Etymonline:
From etymological dictionary to an online digital library

Utilizing LIS notions in transforming a database into a DL

YAELLE KALIFON

© Yaelle Kalifon
Partial or full copying and distribution of the material in this thesis is forbidden without written consent.
English title  Etymonline: From etymological dictionary to an online digital library Utilizing LIS notions in converting a database to a DL

Author:  Yaelle Kalifon

Completed:  2017

Abstract:  The thesis explores the hypothetical turning of an online database of etymology into an online digital library. The reader will be introduced with the principles that constitute a library, as opposed to other sources of information and knowledge, and the contemporary attributes of having those available online.

The thesis argues that by separating metadata from entries and storing it separately, users would be able to gain more information per search, as well as browse through other etymological data in the area of their interest.

The methodology employed is that of characterizing each entry according to categories, each of which composed of various possible tags that fit the existing corpus of entries. A search is then made possible according to specific entry (as is the case currently in the Etymonline database), as well as specific category, specific tag, or any combination of those. These added features also enable random browsing of the collection and the deducing of other etymological ties or points of interest.

Though not without flaws, the suggested methodology proves to retrieve more hits upon searching for a specific attribute of words or related ties to specific entries. Flaws, restrictions and limitations, as well as functions that were not attempted but could improve user experience further, are discussed.

The thesis stresses the importance of searching and browsing functions and their contribution to future etymological research, by using the available information as it is provided in Etymonline.com., in the structure of a DL instead of a database.

Keywords:  categorization, tagging, information organization, data mining, etymology
# Table of Contents

Chapter 1  
**Introduction** ................................................................. 5  
1.1 Background ................................................................. 5  
1.2 Research question and aim ........................................... 5  
1.2.1 Problem ................................................................. 5  
1.2.2 Aim ........................................................................... 6  
1.2.3 Research question ..................................................... 6  
1.2.4 Ideational limitation .................................................. 7  
1.2.5 Etymonline justification ............................................. 7  
1.3 Thesis structure ............................................................. 7  

Chapter 2  
**Literature review** .......................................................... 8  
2.1 Libraries and digital libraries – past, present, future .......... 8  
2.2 Volunteer based digital resources ................................... 9  
2.2.1 Project Gutenberg ..................................................... 10  
2.2.2 Wikipedia ............................................................... 10  
2.2.3 Flickr ....................................................................... 10  

Chapter 3  
**Classification theory** ...................................................... 11  
3.1 Foundations of classification ........................................ 11  
3.2 The power of classification systems ............................... 12  
3.3 Matching a classification system to a corpus – gains and pitfalls .... 13  
3.4 The freedom and difficulties of classifying in the digital world.... 13  
3.4.1 Searching versus browsing ....................................... 14  
3.4.2 Metadata ............................................................... 14  
3.4.3 Misc. ....................................................................... 15  
3.4.4 Different ways of organizing ..................................... 15  
3.4.5 Tagging ................................................................. 16  
3.4.6 Penn tags project ..................................................... 17  

Chapter 4  
**Etymology** ..................................................................... 18  
4.1 English etymology ....................................................... 18  
4.2 Lexical approaches to etymological dictionaries ............. 19  

Chapter 5  
**Conceptual and visual representation** .............................. 20  
5.1 System and interface design ......................................... 20  
5.2 Psychological representation ......................................... 21  
5.3 Text mining ............................................................... 22  
5.4 Text visualization ....................................................... 22  
5.5 Physical versus digital categorization ............................. 23
Figures and Tables

Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Related terms</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Hierarchical categorization</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>Faceted categorization</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>The Indo-European entry *DHEI and its exemplification</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>John Snow’s map of clustering data</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>Screenshots of Etymonline, from start page to specific entry</td>
<td>25</td>
</tr>
<tr>
<td>6.a</td>
<td>Start page</td>
<td>25</td>
</tr>
<tr>
<td>6.b</td>
<td>Focusing on the letter S</td>
<td>26</td>
</tr>
<tr>
<td>6.c</td>
<td>Choosing the random entry “Second (v.)”</td>
<td>26</td>
</tr>
<tr>
<td>6.d</td>
<td>Screenshots of terms related to the entry “Second (v.)”</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>Letter distribution estimation comparison between the English lexicon and Etymonline</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>The entry “Nostalgia (n)” and its related terms, as it appears in Etymonline</td>
<td>37</td>
</tr>
<tr>
<td>9</td>
<td>Comparison between ties found in the suggested model and in Etymonline</td>
<td>46</td>
</tr>
</tbody>
</table>

Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>POS tags</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>Route, foreign sources exemplification</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>Route, English word change exemplification</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>Route, reference exemplification</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>Meaning 2, disputed route entry</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>Meaning 2, change to POS</td>
<td>34</td>
</tr>
<tr>
<td>7</td>
<td>Meaning 3, change to POS</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>Meaning 3, change in meaning</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>Ties’ analysis for the entry “New Zealand (country name)”</td>
<td>38</td>
</tr>
<tr>
<td>10</td>
<td>Compared ties between models for the entry “New Zealand (country name)”</td>
<td>39</td>
</tr>
<tr>
<td>11</td>
<td>Comparison results</td>
<td>45</td>
</tr>
<tr>
<td>12</td>
<td>An analysis of the results</td>
<td>45</td>
</tr>
</tbody>
</table>
Chapter 1
Introduction
1.1 Background

The thesis discusses the intersection of etymology and information organization, in an era of digital information and knowledge storing.

Etymology is the outlining of the origin of words and their historical development, starting from an ancient (and sometimes hypothesized) source, and until present days or the earlier dying out of a language (Gray, 1986). As such, etymology encompasses aspects of language and linguistic patterns, geographical attributes, political, national and historical developments, the effects of trends and fashions, and many more social changes in human culture. I.e., it is a field that treasures much meaningful information to the research community and to understanding of human society, as well as the human mind and consciences.

Information organization too is a field that treasures insights regarding the conscious and subconscious human mind; the way we view the world, the way we approach information and tasks and interpret reality (Ranganathan, 1989). The field of information organization is not limited to one’s own mind, rather is a discipline anchored in computer sciences, mainly to do with information retrieval in response to a user query (Weinberger, 2007).

The two fields combined offer ground for research of how information organization affects the accessing of etymological information, in general as well as in response to specific user needs. Therefore, the thesis will focus on an etymological database --Etymonline-- and suggest an alternative method by which to organize its corpus and metadata such that searchability is increased, attributes of browsing are made possible, and etymological ties may surface, even unintentionally, in relation to user queries.

Namely, by adopting the changes suggested in the thesis, the database would essentially be elevated to being a digital library (hereafter DL) – a reliable, certified source of information for experts and laymen; searchable; browsable; structured and inter-connected; tagged consistently, and allowing entries and tags to be added or removed, altering only metadata whilst keeping the information visible to users in its stable form. For a database as rich in volume as Etymonline is, it will be argued that a DL structure is preferable to the current list formation it occupies.

1.2 Research question and aim
1.2.1 Problem:

Etymonline (www.etymonline.com) is a great contributor to the research of English etymology by uniting information from various reliable and scientific sources, making information available online, and in great multitude (Harper, N.D). Nonetheless, the database is structured like a dictionary, entries being listed according to their alphabetical order.

A recent feature in Etymonline shows “related terms” at the end of each entry, connecting it with similar entries via a link (see figure 1 below). However, the nature of the relationship is at times superficial or unintuitive. As such, it is impossible to browse the database. Rather, a search is restricted to the user’s specific word queries. I.e., if one doesn’t know a word exists, one cannot stumble upon it, but must search for it specifically. In that sense, even the “related terms” function would not help reach a specific entry by means of semantic, syntactic or historic similarity.

Each entry is described in great detail for the various features it exhibits, as well as trivia information about its recorded appearances, it changes throughout history and its suspected or verified sources. However, such information cannot be systematically linked between entries. I.e., one must search for entries one suspects are related in order to validate or refute that suspicion. The “related terms” function does not utilize all the information held in each respective entry, and therefore possible ties between words remain hidden.
1.2.2 Aim:
With the suggested method of segmenting metadata from entries and making them browseable, linguistic attributes are made apparent and comparable, allowing users to make use of a certain property that they find interesting in order to find related terms. For example, searching for a specific word reveals that it originates from source X; the user can then explore all other entries in the lexicon that share a source. Additionally, users can approach a specific category without knowing any of the words it might pertain to. I.e., searching for linguistic source X for an encompassing list of all the entries it includes.

The degree of relatedness between entries would be measured by ties: what properties are shared between entries, how many connections surface and in accordance with what filtering criteria. The ties that will be inspected are those that constitute the categories defining the entries (essentially, the metadata). For example: year of earliest appearance, lexical category, source language, and more. I.e., ties are the number of faceted routes that lead to the finding of a single entry through searching and browsing.

1.2.3 Research question:
Stemming from the problem and aim described above, the research question wishes to find to what degree would a new model of a DL improve making connections between entries, in comparison to the existing database structure. I.e., would a DL be better than a database --in both searchability and browsability-- due to its offering a structure, consistent tagging and metadata segmentation and categorization.

The question will be researched by quantitative measures, comparing the number of linguistic and etymological ties each entry exhibits, once in the current interface and once in the suggested model. The two approaches would be scored and compared, showing which method allows more ties to surface, thereby improving usability of the information stored therein.

The assumption is that the suggested model would enable a wider, more elaborate network of ties than that currently available in Etymonline, and reveal curious findings that might lead users to discovering more connections or etymological properties in the corpus. The null hypothesis is therefore:

\[ H_0 \quad \text{There would be no significant difference in the number of ties generated between words, in existing and the suggested models of organizing the Etymonline corpus.} \]
To put the hypothesis to the test, ten randomly chosen words would be analyzed and compared, to show which system yields greater results when searching for connections between words. The words chosen are: Nadir (n), Nationalism (n), Naval (adj.), Nessie (proper name), New Zealand (country name), Nicker (v), Narcomania (n), Non-concencual (adj.), Nostalgia (n) and Numismatics (n).

These words will be compared according to the following criteria: stem, year of entry, route of entry, slang affiliation and change in meaning. The comparison will then summarize the number of ties surfacing for each approach — a DL versus a database (list) structure —, as well the number of ties shared between the two approaches. The latter meant to inspect if the suggested DL approach misses any connection that the database had spotted.

1.2.4 Ideational limitations

What is considered relevant information is a matter of subjective opinion, dependent on specific information needs, prior knowledge, context and availability of data (Weinberger, 2007). Therefore, it is difficult to assess the success rate of the suggested model in answering to user queries. Additionally, specific information needs are also obscure.

Not much is known about the characterization of the users of Etymonline, and to this date, no contact invitation had been answered by the creator of the site. However, due to the information detailed about entries (relating to date of first appearance, route of entering into English (source language and donor languages), disputes between researchers, change in meaning and usage throughout history, slang affiliation, and more) it is assumed that those are the properties which are of interest to users.

1.2.5 Etymonline justification

Etymonline was chosen to serve as the object of research for several reasons, apart from it being a comprehensive and meaningful database of cultural heritage and knowledge.

The structure of the database is such that its metadata can be segmented from its entries, allowing those same entries to be exhibited in varying constellations, according to the varying needs of users. Thus, user experience can be enhanced, etymological research could be encouraged and improved, and the database could be promoted to being a DL.

The suggested tagging that would be introduced in the thesis would contribute to the browsability and data mining that are currently unavailable in the database’s existing list formation. Despite “related terms” functionality being made available as of recent, the explored methodology introduced herein will grant a wider variety of potential ties between entries, and their structured relatedness — predictable, researchable and subject to user filtering and manipulation.

1.3 Thesis structure

The thesis starts by introducing to the reader the theories and background information for the research objective, and the relation between the various aspects it incorporates:

Chapter 2 will encompass literature review, defining what libraries are and how they are utilized digitally and online, as well as other digital and online resources of knowledge;

Chapter 3 will discuss classification theory by focusing on the writings of Ranganathan (1989), Kumbhar (2011), Spärck Jones (1970) and Weinberger (2007), that constitute the ideational framework of classification and categorization, their practical and ideational meanings. The chapter outlines the concepts that will be used in the thesis and the interaction between them, as well as practical conducts of organizing information;

Chapter 4 will introduce to the reader with etymology, basic notions of its research, and the current state of English etymological research, and its representation in etymological dictionaries;
Chapter 5 will discuss information representation, its effects on the user in respect to ability to utilize the information: awareness to the existence of information, searching through it and browsing it, as well as data mining and the visualization of its results. Further, the chapter will differentiate between physical and digital storing, and classification versus categorization.

Next, an exploration of the specific resources and tools employed in the thesis will be offered, clarifying the practical procedures of conducting the research:

Chapter 6 will introduce the reader with Etymonline, the database which is the object of research and manipulation. The database’s scope, goal, potential and limitations would be presented, complemented by the reasoning to using it as object of research;

Chapter 7 will clarify the method and procedure used for analysis, detailing the categories and tags that were tailored to the dataset, and their reorganization and processing. Excerpts of data would be presented throughout the chapter, giving the reader a glimpse at the corpus and its presentation. A comparison would then be made between the data surfacing in Etymonline and in the suggested DL approach, followed by the reservation and flaws of the method used;

Chapter 8 will report the results of the research and comparison conducted;

Chapter 9 will offer a discussion of the information presented in the thesis, and its meaning in relation to context, metadata, organization approaches, visual representation, user needs and effects, and more. The chapter will clarify to the reader how the theoretical background guided the practical procedure, and directed its interpretation;

Finally, chapter 10 will conclude the thesis, outlining its main contribution and potential for further research.

Chapter 2

Literature review

2.1 Libraries and digital libraries – past, present, future

Libraries first appeared some 25 centuries ago (Witten, Bainbridge & Nichols, 2010), initially as institutions focused on collecting and archiving (mainly written) materials. Nowadays, libraries are facilitators of converse; centers of activities and distributors of knowledge. This social character entails that focus must also be placed on services, rather than purely collecting and maintaining data (Calhoun, 2014).

Although the field of Information Retrieval\(^1\) (hereafter IR) belongs in Computer Sciences, its application and effect on Library and Information Sciences (hereafter LIS) is crucial. Therefore, an effective retrieval theory must be accounted for, concerning mainly the representation and organization of the available data and information resources (Jacob, 2004).

This is of such importance as the presentation of existing knowledge helps spread it, the effectiveness of which contributing to knowledge management and various social processes (Chen, 1999). It is the taxonomy of information which directs its exposure to potential users, and therefore should direct the planning of cultural institutions’ layout (Guin, 2012).

In recent decades there has been a great shift in the field of LIS, namely the switch to digital. More and more libraries are extending their digital services, materials are “born digital”, and some libraries, mostly those of specialized fields, exist only digitally, without a physical location (Wilson & Maceviciute, 2012).

However, as bibliographic metadata is mostly aligned with the physical world of shelf-arrangement and restrictions of space, it often continues to prescribe materials’ physical attributes (Calhoun, 2014). Therefore, the need is to develop intuitive and effective ways of exploring the plethora of information, as well as create an engaging interaction with the digital interface which is devoid of physical attributes (Chen, 1999).

\(^1\) Optimizing the matching between user queries and retrieved results.
One of the problems faced by digital libraries (hereafter DLs) is that they are easily confused with other Web 2.0 applications\(^2\). That is due, among other things, to the two fields’ emerging simultaneously and being developed by the same people (Calhoun, 2014), as well as user consumption patterns (Guin, 2012).

Furthermore, there seems to be a lack of consensus regarding what exactly constitutes as a DL (Henderson, 2005). According to the DELOS manifesto (Candela et al., 2007), a DL is an organization with its own system of maintaining digital content, which is of quality and restricted to polices. The DELOS manifesto states, among other things, that:

- The system should be structured by a reference model - i.e., a hierarchical relation between its different functions, restricted to internal needs;
- Data should be indexed, its metadata outlining the scope of the collection and the paths to reach its content;
- Available services are custom-made to reach a pre-defined goal and a specific section of target users.

Hence, it could be stated that DLs’ collections are restricted and impose harsher standards of inclusion (Calhoun, 2014), and provide a structured and organized set of data (Witten et al., 2010). That, unlike the duplicity and uncontrolled data offered by WWW (Calhoun, 2014).

Furthermore, libraries --digital as well as physical-- have a social agenda: empowering individuals; supporting teaching and advanced knowledge; enhancing collaborations between people, companies and domains; preserving heritage (Calhoun, 2014); and promoting accessibility and a communal experience (Guin, 2012).

Nonetheless, ways of accessing information in DLs are affected by the gain of online social networks: information seeking is becoming more and more social and collaborative in nature. Hence, DLs start to include the usage of tagging, rating, mashups, and more (Ding, Chowdhury & Foo, 2001).

The efficiency of a DL’s indexing will, to a great degree, determine how successful retrieval should be. The way indexing should be carried out, however, is not uniform across all DLs. The method selected for metadata marking depends on each respective library’s intended users and goals. Therefore, it is important that DLs know their intended users, and specify the library’s goals so as to best cater to them (Ding et al., 2001).

The greater the database – the more significant the interface design becomes, as it connects the user with the contents, allowing the manipulation of searching and browsing preferences (Ding et al., 2001). With a clear outline of availability, usability is increased (Dobreva, O’Dwyer & Feliciati, 2012).

In indexing, guiding principles should be simplicity and flexibility, in order to facilitate orientation and allow the introduction or deletion of sub-topics. In such a way, topics that are of greater relevance could have more sub-divisions, easily introducing to users the scope of the dataset and its hierarchy (Garside, 1954).

### 2.2 Volunteer based digital resources

There are many DLs and online library services, either available to the general public or restricted to registered members. The following section will present some of the most notable volunteer-based libraries and online communities, their attributes and precedence, in order to further outline, as well as contrast with the properties which define or are desired in a DL.

---

\(^2\) Web 2.0 is the set of principles that define the application of WWW services. The operation of apps, interfaces and websites adheres to using the web as a platform connecting these varied services according to a shared perception and vision (O’reilly, 2005).
2.2.1 Project Gutenberg
Project Gutenberg is a non-profit DL conceived in 1971 and finally started in 1991 (Witten et al., 2010), that works for the digitizing and spreading of mainly literary works, as well as other arts (Calhoun, 2014). The DL is not driven by any government or an organization, but rather solely by volunteers and donations (Hart, 2004). Although Project Gutenberg is not a scholarly institution as is a university or governmental research facility, it still holds great value to scholars (among other users). The work of scholars is one that relies on the works of their predecessors, and therefore the preservation and sharing of literary works and cultural artifacts is essential (Hedstrom, 1997). Today, Project Gutenberg has more than 54,000 books and their metadata, as it was entered by avid users (Witten et al., 2010). Materials are digitized into different formats so as to enable access to as many users as possible.

2.2.2 Wikipedia
According to Weinberger (2007), Wikipedia, the online encyclopedia, sets an example of how anonymous authors work together and shape and distribute information. The site enables the free creation of entries, their editing and translating; the more an article is edited, the more defined, encompassing and accurate it becomes. One major disadvantage that is promoted by this online encyclopedia is that of anonymity which enables writers to spread invalidated claims, be it purposefully or naïvely. Nonetheless, as soon as another member notices entries that need correction, a change may be implemented at once, thus correcting or eliminating false facts. Another indication of its countless contributors is that Wikipedia constantly grows and improves, including more and more aspects and characteristics in its entries. Long entries reflect a passionate writer or groups of writers, sharing knowledge with anyone who would care to access it. Although Wikipedia is an online encyclopedia and does not pretend to be a DL, it has a great contribution to the topic at hand. Namely, one of the most important lessons to be learnt from the successful implementation of Wikipedia, according to Weinberger, is that such an endeavor is indeed possible, and that communities can operate and grow despite the remoteness and individual work of their members, despite the lack of monetary or other compensation (or even recognition), and despite there being no leadership that directs the growth of this digital institution (Weinberger, 2007).

2.2.3 Flickr
According to Weinberger (2007), Flickr is a community type platform that allows its members to share their digital content. The site does not offer any pre-set system for categorizing the content, but relies solely on user contribution for the tagging of the collection. The site could be compared to Wikipedia and Project Gutenberg as a learning center or a source of producing knowledge by a type of social contact. The tagging, which categorizes the collection, is intuitive, dynamic, ever-expanding and ever-changing, and yet provides an efficient way of exploring the collection. Flickr provides proof that intuitive and uncontrolled tagging works, and that no authority is necessarily required to delimit or nest tags anymore than is already done by users.

According to Weinberger (2007), in all the above examples, the organization, be it a DL, a platform or a site, constitutes a whole which is bigger than the sum of its parts; it presents a collective memory which promotes individual learning.
The three very different presented examples have some central properties that are shared between them, and namely that of being made up of (among others) a flexible group of professionals sharing common interests, engaged in independent tasks of a common purpose, together forming a store of common knowledge (Weinberger, 2007). Such a group is characterized by voluntary and self-regulated participation, tacit understanding of common interest, mutual sources of gain, shared practices and mutual trust (Weinberger, 2007).

The fact that these communities continue to grow and develop is enabled by their respective members’ having similar values and resources (language, conduct, background, interests, etc.), as well as their feeling rewarded by the activities and results achieved by the community (Weinberger, 2007). Weinberger even goes as far as comparing these practices to an organic being that grows and develops, where each respective member leaves a distinct mark, contributing with their own unique imprint.

Although of the given examples, only Project Gutenberg is recognized as a DL, both Wikipedia and Flickr serve as important examples to the current paper in their innovativeness in the digital world and in tagging systems, respectively.

Chapter 3
Classification theory

The following chapter will discuss classification theory: the origins of classification in human cognition, its role in shaping our individual and collective identity, its manifestation in various settings, as well as advantages it offers or disadvantages it causes.

Classification is apparent in almost every field in our lives. However, it remains widely unaccounted for, leaving research that relies upon it to find its own system and the justification for utilizing it (Spärck Jones, 1970).

Classification has come a long way since its early appearance in child development, through records of it in historic human interaction, and to today’s ordering and structuring of data in digital form. Each of these stages will be presented in this chapter, concluding with tagging as a means for classifying, and the application of tagging in a successful computational linguistics project, The Penn Tags Project.

The project illustrates how tagging can perform flexible, intuitive and yet consistent classification, carried out by both experts and laymen. Based on the success of the project, as well as the theory explored in this chapter, the following chapters will exhibit a tailored system of tagging for the corpus analyzed in the thesis.

3.1 Foundations of classification

According to Ranganathan (1989), humans start classifying objects according to their likeness with other items already at the tender age of 12 months. This tendency of sorting things begins at the concrete level of objects and their external properties, and extends to abstract ideas as mental development takes place; as experience becomes richer, so does the degree of subdivision within each respective category. This property not only helps memory load, but also aids thinking processes.

There is written historic evidence that shows how classification has been an important attribute in humans even in prehistoric times; traders would organize their goods according to appearance or function. In the realm of libraries and archives the task of sorting takes a step further, as treasurers needed to arrange articles according to their non-sensual context. This requires the developing of another mechanism of organization – that of numbers, marks and symbols. Hence, the essence of library classification is deciding on a particular order of arrangement and a notation system for its indication (Ranganathan, 1989).
Such a scheme\(^3\) of classification will give greater autonomy to individual classifiers and aid organizing a corpus immediately. When arranged by their classes, objects fall in a preferred helpful order. The science of classification is therefore the result of controversy regarding which is the best order by and which ideas and/or materials it should be arranged; a sub-group of Logic of Relation. As for library classification, it is essentially a product of social forces, as it is a social institution with a social purpose – educating the public and preserving knowledge (Ranganathan, 1989).

In the past there were fewer books and therefore fewer categories for classifying them. Today, however, there is a multitude of knowledge and hence a need for more categories and more fine-tuning that should differentiate between them and their respective documents. Furthermore, digitization allows the creation of new categories that were not previously accounted for, as well as an overlap between them. I.e., sets’ not being mutually exclusive represents a new stage in the evolution of libraries (Ranganathan, 1989).

A certain contradiction arises when considering that on the one hand, the number of specialist-readers increases over time, and it is required that scientific information be arranged such that specialists world-wide may be able to access it. On the other hand, “classification is a logical absurdity” (as noted in Ranganathan, p. 34), since categorization systems were developed by and for librarians, without other sectors in the general public taking part in the process. Nonetheless, there is a tradition of a categorization system, and the fact that it still exists today and enables the entering of new fields without violating its core says a lot about its force (Ranganathan, 1989).

3.2 The power of classification systems

Similarly to Ranganathan’s assertion of classification being a basic human trait (1989), Kumbhar (2011) asserts that list-making too is a fundamental human activity. In fact, lists are one of the oldest records of civilization (Kumbhar, 2011).

Lists are a genre of representation used to coordinate work that spans across time and space, and is organizationally complex. For example, a dictionary is essentially a list, ordering masses of information that took many years to assemble, and detailing its entries according to a chosen pattern (Kumbhar, 2011). However, classifying is a more advanced conduct than list-making, as it entails the following criteria (Kumbhar, 2011):

1. Consistency: regarding issues of origin, descent, hierarchy, etc.;
2. Mutual exclusion of classes: no overlap should be made permissible;
3. Completeness: the system provides total coverage of a given body of data.

According to Kumbhar (2011), values, policies and modes of practice are embedded into classification systems. The actual work of categorizing may cause shifts in the categorization system, raising questions such as how a classification scheme should be constructed and how fine-grained it aims to be.

When designing a classification system, the work stream primarily depends on (Kumbhar, 2011):

1. Comparability: keeping consistent terminology and judgment;
2. Visibility: all classes should be equally shown and accessible, their logic transparent;
3. Control: some entity must oversee the work to assure a successful process.

However, all these factors often compete against each other, and therefore such an ideal system does not exist; all parameters cannot be optimized at once. Even detailed, established and

---

\(^3\) “Scheme” in the sense that a classification method can be regarded as a language, entailing semantic relations between its codes and developing according to a consistent logic.
well thought out systems (e.g., the Dewey decimal classification) are not without flaws and in need of supplements and adjustments, according to specific libraries’ needs (Kumbhar, 2011).

Therefore, any classification system embodies compromise, and cases that fail to fit a certain system pose a problem. This difficulty is worsened when considering how underrepresented classification theory is in the literature and difficult to find (Kumbhar, 2011).

Classification is both material and symbolic, and its impact on our lives is grand - to the point of reflecting and affecting ethical conflicts. Available categories reveal a lot about a person, culture or heritage. Therefore, pragmatic and semantic sensitivity is often called for when dealing with issues of classification (Kumbhar, 2011).

One of the main objectives of classification schemes is making the past accessible by organizing and ordering libraries, archives and other sources of knowledge. If valuable knowledge is encoded well, the need for a detailed recount is diminished. Hence, classifications are powerful technologies (Kumbhar, 2011).

3.3 Matching a classification system to a corpus – gains and pitfalls

Spärck Jones (1970) differentiates between two main types of classification schemes:

- **Ordered classification**: has systematic relationships between classes. This feature increases the permissible connections that may be yielded by the system;

- **Polythetic classification**: allows a single member to belong in more than one class simultaneously. In the current context, this type of classification reflects that a single word has more than one meaning, and can belong in multiple categories.

Ultimately, a classification system should be chosen according to how much less distortion it entails to the data set, or how well it answers to a research question. I.e., what parts of the information it encompasses may be flaunted and what might be veiled (Spärck Jones, 1970).

Therefore, classification is an abstract notion where criteria are put to action in response to a specific goal or requirements. In order to answer to such defined needs one must be fully aware of them, as well as to the contents of the dataset (Spärck Jones, 1970).

However, that is not often the case. Rather, researchers often draw a categorization system from other disciplines (e.g., biology) and impose it onto their data. Had there been a sufficient theory of classification systems, it might have been easier to predict which system would best match which type of dataset, and a choice would be more easily justified (Spärck Jones, 1970).

Also, it is possible that the initial goal of aiming to finding sets is fundamentally wrong if items aren’t matched. I.e., even if criteria are set, there is no guarantee that a categorization system would necessarily emerge. For instance, in a case where there are as many classes as there are items. Therefore, phrasing a classification system is no simple task, even when directed by a specific approach and a specific query (Spärck Jones, 1970).

The above mentioned difficulties are further increased when considering that retrieval tasks are concerned with relevance, which is a subjective and contextual notion. When users are unfamiliar with the contents of a dataset, searching it becomes a greater challenge. The corpus then turns into a collection of hidden, unobtainable facts, only fully accessible when one executes complete control of the classifying principles (Spärck Jones, 1970).

3.4 The freedom and the difficulties of classifying in the digital world

In his 2007 book, “Everything is miscellaneous: the power of the new digital disorder”, Weinberger discusses information organization that is not bound by physical limitations. The multitude of information available digitally is so vast, that strategies must be employed in order to wade through it and locate desired data and information. Such strategies may be divided into searching, browsing, and the use of metadata as the pointer of relevant areas, collections or terms.

The current section §3.4 discusses these, as well as other difficulties in information organization, and some of their suggested solutions.
3.4.1. Searching versus browsing

Unlike searching, which is the purposeful seek of information in order to satisfy a user goal, browsing is the (more or less) random looking through a corpus, getting an impression of its content and further inspecting the items which capture one’s interest. These two activities are related, or even intertwined, and mainly so in the digital world. However, they are governed by different approaches and make use of different strategies available to users, either on the interface or in how to exploit it (Weinberger, 2007).

According to the Weinberger (2007), browsing is the intentional ignoring of information or metadata structure. As such, it is carried out quite differently by different people, or in different occasions, and is characterized by unexpected patterns. These patterns might even surprise users themselves, as the browsing process reveals to them categories or pieces of information they did not previously know existed, redirecting their searching to new domains or with the use of new strategies (Weinberger, 2007).

Often times the layout of a DL or platform can reveal if its creators aim to reach seekers or browsers; i.e., users that approach the search with a purpose and a predefined strategy, or rather users that prefer to get a taste of the contents and decide about their following steps as they become more familiar with it (Weinberger, 2007).

3.4.2 Metadata

Since libraries were initially treasurers of information and not distributors of it (as mentioned in section §2.1 above) a catalogue was used for listing items’ location, so that they may be found by the administrators. The catalogue also includes basic information about items and their characteristics (such as books’ titles, publishers, year of issuing, size, type of cover, etc.). However, as the multitude of information expands, more precise metadata must be noted in order to differentiate items from one another and correctly identify them (Weinberger, 2007).

It might seem counterproductive to add more information on top of the existing mass of information, but in fact, the only way to manage such a scope of data is by adding more data; describing it better and in greater detail. Even though the administrators of the collection must handle such vast amounts of information, it would never turn out to be too much information. The reason for that is that one may choose which information the system should present and which it should conceal; that which is relevant may be easily seen and used, whereas that which at a certain context fails to be beneficial, may simply be hidden from view and ignored (Weinberger, 2007).

Metadata may also be used for other purposes besides the mere locating of items. I.e., metadata, as it is a description of the collection and the items it encompasses, allows for browsing and acquainting one’s self with the contents of a library (Weinberger, 2007).

Furthermore, metadata enables making connections between what might seem to be independent pieces of information. A standardization of metadata in specific realms (e.g., within libraries pertaining to shared topics) would prescribe what properties it should describe. Given that one knows what vocabulary is used in the standard, it enables highly efficient information retrieval. Then, the metadata can be listed or compared in order to efficiently generate outputs of interest. Hence, metadata plays a vital role in the managing of the collection and organizing it according to different sorting schemes (Weinberger, 2007).

As such, metadata makes the collection a library rather than a mass of unstructured information, since structured data may be manipulated without understanding its contents (Witten et al., 2010).

Furthermore, metadata enables the existence of surrogates, browsing, highlighting, accurate display, statistical information, and more. Hence, its arrangement affects display, searchability, interoperability, and more (Witten et al., 2010).
3.4.3 Misc.
When different items do not seem to have common properties between them, the organization of information becomes more difficult (as mentioned in section §3.3). Often times administrators find certain items to be unfitting under any of the available categories, and yet not important or large enough to make their own category (Weinberger, 2007).

In such cases, these independent or unfitting items find their way into the category “miscellaneous”, abbreviated misc., where no organizing principle governs their location within the set. Despite the seeming chaos, mixing items together might actually encourage creativity, reveal surprising facts and even maximize serendipity. The downside of such an approach to information exploration is that discoveries often become more difficult to trace or reconstruct later (Weinberger, 2007).

3.4.4 Different ways of organizing
Organization means the control over the presentation of information, as well as cultural and political views that are manifested in the form of classes, categories, titles and links (as mentioned in section §2.1) (Weinberger, 2007).

It seems that alphabetical order is perceived as the default, or dominant order, for lack of a better alternative. But in fact, alphabetical order is no order at all, as it tells us nothing about the relationships between members of a set (a set, in this case, being all dictionary entries under the letter N, for instance). It actually makes a point of ignoring possible links between items, presenting them in an ascending list. As such, it is merely one step away from projecting no order at all, lists being the most basic way of ordering ideas (as presented in section §3.2) (Weinberger, 2007).

Nonetheless, for the purposes of browsing, if no specific property is called for, alphabetical order can be a very appropriate; a random and yet orderly display of data. However, the administrator cannot predict all the purposes and strategies users might employ in accessing the data. Hence, creating categories that would organize information in chunks would be ideal. A certain ordering would also promote information retrieval when the seeker is not an expert, having only a vague idea about what it is that they are searching for. Another motivation to arrange information in chunks is that it helps users discover or remember points of interest (Weinberger, 2007).

The categorization process may be governed by definitions of objects, or by our relation to those objects; the way we organize things reveals our own judgment about those things, our background and perception. For example, consider the presentation of the world map in atlases purchased in different countries – each would have greater detailing about the country in which it was issued, and may even present the oceans and continents in the same respective perception. I.e., one atlas might have a break in the Atlantic ocean, making Africa appear on the left hand side of the page and the Americas on the right, unlike the way we are used to viewing it here, with a break in the Pacific ocean, and the Americas appearing on the left hand side (Weinberger, 2007).

In essence, every presentation that follows a specific order actually disrupts a different possible order. This could mean that by trying to present something about the world we end up hiding something else in its stead. Therefore, when assigning categories, one should do so according to categories that they believe users would want to search by. As the digital world enables more than one way of organizing information, there is much more flexibility allowed, enabling different users to arrange information according to different properties. No single way could necessarily be claimed as portraying the “best” way of organizing information. Rather, a certain way fits a certain user at a certain time or context (Weinberger, 2007).

When creating categories, administrators should create sets that are open-ended such that they allow for the creation of new entries or the deletion of obsolete ones, without making the entire ideational construction crash. Meaning, if data is categorized in a hierarchical manner, a
change in one category might cascade into daughter-categories, thus promoting greater changes and possible incongruence (See figure 2 below) (Weinberger, 2007).

Unlike hierarchical categorization, faceted categorization operates such that users may choose different paths in getting to the desired item, even without knowing the relation between that desired item and others in its category. This property shows the different interpretations one may assign to reality, according to varying needs (See figure 3 below). In a faceted categorization a set vocabulary is controlled and more than one category may apply to a given item. Hence, searching may only be done in the existing categories. New facets may emerge without disrupting the already existing ones; however, the opposite is not true, and removing facets is not as easily done (Weinberger, 2007).

Figure 2: Hierarchical categorization

Figure 3: Faceted categorization

In the above figure 3, the lexical item “counter” may serve as an example portrayed by the blackened square. “Counter” could belong in the set of nouns (meaning a desk or a heightened surface of some sort); the set of verbs (meaning “to oppose”); and the set of adverbs (meaning a reversed way to another, in context, as in, e.g., “counter clockwise”). A user arriving at “counter” from the perspective of a noun may remain unaware of the other two perspectives, while viewing the item in that given context (Weinberger, 2007).

3.4.5 Tagging

A tagging system means the assigning of category names to items such that they are accumulated and provide a description of the item. Category names may be added or omitted from the search, thereby expanding or restricting the results yielded by the query. Hence, tagging virtually provides filtering (Weinberger, 2007).

As such, tagging is, in essence, a means to an end which is classification. However, it is distinct in that it has no hierarchical relation, and therefore the insertion of new tags or deletion of existing ones may be done easily⁴. This is a great advantage in the DL information management field, as knowledge continues to grow and change our perceptions of previously made ties (Weinberger, 2007).

Furthermore, like in a faceted system, viewing an item in a context of other items that share its tag (be it partial or full likeness) allows users to discover more about properties and ties between items. Unlike in a faceted system, vocabulary is not controlled, which allows users to employ their own searching and browsing experience, as well as to add tags themselves, in some platforms. Changing the tagging relationship⁵ in the query allows users to see the relations that

---

⁴ Granted, tags could be made to embed other tags, such that a hierarchical order is created. Nonetheless, the method enables linear assignment where a given tag stands independently of its neighbors.

⁵ For example, “A and B and not-C” versus “A and C and not-B”.

matter to them, in a specific occasion. Changing the relationship between different searches may encourage further learning of other relations and meanings (Weinberger, 2007).

As for the adding of tags and the creating of new categories, some might claim that that is work best carried out by experts. But in today’s digital conducts, expert opinions are often no longer deemed as necessary, and often users create the tagging variety themselves (as introduced in section §2.2.3) (Weinberger, 2007).

This means that the extent of available varieties is as rich as the extent of users, or even greater. Allowing everyone’s creativity and opinions to show on the surface in the form of tags allows for endless clustering, filtering and presentation possibilities. Allowing user contribution of tags may create “messiness” – inconsistency and constant change. Those, however, do not necessarily have to be viewed as flaws. Depending on the goals of the DL, constant change may be exactly fitting to today’s information consumption patterns (Weinberger, 2007).

Furthermore, the multitude of categorizing options may yield better results when a search is carried out well. I.e., less occurrences of exclusion of relevant results, without the increased inclusion of false positives (Weinberger, 2007).

Although storing and accessing information digitally has been carried out for decades, the usage of tags as we know it now started only by the early 2000’s with the gaining of “Delicious” – a social bookmarking application. Until then, digital organization had been carried out by creating folders. An item would be dragged and dropped into a folder titled by the category name. This means that an item which fits into more than one category has to be copied into more folders, resulting in immense duplicity. With the help of tags, however, branches, relations and links may be exhibited or hidden, with each entry corresponding to a single item. Still, that single entry may correlate to more than one place (tag) at a given time. Furthermore, the connections created by tagging allow reaching a single item by countless routes (Weinberger, 2007).

One great disadvantage of tagging arises when considering the creation of new tags and applying them to older data. Meaning, if a new category is created, how should it be projected to existing data that should fit it, but needs to be tagged? It would entail the tagging and re-tagging of existing data in a never-ending loop in order to maintain an up-to-date and comprehensive network of data (Weinberger, 2007).

### 3.4.6 Penn tags project

An example of a successful tagging system may be seen in the Penn Treebank Tag set. This is a project carried out by the University of Pennsylvania where lexical strings were tagged for their meanings, segmenting string to their respective units and tagging the syntactic role of those units (Santorini, 1990).

When an item had more than one possible meaning or function, a slash was used to allow the notation of further roles. Using the above suggested example “counter” (§3.4.4), it would be tagged NN / VB / RB to denote noun, verb or adverb, respectively (Santorini, 1990).

The project illustrates the benefits of a tagging system by transcribing strings of speech and annotating their possible functions as a short list of symbols and letters that can then be computer-analyzed for further research. It further illustrates that any possible entry can be notated in some way, disallowing any exceptions to rules or words ending up as singletons in their assigned category. I.e., the elimination of exceptions (Weinberger, 2007).

This tagging system is an open ended one, and new tags may be added in case the need arises. Although the symbols and abbreviations for categories are not always intuitive, it does not take an expert to use the system, as each category is explained and exemplifies for easy use by laymen (Santorini, 1990).

---

6 The complete list of tags is available on [https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) (retrieved 2017-10-31).
At the same time as different systems of organizing information offer different advantages and disadvantages, one should keep in mind that no one single knowledge organization system can pretend to capture all intellectual and physical entities. That is due to the fact that natural language plays an important role in the expressing of ideas, and natural language could never be fully and flawlessly translated (Ranganathan, 1989).

Drawing from these properties of an open-ended, intuitive and encompassing system, the current paper will construct a method of tagging an existing pool of data, such that it could later be manipulated, supplemented, edited or removed, as well as fully cover each available entry.

The method used (elaborated on in Chapter 7) reflects a personal interpretation and visualization of the data. That being said, it still offers intuitive understanding and room for change and development once introduced to other contributors.

Chapter 4
Etymology

Etymology is the outlining of the origin and the historical development of words; it reveals how ideas, conceived in the human brain, take form in words that have a metaphoric onset, and develop further into what seems as more arbitrary forms (Gray, 1986).

Etymological research is that which enabled outlining a connection between all the major languages of Europe, India and the Middle East, and showing that the vocabulary of English stems from a variety of different languages. That is thanks to recapturing of Indo-European origins, dating back some 6,000 years. If a root is shared between all daughter-languages, it can be assumed that it dates back (at least) that far (Gray, 1986).

In other cases, where a root is presumed to have dropped out of a language and replaced by a different word, different languages might exhibit respective difference between them (Gray, 1986).

As inconsistent and sporadic as language is, these two neat scenarios are not always the case. Rather, anomalies might be found, and those are more difficult to relate to with certainty. Such cases would often be regarded as pertaining to “obscure origin”, for lack of a founded theory that can account for their appearance (Gray, 1986).

4.1 English etymology

The formation of English as we know it today was initiated in the first half of the second millennia. Until then, English had had a clear West-Germanic affiliation. With the invasion of the Norman-French into England, they imposed their romance language over the population, thereby inserting clear Latin and Greek influences, and even more so in the late renaissance (1500 – 1625). Hence, even though the basic word stock of English is Germanic, the greater part of it (up to ¾ of the lexicon, according to some) is non-Germanic (Gray, 1986).

This brings about the relevance of etymological research, which traces the root of every lexical entry, at least as far back as its Indo-European or Mediterranean origin. In order to account for a proposed analysis, researchers have to back their conclusion with examples of other words that have emerged or developed by a similar path, or from the same root (Gray, 1986).

The results of etymological research can later be edited into etymological dictionaries, where each entry is provided with the historic account and formation of the given word and the general lexical relationships between it and others. But most of all, etymological dictionaries have a dual purpose: they recognize relations between different roots; and establish whether a root is Latin, Greek or other (Gray, 1986).

In the Oxford English Dictionary the editors only provide tracing of entries as far back as their earliest English known source, or, when possible, their Teutonic form. Meaning, entries of a known foreign sources that were adopted into the English language are not explained further.
This approach fails to capture the essence of etymological research and importance, since the origin of foreign words is no less relevant to the understanding of the nature of English, its history and development. Be entries foreign or not, they are what makes the language what it is today, and they reveal, even to the slightest degree, the processes that took place along its history (Gray, 1986).

4.2 Lexical approaches to etymological dictionaries

The segmentation of a single word into its respective units and their respective analyses might make etymological research quicker and more efficient, but it poses a difficulty in cases of scholarly disagreement, or discrepancies between dictionaries (Gray, 1986).

Another difficulty arises when editing all the resulting data into a dictionary. As each entry entails relating to its respective units, explanations become repetitive and tedious (Gray, 1986).

Different dictionaries show different approaches to presenting their data in a way that should best capture the characterization of words, and the relations between them. For instance, in one work the author relates to entries in clusters, which allowed displaying some 12,000 words, but relating to manifolds more. In another, the author categorizes entries according to their semantic meaning, such as body parts, Greek or Latin root, root connection with other languages, semantic relations to words of other roots, and more (exemplified in figure 4 below). In another work, the author includes all relevant information per entry, and therefore uses abbreviations (tags) that account for all of that entry’s properties. However, these tags are provided in German (albeit an English lexicon), making it hard for users to follow. Another method employed by a dictionary is relating strictly to a single lexical category. In this case, a dictionary encompassing some 16,000 English modifiers and their syntactic formation, from noun to adjective (Gray, 1986).

---

**Figure 4: The Indo-European entry *DHEI and its exemplification** (taken from the American Heritage Dictionary, as it appears in Gray, 1986, pp. 12-13):

```
*DHEI
Latin: --IE root meaning to suck
femina meaning "she who suckles."
Examples of English derivatives: feminine, femininity, female, femme savante, or fatale etc.
fetus meaning pregnancy, giving birth
Examples: fetus, fetal, effete, fawn
fecundus meaning fruitful, fertile
Examples: fecund, fecundity
femur meaning hay
Example: fennel
filius, filia meaning son/daughter
Examples: filial, affiliate, affiliation
set (i) are meaning to suck
Example: fellatio
fetes, meaning fruitful, lucky, happy
Examples: felicific, felicitous, felicity, felicitate
Greek:
thel meaning nipple
Examples: endothelium, epitone-thelium
thetus (thety) meaning female
Examples: theelin, thelyblast
```

---

7 In English, those could be prefixes, affixes, stems, conjugation markers, and more.
In addition to these various approaches, some works exclude important information that they regard to as common knowledge, leaving the users to carry out their own additional research (Gray, 1986). Others, as noted in James (2013), are constructed according to their most primal root (namely, Greek or Latin, among others). In effect, this means that in order to be able to locate a word in a dictionary, one should already be familiar with the word, its affiliation or meaning (Gray, 1986). This is, of course, a paradox and a major hindrance for laymen and librarians alike.

Chapter 5
Conceptual and visual representation

“Representation” is having one thing (e.g., a symbol) stand for another thing (e.g., meaning). The symbol does not necessarily need to correspond to the meaning’s medium. For example, when relating to knowledge representation, it is the signifying of information by means of words; expressing ideas through language. In fact, when shifting from meaning to symbol, information may be lost. Hence, knowledge representation is in fact one person’s interpretation and internal processes, rather than actual change in reality (Friedman & Thellefsen, 2011).

How that knowledge is then organized is dependent on semantic tools --indexing, abstracting, classifying and more-- which enable others to evaluate it or enables computer systems to access it. Knowledge management is what enables information retrieval and distribution; constituting the connection between what is treasured in an institution and the way to reach it (Friedman & Thellefsen, 2011).

One of the main difficulties in knowledge organization is the connection (or rather, gap) between language and meaning. I.e., as indexing is the result of interpretation, it also represents a socio-cultural context and an approach to logic (Friedman & Thellefsen, 2011). It also determines what information the system would keep hidden or revealed, as well as the way users would approach information seeking (Vasconcelos, 2008).

Although information management actually produces a lot of documents (data about data), its goal is actually to reduce the volumes of information, as it eliminates duplicates and minimizes retrieval of false positive hits. It also allows the administrators to focus on other issues besides the content of a DL, such as planning for future goals and improvements, budgeting, marketing, preservation, creating information policies and ownership, increased quality, and more. Knowledge management is driven by the following considerations, that determine the tangible representation of abstract knowledge (Vasconcelos, 2008):

- recognizing organizational goals;
- characterizing the information environment and resources;
- promoting receiving the information by those who seek it;
- valuating references and information production;
- defining the strategies of information management.

5.1 System and interface design

There is no necessity in maintaining a lineal relationship between a system’s competence and performance. I.e., the potential offered by a system does not necessarily need to be visible to all users at all times. Some functions available to experts could remain hidden from laymen, and functions available to developers and administrators could remain hidden from experts (Vasconcelos, 2008).

As information systems are made for the facilitating of information retrieval, they should be designed so as to fit specific user goals and strategies. In the process of information seeking, retrieval is the final and most obvious outcome, as well as the only part the user gets to actually participate in. Hence, it is the part that users give most consideration to (Jacob, 2004). Therefore,
the interface design plays a crucial role in information retrieval. Ease of use and reliability were found to be the strongest motivations for using an interface among users (Witten et al., 2010).

Moreover, it was found that as the scope of a collection is obscured from users, they find it more difficult to navigate. As opposed to when the organizational system is familiar to users, its specific terminology and structure make it more approachable as well as effective in obtaining expected results. Of course, finding information is easiest when users know what it is that they are searching for. But for most cases, which are less focused than that of an exact term, browsing has been shown to be more effective than searching (Clough, 2012).

5.2 Psychological representation

Different users have different structures of subject perception in their mind’s eye. Meaning, a subject can be comprised of various bits and pieces of information, background, assumptions, and connections with other subjects. The scope of this metaphorical network is dependent on individual familiarity with the given subject. Therefore, in accordance with each individual model, users expect retrieved results to be displayed in a certain order that reflects that perception. Some research engines are built to cater to that expectation, relying on psychological research to assess user needs and goals (Hjørland, 1992).

Hence, the describing of a subject or document should also take into consideration the aspect of user-psychology, and not assume that the concept subject is a shared knowledge between writers and readers. This can be achieved if we know something about the users (language, background, etc.) and their expectations (Hjørland, 1992).

At the same time, it is the representation of knowledge within documents that is the primary concern. Therefore, one should expect individual user responsibility in their making themselves aware of specific subject terminology, categories and classes, in order to assign well formed queries (Hjørland, 1992).

A document may be characterized by its properties, which are all the true statements that can be made about it. The register in which the document is displayed also plays a role; if it is valid, representative, superficial, etc. The different properties have different meanings to different users and in different occasions or under different disciplines. Therefore, defining ‘the’ central property of a document is as pragmatic as its usage, which depends on the contextual information need. Hence, a subject is how one perceives the contents of a document (Hjørland, 1992).

Given the gravity of context, a subject could thus be claimed to be merely the potential embedded in a document. This is why subjects are at times elusive and difficult to define (Hjørland, 1992).

The potential of the document too is fluid, since it depends on the stage of society where it exists, its development and worldviews. Hence, the perception of subject may change over time. Therefore, the more objective a description of a document is – the better the use of subject system can be made, as well as individual interpretation (Hjørland, 1992).

It has been shown that the visual representation of organization plays a critical role in information retrieval systems, and in affecting personal internal knowledge representation in users (Shera, 1956/1965, as it appears in Jacob, 2004). Grouping search results into topical categories enables effective information representation to end-users (He, 2012).

A topic is defined as the “main theme or subject contained in a (set of) document(s)” (He, 2012 p. 84), whereas a structure of topics shows how a set of documents is focused on a certain topic, measuring topical diversity among documents and semantic relatedness of their topics (He, 2012). Meaning, the structure of topics outlines how alike a set of documents can be claimed to be.

According to Jacob (2004), several main issues need to be considered in order to allow effective communication between the information system and its users:
1) That the communication between systems and users is not harmed by the representation of data;
2) That the organizational structure of the system does not conflict with the internal cognitive structure representation in users;
3) That the organization of data contributes to creating meaningful contexts of information;
4) That the meaning of data is not compromised by the organizational structure in which they appear;
5) Different organizational structures can be applied in order to enhance a single collection of organizational resources.

Following these important points to consider, it is clear to see the importance of the role of representation and organization in the dynamics of information retrieval.

5.3 Text mining

The term text mining relates to the outlining of new knowledge found in a collection, as well as the discovering of latent information in a collection; tracking down trends, relations and patterns between pieces of information through parsing, clustering, tokenizing and more. I.e., it is not the finding of pre-defined relations, but rather the locating of new relations, with the help of existing data (Stavrianou, Andritsos & Nicoloyannis, 2007).

This also helps make up for gaps of missing data, as unsuspected facts may emerge and further enrich the scientific community by facilitating the work of the researchers and minimizing the amount of documents yielded by a single query (Stavrianou et al., 2007). Further, text mining improves navigation and browsing methods (Bruley, 2016).

Text mining is tightly tied with information retrieval methodologies, as well as with linguistics, as it incorporates dealing with natural language peculiarities. Initially, text mining was aimed at the biomedical field, associating terms with ideas (Stavrianou et al., 2007). As the amount of data for analysis is so very vast, computerized tools are necessary to filter it. Otherwise, data might get lost in the sense that it would be overlooked by humans, or simply not seen in the correct context to be able to make sense (Bruley, 2016). The better the text representation is – the more data manipulation is made possible (Stavrianou et al., 2007).

5.4 Text visualization

In information retrieval, it is customary to visualize results in the form of clusters. Users can then enhance the focus over a certain area of the visualization, to gain a better understanding of specific relations, while still maintaining the overarching context (Bruley, 2016).

The first usage of a map in clustering data in the context of research was done in 1844 by Dr. John Snow when analyzing the spread of death cases of a Cholera epidemic in London. Viewing the cases in relation to one another eventually lead to finding their common source, which had been a contaminated water source in a south London neighborhood (Bruley, 2016).
Hence, a visual representation of facts helps draw insight from massive, dynamic, ambiguous, and often conflicting data, and especially so in text documents and in temporal metadata (Bruley, 2016).

### 5.5 Physical versus digital categorization

“It is a case of arranging books and kindred materials in such a way that one kind of arrangement presents itself to one person and another kind to another person. To secure this by pressing a button is obviously possible only in the world of fancy; it is not possible in the world of reality”.

S.R. Ranganathan, 1989, pp. 43-44

Quite unlike the above quote, postulated in 1951\(^8\), the digital reality of today allows just that; the reorganization of information and a customized display at the push of a button.

In the traditional conduct of librarianship, the rule of “one book – one card” was that which drove categorization and organization schemes (Ranganathan, 1989). Nowadays, however, this rule no longer applies, since digitization allows for the exact opposite; there is no limit to how many classes or categories a single item can belong to.

Still, a collection needs to be organized according to some principle that would allow users to view it, browse or search through it, in some sort of topical “shelving”. This is in order to answer to users’ expectations, which are nurtured by the traditional organization of libraries, as well as of other resources of knowledge.

Also, it is because of the vast scope of information available online that there exists a need to delimit it somehow, organize and construct entries, topics and items, in order to be able to find them. Their being digital and accessible alone does not make them any more available to the public, which needs to filter through all manner of data.

\(^8\) according to [http://www.worldcat.org/title/philosophy-of-library-classification/oclc/1409443](http://www.worldcat.org/title/philosophy-of-library-classification/oclc/1409443)
5.6 Classification versus categorization

Classification and categorization are often mistakenly used in synonymous distribution, but they do not stand for the same concept (Jacob, 2004).

5.6.1 Classification

Classification is a taxonomy delimiting groups and the entities that they encompass, according to a predetermined set of principles; a representational tool used to organize information (Jacob, 2004), translating the name of a specific object from natural language into a classificatory language (Ranganathan, 1989). Hence, classification systems regularize the movement of information from one context to another, and the accessing of that information (Kumbhar, 2011).

Each class in the taxonomic scheme is given a label that applies to all of its composing entities (Jacob, 2004), this label owing to a definite ordinal value in relation to the others (Ranganathan, 1989). The label or symbol chosen have a semantic value which remains constant over time, unlike natural language which evolves beyond the control of the creator (Ranganathan, 1989).

Stemming from this, a close inspection of each entity becomes unnecessary, as its class affiliation already reveals a lot of information about it (Jacob, 2004). A classification representation (i.e., class title) is easy to write, brief, subject to the canon of reality (Ranganathan, 1989). Therefore, this taxonomy helps us construct logic, which in turn aids cognitive storage and retrieval of information (Jacob, 2004).

According to Ranganathan (1989), classifying and cataloguing are inseparable.

5.6.2 Categorization

Categorization is the grouping of entities according to similar properties in a given context. It allows simplification of the interpretation of the world around us, as well as shapes our interaction with it, revealing a person’s perception of relations between entities (Jacob, 2004). Moreover, as categories stem from actions and relationships, they are continually refreshed (Kumbhar, 2011).

For example, consider the historic period categorized as the English Renaissance (1500–1650) as opposed to the Middle ages (ending in the 14th century); the former focusing on the similarities between the 14th and 15th century and the latter focusing on their differences. Meaning, categorization systematically divides the world of experience into a structure of categories, each attributed a unique set of features (Jacob, 2004).

Text categorization is the organization of an unstructured text into a structured repository, facilitating processes as storing, searching and browsing, as well as analyzing (Stavrianou et al., 2007).

Classification is rigorous in the sense that it is mutually exclusive and does not allow overlap between classes. Categorization, however, is flexible in the sense that it is context dependent. As context may change, so might the categories, as well as the entities that they include, given that the process of categorization is creative, non binding and concerns immediate similarities between entities. Hence, entities can easily migrate from one set to another, or belong to more than one set at the same time (Jacob, 2004).

Other differences between classification and categorization, besides the processes in their creation, involve their set boundaries, criteria for assignment, permissible or impermissible hierarchy within sets (Jacob, 2004).
Chapter 6
Etymonline

The current thesis will explore the above introduced concept by analyzing Etymonline (http://www.etymonline.com/), an online etymological dictionary (Harper, 2001-2018). The site is not a researching project but rather a compilation of scientific sources and scholarly explanations about the origins of words (according to http://www.etymonline.com/sources.php). The site creator, Douglas Harper, is no linguist or historian; just an avid lover of words, and a supporter of spreading knowledge.

The site is built such that all entries appear in alphabetical order, and various entries entertain links to other entries to which they are related (see figures 6a-6d below). Next to each entry appears a year which represents its earliest written appearance. Doubtlessly, words had been used centuries before they made it to writing, and probably even longer before a written piece of evidence was preserved and made it to our time. The absence of a time indication means unknown year of appearance.

Next, an account for the etymological source(s) appears, the changes a word underwent through the years (e.g., spelling, pronunciation, semantic meaning or syntactic function, and more) as well as links to related words. If no source-language is noted, the default is English.

The creator of the site invites other users to write to him with suggestions for corrections, additions or contradicting information, as long as those are supported by reliable and scholarly explanation. He does not judge for himself in cases where he finds an explanation to be unconvincing, if it is a scholarly explanation. I.e., he removes himself from the researching front and takes only the position of a collector and mediator of knowledge, contributing to a uniformity of data representation along the process (as different entries originate from different sources, each source with its own respective style).

As such, the site is of great contribution to the scientific community, as well as laymen and even random users with a passion for words.

Figure 6: Screenshots of Etymonline, from start page, to a specific entry (retrieved 2017-11-02)

6.a Start page
6.b Focusing on the letter S:

- (s)  suffix denoting all Modern English place names, generally neutered in Middle English, from Old English ge-, the nominative plural and accusative plural ending of certain ‘setting’ masculine nouns (such as earth, earthen).
- (a) third person singular present indicative suffix of verbs, it represents Old English -an, -an, which began to replace -ian, -ian as Old English -an, and generally spread south until by Shakespeare’s time it had survived.
- ean - zoean element used in forming determined names, from Latinized form of Greek noun “linoz,” a word of unknown origin; possibly related to noun phrase “observing, overseeing.”
- -er - this letter group can represent five distinct sounds in English; it first was used by Middle English writers to render Old English -ere, the pronunciation of which was simplified to -ere (an evolution that顺畅ook.
- -scope word-forming element indicating “an instrument for seeing,” from Late Latin -spectus, from Greek -skopein, from skopein “to look at, examine” (from Hittite root -spake “to observe”).
- -scrip word-forming element meaning “writing, counting, observing,” from Middle Latin -scriptos, from Greek -skriptos, from skriphtos “to write at, examine” (from Hittite root -spake “to observe”).
- -ship word-forming element meaning “ability, condition; act, power; skill, office, position; relation between,” Middle English -ship, from Old English -scip “state, condition of being,” from Proto-
- -sien suffix in Greek-derived names denoting action, process, state, condition, from Greek -oia, which is identical in meaning with Latin -sectis (see sectis).
- -sk relative suffix to words in Old English origin (such as tank; literally “chaff the mouth”), contrived from Old Norse sk, reflexes present corresponding to Gothic sk, Old High German sk, German sch “sound, vocal, sound.
- -some (s) word-forming element meaning “the body,” Middle Latin, from Greek noun “the body” (see semeios).

6.c Choosing the random entry “Second (v.)”:

**second (v.)**

1590s, “to support or represent in a duel, fight, etc.,” from Middle French seconder, from Latin secundare “to assist, make favorable,” from secundus “assisting, favorable, following, second” (see second (adj.).) The parliamentary sense is first recorded 1590s. Related: Seconded; seconding.

The creator’s strictness about fact-checking and reliance on written scientific research make its content to be of great value, unlike Wikipedia, for instance, where any user may add any piece of information without scholarly justification.

Also, the creator’s ambition to smooth out the differences between the different sources and provide a uniform representation of all etymological explanations, as well as accurate linking between related entries, make a worthy improvement in the field of linguistic research.

A recent development to the site is that of “related terms” appearing at the bottom of many of its entries. Those represent links to other entries in the site that share a stem, show related semantic meaning or include the current entry in a compound word. (See figure 6d below).
This “related entries” feature had not been present in the start of the writing of the current thesis. Nonetheless, its application may only be used to further stress the contribution of the methodology introduced herein (chapter 7 below), which offers a structured, comprehensive and consistent network of ties that enables a thorough investigation of etymological data.

The reasons this site was chosen to serve as the basis for the thesis are:

1. **Turning the database into a DL**
   Owing to metadata, a change in the structure of the site is all it takes to turn it from a dictionary or a database into an online DL. The contents of the site does not need to change, but merely its underlying construction that would turn it into an ordered, catalogued library.

   Having the same data accessible under the framework of a library is beneficial to users and administrators alike:
   - To users, as it enables them a better searching and browsing experience, as well as the confidence of a scholarly source rather than a website that the validity of which needs to be justified; and to administrators, as metadata and the way in which it is stored make for a much smoother manipulation or changing of features, representation, data migration, server updating, and many more such “behind the scenes” maintenance routines.

2. **Separating entries from metadata**
   Since the foundation of the metadata is already in existence, the site enables a simple transformation of structure; there is no need to conduct more etymological research or import information from another source, but rather only tease apart the entries from their metadata, storing each in a separate respective “layer”.

   Having metadata searchable, be it available to the user at the interface level or not, allows maximal usage of the contents and more varied alternatives for researching etymology and making links which are currently impossible to find.

   As metadata is the constructed representation of additional information about the existing entries, information can be easily added or removed from it without causing any apparent change to the entries, as they are presented to users.

   The structure of metadata also enables a more efficient search in its contents, which is not currently available in searching through entries. Since it is the metadata which holds information about the contents, it can be made use of for in-depth searches, such as shared roots, shared paths of change between words, year of first appearance, and more.

3. **Improving user experience**
   Having metadata segmented from entries allows users to not only search the catalogue, but also to browse it. The currently available searching is restricted to alphabetical order, where a user is expected to approach the site with a specific goal (term) in mind. Unlike
it, browsing may be done according to other criteria and may lead users to information that they did not plan for, and possibly didn’t even know existed.

For instance, browsing the catalogue through word ending rather than beginnings (e.g., -phobia, -ism, -ship); browsing lexical categories (e.g., entries that are both noun and adjective); year of first appearance (e.g., prior to 1200 or restricted to the first half of the 16th century); complex routes (e.g., words that have entered English through several languages, like links in a chain), and many more.

Browsing might prove not only more enjoyable and attractive to users, but might also fill a greater role, enabling making new discoveries, constituting a novel feature in the field of etymological research.

4. Supporting etymological research

Improving the site and its up-keeping routines, granting a better user experience and promoting etymological research would offer further support to the creator of the site. The approach of presenting words’ histories for as far back as research enables offers a more comprehensive experience than other dictionaries; the dictionary being online further promotes access to these historic recounts. Therefore, improving the structure of the site and its interface would contribute even more to the goals of Etymonline and its creator.

In a quest for the spreading of knowledge in the field of etymology, this is a great place to start; such a contribution can really make a difference and improve etymological research, as well as attract potential users and increase the site’s funding, influence and scope.

As the situation currently is, having each entry stand on its own --despite the detailed explanation of its origin-- does not answer to searching and browsing expectations, and constitutes a scope too wide to be able to capture in a list formation. Hence, for the current thesis, the goal is to associate terms with properties (periods, patterns of change, origins of words, and other common denominators) by using a custom filtering mechanism operated by tags. By doing so, the information would be searchable according to predefined criteria. In a dictionary one has to know the term to look up in order to find information about it. However, in a library, one only needs to know one attribute which is of interest in order to generate a list of hits that could be relevant.

In both a dictionary and a DL one enters a term into the search field and is answered by details pertaining to that term. But in a DL, those hits are structured, interconnected, and offer the user much more related relevant information (e.g., in the form of metadata). The same could not necessarily be said about a dictionary, where one needs to be aware of connections between words in order to locate and verify them.

Hence, the research question driving the current thesis --exploring how a model of DL would improve making connections between entries, in comparison to the existing database--, was applied to Etymonline. Etymonline constitutes ideal material for this research as it encompasses great and diverse data, which nonetheless has many shared attributes and presents links of different dimensions (i.e., connections between words in respect to temporal, phonological, geographical, and other aspects).

The database being available online makes it easy to reach and analyze, and the information it holds allows for the segmentation of data and metadata to be carried out systematically. Following these, it is possible to compare between the two structures (the existing list formation and the suggested tagging) and the information they reveal. Namely, it is possible to check and quantify hits per search query and number of ties offered for browsing upon filtering for specific attributes (categories or tags).
The research will therefore take on a quantitative approach, comparing the number of linguistic and etymological ties each entry exhibits – once in the current interface, and once in the suggested model. The ties that will be inspected are those that make the categories defining the entries (metadata). I.e., the thesis will show the number of faceted routes that lead to the possible finding of a single entry, through searching and browsing.
Chapter 7
Method and procedure
7.1 Contents and set selection

Etymonline covers more than 47,000 entries, alphabetized into 26 sections, according to their starting letter. Clearly, such a scope is too wide for a single person to cover under the framework of a master’s thesis, and therefore there was a need to delimit the set of material to be analyzed.

A random selection of any one letter would have sufficed in showing the principles introduced in the thesis, but it would have made it difficult to predict the amount of work required for the analysis. For instance, it was found that the letter S has more than 5,000 entries, which would have made the process extremely long, and not necessarily provide more informative results. Additionally, a bias might have been created if choosing a letter that the entries of which mostly derive only from a single source, or a limited scope of diversity. For example, the letters X and Z that mostly contain entries derived of Greek and Latin, respectively.

Therefore, an estimate of the set-size of each letter was done, followed by a comparison to the words distribution in a general English language dictionary. It was found that the ratio between set sizes in Etymonline was consistent with the set sizes of a lexical dictionary. I.e., Etymonline covers English entries in accordance with their scope in a lexical distribution; although it does not cover the whole of the lexicon, it maintains a fair representation of the English vocabulary (see figure 7).

In both the general dictionary and Etymonline, N represented a close to median value of entries (between S, P and C with most entries and X, Y and Z with the fewest), comprising of 978 entries.

Figure 7: Letter distribution estimation comparison between the English lexicon and Etymonline

9 Value is the result of own estimate, further supported by https://journal.orton-gillingham.com/qa-with-douglas-harper-creator-of-the-online-etymology-dictionary/  
10 Available as Appendix 1.  
12 The English lexicon estimation constitutes the average results of the two online sources used.
7.2 Method

The 978 entries of the letter N were collected and processed with the help of Excel – columns representing classes and categories, and rows representing entries. The actual process of assigning categories and tags to the corpus took place in two rounds:

As the total of sets’ requirements were unknown at first (i.e., categories and tags could not be predicted prior to processing entries), they were constructed and added as analysis progressed and more data was being inspected and characterized.

After completion of analysis of the corpus, certain set-titles needed to be changed in order to better capture their essence and to include or exclude members; typos and spelling mistakes were spotted and corrected, additional routes of entry were added, and more. Therefore, a second inspection of the corpus took place, to assure that no item was neglected or improperly tagged.

In order to keep track of the rate of progress and to more easily spot inconsistencies between the Excel collection of data and the existing Etymonline database, entries were ordered exactly according to their appearance in the site. I.e., as the site allows the presentation of twenty entries per page, a new Excel sheet was started after every 20 entries. That way, a word appearing on the site on page 3, for instance, appeared in the analysis on sheet 3, accordingly. This display made it easy to double-check the data, spot incongruity and keep track of the scope of work. Later, to allow smooth analysis of the corpus in its entirety, all sheets were combined into one long list of 978 rows which make the full corpus of all entries under the letter N. Then, inspecting, filtering and categorizing the corpus in one complete context was made possible.

7.2.1 Categories

The 978-word corpus is listed in a single column, A, according to its alphabetical order of appearance. The following columns were used to denote categories which form the organizational scheme of the corpus:

- **Alt. or corr. (*) spelling**
  Cases of alternative spelling, denoting two types: the first owing to minor regional or stylistic differences, e.g., “narrow minded” versus “narrow-minded”; the latter owing to corrected spelling or terminology, marked with an asterisk (*) to cater to search queries that are logical, and yet incorrect. For example, users searching for the term denoting alternative medicine by using all-Greek components might search for *physopathy*, whereas the English term is “naturopathy” (the word “nature” originating PIE through Latin and French; “-pathy” originating from Greek through Latin).

- **Slang**
  This is a binary yes/no affiliation. If the entry is considered slang at the time of its assignment then it was tagged “y” in the slang column. Otherwise, left blank. Essentially this means that words that in previous times were slang, but by now are not considered so anymore, do not bear the tag “y” in the slang category. Rather, their origin as coming from slang is detailed in the category “route”.

- **Stem**
  When possible, a stem was noted for words with a shared origin. This mainly applies for different formations or parts of speech that are the conjugation of a single stem. For instance, “nouvelle” (n), “novel” (n), “novelist” (n), “novelize” (v), “novella” (n) and “novelty” (n) have a shared stem “novel”.

---

13 Available as Appendix 2.
For compound words (e.g., “nutcracker”) and for phrases (e.g., “ne plus ultra”), no stem was chosen, as that would entail further etymological research and reasoning, which are not the object of the current research. Similarly, in cases where a stem was not specifically noted by the creator, no attempt to research it further was done. Therefore, the category allows exclusion from it in the form of blanks.

- **POS (part of speech)**
  Signifying the lexical category of an entry, according to the following 25 available tags:

<table>
<thead>
<tr>
<th>POS tag</th>
<th>interpretation</th>
<th>Comments and examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Abbreviation</td>
<td>Abbreviation</td>
</tr>
<tr>
<td></td>
<td>N.B., originating from Latin. In cases where the abbreviation derives from English, the full form of the word or phrase was provided as “route” (e.g., “newsie”, from “newsboy”)</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>Acronym</td>
<td>NATO</td>
</tr>
<tr>
<td>c</td>
<td>Adj</td>
<td>“native”</td>
</tr>
<tr>
<td>d</td>
<td>Adj, Adv</td>
<td>“northerly”</td>
</tr>
<tr>
<td>e</td>
<td>Adj, N</td>
<td>“nudist”, “Norwegian”</td>
</tr>
<tr>
<td>f</td>
<td>Adv</td>
<td>“nearly”</td>
</tr>
<tr>
<td>g</td>
<td>Affix</td>
<td>“nano-“</td>
</tr>
<tr>
<td>h</td>
<td>Conjugation</td>
<td>“neither”, “nor”</td>
</tr>
<tr>
<td>i</td>
<td>Contraction</td>
<td>“nonetheless”</td>
</tr>
<tr>
<td>j</td>
<td>Interjection</td>
<td>“nertz”</td>
</tr>
<tr>
<td>k</td>
<td>Mathematical term</td>
<td>“nth”</td>
</tr>
<tr>
<td>l</td>
<td>N.1</td>
<td>noun with more than one meaning</td>
</tr>
<tr>
<td>m</td>
<td>N.2</td>
<td>noun with more than one meaning</td>
</tr>
<tr>
<td>n</td>
<td>Nursery talk</td>
<td>“night-night”</td>
</tr>
<tr>
<td>o</td>
<td>People’s names</td>
<td>A national group</td>
</tr>
<tr>
<td>p</td>
<td>Phrase</td>
<td>“ne plus ultra”</td>
</tr>
<tr>
<td>r</td>
<td>Place name</td>
<td>“Nebraska”, “north sea”</td>
</tr>
<tr>
<td>s</td>
<td>Preposition</td>
<td>“notwithstanding”</td>
</tr>
<tr>
<td>t</td>
<td>Proper name</td>
<td>“Neil”, “Nigel”</td>
</tr>
<tr>
<td>u</td>
<td>Reduplication</td>
<td>“no-no”</td>
</tr>
<tr>
<td>v</td>
<td>Trademark name</td>
<td>“Naugahyde”</td>
</tr>
<tr>
<td>w</td>
<td>V.1</td>
<td>verb with more than one meaning</td>
</tr>
<tr>
<td>x</td>
<td>V.2</td>
<td>verb with more than one meaning</td>
</tr>
<tr>
<td>y</td>
<td>V.1</td>
<td>Relating just to meaning number one. E.g., “net” relating to capturing with a net</td>
</tr>
<tr>
<td>z</td>
<td>V.2</td>
<td>Relating just to meaning number two. E.g., “net” relating to gaining a sum</td>
</tr>
</tbody>
</table>
The year notation relates to the earliest written evidence of the entry (though it may have been in use centuries preceding it).

(a) When the exact year was known, it was written in its four digits form (e.g., “1993” for “netiquette”)
(b) When the tracing was as accurate as a decade, the year was followed by “-’s” (e.g., “1980’s” for “nad”)
(c) When the tracing was as accurate as a century, the year was followed by a “c”, denoting century (e.g., “12c” for “nurse”)
(d) When the tracing estimates a century (or a specific year, in some cases), the year was preceded with a “ca”, denoting approximation (e.g., “ca 2000” for “noob”)
(e) When the tracing estimates a period, an indication of the part of century was noted (e.g., “mid 13 c” for “not”)

Route
Route indicates the route of entry into the English language. When there is more than one language as source, the chain of contact is detailed in sequential columns such that the left-most is the language which gave English the entry, and the right-most is the ultimate source language. For example, for the entry “naïve”, the route is French, Old French, Latin; meaning, English was introduced with the entry by French, which inherited it from Old French, which assimilated it from Latin, which makes the most ancient known source.

Table 2: Route, foreign sources exemplification

<table>
<thead>
<tr>
<th>ENTRY</th>
<th>MEANING 1</th>
<th>ROUTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>naive</td>
<td>adj</td>
<td>1650's</td>
</tr>
</tbody>
</table>

Under the category “route” other sources may be available, not necessarily foreign languages. For example, “imitative origin”, “nursery talk” and “child’s word” indicate categories of English word-change and emergence. Other available tags are specific words which are the known origin of an entry (e.g., “narcotics agent” for “narc” (slang)). In cases where the origin is uncertain or lacks consensus, tags applied were “unknown” and “disputed”, respectively.

Table 3: Route, English word change exemplification

<table>
<thead>
<tr>
<th>ENTRY</th>
<th>MEANING 1</th>
<th>ROUTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>nicker</td>
<td>v</td>
<td>1774</td>
</tr>
<tr>
<td>night-night</td>
<td>phrase</td>
<td>1896</td>
</tr>
<tr>
<td>nanny</td>
<td>n</td>
<td>1795</td>
</tr>
<tr>
<td>narc</td>
<td>n</td>
<td>1967</td>
</tr>
<tr>
<td>nibble</td>
<td>v</td>
<td>ca 1500</td>
</tr>
<tr>
<td>nome</td>
<td>place name</td>
<td>1898</td>
</tr>
</tbody>
</table>
When the same route was shared between entries pertaining to the same stem, the route was only detailed once for the stem word (e.g., “natural” (adj) coming from Old French, and ultimately Latin). In related entries, the route details start with the reference “see” and an indication to the stem origin (e.g., “natural” (N) detailed as “see natural” (adj); or “naturally”, detailed as “see natural (adj) + -ly”).

Table 4: route, reference exemplification

<table>
<thead>
<tr>
<th>ENTRY</th>
<th>ALT. or CORR. (*) SPELLING</th>
<th>SLANG</th>
<th>STEM</th>
<th>MEANING 1</th>
<th>ROUTE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>natural</td>
<td>natural</td>
<td></td>
<td>natural</td>
<td>adj</td>
<td>ca 1300</td>
</tr>
<tr>
<td>natural</td>
<td>n</td>
<td></td>
<td>natural</td>
<td>n</td>
<td>1925</td>
</tr>
<tr>
<td>naturally</td>
<td>adv</td>
<td></td>
<td>natural</td>
<td>adv</td>
<td>late 13 c</td>
</tr>
</tbody>
</table>

There is no required threshold to be crossed in order to justify a category; i.e., a single item can populate a category as a singleton set, as long as its source is validated. There is no need to spare on tag affiliation, and group items together unnecessarily.

- Meaning 2
In cases where a single entry has a changed or added meaning, the same categories were iterated, this time with different information to signify the change in meaning (to either POS, year or route).

For example, the entry “narcotic” is assigned POS “adj”, year c 1600 and route Middle French, Medieval Latin and Greek. In the category Meaning 2 it is assigned the route German, Medieval Latin and Greek, leaving all other cells unchanged. The difference signifies that both routes of entry into English are possible, though there is no contesting over the POS and year data.

Table 5: Meaning 2, disputed route of entry

<table>
<thead>
<tr>
<th>ENTRY</th>
<th>STEM</th>
<th>MEANING 1</th>
<th>MEANING 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>POS YEAR ROUTE 1</td>
<td>POS YEAR ROUTE 2</td>
</tr>
<tr>
<td>narcotic</td>
<td>narcotic</td>
<td>adj ca 1600 middle french</td>
<td>adj ca 1600 german</td>
</tr>
<tr>
<td></td>
<td></td>
<td>medieval latin greek</td>
<td>medieval latin greek</td>
</tr>
</tbody>
</table>

Another example is “nark”, where the entry is assigned POS “V”, year 1859, route Romany, Hindi and Sanskrit. The category Meaning 2 details POS “N” and year 1860, leaving the rest of the cells unchanged. This signifies no change to the postulated route of entry, but rather that the word has changed its meaning, entering English as a verb and later appearing as a noun.

Table 6: Meaning 2, change to POS

<table>
<thead>
<tr>
<th>ENTRY</th>
<th>STEM</th>
<th>MEANING 1</th>
<th>MEANING 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ROUTE 1</td>
<td>ROUTE 2</td>
</tr>
<tr>
<td>nark</td>
<td>v</td>
<td>1859 roman</td>
<td>1860 roman</td>
</tr>
<tr>
<td></td>
<td></td>
<td>hindi sanskrit</td>
<td>hindi sanskrit</td>
</tr>
</tbody>
</table>
As is in Meaning 2, where another speculated route of entry or change in meaning was recorded, the same entry has detailed information in the categories that indicate a difference or change.

For example, the entry “narc” presented in Table 3 above appears also in Meaning 3, where it is assigned to POS “adj”, and year 1958. Since it is already introduced as a noun in Meanings 1 and 2, the information entails a change to the POS adjective, and the year of recording that change. As all other cells in Meaning 3 are unchanged, it indicates no alternative route of entry.

Another example is in the entry “nationalism”, where all three meanings pertain to POS “N”, but in different years: 1844, 1836 and 1892. The difference between each assignment indicates a change to the meaning of the word throughout these three date indications.

It is worth noting that no alternative routes of entry were recorded under Meaning 3, rather only changes to the categories “POS” or “year”, the latter indicating a change in meaning through time.

It might also be worth mentioning that changes in meaning between the categories Meaning 1, Meaning 2 and 3 are not ordered chronologically, but rather in the same order in which they appear in Etymonline.

### 7.2.2 Tags

Within each of the categories (presented in the Excel sheet as columns), there are various tags. As detailed above, the tags are made up by either a binary value for “slang” (“y” or blank), a stem name, lexical categories (POS) or fixed terms (e.g., “proper name”), years and their approximations – according to specific patterns (i.e., decade, century, etc.), and languages participating in the entrance route into English.

Meaning, the scope of the corpus is characterized and defined by these properties, which are open for additions and change, and exhibit the connection between words to their language(s) of origin and influence (the latter, in cases of change within English, after assimilation).

At the head of each column (i.e., each category) there is a filter which allows narrowing down the presented data, according to interest within a given context. Hence, filtering the category “slang” by the tag “y” will remove all rows with a blank cell, resulting in a presentation of only slang words.

---

14 Total of 14 words, available as Appendix 3.
Within a given set of results, further filtering may be done. For instance, in the set of slang-word entries, filtering away all blank cells from the category Meaning 2 would yield a narrow set of three entries, where words exhibit a change in meaning and POS, re-entering the language in different years.15

The same process can be done in a different category, for instance “route”, where a filter “Latin” is chosen. In order to include all variations within Latin, the filtering was iterated and supplemented with the tags “Church Latin”, “Greek and Modern Latin”, “Late Latin”, “Latin + - [suffix]”, “Latinized [language name]”, “Medical Latin”, “Medieval Latin”, “Modern Latin” and “[prefix]- + Latin”.16

Further filtering in the category “route”, could, for instance, focus on the most ancient donor language, regardless of the ultimate language of contact that had transferred the word into English. Meaning, filtering the third link in the chain of “route” (since the fourth link is composed of all blanks) and choosing “PIE” as source would yield a narrow set of eight entries, and a limited variety of contact languages: “Church Latin”, “Latin”, “Medical Latin”, “Medieval Latin” and “Modern Latin”. Even more narrowing down and filtering may be carried out according to stems, POS, years of entry, contact languages and additional meanings, resulting in a more and more focused set of results with each query.17

The tags appearing in the analysis are those evoked by the inventory available. I.e., a greater corpus might have yielded more tags, or more items per tag.

If expanded to include more entries, tags could easily be added without a need to change the existing ones, as they already capture the existing corpus.

A cell left blank indicates no tag associated with it, and therefore no category affiliation.

This connection too may be added or expanded if necessary, to allow more information to surface and relate an entry to categories, and therefore to other entries.

7.3 Hypothesis and analysis

In order to illustrate the power of the suggested methodology, it was put to the test by generating a network of ties between words, and comparing it to the ties offered by the dictionary.18 The assumption is that the suggested model, pertaining to the principles of a DL (in systematic, transparent and reliable tagging) would enable a wider, more elaborate network of ties than that currently available in Etymonline, and reveal curious findings that might lead users to discovering more connections or etymological properties in the corpus. The reader is therefore reminded that the null hypothesis is:

\[ H_0 \quad \text{There would be no significant difference in the number of ties generated between words, in the existing and suggested systems of organizing the corpus} \]

To put the hypothesis to the test, a designated online site19 generated ten random values between the numbers 1 and 978 (the number of entries and their respective row number in the corpus), that should represent ten randomly chosen words for comparison. The numbers picked were: 15, 129, 172, 365, 431, 186, 81, 684, 798 and 932. These numbers correspond to the following entries, respectively: Nadir (n), Nationalism (n), Naval (adj.), Nessie (proper name), New Zealand (country name), Nicker (v), Narcomania (n), Non-concencual (adj.), Nostalgia (n) and Numismatics (n).

15 Also available in Appendix 3.
16 Total of 133 words, available as Appendix 4.
17 Also available in Appendix 4.
18 The complete comparison and its analysis are available in Appendix 9.
Only ten values were selected for the analysis (~1%) as the analysis served only to clarify the added features of the methodology, and the potential usage of it. As Etymonline does not offer any browsing options at all, nor does it detail a network of ties or their justification, its added features’ contribution amount to near non-existent. Contrastively, the added features introduces in the current thesis optimize searching and browsing the corpus through various paths, in addition to the default alphabetic list.

I.e., any searching and browsing available in the DL structure should be better than that of the currently available list formation. Hypothetically, no comparison needs to be made in order to prove that a DL is preferable to a database or list, in the options of searching and browsing that it entertains (given that it still, by default, allows the exact same list formation as its competitor database).

Nonetheless, in order to quantify and contrast the results yielded by each model, ten words were chosen to exemplify the strength of the suggested methodology, advocating for a DL construction. The exact procedure of analysis is detailed in the next segment §7.3.1, its results in Appendix 9.

### 7.3.1 Etymonline related ties

Each word was inspected separately in its Etymonline value. I.e., an entry was clicked and viewed in full detail (as opposed to a list view which might obscure data). Each respective document listed related terms to the entry, and those were collected and counted. I.e., how many related terms were suggested in Etymonline per entry.

Additionally, the text describing each entry was carefully read, and in cases where terms within the text were linked to other entries, those were collected as well. (Meaning, terms that serve as hyperlinks one can click and explore further for more information). See figure 8 below:

*Figure 8: the entry “Nostalgia (n)” and its related terms, as it appears in Etymonline (retrieved 2017–11–12)*
In cases where links were available but were not related to the entry under inspection, they were not included in the analysis. For example, a case where an entry was negated with its antonym, and the antonym gave rise to more links and ties that characterize only itself.

### 7.3.2 Suggested model related ties

The same ten words were looked up in the Excel sheet which makes the corpus of the thesis, and copied together with all the information that they encompass. I.e., an entire row was copied, detailing the tags of each category that characterizes the entry. Then, in relevant tags (non-blanks), further searching was carried out, to see if other entries bearing the same tag could be associated with the chosen entry in a meaningful way (meaningful in the sense that a relation can point out to interesting research questions, a shared way of coming into formation, or anything beyond the mere coincidence of the words pertaining to the same lexical class, for instance).

Hence, for each entry a small table was constructed, containing three columns: category name, tags within that category, and number of entries which share the same property. See table 9 below:

<table>
<thead>
<tr>
<th>Category</th>
<th>Tags within category</th>
<th>Number of entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Year</td>
<td>1647</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>17th century</td>
<td>0</td>
</tr>
<tr>
<td>Route</td>
<td>Dutch</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Dutch and place name (both in ca 1600)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>POS and “new”</td>
<td>5</td>
</tr>
<tr>
<td>Slang</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Meaning 2</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Meaning 3</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

| Thesis number of related entries: 8 |

In the above example, it shows that for the category “stem” there is no available tag. In the category “year” there is a tag 1647, which is not shared with any other entry in the corpus. In order to try and broaden the search, it was iterated with the tag 17th century (which includes the year 1647). That search yielded 113 hits with a shared tag.

However, these 113 entries ended up broadening the search too much to be able to make sense of, or find a meaningful relationship. Rather, it is quite likely that the entries included in the hits have no relation to the search term (“New Zealand”) and are merely the result of coincidence. Therefore, that row had been grayed out and excluded from the analysis.

In the category “route”, the tag Dutch was searched for and showed two other hits with the same origin. Of those two, one showed another relation, sharing both the tag for place name as well as source-language.

Another curious finding was that of the shared properties POS (in this case, place name) and the prefix “new”. Of the 31 other entries that are tagged with place name, five had a name that started with “New”. This information could elicit further research as to the founding of sites, naming-trends, influential sources of power throughout history, and more.

In the following categories, slang, Meaning 2 and Meaning 3, no search was carried out as the entry does not bear these respective tags.
7.3.3 Comparison

After having analyzed the related terms and their relationship with the inspected entries, the ties were compared and counted between those of Etymonline and those of the suggested model.

Since the current thesis has inspected only entries under the letter N, the scope of the related ties it offers is much more limited. Hence, the comparison restricted Etymonline in the same way, taking into consideration only those ties that start with the letter N.

The comparison therefore shows how many ties Etymonline provided in total, how many of those start with the letter N, and how many were shared with the ties found by the current suggested methodology. The latter was a means to inspect and verify that the suggested methodology did not miss any ties that it should have been able to spot, if those were within the scope of the corpus.

An example for the comparison of the entry “New Zealand” is provided in table 10 below.

| Etymonline ties: | 0 |
| Thesis ties: | 8 |
| Etymonline ties starting with N | 0 |
| Shared ties: | 0 |

As can be seen, while Etymonline spotted no ties worthy of mentioning for “New Zealand”, the suggested methodology spotted eight entries that share a route of entering into English with that could be interesting for users. The comparison therefore yields the score 8:0 in favor of the suggested model.

7.4 Reservations and flaws

7.4.1 Reservations

(a) Correctness of contents

Since the current thesis is solely geared towards advocating for a categorization system, no attention was given to the correctness of contents. Claims are not refuted or strengthened with the tagging and categorizing, but merely represented in a different way, signifying the same meaning as is currently given on the site.

(b) Stemming

Stems were noted in cases where several conjugated entries were available and connected with a single entry that reflects their origin (see example “novel” in §7.2.1 - Stem).

However, the base entry that exhibits the root’s history does not necessarily make for the actual stem of the word. (In this case, “novel” (adj) comes from the latin “novelus”, which is in fact a diminutive of “new”, the root of which is actually PIE “*newo-“). The further one explores word histories, the further from English one gets in revealing the roots of words, as they are believed to be in the source languages (namely PIE, Latin and Greek).

Therefore, what is tagged as “stem” herein, does not necessarily reflect entries’ true etymological root. Nonetheless, it does clarify to the user word connections and allows their grouping according to semantic derivation.

It also serves to illustrate how words can be categorized and tagged according to etymological root, and as such it demonstrates the advantage of the suggested model.

In cases of compound words the need to choose a single root arose. Alternatively, a new tag could have been created to allow two roots for an entry which is comprised of two semantic entities (or more, accordingly).
Nonetheless, due to the scope of the corpus, this was not pursued further. Creating a new tag would have entailed the need to go over all the entries for a third time, which was beyond the framework of the thesis. Given that the point of pursuing lexical roots was made, a single tag to denote stem was deemed sufficient.

It would also be worth mentioning that affixes (neo-, non-, etc.) are not considered stems and therefore words that are comprised of two semantic units where one of which (or more) is an affix, are not considered as compound words.

(c) Affixes

As affixes are not considered stems, they were not labeled. However, ignoring the syntactic role of affixes led to the difficulty of not being able to group together words with shared suffixes.

Had there been a tag dedicated to suffixes (e.g., -phobia, -ism), the entries containing such lexical units could have been filtered to possibly show a certain trend (e.g., pertaining to words of a specific field, sharing a route of entry into English, sharing a historic period of forming, etc).

That would also have enabled searching within a certain tag, to constrict the scope of hits and possibly find interesting information. For example, in the category POS, filtering for the tag “adj.” and then adding a tag for suffix (e.g., “-al”, as in nautical, normal, naval; “-ic”, as in negritic, Neolithic, nictonic; etc.).

Again, as achieving this would have entailed going over the entire corpus for the third time, and as the point of grouping according to a set criteria is made via other tags, this addition was not executed.

(d) More categories

Another property that might have enabled strengthening ties is having more category names that could relate to PIE root (an ancient and speculated root), semantic meaning, translation (in case of loan words or phrases), antonyms, synonyms, and more.

Having included such information would have enriched the semantic depth reached in the corpus, and would have allowed making more ties that are anchored in meaning, rather than form (i.e., POS, spelling).

(e) Number of alternative routes and changed meanings

Only three possible routes of entry or changes in meanings were noted (by the categories Meaning 1, 2 and 3). This too, is due to the scope of work required to cover all the data. In essence, this means that entries which exhibit more data were analyzed as far as the third alternative provided in Etymonline. The three alternatives were chosen according to their order of appearance on the site, which does not necessarily correspond to their chronological order.

(f) Typos

The analysis conducted in the thesis may still contain typos or other small mistakes. Inconsistencies, if there are such, are due to the sheer scope of work, and should not be regarded as disturbing the validity of the claims made herein.

---

20 As it appears in Appendix 2
7.4.2 Flaws

(g) Absence of nesting features
As stated earlier (section §7.2 above), each column in the Excel chart represents a category. However, there was a need to tie several categories together (namely POS, year, route) under an overarching category (namely Meaning 1, 2 and 3) in order to differentiate between them as characterizing differences within a single given entry. Although visually they appear as if they were nested under the overarching category, in fact they are not, since there is no searching function that can utilize this feature and bind the categories as a single cohesive unit.
Hence, the row indicating meaning should only be regarded as a visual aid.

(h) Timeline continuity
Since the tags are literal (as opposed to symbols), they appear in each category according to their alphabetical order. This means that their order does not represent a chronological order nor a nested construction.
Therefore, if one wishes to view only the set pertaining to a chronological period earlier than 15c, one must choose to include all the tags (5c bce, 9c, 12c, 13c and 14c), their approximations (e.g., ca 12 c, mid 13 c, late 14 c, etc.) as well as individual years contained in the restricted time frame (namely, 1066 and 1296) in order to achieve the desired presentation.
Since tags are independent, the system does not automatically deduce the nested representation of periods, where, for example, the 12th century is included in the period of “up to the 15th century”.

(i) Inclusive / exclusive sets
Similarly to the previous difficulty, the inclusion of two languages in a single tag is not further translated by the system as the two independent languages. For example, a tag “Yiddish and Polish” is not interpreted as both “Yiddish” and “Polish”, but rather a single cohesive and unique unit.
In effect, this means that searching for all entries that include Yiddish requires the adding of two tags, one for “Yiddish” and one for “Yiddish and Polish”.

(j) Befitting program for analysis and representation
Due to lack of advanced technical skills, Excel was chosen to be the program in which to conduct the analysis as well as representation of the thesis and its results.
Ideally, a better graphical representation (of clusters or network, like in VOSviewer, for example), would have illustrated the model’s benefits better.
Also, a different program might have allowed nesting categories such that the difficulties discussed above could have been smoothly resolved.
Given the relatively limited scope of the corpus, the manual assigning of multiple tags when filtering (for either browsing or searching purposes) does not pose a hindrance. Had the corpus been larger or more varied, however, a different program would have had to have been used.

(k) Space limitations
In the space allotted for tags’ description, brevity is a virtue (Yannakoudakis, 1989). Hence, complex notions that could not be expressed in the form of a simple and concise tag may not be allowed. In the future, if more tags, and more complex tags are created, there might need to be a revision to all tags in turning them into a code language.
Differentiating between disputed and new meanings
Since all cells are filled in all columns depicting Meaning 1 and Meaning 2, it becomes difficult to discern which row shows a change in meaning and which shows disputed origins.
Initially, the methodology was such that a property which is undisputed should remain blank in Meaning 2; i.e., only changes (of any sort) get transferred into the category. However, in practice this meant that within the category it becomes impossible to filter further, as the majority of the fields are blank.
Therefore, all relevant information was restored, and when inspecting a change one must be meticulous so as not to miss any small differences.

Chapter 8
Results
8.1 Testing the methodology in producing sets of interest
In addition to testing and analyzing the ten randomly chosen words (introduced above in section §7.3), filtering according to various points of interest was done in order to explore the model and exhibit the data mining potential that it offers. Various searching and browsing queries were tested, for each of which a filter was applied, resulting in set of data that was then imported to a separate Excel sheet and saved under the name of the point of interest it represents.
For each hit yielded, all information regarding that hit was retrieved and imported together with the filtered results. For example, if setting up a filter in the category POS (e.g., for a query regarding the tag Adj.), other data such as Route 2 were retrieved and presented alongside it. All data were left in the results sheets, as they might provide other curious insights or point at other directions that should be inspected further.
In the following sections, some of the various tested filters will be introduced together with their resulting sets and interpretation.

(a) Category: “Route 1”; tag: “German”
This tag groups together all words that have entered English via German as the immediate contact language.
The set includes 17 members. The filtering was done according to the contact language, i.e., the source from the perspective of the English language. However, the various members of the set show greater diversity in entries’ more ancient sources (e.g., Greek, Latin, Middle High German, and more).
This function shows that with a simple filter of just one tag (“German”) it becomes evident that only seven out of the 17 entries can be traced back to pure Germanic roots, whereas the rest are derived of other language lineages, or of unknown source. Namely, this finding could support the claim that the basic word stock of English is Germanic, and yet the majority of the lexis (and up to ¾ of it) is non-Germanic (Gray, 1986).
The set of results is available as Appendix 5.

(b) Category: “Slang”; tag: “y”
This tag groups together all words that are derived of slang.
The set includes 14 members, of which eight are nouns, and ten are less than a century old. This set allows the characterization of slang derivation; i.e., how long into the assimilation process words take, before their status as slang is removed, and they are perceived as an integral part of the normative lexicon.
Also, it allows the inspection of what POS is more prone to becoming slang. The yielded distribution shows eight nouns, two verbs, one adjective and one adverb, one abbreviation and one proper name. I.e., almost 60% are nouns. Furthermore, three entries exhibit other meanings across time, and in one entry even the repeated change of POS (“nerf”; from verb, to trademark name, to adjective). The set of results is available as Appendix 3.

**c) Category: “Meaning 1”; tag: “Latin”**

As introduced above (section §7.4.2), in order to encompass all of the entries that are derived of Latin, the search needed to be carried out in several steps, each focused on another characterization of Latin.

All entries that include a variety of Latin as their contact language (i.e., the most recent link in the contact chain) were therefore copied into a new sheet which, at the end of the process, included all varieties of Latin.

The set includes 133 members ranging from c 1200 to 1966, which is very interesting considering that Latin had been a dead language for many centuries back.

These results might support the claim that Latin was, and continues to be one of the most influential languages in English, in all POS and periods.

Also, this filtering also shows the disadvantage of working with Excel as a search engine, as the initial filtering for contact language needed to be iterated for all varieties of Latin.

Additionally, if wishing to deepen the search further, the same process could be carried out also for the second and third links in the chain of contact, where Latin might have influenced English via an indirect channel of contact. For example: “natural” entering English from Old French, but originating in Latin.

The set of results is available as Appendix 4.

**d) Category: “Meaning 1”; tag: “Latin”, “PIE”**

In continuation to the previous search, a second tag “PIE” was added to the filtering. Meaning, of the set of 133 members pertaining to Latin, those that also included the tag “PIE” as the next link in the chain were retrieved, and resulted in a set of 25 entries, from the 14c and to 1905.

An iterated filtering was done, positioning PIE as the third link in the chain. This search enables any other language to stand in the second place, linking between PIE and Latin.

This search yielded seven entries, owing to one of the following routes:

- PIE → old Latin → Latin
- PIE → Greek → Latin
- PIE → Greek → Medical Latin
- PIE → Greek → Church Latin
- PIE → Latin → Medieval Latin
- PIE → Latin → Modern Latin

The set of results is available in Appendix 4.

**e) Category: “Meaning 1”; tag: “4th link non-blanks”**

This tag does not have a specific name, as it allows any language that stands in the place of the 4th link to occupy it. Meaning, whichever language is the most ancient source-language traced, in cases where a single entry has four different influential language that directed its entering into English.

This set includes 17 members, the most ancient origins of which are PIE (11), Latin (3), Old Latin (1), Proto-Germanic (1) and Greek (1).

This distribution shows that the greater contributor of words into various other European languages is PIE. This is not surprising, as PIE is assumed to be the source of all
European languages and their smallest shared denominator, according to etymological research (Gray, 1986).
The set of results is available as Appendix 6.

(f) Category: “Meaning 2”; tag “1st link” non-blanks”
When removing all blanks from “Route 2”, the retrieved set of results focuses on entries which are argued by different researchers to stem from different donor languages. I.e., disregarding the tags “POS” and “year”, the search allows filtering only the entries for which there is a dispute over their origin, rather than a reflected change in meaning throughout time. The set includes 26 members.
When comparing between the claimed Route 1 on the left-hand side and Route 2 on the right hand side, small differences are noticeable between the argued routes of entry:
- year of entering (e.g., “nebula” as either entering in the early 15c or at 1660);
- length of the chain of contact (e.g., “negatory” as either derived of Middle French, or of Latin and Medieval Latin);
- reference (e.g., “nudity” as either “see Nude (adj) + -ity” or Latin and Late Latin)
This function shows how the suggested tags and filtering allow for more than one alternative route to be exhibited, and enables a comparison between two suggested routes to further investigate the source of controversy.
The set of results is available as Appendix 7.

(g) Category: “Meaning 1”; tag: “1500’s”
This search too needed to be carried out in several steps, each of which pertaining to a different characterization of the 1500’s (e.g., a specific year, a decade, etc.). The set includes 104 members.
A second filtering in Meaning 2 to eliminate all members of changed meaning (by removing all entries with a specification of a POS and year, also in the category Meaning 3) yields all the words that did not change their meaning from the time of entering the English language and to present days. This set includes 63 members.
Namely, this filtering reveals a set of 63 words that have remained stable in more than 500 years of usage (or longer), since their entering into the English language.
The set of results is available as Appendix 8.

8.2 Testing the DL model in comparison to Etymonline
Upon analyzing and comparing the ten randomly chosen words between the related ties offered in Etymonline and the related ties found by the suggested methodology, it was found that a DL model yields a significantly greater number of ties between entries. I.e., a single entry shows a broader network of ties to other entries in the corpus. The results and their interpretation are provided in tables 11 – 12 below:
When comparing the two right-most columns, “Etymonline ties starting with N” and “Shared ties between Etymonline and thesis”, it becomes apparent that only in rows 3, 6, and 9 was there a related entry that started with N, and yet was not picked up by the suggested model as relevant. The detailed explanation for each occasion is available in Appendix 9.

Table 12: an analysis of the results

<table>
<thead>
<tr>
<th>Thesis ties</th>
<th>Etymonline ties starting with N</th>
</tr>
</thead>
<tbody>
<tr>
<td>average</td>
<td>15.3</td>
</tr>
<tr>
<td>t-test</td>
<td>0.00055777</td>
</tr>
</tbody>
</table>

Additionally, the results show that the suggested methodology yielded an average of 15.3 hits for related entries to a given term (per search), whereas Etymonline yielded, on average, 1 hit for a related term. In order to test if this difference is significant, under the scope of the 978 entries of the corpus, a one-tailed t-test was conducted.

A one-tailed dependent samples t-test was chosen as it measures the difference between the two distributions (number of related ties retrieved in the DL structure, and number of related ties retrieved in the database structure) in one direction of interest. Meaning, the only assumed or permissible relationship is that where the value of the results of the DL is higher than that of the database. The only question is – higher to what extent; significant or insignificant.

As mentioned in Chapter 6, Etymonline does not offer any browsing options at all, nor does it detail a comprehensive network of ties or their justification. Therefore, any contribution by the added features of a DL’s structure would constitute an improvement in hits’ quantity upon searching for ties in the corpus. Hence, the direction of the relationship is known, and a one-tailed t-test would reflect its strength.

A dependent samples t-test was chosen since the random selection of words was the same across both test conditions (the analysis proposed by the thesis’ methodology and Etymonline methodology).

The test’s score of 0.00055777 is much smaller than the commonly used parameter of testing significance (p < 0.05). Hence, it falls safely within the area of significant results, indicating that the number of ties generated by the DL model is significantly greater than that of the database list even when measured with a relatively small sample size. The same can be seen in figure 9 below:
The null hypothesis\textsuperscript{21} is therefore refuted, as the suggested methodology enables significantly greater searching and browsing options, ultimately allowing higher retrieval rates of related terms per query.

**Chapter 9**

**Discussion**

By categorizing the Etymonline corpus, different angles of analyzing its content may surface. Such angles may include relationships between words or stems, trends affected by geography, time periods or historic events, changes in words and their POS across time, and other processes of word assimilation and change in the English language.

Unlike other dictionaries, Etymonline does not take Latin to be the starting point of word origin; rather, it acknowledges Latin as a link in a chain of influence over the lexicon (as well as structure) of English (Gray, 1986).

This recognition allows the researching of word-histories in several steps across an etymological trace line, and segmenting the influence that each step has imprinted upon English. I.e., if a word is shown to have entered English originally from PIE, through Latin and German, each link in that chain of contact plays a certain role that can be related to separately, and not equivalent to a word that entered the language directly and solely from German, for instance.

These different layers of impact can be grouped together or analyzed in a layered manner that recognizes their respective contribution. Viewing each layer of contributing language separately might enable researchers to deduce rules about word creation or change, locate patterns, trace the crystallization of spelling rules and their exceptions, and many more linguistic features and considerations.

The following discussion will inspect the contribution of structuring Etymonline as a DL instead of a database, according to different aspects, and offer insights regarding the contribution of the current thesis to the study of DLs, data mining and knowledge organization.

\textsuperscript{21} “There would be no significant difference in the number of ties generated between words, in existing and the suggested systems of organizing the Etymonline corpus”.
9.1 Context

Changing the linguistic context by which to appraise a word or a phenomenon could spike new inspiration or reveal new connections that have not yet been considered, due to their obscurity. The suggested model has shown that the corpus of 978 entries may be presented in various contexts and constellations, revealing and masking data according to foci of interest. The same data may be grouped together and filtered through according to a specific query, retrieving a limited set which is easily analyzed, revealing interesting facts about the properties of the entries it includes.

For instance, in the query of slang, the filter retrieved 14 members that could not have been otherwise collected, if needing to sort through the entire Etymonline corpus according to its alphabetical order. The number of entities in the retrieved set in comparison to the whole of the thesis-corpus (14 out of 978 entries) also says something about the scope of slang in contemporary English, and the speed of its creation, assimilation and adoption into the normative lexicon. Inspecting the age of slang words, their origin, their common attributes, POS, environment of usage (e.g., linguistic register, pragmatic context, semantic gradations, etc.) and more, within the context of other slang words, gives a different angle than if considering each individual word. Also, the filtering allows the immediate retrieval of all words that are tagged as slang, without having to have a specific word in mind initially.

Other criteria of searching allow similar advantages (provided in the results in Chapter 8) that enable further in-depth analyzing, focused browsing, and the surfacing of relevant hits in relation to a specific query. Within the given set of results, further filtering can be done to hide superfluous information and reveal points of interest.

I.e., simply changing the context of the object of research may show it in a completely different light and provide new paths of research that have yet been considered. As one of the main objectives of classifying is facilitating access to our past and heritage (Bowker & Star, 2000) – the suggested model does just that, increasing the accessibility of the information treasured in Etymonline.

“A word is characterized by the company it keeps” (Stavrianou et al., 2007, p. 27); following this claim, words can be characterized by various means, by changing their context of appearance. By assigning different categories and tags for the researching of words, one may discover new angles (i.e., facets) for the appraisal of the word of interest and its role in the lexicon.

9.2 Contextual needs

Not only does the context of appearance cause fluctuations in the valuation of an entry, but also the contextual needs of users. As the use of a document (or in this case, an entry) is pragmatic, certain pieces of information may be deemed more or less relevant in different occasions (Hjørland, 1992).

The purpose of the categories and tags is to allow as much variability within the construct of uniformity of data categorization, in order to allow different search criteria to be applied on the corpus (Ranganathan, 1989). The chosen array of categories and tags attempts to answer to potential needs and expectations of potential users, and their research foci. However, since this work had been executed alone, it may not sufficiently represent potential user needs.

According to Ranganathan (1989), classification of a database should ideally incorporate the following:
- be conducted by a team of classifiers, reflecting as much as possible of the given field of knowledge (in this case, linguistics and etymology);
- continuously monitor and renew the collection for changes and developments;
- include a team of both technical orientation (the classifiers) and intellectual orientation (linguists);
- keep constant contact with the community of users to gain further insights for the development of the library;
- and engage in research for the investigation and unveiling of new fields of interest, new categories and new data to consider.

Given the above, it might be argued that the suggested categories are insufficient in catering to the entirety of the aimed user population, as they represent only one world view. The more participants in the project – the more diversity would be allowed to surface and more (potential) user needs can be answered. I.e., greater variety of categories, more tags per category, continued surveying of the database and continued improvement in its accessing, appearance and relevance, among other advantages.

Unfortunately, that is a kind of development which is beyond the scope of the current thesis.

9.3 Metadata
This changing of context is enabled by the teasing apart of metadata from corpus. Currently, Etymonline is organized such that each entry individually encompasses all of the data known about its roots and route of entry, as well as links to related words (and especially when they share a stem or an affix).

However, if metadata is organized separately from entries, it can be labeled and structured, allowing queries to be phrased in more than one way and not subordinate to the alphabetical order of the current presentation (Weinberger, 2007).

When the metadata is coded according to tags, changes, updates and re-arrangements, statistical information and more actions may be executed with ease and without having to reconstruct the entire corpus. Additionally, natural data clusters may emerge by manipulating metadata codes, instead of manipulating corpus entries (Yannakoudakis, 1989).

Meaning, the process of coding need only be carried out once; in the current case it is the constructing of categories and tags, and distributing the corpus in accordance with those. After it is completed, changes may be easily applied over the whole corpus in a matter of several keystrokes.

Similarly to how a structured text is better than an unstructured text (Stavrianou et al., 2007), so is a structured dataset better than “sack of words” collections, albeit arranged alphabetically. The benefit of presenting entries according to different organizing principles rises above the advantages of having no structure at all, save the randomness of the alphabetic order (Weinberger, 2007).

Metadata enables the effective information management that prevents duplications, miscategorization and loss of information (Vasconcelos, 2008).

9.4 From database to DL
Metadata, organization and cataloguing, and adhering to specific aims and services, is what makes a library, as opposed to a mere database or platform. Additionally, the methodological way by which to access materials, find and retrieve information is of the pillars of a library, in contrast to a database (Witten et al., 2010).

Referring back to the DELOS manifesto (introduced in section §2.1), a DL is an organization with its own system of maintaining digital content, which is of quality and restricted to polices. The data available in Etymonline is already of quality, thanks to the efforts of its creator and his standards of scholarly references. Additionally, unlike a database, Etymonline strives to reach as many users as possible, not limited to a restricted group of users (as could be the case in portals, for example) (Calhoun, 2014).

However, what is missing in Etymonline, preventing it from being a DL, is structure:
The system should be structured by a reference model - i.e., a hierarchical relation between its different functions, restricted to internal needs;

Data should be indexed, its metadata outlining the scope of the collection and the paths to reach its content;

Available services are custom-made to reach a pre-defined goal and a specific section of target users (Candela et al., 2007).

As was shown, the suggested model adheres to the criteria of constituting a library; by categorizing and tagging Etymonline, it can be transformed into a DL. With the change of organizing principles, the face of Etymonline is changed, making it more searchable, and therefore have increased usability (Dobreva et al., 2012). With the new organizing principle, the library should hopefully attract more users and researchers, as well as gain more contributions (volunteer work, ideational or monetary support).

Currently, and unlike a traditional library, the Etymonline does not exhibit a certain shelf arrangement. However, like other DLs, the suggested grouping and filtering options introduced make for a virtual shelf arrangement that is not constant. I.e., viewing an item in a context of other similar items, the factors determining similarity being flexible and adjustable from user to user and from query to query.

Another factor shared with other DLs is that Etymonline has an administrator who directs the decisions and future of the library (Calhoun, 2014). Unlike, e.g., Wikipedia and Flickr, although user contribution is appreciated, it does not rely on users for the construction of the database, rather the final word remains in the hands of the initiator and guiding authority.

The type of organization affects the production of new knowledge. In a DL, the whole is bigger than the sum of its parts; the DL offers a collective memory and the promotion of individual learning, amongst other things (Vasconcelos, 2008). With the enhanced features suggested in the thesis, Etymonline can enhance learning, evoke curiosity and gain loyalty from its users, as well as spread to a greater community than the current. According to Chen (1999), exactly those are the challenges of a DL. Hence, the change offered herein would promote the database which is today Etymonline into constituting a DL.

9.5 Original titles for categories and tags

It was decided to create custom categories and tags for the dataset and not just rely on existing familiar ones (e.g., POS, slang) in order to best fit the corpus and its respective entries’ distribution (Kumbhar, 2011). Opting for existing categories and tags adopted from another system, discipline or project might have ultimately led to a loss of information.

In the future, new tags would be created or existing ones could be adjusted in order to fit new findings and new entries. The work of tagging is that of constant revision (Kumbhar, 2011). If more of the Etymonline corpus had been covered under the framework of the thesis (beyond the 978 entries analyzed), more categories and tags might have been generated.

Creating a more diverse classification system would enable more ways in which to view the material and more groupings to form, revealing new facts about the existing entries (thus enhancing data mining).

Some future tags to consider could be: occupations, body parts (as was mentioned in Gray, 1986), semantic meanings, phonological attributes, affixes, and more. However, due to the limitations imposed by the scope of the thesis, these were not executed herein.
9.6 Why tagging

Of all the available organizing methods (e.g., hierarchy, nesting, etc.), tagging was chosen as optimal due to its enabled flexibility. A tagging system allows the immediate and successful incorporation of new entries and new categories, or the deletion or change of existing ones, without the cascading restructuring of the entire categorization system (Weinberger, 2007).

This property of tagging is especially attractive when considering the field which is at the heart of analysis – Etymology; a field of knowledge and language, that by definition is subject to change and development (Weinberger, 2007).

Secondly, as there is no single objective “best” way in which to classify the data of the DL (Spärck Jones, 1970), tagging allows users to choose and adjust their preferred presentation, according to ad-hoc needs.

Further, it could be argued that a good categorization system is that which causes less disruption to the dataset (Spärck Jones, 1970), not forcing it into specific patterns that might be, for instance, drawn from other scientific fields, or may suit a certain scope, failing to relate to developments and additions. Tagging allows that fluidity, adjusting tags’ names and variety to the existing inventory.

Unlike the Penn Tags Project (introduced in section §3.4.6), the suggested tags in the current thesis appear as whole words or phrases, and not as symbols. Even tough “symbols are easier and quicker to perceive” (Weinberger, 2007, p. 151), with the current array of tags (i.e., not a variety big enough to elicit confusion), their names do not pose a particular challenge. If anything, it is quite possible that abbreviations or symbols in their stead would be cause for confusion, since the titles as they are now are quite intuitive and concise.

A future development to the tags could be that of nesting, which would group together entailed meanings of tag usage. Meaning, searching for entries appearing in the 14th century would automatically include the tags “early 14 c”, “late 14 c” “c 14”, etc., easing the filtering-work of the user.

9.7 Results’ representation

Presentation, however, is not the strong suit of the suggested model, as it is constructed in Excel, which is primarily concerned with calculating and formulating figures.

For technical reasons, Excel was chosen as the most appropriate program in which to gather and manipulate data. I.e., lack of familiarity with a better suited program, and/or incongruence between properties of existing programs (e.g., WEKA, VOS, vectorial representation, and more) and the data-set.

However, had there been a more graphically-oriented program, the presentation of results might have revealed even more interesting facts about the data they encompass, or aided in interpreting results and highlighting trends. Such graphic features might include, for example:

- Map representation:
  Viewing results in a map that outlines words’ travel through space and time, or their spreading through the migration of peoples.
  There can be use of colors to denote usage across a certain time span, indicating years or centuries.
  The graphical representation allows users to actually and physically see how words travel, spread and influence the language and its speakers, or are influenced by them, in a framework that is easier to envision and grasp.
  A map representation may also clarify certain contacts between languages, in cases where wars or occupation of new territories inflicted a sudden change in the lexicon or structure of a language, changed its characteristics, status and scope (i.e., prestige and linearity or spreading of usage).
- **Network representation**
  Given that the categorization aims to outline connections between entries, network representation might provide an illuminating representation depicting connections of various levels: between different languages of contact, specific entries or general routes of change.
  Such a network representation may focus on a “family” of words derived of a shared stem, or alternatively, a single term (positioned at the center of the network), where each branching out stands for a postulated route of entry, in cases where more than one alternative is available.
  By zooming in, the user could inspect each suggested route and the differences between them, as well as overlapping that they might share (in origin, time of entering, POS, etc.).

  However, for the time being, these graphical additions are merely hypothetical, and users must rely strictly on their own interpretive skills by analyzing the columns and rows of Excel. In order to facilitate the analysis procedure, results may be imported into new sheets where they can be filtered further and compared. In this case, each new sheet represents the “zooming in” where a query is inspected in the micro level.

  It is possible that the graphic representation is more practical and appealing for laymen than it is for professionals, who can filter through results just as well with the available tools of Excel.

### 9.8 Community or experts’ control

One way of broadening the extent and rate of development of the DL would be to turn it into a community-based project, similarly to Project Gutenberg, Wikipedia or Flickr. In these three unique platforms (introduced in section §2.2) the general public, expert as well as enthusiasts or laymen, is invited to share and contribute to the welfare of the entire community. Allowing contributions of expert, their judgment regarding users’ suggestions and their confirmation or discrediting of a source, would turn the DL into a professionalized and exclusive institution.

However, in the eyes of many, experts’ opinion might be helpful, but not necessary for the successful development and maintaining of such specialized communities, institutions or DLs (Weinberger, 2007). Another point in support of that is that the original Etymonline site had been created by a layman22, and is nonetheless one of the best known and highly appreciated online etymology resources to date23. Therefore, if the method is applied, there is room to broaden it by adhering to users’ suggestions.

Judging by the immense importance and contribution of Project Gutenberg, Wikipedia and Flickr, it seems that Etymonline has great potential to also becoming a leading DL with a broad public of supporters and contributors, including affecting online research and the scientific community.

### 9.9 Value of Etymonline

It is difficult to valuate that which is not tangible (Currall & McKinney, 2006). Etymonline is not only a digital service, relying on technology that needs to be maintained and updated, but it also represents a mass of information, and information is not without worth (Currall & McKinney, 2006).

---


Furthermore, that worth does not remain constant over time, rather changes and fluctuates according to societal changes (Currall & McKinney, 2006).

However, it is society that wishes to preserve and exhibit its knowledge and characteristics, and pass them along to future generations, as well as to spread them in present-day time amongst different cultures (Hedstrom, 1997). Therefore, the developing of Etymonline is of great importance and contribution to the whole of society – research communities as well as private sectors, English speakers and others.

Hopefully, if the suggestions of the thesis are applied, they will cater to this need of preserving, spreading and improving the heritage of English etymology. The scope of the contribution could be even greater if applied to other resources pertaining to other languages or other aspects in linguistics.

**Chapter 10**

**Conclusion**

The current thesis inspected an online database of etymology, Etymonline.com, and attributed it with the potential of turning into a DL if implemented with a change in its construction. The thesis pointed at the differences between DLs and other online sources of knowledge, and the connection (and often reasons for confusion) between the two.

It was argued that by segmenting the metadata from the listing of entries, the online dictionary can be turned into an online DL, enabling improved searching, browsing and managing of data. By keeping metadata separately, the information held in the database could be catalogued and characterized, easily changed, manipulated or enhanced, and thereby serve the science community to a greater degree, as well as layman and language enthusiasts.

First, the power of the suggested model was shown by creating example information need and browsing scenarios (section §8.1). Secondly, a comparison was made between the number of retrieved ties (hits) provided by each model --the currently existing Etymonline database versus the suggested DL--, to show which of the two reveals a more comprehensive network of relations between entries (section §8.2).

A one-tailed t-test showed that the number of ties retrieved per entry in the suggested model was significantly higher than that provided by Etymonline (p < 0.05), thereby refuting the null hypothesis and supporting the claim that the suggested model enables greater searching and browsing.

Although claimed to be better, the suggested model is not without flaws, as those were surveyed and elaborated upon in section §7.4. The limitations and problems encountered by the thesis are, nonetheless, not crucial, and they could be addressed in the future, in a project with a greater scope than the current one.

Ultimately, Etymonline is a true treasure in the sense that it encompasses great and valuable information about etymology, drawing its data from credited sources and unifying and simplifying their presentation.

As such, it has the potential of changing the way we research and approach etymology and etymological research. The suggested methodology wishes to enhance the potential of the database by allowing greater functionality and usability. Like the introduced Project Gutenberg, Wikipedia and Flickr, Etymonline could be a site where users visit regularly, contribute to and gain knowledge from.

Other properties that might be attractive to users and could be implemented in the future are those of personalization, as is widely done in many online resources. Those include, among others, properties of user-specific logging in, marking of favorite entries, keeping score of contributions, receiving notifications or news feeds, and more.
The discussed methodology and suggested possible developments could broaden the tested model and attract a community of etymologists and language enthusiasts, as well as encourage similar projects in other languages, or in other resources of language treasuring or teaching.

There is a tight connection between language and information studies and librarianship (Dickson, 1979), and the current thesis touches just upon that intersection. Hence, the introduction of linguistic characteristics into DLs and information organization and retrieval systems strengthens that relationship and invites the cooperation of specialists and enthusiasts of all respective fields to work together.

Ironically, the thesis is about segmenting, categorizing and distributing information, whereas at the same time the field which it pertains to is that of mixed disciplines. In the academia it is this segmentation which actually might hinder collaborations and mutual understanding between researchers of different fields.

But alas, perhaps it is inevitable to segment and regroup the topics which occupy our minds, only to segment them again, and regroup them differently, in order to answer to our changing needs. As reviewed (Chapter 3), categorizing is, after all, a most basic human trait, constantly shaping and being shaped by our environment.
Bibliography


Appendices

All appendices are available in the following DropBox folder:
https://www.dropbox.com/sh/l4me6x9xh0y13qe/AADi5PYpceBM9m7MsyHCvtZ7a?dl=0