding and improved handling of new MeSH descriptors we investigate their origin. That is, whether they have been covered by descriptors (PHs) in the corresponding *version 0*, and if so, what their current relation to each of these descriptors is. In order to better quantify and organize these cases,

we define types of "conceptual provenance" for the new descriptors.

4.2 Previous Hosts (PHs)

A PH of a new descriptor is defined as a descriptor that was used to annotate articles for the topic of the new descriptor in the *version 0* of the new descriptor. In that sense, we say that the PH used to cover the topic of the new descriptor for the purposes of literature annotation in *version 0*. However, it is not required that a PH used to cover the topic exclusively. That is, a PH may have been used for indexing other topics as well, apart from the topic of the new descriptor. Therefore, several new descriptors may share the same PH. In addition, it is not required that a PH used to cover the topic of the new descriptor entirely. That is, a PH may have been used for indexing only part of a topic, for example in cases of a new high-level descriptor added to provide a new grouping of related topics.

A formal definition for a PH descriptor *d0* for a new descriptor *d1* can be based on the condition of *topic-overlap* as follows:

- The ***topic-overlap(d1, d0, v)*** is true when articles for the main topic of *d1* used to be indexed under *d0* in the MeSH version *v*. In all other cases, *topic-overlap(d1, d0, v)* is false.
- The ***previous-host(d1, d0)***, denoting that *d0* is a PH of *d1*, is true when *topic-overlap(d1, d0, v)* is true, where *v* is the *version 0* of *d1*. In all other cases, *previous-host(d1, d0)* is false.

This definition of a PH focuses only in the MeSH version that precedes the introduction of the new descriptor (*version 0*). Descriptors that used to cover the topic in older versions can be recursively described as the PHs of a PH and so on. However, our original motivation is to characterize each new descriptor based on whether its topic was already covered by MeSH, at the time of its introduction (*version 1*), or not. Therefore, in this work we do not track the history of each new topic in the distant past.

As already discussed in subsection 4.1 there are two types of coverage for a new descriptor in a previous version of MeSH. a) Explicit coverage, which is based on the conceptual structure of MeSH descriptors and SCRs into concepts, and b) implicit coverage, that can be identified based on the PI information. Based on the coverage type, we also characterize the corresponding PHs. That is, an *explicit* PH used to host a subordinate concept or used to be mapped from an SCR, that corresponds to the new descriptor. On the other hand, an *implicit* PH was used by the indexers for annotating articles that correspond to the topic of the new descriptor, without any explicit link with the latter in its conceptual structure.

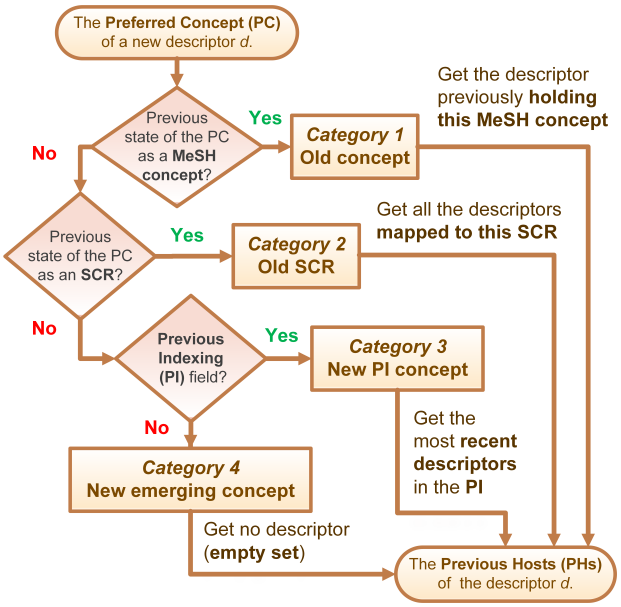

Fig. 3 Identifying the provenance category and the Previous Hosts (PHs) for a new descriptor.

Explicit PHs are of primary importance, as they provide strong conceptual links to the new descriptors. In our quest for a conceptual link to PHs, we focus on the preferred concept of each new descriptor. This is because the preferred concept is the dominant entity that represents the main meaning of a descriptor, as well as the vast majority of articles indexed with the descriptor. For descriptors that are new to the vocabulary, in the absence of any explicit PH, we exploit the PI field to identify any implicit PH. The "Tauopathies" descriptor, shown in Fig. 1, is such a case of a new descriptor without any explicit PH, where the PI field is exploited to identify the implicit PH "tau Proteins".

4.3 Provenance Categories

For the purpose of identifying the PHs of a new descriptor *d1*, we seek its preferred concept in the corresponding *version 0* of MeSH. Based on whether and how we identify it in existing descriptors, we define four cases (*categories*) of conceptual provenance, as depicted in Fig. 3 and described below:

Category 1. Old Concept: Although *d1* is a new descriptor, its preferred concept is available in the previous version of MeSH (*version 0*) as a subordinate concept of another descriptor *d0*. In this case of explicit

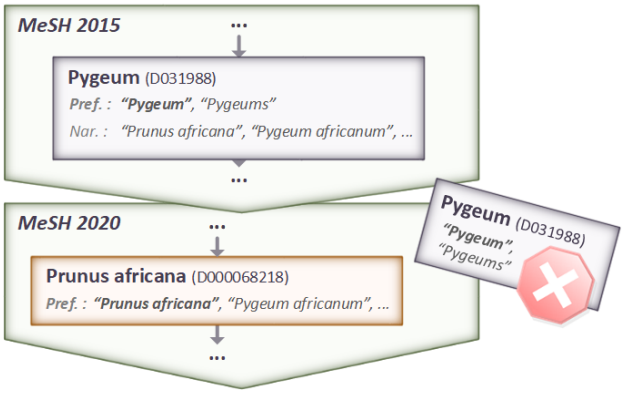

Fig. 4 An example of descriptor *succession*.

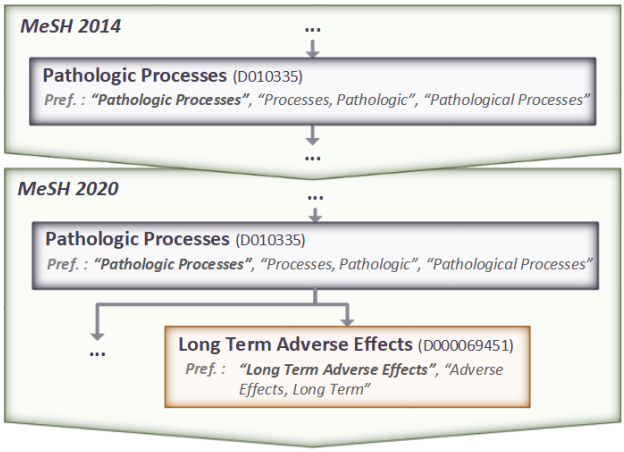

Fig. 5 An example of descriptor *emersion*.

coverage, since *d0* used to hold the preferred concept of *d1*, *topic-overlap(d1, d0, version 0)* is true. The descriptor *d0* therefore, is an explicit PH of *d1*. In addition, as each MeSH concept can only belong to a single descriptor in a given version of MeSH [12], *d0* is the unique PH of *d1*.

For example, "Prunus africana", shown in Fig. 4, introduced in 2016 as a descriptor, was a subordinate (narrower) concept of the "Pygeum" descriptor, which is not included in MeSH anymore. In this case, the unique PH of "Prunus africana" is the "Pygeum" descriptor, which explicitly included the concept "Prunus africana" in the version prior to the introduction of a dedicated descriptor for it.

Category 2. Old SCR: Alternatively, since the SCRs account for a large volume of domain concepts that are not included in MeSH descriptors [4], the preferred concept of *d1* may have been available as a concept in an SCR *scr*, prior to the introduction of *d1* (*version 0*). In this second case of explicit coverage, for each descriptor *d0* mapped from *scr* holds that the literature indexed under *scr* was also indexed under *d0*. Therefore, *topic-overlap(d1, d0, version 0)* is also true. As a result, each descriptor *d0* is an explicit PH of *d1*. For example, the descriptor "Adenocarcinoma of Lung" introduced in 2019, was previously available as an SCR mapped to the descriptors "Lung Neoplasms" and "Adenocarcinoma" (Fig. 2). These two descriptors are the explicit PHs of the new descriptor.

Category 3. New PI Concept: The preferred concept may be new, introduced together with the new descriptor *d1*. For such new descriptors, if previous-indexing (PI) information is available, this means that some other descriptors were previously used to index articles for the topic of *d1* (new PI concept). Therefore, the preferred concept of *d1*, though new in the MeSH thesaurus, was previously indexed under some older descriptors with other concepts, hence implicitly covered by them. In such cases of implicit coverage, the PI descriptors that were used until the introduction of *d1* are the ones with the most recent ending year in the accompanying period (*version 0*). Therefore, for each descriptor *d0* that was used until the introduction of *d1*, we have that *topic-overlap(d1, d0, version 0)* is true. As a result, the most recent PI descriptors are the implicit PHs of the new descriptor *d1*.

For example, "Zika Virus Infection" was introduced as a descriptor in 2015, and was not previously present as a concept in MeSH. However, it is annotated with a PI note, revealing that the descriptors named "Arbovirus Infections" and "Flavivirus Infections" have been used for indexing literature relevant to "Zika Virus Infection" until 2015. Therefore, these two descriptors are the PHs of "Zika Virus Infection".

Category 4. New Emerging Concept: On the other hand, there exist new descriptors where no PI information is available, no PH can be identified and their PHs is an empty set. Such totally new descriptors are expected to include emerging domain concepts without any significant presence in prior literature. Therefore, the curators begin indexing articles for a domain topic previously not indexed under any specific MeSH descriptor.

For example, "Long Term Adverse Effects" was introduced in 2015 as presented in Fig. 5. No "Long Term Adverse Effects" concept was previously present in MeSH and no PI information is available to report that articles for "Long Term Adverse Effects" were indexed under some particular descriptor until 2015. Therefore, this totally new descriptor has no PHs at all.

4.4 Provenance Types

Having identified the PHs and the provenance category of each new descriptor, next we investigate the hierarchical relation of the new descriptor with each one of its PHs. This relation starts with the introduction of the new descriptor in the MeSH hierarchy but may change in the course of the years, as the hierarchy evolves further. Therefore, to characterise the relation of a new descriptor *d1* with a PH *d0* in the context of a given *reference version* of MeSH, we focus on two basic properties of this relation in the corresponding hierarchy. Namely the *relation type* of *d1* with *d0* and the *distance* between them in the hierarchy.

- The ***relation_type(d1, d0)*** is: (a) *ancestor* when *d1* has at least one *tree number* that includes a tree number of *d0*, (b) *descendant* when *d0* has at least one *tree number* that includes a tree number of *d1*, (c) *unrelated* when none of the *tree numbers* of *d0* includes or is included in any of the *tree numbers* of *d1*, and (d) *undefined* when *d0* is not present in the *reference version* of MeSH.[8]
- The ***distance(d1, d0)*** is the number of other descriptors included in the shortest path connecting *d1* and *d0*. A path connecting two descriptors comes from any pair of overlapping *tree numbers* of the two descriptors. Two *tree numbers* are overlapping if one of them includes the other, or if both include the same *tree number*, which corresponds to a common ancestor. If no such pair of *tree numbers* exists for *d0* and *d1*, then they are located in positions of the hierarchy that are not connected. In this case, the *distance(d1, d0)* can be considered to be infinite. If *d0* is not present in the *reference version* of the hierarchy the *distance(d1, d0)* is *undefined*.

For example, the relation between the new descriptor "Adenocarcinoma of Lung" and its PH "Lung Neoplasms" has *ancestor relation_type* and *zero distance* with MeSH 2019 as *reference version* (see Fig. 2). On the other hand, the current relation of "Prunus africana" and its PH "Pygeum", in the context of MeSH 2020 *reference version*, has *undefined relation_type* and *undefined distance* (see Fig. 4). Based on the *relation_type* and the *distance* of the relation between a new descriptor *d1* and a PH *d0* we also define some cases of interest, which we call conceptual provenance *types*.

Type 0. Emersion: No PH found. For new descriptors in category 4, where no PH can be identified. In these cases there is no PH for which to investigate the current relation, therefore we define the trivial type of provenance *emersion*, which includes all descriptors of provenance *category 4* and only descriptors of *category 4*.

This exceptional type of provenance does not reflect the relationship with any PH, therefore it is not based on *relation_type* and *distance* in a specific *reference version* of MeSH. The meaning of such a completely new descriptor is emerging when the new descriptor is introduced and is characterized as emergent hereafter. The "Long Term Adverse Effects" descriptor introduced in 2015, is an example of *emersion* (see Fig. 5).

Type 1. Succession: relation_type(d1, d0) = undefined and distance(d1, d0) = undefined. For some new descriptors a PH can be no longer present in the *reference version* of MeSH. In this case, *d1* is considered one of the successors of *d0*, because at least some of the articles that used to be annotated with *d0*, in *version 0* for *d1*, are annotated with *d1* instead, in the *reference version* of MeSH. In the example of Fig. 4, the new descriptor "Prunus africana" is a case of succession, as its PH is not available in the context of the *reference version*, MeSH 2020.

Type 2. Subdivision: relation_type(d1, d0) = ancestor and distance(d1, d0) = 0. A new descriptor *d1*, whose PH *d0* has become its parent. In this case, *d0* covers the topic of the new descriptor entirely, but *d1* supports the partition of the corresponding literature into more fine-grained conceptual sets. This is the most expected type of relation between new descriptors and their PHs, as the vocabulary evolves towards more detailed descriptors to support more precise topic annotations. In the *subdivision* example of Fig. 6, "Regulated Cell Death" introduced in 2020, used to be indexed under "Cell Death" until 2019, which became its parent.

Type 3. Submersion: relation_type(d1, d0) = ancestor and distance(d1, d0) > 0. A new descriptor *d1*, whose PH *d0* has become an ancestor, but not a parent. This is similar to *subdivision*, as they both are characterized by *ancestor relation_type*, but at least one other descriptor appears between *d0* and *d1* in the hierarchy. This is also in accordance with the evolution towards more detailed descriptors, as the *d0* keeps covering the topic of the new descriptor entirely. However, the distance between

[8] Although the four relation types are expected to be mutually exclusive as a rule, the types (a) *ancestor* and (b) *descendant* can coexist in exceptional cases. In particular, analyzing the last fifteen versions of MeSH we found six such descriptor pairs where one descriptor is both an ancestor and a descendant of the other descriptor. All these pairs were introduced prior to 2014 and none of them corresponds to a descriptor and its PH.

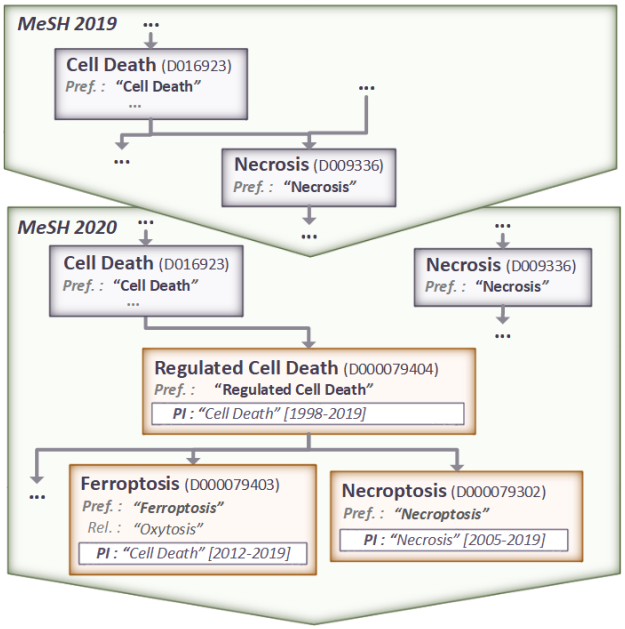

Fig. 6 "Regulated Cell Death" as a *subdivision* of "Cell Death", the *submersion* of "Ferroptosis" and the *detachment* of "Necroptosis".

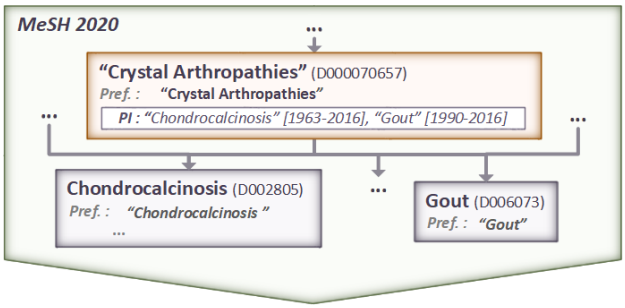

Fig. 7 The "Crystal Arthropathies" *overtopping* its PHs.

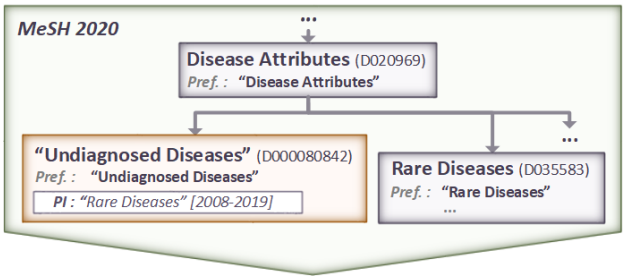

Fig. 8 The *detachment* of "Undiagnosed Diseases" from "Rare diseases".

them suggests that intermediate levels of detail are also available.

"Ferroptosis", introduced in 2020 (Fig. 6), is an example of *submersion*, as it was indexed under "Cell Death" until 2019, which is now an ancestor but not a parent of it. In this example, "Regulated Cell Death" which acts as the intermediate level of detail, was also introduced together with "Ferroptosis", which explains why "Ferroptosis" articles were previously indexed under "Cell Death" instead of "Regulated Cell Death".

Type 4. Overtopping: relation_type(d1, d0) = descendant. A new descriptor *d1*, whose PH *d0* has become its descendant. In this case, although literature for the new topic used to be indexed under *d0* in the past (*version 0*), *d1* is an ancestor of *d0* in the *reference version* of MeSH, hence broader than it. Such new descriptors provide a new grouping of the old topics, potentially enhanced with additional terms for the aggregate topic. In the example depicted in Fig. 7, "Crystal Arthropathies", introduced in 2017, has two implicit PHs, as it was indexed as "Chondrocalcinosis" and "Gout" until 2016. Both of them are children of "Crystal Arthropathies" in 2020, hence overtopped by it.

Such cases seem less expected than the ones with *ancestor relation_type* (*subdivision* and *submersion*), as this situation suggests that *d0* used to cover only a part of the topic of *d1*. In addition, *overtopping* is less interesting from a practical point of view, as the use of the narrower descriptor that covers a topic is a common MeSH-indexing practice [12]. Therefore, though different levels of detail may exists between the new descriptor and its descended PH, splitting this small group of cases based on the *distance* would not add any particular value.

Type 5. Detachment: relation_type(d1, d0) = unrelated. A new descriptor *d1* that is not related to its PH *d0* with any of the above relations. In this case, *d1* is detached from *d0*, placed in a position without the one being included by the other. In the example of Fig. 6, "Necroptosis", introduced in 2020 as a child to "Regulated Cell Death" in the "Phenomena and Processes" MeSH category (G), was previously indexed as "Necrosis". Although "Necrosis" used to be a child of "Cell Death" in 2019, in 2020 it belongs only to the "Diseases" MeSH category (C) and is not directly related to "Necroptosis". Therefore, we consider the "Necroptosis" descriptor to be detached from its PH "Necrosis" in 2020.

Detached descriptors may be positioned quite close to their PH in terms of *distance*, but are not related as ancestors or descendants to it. In the example of Fig. 8, "Undiagnosed Diseases", introduced in 2020 as a child descriptor to "Disease Attributes", was previously indexed under "Rare Diseases" which is also a child of "Disease Attributes". However, we consider that "Undiagnosed Diseases" is detached from its PH "Rare Diseases", as their topics are effectively disjoint. That is,

Table 1 Provenance codes characterizing the relationship of a new descriptor with a PH, encoding categories and types as prefixes and suffices respectively. The exceptional case of *emersion* type corresponds to code 4.0.

Provenance Type		Properties		Provenance Category		
		relation_type	distance	1. Old concept	2. Old SCR	3. New PI concept
.1	**Succession**	undefined	undefined	1.1	2.1	3.1
.2	**Subdivision**	ancestor	0	1.2	2.2	3.2
.3	**Submersion**	ancestor	> 0	1.3	2.3	3.3
.4	**Overtopping**	descendant	≥ 0	1.4	2.4	3.4
.5	**Detachment**	unrelated	≥ 1	1.5	2.5	3.5

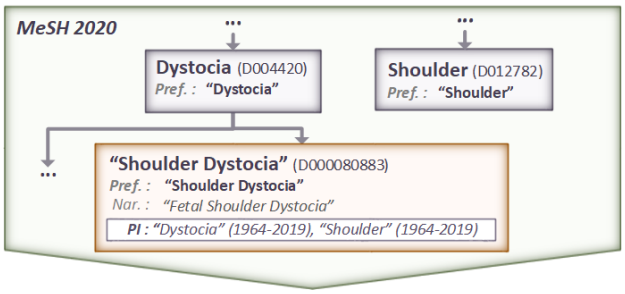

Fig. 9 The detachment of the "Shoulder Dystocia" descriptor has two provenance codes, namely, code 3.2 for the subdivision of the PH "Dystocia" and code 3.5 for the detachment from the PH "Shoulder".

neither of the two topics includes the other in the *reference version* (MeSH 2020).

Provenance codes: In order to easily refer to both category and type of conceptual provenance, we adopt a composite *provenance code*, with a prefix indicating the category of a descriptor and a suffix indicating the *type* of its relation to some PH, separated by a dot, as shown in Table 1. For example, the *provenance code* for "Necroptosis" (Fig. 6) is 3.5 indicating a provenance *category 3* for "new PI concept", as the PH has been identified based on PI information, and a provenance type 5 for *detachment* from "Necrosis". Similarly, the provenance code for "Prunus africana" (Fig. 4) is 1.1 with *category 1* for "old concept", and type 1 for *succession* of "Pygeum". In the special case of type 0, *emersion*, the preferred concept of the new descriptor is by definition new (*category 4*), hence, all *emersion* cases have the trivial provenance code 4.0.

As some new descriptors can have more than one PH, the provenance types described above are not mutually exclusive.[9] Therefore, a new descriptor can have multiple *provenance codes*. This is not true for the provenance categories, therefore all *provenance codes* of a specific new descriptor begin with the same prefix. For example, "Shoulder Dystocia" depicted in Fig. 9, was introduced in 2020 as a child descriptor to "Dystocia". Articles for shoulder dystocia were indexed as both "Dystocia" and "Shoulder" until 2019, hence it is both a case of *subdivision* (3.2) of the PH "Dystocia" which became its parent and a case of *detachment* (3.5) from the PH "Shoulder" which is not directly related with the new descriptor.

[9] In addition, multiple provenance types could also be the result of exceptional cases, where a new descriptor is related to a single PH, both as an ancestor and as a descendant.

5 Computing provenance of MeSH descriptors

In this section, we describe the computational tools developed for the automated identification of new MeSH descriptors, their PHs, and *provenance codes*, in the context of the conceptual model introduced in section 4. These tools, access the original source files of MeSH[10], as provided by NLM, in the MeSH XML format[11]. Therefore, all available information is accessible by the tools and any new versions of the hierarchy can be directly incorporated upon release. Figure 10 illustrates the sequence of processing steps that are involved in relating new descriptors to their PHs. The source code of the tools is openly available on GitHub.[12]

5.1 Harvesting MeSH versions

As mentioned in previous sections, we focus our analysis on the provenance of descriptors that are present in a *reference version* of MeSH, namely the latest one. Therefore, we do not process descriptors that appear and disappear in various older versions. However, we are still interested in annotating descriptors that appear in older versions and remain available in the *reference version*. As a result, we need to process older versions as well, covering a period from *year_0* to *year_N*.

In particular, the process starts with the harvesting of MeSH files for different versions of the hierarchy. This step begins with parsing the basic XML file for each

[10] `https://www.nlm.nih.gov/databases/download/mesh.html`
[11] `https://www.nlm.nih.gov/mesh/xmlmesh.html`
[12] `https://github.com/tasosnent/MeSH_Extension`

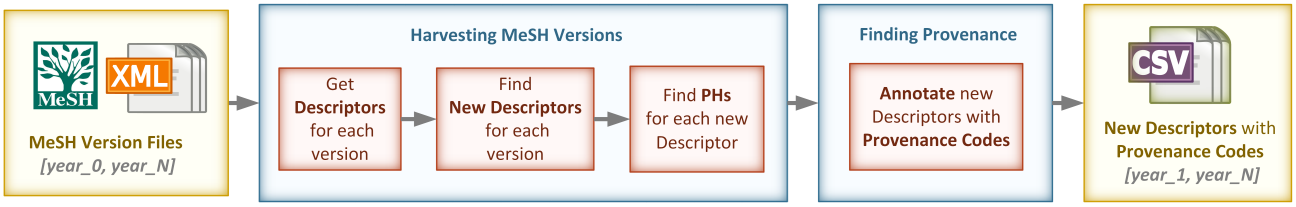

Fig. 10 The computational process for identifying new descriptors and annotating them with provenance codes.

year to extract the descriptors available in this version. This set of descriptors is then compared to those of the previous year to identify the new ones. The same process is repeated for each year, with the exception of the very first one, for which no previous version is available. Apart from the basic file comprising the MeSH descriptors, the XML file of the SCRs is also parsed for each version, to extract the corresponding set of available SCRs. These are needed for the identification of provenance categories and types.

Extracting descriptor attributes: For the descriptors of interest, a number of attributes need to be extracted, in order to help us trace its provenance. The most important attribute is the MeSH code of the descriptor, which is the unique identifier considered for checking descriptor existence and identity. Other relevant information includes the positions of a descriptor in the hierarchy (*tree numbers*), its preferred concept, and the content of the PMN and the PI fields. Most parts of this attribute extraction step are quite straightforward, as we primarily rely on the unique identifiers of the entities involved in the analysis. For example, the information needed for identifying the earlier status of a descriptor as a subordinate concept in its *version 0*, is the unique concept identifier of its preferred concept. This is because we need to compare this identifier with the identifiers of subordinate concepts of any descriptor in *version 0*.

However, automated extraction of information from the PMN and PI fields proved more challenging as these fields contain information in semi-structured text, meant to be read by humans. Therefore the structure of this text is inconsistent, while descriptors and SCRs are mentioned with their current preferred terms, instead of the corresponding unique identifiers. Consequently, we adopted a semi-automated approach, based on regular expressions, in order to extract information from these fields. In the large majority of cases, we managed to minimize the required manual effort as described below.

Extraction from the PMN field: The PMN (*Public MeSH Note*) field of a MeSH descriptor typically consists of sentences separated by semicolons and may provide varying information, such as the year the descriptor was introduced and changes in the preferred term. Of particular interest for this work, are PMN sentences that report the earlier status of the descriptor as an SCR. This is done with expressions of the form "*X was indexed under Y*", where *X* is the SCR and Y comprises one or more descriptors together with the corresponding time periods, as shown in the example of Fig. 2. This is useful as in some cases an SCR that gets "promoted" to descriptor may undergo some minor term modifications and receive a new identifier. In such cases, exploiting the PMN is the only way to identify the old SCR for the new descriptor, which would otherwise be considered totally new.

Therefore, when attempting to associate a new descriptor to an earlier SCR, we start by comparing the identifier of the preferred concept of the descriptor to the concept identifiers in earlier SCRs. If this exact-match search fails, we resort to the use of the PMN expressions mentioned above. In particular, we first use regular expressions to extract from the PMN field the preferred term (X) of the old SCR and map it to some SCR identifier in the corresponding version of MeSH. In our analysis, this method managed to automatically identify the missing links for the majority of cases (74%) where the PMN field matches the "*X was indexed under Y*" expression and the exact-match search fails.

For the few remaining cases, we calculated the similarity of X and the current descriptor name to earlier SCR terms. Based on this similarity, the system produced best-match suggestions, which were confirmed manually. More details about this method are available in a technical report available online [13]. There is also a small number of cases where more than one old SCRs is reported in the PMN field. In such cases, only the first SCR was considered, as this usually corresponds to the preferred concept of the new descriptor, representing its central meaning.

Extraction from the PI field: The PI (*Previous-Indexing*) field of a MeSH descriptor can be used to link a new descriptor to old ones when such a link is not provided explicitly, that is by a previous state of

the descriptor as a subordinate concept or SCR. The PI field contains a list of semi-structured notes in English. Each note usually consists of the relevant descriptors for a previous period, often followed by the corresponding time period in parentheses (Fig. 1). Exploiting this pattern we used regular expressions to extract the terms and the corresponding time periods.[13] In cases where the PI field consists of multiple notes, all the descriptors with the most recent end year are considered as PHs, as done for "Shoulder Dystocia" in the example of Fig. 9. Any older PI elements are neglected.

Selecting provenance type: In the last part of the MeSH harvesting step, each new descriptor is annotated with conceptual provenance codes. In particular, the first step is to select the provenance category based on the previous state of the current preferred concept as a subordinate concept or an SCR concept in the corresponding *version 0*, as depicted in the schema of Fig. 3. Then, the provenance type is selected, based on the current relation of the new descriptor to its PHs, which have been identified by the extraction process. Combining the provenance types with the category, the complete set of provenance codes is formed. The end result is a collection of all the new descriptors that have been introduced during the period considered and remain available in the *reference version* of MeSH. These descriptors are annotated with their basic information and provenance annotations and stored in CSV files named after the year that corresponds to the $vesrion_1$ for each descriptor.

6 Analysis of MeSH versions

6.1 Analytical setting

In this work, we applied the process presented in section 5 on the source files of all versions of MeSH published in the last fifteen years. In this manner, we identified and annotated all the descriptors introduced during this period, considering MeSH 2020 as the *reference version*. In other words, we are interested in the current status of the descriptors, but we use the year of their introduction *version 1*, in order to identify their previous hosts (PHs) and provenance category. The result of the computational processing is a CSV file for each MeSH version, comprising the new descriptors introduced this year and their provenance annotations.

[13] Some exceptions not fitting the patterns were identified and handled manually.

Table 2 The distribution of the 6,915 new descriptors (2006 - 2020) into provenance codes. The total per category can be lower than the sum of distinct type counts as the types are not mutually exclusive.

		Prov. Category			
Prov. Type		**1. Old con.**	**2. Old SCR**	**3. New PI con.**	**Total /type**
.1	**Succession**	21	12	84	117
.2	**Subdivision**	276	967	1,603	2,846
.3	**Submersion**	47	535	506	1,088
.4	**Overtopping**	24	7	91	122
.5	**Detachment**	151	364	1,313	1,828
Total/category		519	1,616	3,060	**5,195**

The total for *category 4*, Emersion (4.0), is 1,720.

As a final step, these files[14] are parsed and analysed to produce statistics and diagrams that provide alternative views of conceptual provenance in the course of MeSH expansion in order to answer the basic questions driving this study. In particular, the diagrams that are generated present the frequencies of provenance categories, types, and codes per year of introduction and in total. Based on these diagrams, we attempt to answer the basic questions driving this study and identify patterns and observations that may be of interest for understanding the dynamics of the extension of MeSH.

6.2 Overview of new descriptors and their provenance

Table 2 presents the distribution of new descriptors into provenance categories and types. In total, 6,915 descriptors were introduced in MeSH since 2006 and were retained until 2020. This corresponds to an extension of about 30%, compared with the 22,997 descriptors available back in 2005, and indicates that about 23% of all current descriptors have been introduced during the last fifteen years.

The new descriptors introduced for new concepts that have been implicitly covered in their *version 0* by old descriptors (*category 3*) is the most frequent provenance category, accounting for about 44% of all new descriptors. New descriptors for old concepts that have been explicitly covered in previous versions account for about 31% of all new descriptors, with the majority of cases covered by SCRs (*category 2*, ∼23%) and only a small portion having a previous status as a subordinate concept (*category 1*, ∼8%). This suggests that new descriptors for concepts already covered explicitly by older descriptors are primarily added for promoting SCRs to descriptors (*category 2*), rather than for

[14] `https://raw.githubusercontent.com/tasosnent/MeSH_Extension/main/NewDescriptors_2006_2020.csv`

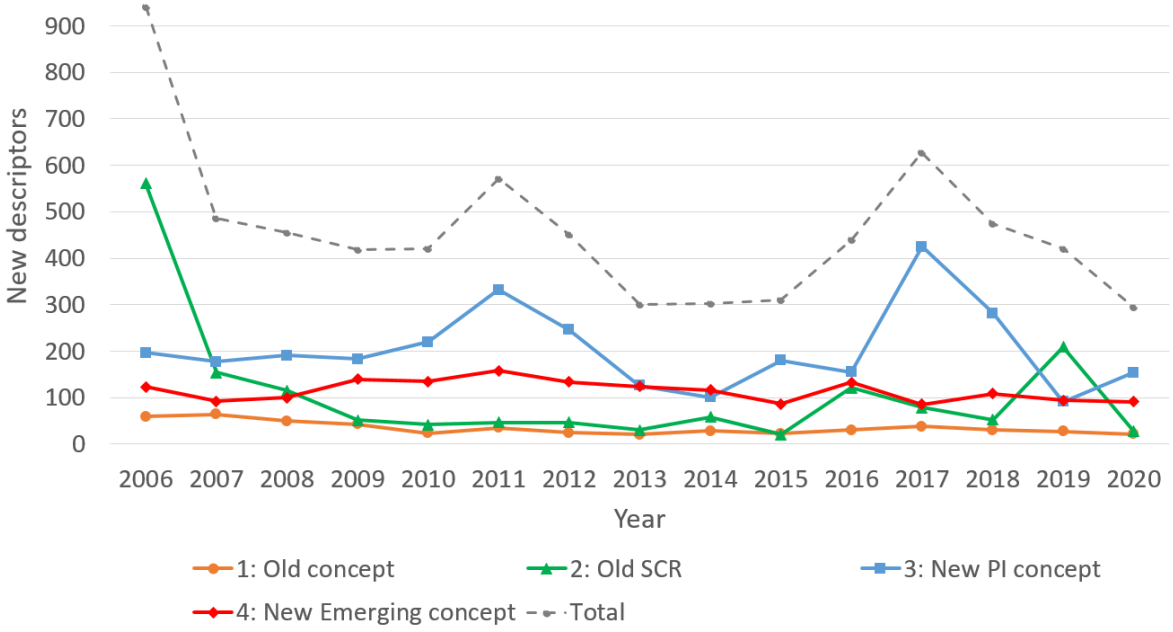

Fig. 11 Frequency of provenance categories for new descriptors, per year of introduction.

promoting subordinate concepts restructuring old descriptors (*category 1*).

On the other hand, new descriptors for emerging concepts (*category 4*), which are totally new for the MeSH vocabulary, account for 25% of all new descriptors. This relatively low frequency of *Emersion* suggests that in most cases new descriptors are linked to domain entities that are already covered by other descriptors either implicitly (*category 3*) or explicitly (*category 1* and *2*). Therefore, the new conceptual entities that are very often introduced (*category 3* and *4* account for 69% of new descriptors) are not completely novel, but they usually offer dedicated descriptors to known concepts (*category 3*).

Furthermore, the annual distribution of new descriptor categories, shown in Fig. 11, confirms the consistently high frequency of *categories 3* and *4* throughout the years. In particular, both the introduction of descriptors for new PI concepts and new emerging concepts accounts for at least around 100 cases annually for the whole period of study. However, *category 4* is more stable around its mean value (AVG) of almost 115 cases per year, with a standard deviation (SD) of 22 cases, whereas *category 3* presents more variation around its mean of 204 cases (SD ∼85 cases), reaching up to 300 and 400 cases in certain years.

On the other hand, the promotion of existing SCRs into descriptors (*category 2*) seems the less predictable category with an AVG around 108 and a SD of around 131 cases per year. In particular, in certain years (e.g. 2006, 2019) there seems to be a surge of such cases, while in others the number is much smaller. Finally, the evolution of existing subordinate MeSH concepts into independent descriptors (*category 1*) seems the least frequent and the most stable category with an AVG of around 35 and a SD of around 13 new descriptors per year.

The extreme peak of more than 900 new descriptors observed in 2006, may be the result of an effort at NLM to restructure descriptors for chemicals that combined meanings for activity and structure. This effort, which has been spanning across many years, was continued in 2006.[15] In addition, promoting SCRs to descriptors was particularly encouraged this year in NLM[16], which is in agreement with the fact that this peak seems to be almost exclusively attributed to promoted SCRs (*category 2*), which are known to represent mainly chemicals. This is also confirmed by the distribution of new descriptors into MeSH categories (Fig. 12), as 73% of the new descriptors introduced in 2006 belong to "Chemicals and Drugs" (D). This relative frequency for 2006 far exceeds the overall relative frequency of category D for the whole period considered, that is around 41%.

Two less extreme peaks are also observed in 2011 and 2017, with the introduction of about 600 new descriptors each. In contrast to the 2006 peak, these ones

[15] `https://www.nlm.nih.gov/pubs/techbull/nd05/nd05_2006_MeSH.html`

[16] Cho, Dan-Sung (NIH/NLM) personal communication

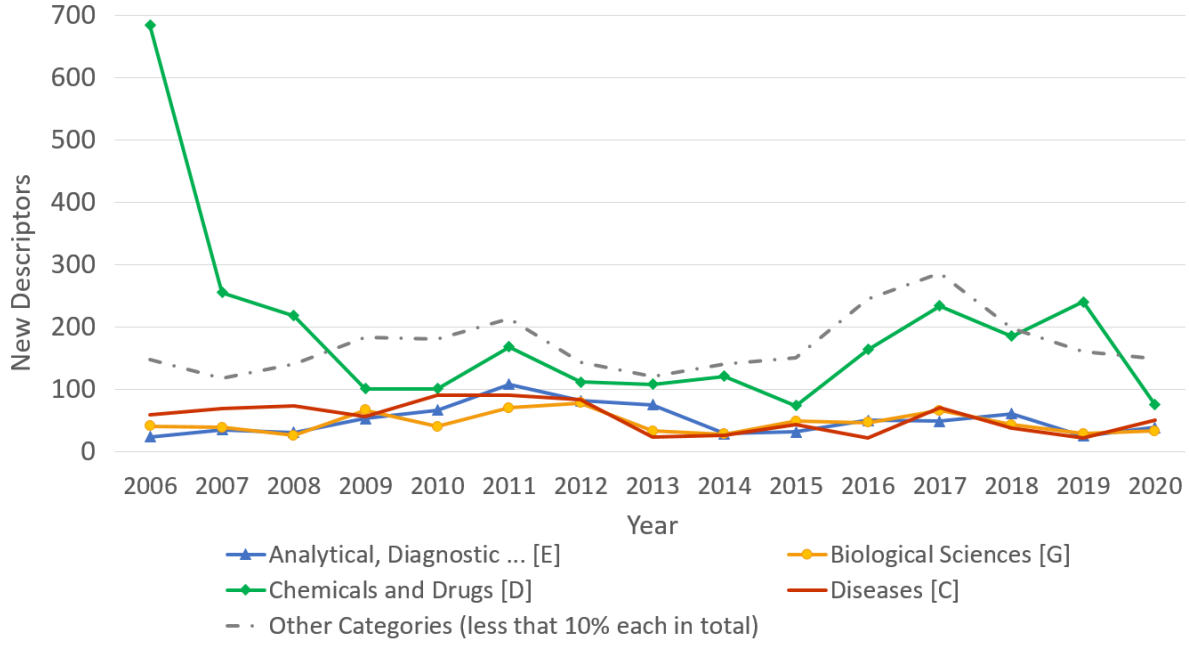

Fig. 12 Frequency of MeSH categories for new descriptors, per year of introduction. The four MeSH categories accounting for at least 10% of new descriptors each are presented independently. The remaining twelve cases, that have an overall frequency of less than 10% of new descriptors each, are collectively presented as "Other Categories".

seem to be primarily attributed in *category 3* cases, as other categories present frequencies close to the ones of the adjacent years. In addition, the distribution of the corresponding new descriptors into MeSH categories suggests that, though the chemicals category D has relatively high frequencies these years, other MeSH categories also have a considerable contribution to these peaks. In other words, these peaks of new descriptors for new PI concepts (*category 3*) seem to be more evenly distributed across MeSH categories, than the 2006 peak of *category 2* cases.

For 2011, this is in agreement with a focus in MeSH on projects related to the categories "Biological Sciences" (G) and "Analytical, Diagnostic and Therapeutic Techniques, and Equipment" (E) in MeSH.[17] The peak of 2017, on the other hand, seems to be affected by the "MeSH Protein Project"[18], as part of which, almost 290 new descriptors were added. The aim of this project was to achieve alignment of gene families, as described by the Human Genome Nomenclature Committee (HGNC), with protein classes in MeSH. In addition, more new descriptors than usual are introduced in 2017 for some less frequent MeSH categories, such as "Health Care" (M) and "Persons" (N).

Regarding the provenance types of new descriptors, *Subdivision* (.2) is the most common case (41%), followed by *Detachment* (.5, 26%) and *Emersion* (.0, 25%). *Submersion* has also a considerable frequency of 16%, but *Succession* (.1) and *Overtopping* (.4) are quite scarce, accounting for about 2% each. This distribution seems to be in agreement with the expected low frequency of new descriptors being broader of their PHs (*Overtopping*) or having their PHs removed from the vocabulary (*Succession*). However, the frequency of new descriptors that are no longer covered by any of their PHs (*Detachment*) seems quite notable, representing 35% of non-emerging new descriptors (*categories 1, 2* and *3*). This implies that the addition of dedicated descriptors for concepts that used to be covered by older descriptors (PHs), often serves the removal of these subordinate, supplementary, or implicitly covered concepts from these PHs, improving the specificity of the latter.

On the other hand, the majority of new descriptors appear to be still covered by their PHs, offering subtopics to the latter. In particular, about 55% of all the new descriptors have at least one *ancestor* in their PHs, that is they belong to *Subdivision* or *Submersion* cases, with the last being far less frequent as expected (16%). This suggests that only half of the new descriptors end up as descendants of their PHs. However, focusing on the 5,195 non-emerging new descriptors, that actually have at least one PH (*categories 1, 2* and *3*), this relative frequency increases to 73%, with *Subdivision* accounting for 55% of the cases and *Submersion* for only 21% of them. This is in agreement with

[17] Cho, Dan-Sung (NIH/NLM) personal communication
[18] `https://www.nlm.nih.gov/pubs/techbull/nd16/nd16_mesh.html`

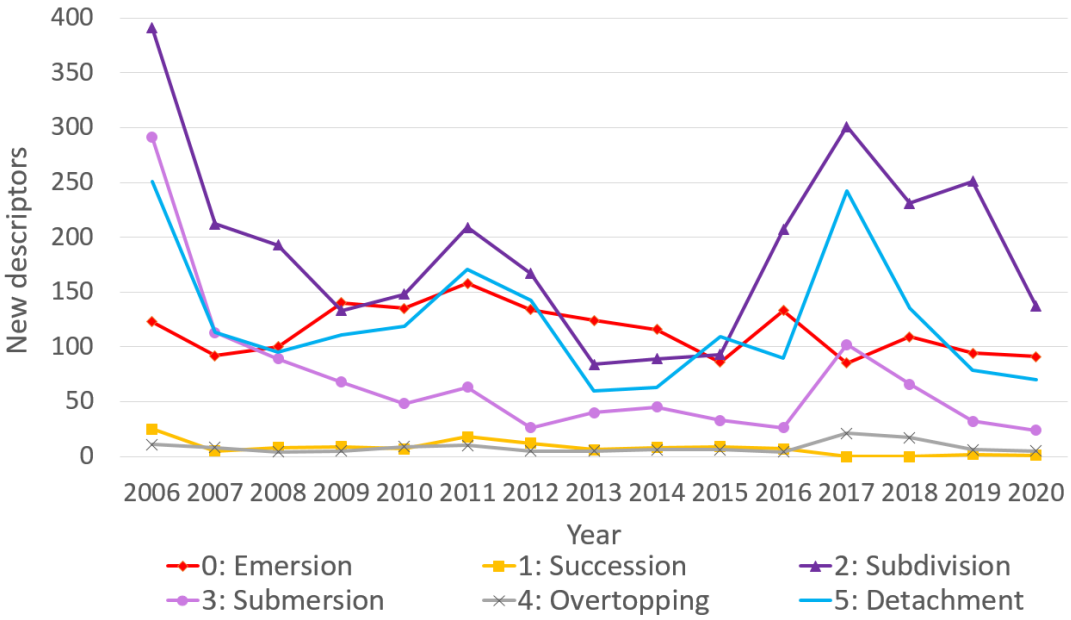

Fig. 13 Frequency of provenance types for new descriptors, per year of introduction.

the expected evolution of the topic vocabulary towards more fine-grained descriptors. The latter support more precise topic annotations and retrieval, especially when more documents are accumulated for some descriptors during the years.

Figure 13 presents the annual distribution of new descriptors into provenance types. Despite annual fluctuations, there seems to be a clear separation of the frequent types (*Emersion*, *Subdivision*, and *Detachment*), from the infrequent ones (*Succession* and *Overtopping*) throughout the period of study. Finally, the *Submersion* type seems to fall in-between the two groups. In addition, it seems that the infrequent types of *Succession* and *Overtopping* vary the least through the years (SD 7 and 5 respectively). The more frequent types of *Subdivision*, *Detachment* and *Submersion* seem to be the less predictable (SD 81, 56 and 66 respectively), whereas the trivial type of *Emersion*, though quite frequent as well, appears to be relatively stable, as already noticed for *category 4*.

As with MeSH categories, the surge of cases in certain years is not evenly distributed across all provenance types. Although the representation of all provenance types appears to be close to their overall relative frequency in the peak of 2011, this is not always the case. In 2006, *Submersion* seems to be over-represented, accounting for 31% of the cases, which is more than double its overall relative frequency for the period of study (16%). This could be related to the complex organization of chemical SCRs into groups and subgroups. For example, "Receptors, Scavenger" as well as the six classes of them ("Scavenger Receptors, Class A" etc) used to be SCRs indexed under "Receptors, Immunologic" until their promotion into descriptors in 2006. Although "Receptors, Scavenger" was added as a child (2.2) to their PH "Receptors, Immunologic", the six classes were added as children of "Receptors, Scavenger", hence more distant descendants of "Receptors, Immunologic" (2.3).

On the other hand, *Detachment* seems to be over-represented in the peak of 2017, accounting for 39% of the new descriptors, whereas its overall relative frequency for the whole period is 26%. Some of these *Detachment* cases are new descriptors for protein domains or motifs detached from the corresponding protein descriptors, which can be related to the "MeSH Protein Project". For example, the new descriptor "Methyl CpG Binding Domain" detached from its PH "DNA-Binding Proteins". In addition, several new descriptors in the MeSH categories "Health Care" (M) and "Persons" (N) appear to represent medical professions detached from the corresponding medical domains. For example, the new descriptor "Nephrologists" was detached from its PH "Nephrology".

Some of the types, in particular *Subdivision* (.2) and *Detachment* (.5), seem to be correlated in the way they increase or decrease over the years. It would, therefore, be of interest to investigate whether the correlation of their annual frequencies observed in Fig. 13 should be attributed to the addition of descriptors that exhibit both these provenance types simultaneously. This is mainly possible in *category 2* and *category 3* where the

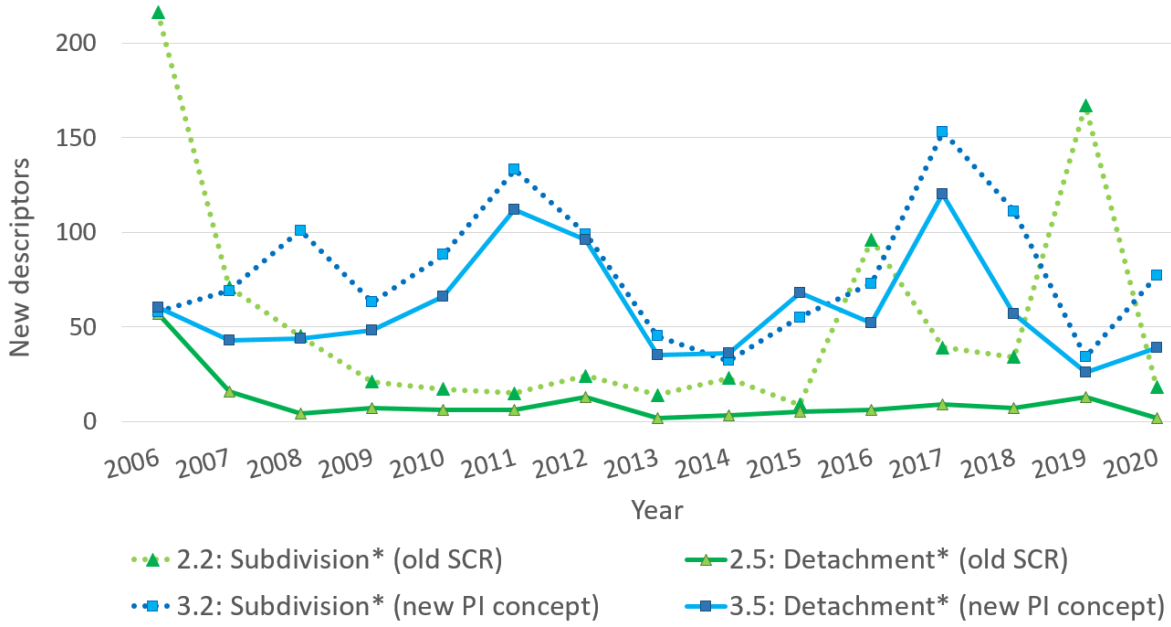

Fig. 14 Frequency of type *Subdivision* (.2) and *Detachment* (.5) in new descriptors introduced during the last fifteen years, per provenance category. The asterisk (*) denotes that only descriptors with a single type are considered, excluding descriptors combining more than one type.

availability of multiple PHs for a new descriptor can lead to multiple provenance codes. In practice, however, new descriptors with multiple provenance codes are not very common, representing almost 17% of all new descriptors in these two categories.

Focusing on the majority of new descriptors that have a single provenance type, we compare the annual frequencies of the *Subdivision* (.2) and *Detachment* (.5) (Fig. 14). The correlation of the frequencies seems to be preserved in the frequent *category 3* (blue lines with square markers). In other words, even when looking at distinct new descriptors that share no common provenance types, *Subdivision* (3.2) and *Detachment* (3.5) seem to fluctuate in the same way across the years. For *category 2* on the other hand (green lines with triangle markers), *Detachment* (2.5) doesn't seem to keep-up with *Subdivision* (2.2) which presents some high peaks (2006, 2016, 2019). This is reasonable, as the link of the new descriptors to their PHs is stronger in *category 2*, which is based on explicit coverage, compared to *category 3* where the PHs used to cover the new descriptors only implicitly.

It appears that in *category 3*, the amounts of new descriptors that are added as children of their PHs are usually comparable to the ones that are detached from their PHs. This observation could be the effect of an internal procedure in the maintenance of MeSH and may warrant further investigation. On the other hand, the frequency of emerging descriptors without any PHs (Emersion 4.0) (Fig. 13) exhibits fluctuations that are not particularly correlated to the other frequent types of provenance. This suggests that the addition of descriptors with totally new preferred concepts forms a distinct subset of the new descriptors added each year.

7 Conclusion and Future Work

In this work, we proposed a novel conceptual framework for organizing and studying the conceptual provenance of new descriptors in the Medical Subject Headings (MeSH) Hierarchy. In particular, we defined the notion of the previous host (PH), as a descriptor covering the main topic of a new descriptor prior to its introduction, and suggested an approach to identify such PHs for a new descriptor. Then, based on the current relationship of the descriptor with its PHs we also defined a set of provenance types and codes. In addition, we developed an open-source computational process for the automated extraction, annotation, and analysis of new descriptors, using the raw files of different versions of MeSH as distributed by NLM. Employing this approach, we investigated the conceptual provenance of new MeSH descriptors for the period 2006-2020.

The results reveal that about 115 new descriptors for emerging concepts (*category 4*) are introduced each year quite steadily. These descriptors represent about 25% of all new descriptors of the study period, indicating that the majority of the new descriptors cover

non-emerging domain concepts that are not really new for the MeSH thesaurus. Less than half of these non-emerging concepts were explicitly covered in MeSH prior to the introduction of dedicated descriptors for them (*category 1* and *category 2*). The majority of non-emerging concepts, though not explicitly included in older versions of MeSH, used to be indexed under specific older descriptors (PHs) that covered their meaning implicitly (*category 3*).

This suggests that the main force which is consistently driving the extension of MeSH during this period is the need to explicitly cover more conceptual entities. Namely, a stable annual amount of new emerging concepts (*category 4*) and a similar or greater amount of new PI concepts (*category 3*), that used to be implicitly covered by MeSH. The need to introduce descriptors for reorganizing concepts that are already explicitly covered (*category 1* and *category 2*) appears to be auxiliary, with low amounts of new descriptors for most years. However, in certain years, we also observed a surge in the promotion of existing SCRs into descriptors (*category 3*), particularly for chemicals. Such surges in *category 2* and *category 3*, seem to be related with internal MeSH projects and resource allocation in NLM.

In addition, the results on conceptual provenance types reveal that more than 70% of all non-emerging new descriptors (*categories 1, 2* and *3*) become subtopics of their PHs' topics. That is, they remain under the coverage of the latter, usually as children of them (.2, *Subdivision*) and less often as more distant descendants (.3, *Submersion*). However, the amount of new descriptors that are detached from their PHs (.5, *Detachment*) is also considerable, particularly for implicit PHs (*category 3*). These observations suggest that the extension of MeSH primarily serves the need to enrich the MeSH thesaurus with more detailed subtopics, supporting the annotation of articles with new fine-grained topic labels. Nevertheless, it appears that a notable amount of new descriptors also serve to rid the PHs of some implicitly covered topics, rendering the PHs more precise as well.

This grouping can be particularly useful for improving semantic indexing models for new descriptors. For example, the articles annotated with their PHs can be a source of weakly labeled data for topical annotations. In addition, the provenance types can provide indications for the prevalence of such weak labels. In the case of *Detachment* for example, we may expect that only a small part of the articles annotated with the PHs will be relevant to the new descriptor. In the case of new descriptors for new emerging concepts (*category 4*), on the other hand, Zero-Shot Learning approaches may be more appropriate as no PHs are available as a source of weak labels.

Although our findings primarily provide insight to researchers working with MeSH, we believe that the proposed approach is of more general interest. In particular, it can be adapted for analyzing the extension dynamics of other similar topic hierarchies. In the lack of appropriate fields, such as subordinate concepts, PMN and PI, for identifying explicit or implicit PHs in other hierarchies, one could explore term-matching approaches in different hierarchy versions. The annotations of conceptual provenance produced by the proposed method capture the hierarchical relationship of a new topic with the topics that were previously used in its place. Such information can be used to characterise and group the topics, facilitating the process of maintaining topic hierarchies.

Our future plans include the investigation of further uses of the provenance information provided by the proposed method. In particular, we are examining whether new descriptors with the same provenance category, types or codes, present similarities that can be exploited in the semantic indexing of documents with newly introduced labels. Additionally, we are looking into the use of provenance information for predicting ontological expansion. Last but not least, we would like to explore the use of the conceptual framework and computational procedures for tasks related to the maintenance of the hierarchy itself, such as identifying special cases and inconsistencies in textual descriptive fields.

Acknowledgements This research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI Ph.D. Fellowship grant (Fellowship Number: 697). We are grateful to James Mork and Dan-Sung Cho from the National Library of Medicine (NLM) for kindly providing valuable feedback on this work.

References

1. Abcckcr, A., Stojanovic, L.: Ontology Evolution: MEDLINE Case Study. In: Wirtschaftsinformatik 2005, pp. 1291–1308. Physica-Verlag HD, Heidelberg (2005). DOI 10.1007/3-7908-1624-8_68
2. Balili, C., Lee, U., Segev, A., Kim, J., Ko, M.: TermBall: Tracking and Predicting Evolution Types of Research Topics by Using Knowledge Structures in Scholarly Big Data. IEEE Access **8**, 108514–108529 (2020). DOI 10.1109/ACCESS.2020.3000948
3. Balogh, S.G., Zagyva, D., Pollner, P., Palla, G.: Time evolution of the hierarchical networks between PubMed MeSH terms. PLOS ONE **14**(8), e0220648 (2019). DOI 10.1371/journal.pone.0220648
4. Bushman, B., Anderson, D., Fu, G.: Transforming the Medical Subject Headings into Linked Data: Creating the Authorized Version of MeSH in RDF. Journal

of Library Metadata **15**(3-4), 157–176 (2015). DOI 10.1080/19386389.2015.1099967
5. Cardoso, S.D., Da Silveira, M., Pruski, C.: Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies. Knowledge-Based Systems **194**, 105508 (2020). DOI 10.1016/j.knosys.2020.105508
6. Cardoso, S.D., Pruski, C., Da Silveira, M.: Supporting biomedical ontology evolution by identifying outdated concepts and the required type of change. Journal of Biomedical Informatics **87**(August), 1–11 (2018). DOI 10.1016/j.jbi.2018.08.013
7. Castillo, S., Naacke, H., Amann, B., Chavalarias, D.: Exploring the evolution of science through interactive phylomemetic topic maps. BDA 2016 Gestion de Données–Principes, Technologies et Applications 32 e anniversaire 15-18 novembre 2016, Poitiers, Futuroscope p. 89 (2016)
8. Da Silveira, M., Dos Reis, J.C., Pruski, C.: Management of Dynamic Biomedical Terminologies: Current Status and Future Challenges. Yearbook of Medical Informatics **24**(01), 125–133 (2015). DOI 10.15265/IY-2015-002
9. Eljasik-Swoboda, T., Engel, F., Kaufmann, M., Hemmje, M.: Word embedding based extension of text categorization topic taxonomies. In: CERC, pp. 15–26 (2019)
10. Fabian, G., Wächter, T., Schroeder, M.: Extending ontologies by finding siblings using set expansion techniques. Bioinformatics **28**(12), 292–300 (2012). DOI 10.1093/bioinformatics/bts215
11. McCray, A.T., Lee, K.: Taxonomic Change as a Reflection of Progress in a Scientific Discipline. In: Evolution of Semantic Systems, pp. 189–208. Springer Berlin Heidelberg, Berlin, Heidelberg (2013). DOI 10.1007/978-3-642-34997-3_10
12. Nelson, S.J., Johnston, W.D., Humphreys, B.L.: Relationships in Medical Subject Headings (MeSH), pp. 171–184. Springer Netherlands, Dordrecht (2001). DOI 10.1007/978-94-015-9696-1_11
13. Nentidis, A., Krithara, A., Tsoumakas, G., Paliouras, G.: Harvesting the Public MeSH Note field. Identifying the previous state of new descriptors in the MeSH thesaurus as Supplementary Concept Records. Tech. Rep. arXiv:2106.00302, National Center for Scientific Research "Demokritos" & Aristotle University of Thessaloniki (2021). URL `https://arxiv.org/abs/2106.00302`
14. Oliver, D.E., Shahar, Y., Shortliffe, E.H., Musen, M.A.: Representation of change in controlled medical terminologies. Artificial Intelligence in Medicine **15**(1), 53–76 (1999). DOI 10.1016/S0933-3657(98)00045-1
15. Sari, A.K.: Mapping of change operations from gene ontology into medical subject headings. International Journal of Intelligent Engineering and Systems **13**(4), 44–55 (2020). DOI 10.22266/IJIES2020.0831.05
16. Tsatsaronis, G., Varlamis, I., Kanhabua, N., Nørv, K.: Temporal Classifiers for Predicting the Expansion of Medical Subject Headings. Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13) pp. 98–113 (2013). DOI 10.1007/978-3-642-37247-6-9
17. Yu-Wen Guo, Yi-Tsung Tang, Hung-Yu Kao: Genealogical-Based Method for Multiple Ontology Self-Extension in MeSH. IEEE Transactions on NanoBioscience **13**(2), 124–130 (2014). DOI 10.1109/TNB.2014.2320413