Exact versus Estimated Pruning of Subject Hierarchies

Charles-Antoine Julien
School of Information Studies
McGill University
3661 Peel St., Montreal, Qc. H3A 1X1
charles.julien@mcgill.ca

Pierre Tirilly
Laboratoire d'informatique fondamentale de Lille
Université de Lille 1
Parc Scientifique de la Haute Borne
50, avenue Halley, B.P. 70478
59658 Villeneuve d'Ascq, France
pierre.tirilly@lifl.fr

ABSTRACT
Many large digital collections are organized by subject; these useful information organization structures are large and complex, thus difficult to browse. Current online tools and visualization prototypes show small localized subsets and do not provide the ability to explore the predominant patterns of the overall subject structure. This study builds on existing work concerning automatic subject hierarchy modification techniques that aim to facilitate browsing for documents by capitalizing on the highly uneven distribution of real-world collections. Specifically, previous work used an estimation of the number of accessible documents offered by each subject term, while the current study uses the exact number of accessible documents. The impact is demonstrated on a large collection organized using Medical Subject Headings (MeSH). Results show that, although computationally more demanding, pruning the MeSH hierarchy based on the exact access produces a different subject hierarchy under some conditions. The visual impact is demonstrated using traditional outline views. This study has implications for the development of information organization theory and human-information interaction techniques for subject hierarchies.

Keywords
Organization of information, subject hierarchies, trees, browsing, information visualization, human-information interaction.

INTRODUCTION
Subjects and their relations form a network (i.e., ontology) that users can browse to discover domain vocabulary and retrieve documents (Belkin, Oddy, & Brooks, 1982; Lancaster, 1986; Marchionini & Shneiderman, 1988). Representing a collection’s subject structure as a hierarchy is a promising approach for improving subject browsing (Zhu & Chen, 2005); however, real-world CV structures are large and complex. These can overwhelm users, making exploration difficult. This especially concerns users who are not familiar with the knowledge domains covered by the collection. Julien, Tirilly, Dinneen, and Guastavino (2013) showed that the size of the Medical Subject Heading (MeSH) hierarchy could be significantly reduced while still providing access to the vast majority of the collection. For computational efficiency, they used an estimation of the number of accessible documents offered by each subject term. This study extends that research by using the exact number of accessible documents, and evaluating the impact of the high prevalence of multiple subject access points per document that is systematically present in organized collections.

The following section describes subject trees and how they can improve the navigation of subject hierarchies. The general behavior of users browsing a subject tree is described based on the information foraging model (Pirolli & Card, 1999). Multiple subject assignment per document is described as a distinctive characteristic of document collections organized using a subject hierarchy as compared with digital file folder organization. The work of Julien et al. (2013) is a novel line of research that aims to facilitate browsing for documents in large and complex subject hierarchies. Finally, the current study is shown to be an alternative method of indicating the subjects that can be pruned with minimal effect on collection access.

Subject Trees
A tree is one of the most recognized ways to represent a hierarchy: for example, the vast majority of computer users organize their files in a visual tree structure represented as an outline view. Ubiquitous hierarchies, such as the trees shown in computer file explorers, follow sets of specific structural rules. Specifically, trees consist of a set of nodes connected by directed edges, including a root node that is a common ancestor to all other nodes (Di Battista, 1999). Nodes can be linked to other nodes, which are called its children, and nodes without children are called leaf nodes. Each node is contained by a single other node, which is called its parent, except the root node, which has no parent. Hierarchies adhering to these structural rules are generally

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.
Copyright notice continues right here.
considered easier to navigate since they facilitate their visual representation and interaction using traditional outline views by specifying a single parent per node; as a result, users can perform systematic top-down navigation (also called drill-down) from root to leaves.

The MeSH subject hierarchy is not a valid tree structure since MeSH terms can have multiple parents. This issue of multiple inheritances is addressed by duplicating child nodes, and all their descendants if any, under all their parents (Katifori, Halatsis, Lepouras, Vassilakis, & Giannopoulou, 2007).

**Browsing Subject Trees**
Information foraging (Pirolli & Card, 1999) describes searchers as informavores consuming information located in patches or groups of documents, and constantly scanning the environment for other more promising patches. Information foraging systems are those that facilitate the extraction of value from a patch and reduce the cost of finding and switching to other more promising patches. Within the information foraging framework, users browsing a subject tree are constantly assessing the value offered by surrounding subject nodes as compared with the perceived remaining value of the current node.

Browsing a subject tree is thus described as a series of decision points where the options are either to expand a subject node (i.e., reveal narrower terms), return to a previously visited node, or continue extracting value from the current node. Although browsing and its outcomes are difficult to constrain (Chang, 2005), when users browse a subject tree, two kinds of outcomes are considered: 1) the user finds valuable documents, and 2) the user learns about a subject by inspecting the relations between neighboring subjects. A modified version of the subject tree may facilitate document retrieval while respecting the logic of the subject hierarchy.

**Facilitating Subject Tree Browsing**
Julien et al. (2013) describe an algorithm that aims to facilitate the discovery of documents in subject trees. Their approach opportunistically modifies a subject tree based on the power law distribution of the collection; a small group of subjects provides access to most of the documents while most subjects are assigned to very few documents. Power law distributions are common in organized information (Julien, Tirilly, Leide, & Guastavino, 2012a; Liu et al., 2005; Wang & Lee, 2007), and they are found in other knowledge domains (Egghe, 2005; Newman, 2005); thus, there are strong indications that organized collections in general follow the same kind of distribution within their respective organizing structure. Capitalizing on this trend, Julien et al. (2013) reduce the size and depth of a collection’s subject tree by pruning the subjects that are not representative of a collection.

**Browsing Complexity**
To facilitate comparisons between subject hierarchies and modification approaches, Julien et al. (2013) describe a combined measure that aims to express the overall browsing complexity of a subject tree. Specifically, browsing complexity (BC) expresses the predictability and number of narrower term choices that users can expect from an arbitrary subject term, and the distance from this term to the boundaries of the subject tree. A subject tree having a lower BC score may facilitate browsing since users must navigate fewer levels to discover the boundaries of the tree, and each browsing decision is made from a reasonable and predictable number of narrower term options. Alternatively, a higher BC score indicates a deep subject tree (e.g., fifty levels of depth) that could be mistaken as being an endless network, where users are sometimes asked to choose from a prohibitively demanding number of narrower terms (e.g., dozens or more than a hundred).

**Subject Tree Pruning**
To reduce the number of narrower term choices and increase its predictability, Julien et al. (2013) describe a pruning operation that removes the subject terms that provide access to a statistically marginal proportion of the collection documents. They describe a browsing interface where users can choose the amount of pruning according to their task requirements by selecting the appropriate pruning threshold factor $T_s$. They test their approach on the MeSH tree containing a collection of over 230,000 documents. The pruned MeSH tree is shown to be significantly smaller, both in terms of size and depth, while still providing access to the vast majority of the collection: between 98% and 100% access depending on the $T_s$ selected by the user.

**Number of Accessible Documents**
Before pruning can be performed, each set of children must be sorted in descending order based on their number of accessible documents. Julien et al. (2013) calculate the number of accessible documents per subject heading by summing the numbers of documents assigned to each of its descendants, which assumes documents appear only once in the subject structure. This is intuitive in a hierarchical file management context (i.e., a folder tree) where a file is generally placed in a single folder; however, in collections organized by subject, documents are often assigned to multiple subjects. Figure 1 illustrates this difference, which is akin to creating copies of a document for each assigned subject access point; therefore, summing the number of documents assigned to each subject term creates an error in the case of collections organized by subject since it inflates the total number of documents to equal the total number of subject assignments. In fact, a subject heading’s exact number of accessible documents is the number of unique documents assigned to the set of its descendants: a much more computationally demanding calculation. Describing the size and impact of this error on the pruned subject tree is the aim of this study.

**Related Studies**
To our knowledge, the problem of modifying a subject hierarchy for the purpose of human-information interaction has never been addressed. Wang and Lee (2007) reconstruct
the Dewey Decimal Classification (DDC) for automatic text classification. They balance the original DDC hierarchy so that the documents are evenly distributed among the classes. Although their initial data and objective differ from this study, they provide a promising research direction by suggesting that a subject hierarchy should be adapted to the task being performed. This contrasts with prior subject hierarchy browsing applications (Hearst & Karadi, 1997; Korn & Shneiderman, 1995; Lin, 1999; Sheth & Cai, 2003; Yi & Chan, 2008) that assume searchers will use the same subject structure as professional subject indexers. The current research differs from those works since it modifies an existing subject hierarchy to facilitate browsing for the purpose of finding collection documents.

METHOD
The objective is to compare a subject tree pruned based on the sum of assigned documents (Julien et al., 2013; Julien, Tirilly, Leide, & Guastavino, 2012b), henceforth referred to as the estimated access, with the same subject tree pruned based on the exact number of accessible documents per subject heading calculated specifically for this study, henceforth referred to as the exact access. This is accomplished by using the same MeSH organized collection of 233,388 bibliographic records used by Julien et al. (2013), and replicating their experiments using the exact access. Although they also prune a Library of Congress Subject Headings (LCSH) tree, only MeSH is chosen for this study because its smaller size and highly taxonomic nature make it a more computationally tractable test case.

Sorting Narrower Terms
Pruning the MeSH tree requires that its sets of children be sorted in descending order based on their exact access. This must be done for all MeSH terms that have two or more children (henceforth referred to as parent terms). This is a highly demanding computational process since, for each child being considered for sorting, it involves calculating the union between the set of documents accessible from the previously sorted children and the respective sets of documents accessible from all remaining unsorted children.

Given a node $N$, we define its sorted children $SC_N = (C_1^i, C_2^i, \ldots, C_k^i)$ as the list containing the children of $N$ in descending order of their number of accessible documents, where $C_i$ identifies a specific child of $N$ (i.e., the $i^{th}$ child of $N$). $SC_N^i$ is the list containing the first $i$ elements of $SC_N$. $\text{Asg}(C_i^1)$ is the number of documents assigned to child $i$ or one of its descendants; note that this number includes duplicate documents resulting from multiple subject access points. $\text{Acc}(SC_N^i)$ is the set of unique documents (i.e., excluding duplicate documents resulting from multiple subject access points) that are accessible from the first $i$
children of the sorted list of children or one of their descendants. Finally, \( C(S) \) is short-hand for the count of the number of documents in a set of documents \( S \) (i.e., the cardinality of the document set). Using these notations, the estimated number of accessible documents used by Julien et al. (2013) is defined as:

\[
C\left(\text{Acc}\left(SC_i^j\right)\right) = \text{Asg}(C^1) + \text{Asg}(C^2) + \cdots + \text{Asg}(C^i)
\]

Given \( \text{Acc}(C^i) \) as the set of unique documents that are accessible from child \( i \) or one of its descendants, the exact number of accessible documents calculated for this study is defined as:

\[
C\left(\text{Acc}(SC_i^j)\right) = C\left(\text{Acc}(C^1) \cup \text{Acc}(C^2) \cup \cdots \cup \text{Acc}(C^i)\right)
\]

Although simpler forms of these formulas can be derived, their presented forms are chosen to highlight that:

- calculating the exact access requires numerous unions between potentially large sets of documents, which is a computationally demanding operation.
- calculating the estimated access requires a series of simple additions, which is a simple computational operation.

For example, given a parent term with ten children, sorting the children based on their exact access executes the following steps, which are illustrated by Figure 2:

1. Build the sets of unique documents accessible from each child and all the narrower terms to which it provides access, no matter the distance (i.e., its descendant MeSH terms). Note that there could be thousands of descendant terms in the case of a broad MeSH term; as a result, a set of accessible documents can contain a large proportion of the collection. The child term having the largest number of accessible documents is placed first in the sorted list of children.

2. Calculate the union between the set of accessible documents offered by the first sorted child and the sets of accessible documents offered by all other remaining unsorted children. This crucial step eliminates duplicate documents; however, it involves comparing each document of the first child with all documents of all other children. Assuming that broad MeSH terms and their descendants provide access to an average of 1,000 documents, comparing the accessible document sets of two children entails one million comparisons (i.e., \( 1,000 \times 1,000 \)). In our example with ten children, there are 45 possible pairs of children; thus, finding the second sorted child would require 45 million comparisons.

3. The child that adds the largest number of unique additional documents (i.e., the cardinality of its union with the first sorted child) is placed second. The set of accessible documents from these two sorted children becomes the next set of documents with which to calculate unions with the accessible document sets of the remaining unsorted children (i.e., eight in our example).

4. Repeat steps 2 and 3 until all children are sorted. In our example with ten narrower terms, this requires 164 unions between potentially very large sets of accessible documents. To illustrate, again assuming an average of 1,000 accessible documents per subject term, sorting ten children based on exact accessible would require 164 million comparisons.
Figure 2 illustrates the fact that the first sorted child must be known before the second sorted child can be determined. In the same manner, the first and second sorted children must be known before the third sorted child can be determined, and so forth (i.e., the $i$th child must be known before the $(i+1)$th sorted child can be determined). For this reason, the computationally demanding selection sort algorithm (Knuth, 1973) is the only option.

Sorting children based on exact access is computationally demanding when sorting parent terms that have dozens or sometimes over a hundred children (e.g., the MeSH term Antigens, CD has 115 narrower terms, Proteins has 94); it becomes prohibitive when sorting broad children that provide access to most of the collection (e.g., documents sets containing 100,000 documents). This study is a test case that evaluates if these added computational demands produce a quantitatively and qualitatively different subject tree as compared with the MeSH tree modified using the estimated access (Julien et al., 2013).

**Estimated vs. Exact Sort Order**

Measuring the differences between two sorts of a set of items is done using the normalized Kendall tau rank distance (Fagin, Kumar, & Sivakumar, 2003). This metric counts the number of pairwise disagreements between sorted lists; larger values indicate higher dissimilarity between the lists. This metric is used to measure the differences, if any, between the sets of children sorted based on the estimated access reported by Julien et al. (2013), and the equivalent children sets sorted based on the exact access calculated for this study. For example, a zero distance implies that the children are sorted identically based on the estimated and the exact access; alternatively, a distance equal to one implies the children sets are in reverse order. The result is the list of all parent terms, and their associated normalized Kendall tau distance. The proportion of parents whose children are sorted differently (i.e., have a non-zero Kendall tau distance) will indicate if the pruning operation might produce different pruned MeSH trees depending on the access calculation method (i.e., estimated or exact). For example, a very small proportion will suggest that the estimated access provides an adequate sorting of children, while a large proportion of differently sorted children sets will suggest that the pruned trees may be significantly different.

**RESULTS**

Of the 4965 parent MeSH terms, 24.65% (1224/4965) of their children sets are sorted differently using the estimated versus exact access. This shows that using the estimated access incorrectly sorts the children of just under a quarter of parent terms.

Some sorting differences can be explained by duplicate documents created by multiple subject assignments. Figure 3 shows an example of the sorting differences for the parent term Information Systems, which has fourteen children. The MeSH term label fonts in Figure 3 illustrate the power law distribution of documents within the MeSH hierarchy: bold labels are subjects that provide access to at least 20% of all documents accessible from the children, normal font labels provide access to between 2% and 20%, and grey labels provide access to less than 2%. The fonts in Figure 3 clearly show that most subjects provide access to relatively few documents (i.e., grey labels); therefore, most subjects could be pruned with little or no impact on collection access.
Figure 3 also shows that the sort order begins to differ beyond the second child Management Information Systems. For example, Online Systems is wrongly placed in the third position when children are sorted based on estimated access; this child is actually placed in the fourth position when sorted based on the exact access. This implies that although Online Systems and its descendants offer a larger number of assigned documents, Medical Records Systems, Computerized actually adds a higher number of additional unique documents to those offered by the first two children (i.e., Databases as Topic; Management Information Systems).

Sorting differences are also explained by duplicated terms resulting from multiple inheritances; indeed, many MeSH terms have multiple broader terms. This is illustrated by Radiology Information Systems, which is placed in the eighth position when using the estimated access as compared to the fourteenth (i.e., last) position when using the exact access. This multidisciplinary MeSH term is a child of both Information Systems (as shown in Figure 3), and Management Information Systems (not shown in Figure 3). This explains why Radiology Information Systems appears in the last position when sorted based on exact access: all the subject’s documents are already accessible from Management Information Systems.

Given that almost a quarter of children sets are incorrectly sorted by the estimated access, we wish to examine the qualitative differences in the pruned MeSH hierarchies using the estimated access as compared with the exact access calculated for this study.

**Pruned MeSH Tree**

As described by Julien et al. (2013), the pruning operation reduces the number of children per parent by removing the children that provide access to a relatively small proportion of the collection. Figure 4 presents an example of the qualitative differences when pruning using the estimated versus exact access for the MeSH term Information Science. It shows that the pruning operation using the exact access (right column of Figure 4) removes additional children as compared to the estimated access. Specifically, Information Science initially has 21 children: eleven are pruned using the estimated access as compared with the fourteen that are pruned using the exact access. This is intuitive given that the estimated access artificially inflates the number of accessible documents to equal the number of subject assignments, which counts the same documents multiple times. For example, the estimated access of the root node would be 2,750,199 (i.e., the total number of subject assignments in the collection) while the collection actually contains 233,388 unique documents: this translates to a 1078% error. This error is corrected when using the exact access; as a result, each additional child has progressively less potential to provide access to documents that are not already accessible from one of its larger siblings.

---

1Note that the root node is likely to have the largest error.
Figure 5 presents a pruned portion of the MeSH tree related to Humans. It shows that although there are various kinds of Organisms, the vast majority of the literature concerns Animals, although there are different kinds of animals, the vast majority of the literature concerns Mammals, of which there are various kinds but the vast majority of the literature concerns Humans. This illustrates how pruning of the MeSH tree facilitates browsing for documents concerning humans by reducing its depth from eleven levels in the original hierarchy, down to three in the pruned version. Notice that the pruned MeSH tree still provides access to 100% of the collection since documents are assigned to multiple MeSH terms, and pruned terms are those that necessarily provide access to relatively few documents. Although salient, the examples presented in Figures 4 and 5 are anecdotal; the following section describes the overall effects of pruning using estimated versus exact access.

**Number of Children per MeSH Term**

The overall effects of pruning the complete MeSH tree are described in terms of the change in the mean and standard deviation of its number of children (NoC) per MeSH term.

The mean NoC describes the expected number of narrower term options users must choose from at any given MeSH term; a low mean NoC indicates users must generally choose from smaller more manageable lists of narrower terms. The standard deviation (SD) of the NoC describes the predictability of the number of children per MeSH term; a high SD indicates that there are large variations in the numbers of narrower term options per MeSH term (e.g., a hundred narrower terms, followed by a single narrower term), while a lower SD indicates a more balanced subject tree, which is critical for effective narrowing of the collection (Smith et al., 2006) so that the user is not led into areas of the subject tree that contain limited amounts of information.

Figure 6 shows that pruning using the exact access produces a sharper reduction in the mean NoC and SD as compared with the estimated access. This is shown by the distance between the solid dark curves as compared with the dashed grey curves. The numbers of children per MeSH term are subjected to a two-way ANOVA having two levels of access measurement types (Estimated, Exact) and the twenty levels of pruning threshold $T_r$ shown in Figure 6 (i.e., excluding the initial unmodified MeSH tree). The main effect of the pruning threshold $T_r$ is significant [$F(1, 397495) = 121.78$, $p < 0.001$]; post hoc comparisons are not conducted since this is the expected effect of pruning and
not the interest of this analysis. The main effect of access measurement type is also significant [F(1, 397495) = 71.37, p < 0.001]. There is a significant interaction between the two independent variables [F(1, 397495) = 3.95, p = 0.047 < 0.05], which is no longer significant when \( T_r = 0.2 \) is excluded. These results are consistent with the examples shown in Figures 4 and 5: exact access based pruning generally removes more children per parent. As suggested by the interaction effect, this is true until the pruning threshold \( T_r \) is increased adequately so as to permit equivalent pruning based on estimated access.

**ANALYSIS**

Figure 6 and the ANOVA results show that revealing the few most populated subjects of the collection (i.e., using a high \( T_r \) value) is adequately performed using the estimated access. This suggests that the added computational costs required for pruning based on exact access are justified when the user task requires fine-grained incremental pruning.

Given that the children of 24.65% of MeSH parents’ children are sorted differently, over 75% are sorted identically using the estimated and exact access. This implies that for most MeSH parents, pruning based on estimated access may be equivalent to pruning based on the much more computationally demanding pruning based on exact access. This is in part explained by the varying degrees of overlap between subject branches. Many subjects are generally self-contained: they offer access to documents not assigned to any other subject branches. In contrast, the multi-disciplinary MeSH terms shown in Figures 4 and 5 provide access to documents assigned to multiple MeSH terms. To our knowledge, there is no standard measure of the multi-disciplinarity of a subject hierarchy. Assuming such a measure can be developed, it may indicate those few subject branches that can benefit from the added computational costs needed to calculate their exact number of accessible documents.

Beyond the study by Julien et al. (2013), the question of modifying an existing subject hierarchy for the purpose of document browsing by humans is unexplored territory. The current study furthers this research direction by showing that various kinds of pruning can be applied to subject terms depending on factors such as their amount of overlap with other subjects (i.e., multi-disciplinarity).

**Subject Tree Redundancy**

The MeSH hierarchy is not defined as a valid tree since, among other issues, nodes can have multiple parents. This issue of multiple inheritances is addressed by duplicating child nodes, and all their descendants if any, under all their parents. Duplicating a leaf node means a single node is
added to the tree. Duplication of a broad subject that provides access to hundreds or even thousands of narrower terms creates copies of all descendants; this can significantly increase the size of the subject tree as compared to the number of unique subject terms in the hierarchy.

To facilitate measurement and comparison of subject trees, Julien et al. (2013) define the subject tree redundancy as \( \frac{\text{Total tree nodes} - \text{unique headings}}{\text{Total tree nodes}} \), or the number of duplicate nodes divided by the total number of tree nodes. They report that the unmodified MeSH tree is 64.3% redundant, meaning that almost two out of every three nodes have at least one copy elsewhere in the MeSH tree. When browsing a redundant subject tree, users can find a single subject via different routes. This may be desirable when searching for one of these duplicated subjects; however, it may be detrimental if the user is searching for a single non-duplicated subject buried within a highly redundant subject tree.

Figure 7 shows the evolution of the redundancy of the MeSH tree pruned using the estimated versus exact access, with respect to the proportion of the browsing complexity \( BC \) as the pruning threshold increases from left to right. It shows that, in both cases, pruning reduces the redundancy of the MeSH tree. When pruning based on the estimated access (i.e., dashed gray line in Figure 7), the MeSH tree redundancy is gradually reduced until just under 60% of its initial \( BC \), where it becomes erratic. In contrast, using the exact access (i.e., dark line in Figure 7), redundancy is gradually and consistently reduced.

These results show that pruning based on the exact access consistently removes branches that contain a higher proportion of duplicated subjects. This is explained by the fact that pruning using the exact access ensures that each retained child provides access to the highest number of documents not accessible from those already retained. In contrasts, pruning based on the estimated access may retain nodes that contain duplicated documents or share nodes with other retained children. This shows that pruning using exact access reliably produces subject hierarchies containing fewer duplicated subjects. Although this could justify the added computational costs associated with using exact access, as far as we can gather, there are no obvious preferred values of subject tree redundancy, and its proper optimization for a usable subject tree is an open question.

**Pruning Subject Trees and Faceted Browsing**

The current trend towards faceted browsing (La Barre, 2010) does not exclude the existence and usage of hierarchical subject structures; indeed, facet analysis should include a hierarchy of values for each facet (pp. 249-250). Adequately coupling faceted browsing and facet value hierarchies is an open question for online search interface designers, and current online systems vary broadly. For example, next-gen OPACs (e.g., NCSU) flatten these hierarchies by ignoring the relations between subjects; at the other extreme, hierarchical faceted metadata, or HFM, is the primary browsing feature of the open-source Flamenco testbed (Hearst, 2006), one of the best known implementations of facets (Hearst, 2007). Hearst’s research indicates that users prefer HFM interfaces, and an increasing acceptance of tightly coupled browsing and searching (La Barre, 2010, p. 261). Therefore, the question of pruning and simplifying subject hierarchies to facilitate browsing for documents is relevant to faceted browsing as the size of facet value hierarchies increases.

**CONCLUSION**

Existing subject trees currently organize sizable portions of published knowledge; however, these useful structures are large and complex, making them difficult to browse. Building on the work of Julien et al. (2013), we have proposed an alternative way of designating the subjects that can be pruned from a subject hierarchy while maintaining access to the collection. Experiments performed on the MeSH hierarchy showed that the added computational costs required for the calculations of the exact numbers of documents accessible from each subject can be justified when the user task requires fine-grained pruning of the structure. Otherwise, the more computationally tractable estimated number of accessible documents used by Julien et al. (2013) was shown to be equivalent.

By capitalizing on existing information organization investments, this line of research explores techniques that allow searchers to trade data fidelity in exchange for a reduction in subject tree browsing difficulty. Specifically how and how much this helps users to browse and search remains to be tested with human participants. Future work includes investigating the effect of subject tree redundancy on browsing behavior, and calculating the exact number of accessible documents with other hierarchically organized collections (e.g., Library of Congress Subject Headings). We also plan to investigate how our approach based on local document statistics relates to more formal probabilistic frameworks (e.g., Tirilly & Julien, 2012).