information

Article

# Towards the Representation of Etymological Data on the Semantic Web

**Anas Fahad Khan**

Istituzione di Linguistica Computazionale "Antonio Zampolli" (ILC)—CNR, Via Giuseppe Moruzzi, 1-56124 Pisa, Italy; anasfkhan81@gmail.com

Received: 15 September 2018; Accepted: 12 November 2018; Published: 30 November 2018

**Abstract:** In this article, we look at the potential for a wide-coverage modelling of etymological information as linked data using the Resource Data Framework (RDF) data model. We begin with a discussion of some of the most typical features of etymological data and the challenges that these might pose to an RDF-based modelling. We then propose a new vocabulary for representing etymological data, the **Ontolex-lemon Etymological Extension** (***lemonETY***), based on the ontolex-lemon model. Each of the main elements of our new model is motivated with reference to the preceding discussion.

**Keywords:** linked data; ontolex-lemon; etymologies; historical linguistics

## 1. Introduction

In this article, we propose a new RDF vocabulary for the representation of etymological information as linked data. This vocabulary is a minimal one and our intention is that it should serve as the foundation for further work in this area, that is, as the basis for other, more extensive, and more specialised RDF etymological vocabularies. Etymological information is, as we will argue below, very often heterogeneous, both in terms of the different subject areas it touches upon (and in particular the different subfields of linguistics which are often involved in etymological descriptions) as well as the kinds (and forms) of data which it incorporates. It would therefore be very difficult to come up with a comprehensive, wide coverage such vocabulary, at least on a first attempt such as this is. However there are a small number of elements, which as we will, once again, argue below, are frequently necessary in encoding etymological information—and in our particular case in encoding such information using the RDF framework—in a way that remains (reasonably) faithful to the original representation and elaboration of such information (given certain constraints of efficiency and usability), and in particular to its often highly speculative nature, and that also renders salient aspects of such data accessible for querying or for further processing: at least, that is, if we intend to carry out what we will term a 'deep' encoding of etymological data. What we mean by a 'deep' encoding should (hopefully) become clearer in what is to follow. Here we will only mention that the current move towards the publication of digital editions of legacy dictionaries using such standards such as the Text Encoding Initiative (TEI) and RDF (one big push in this direction recently has been given by the European Horizon 2020 project ELEXIS which strives to create a suite of tools for publishing as well as sharing and linking together both born digital and retrodigitized lexicons in RDF. Other interesting recent work in this area includes the Nénufar project to digitalise consecutive editions of the Petit Larousse Illustré [1].), as well as the creation of more innovative electronic editions of digital lexical resources (Of particular interest for our purposes is the emerging recognition of the importance of digital lexical resources for historically significant languages; one interesting project in this respect is DITMAO in which a digital RDF lexicon has been developed for describing medical terminology in Old Occitan, and which also links together terms in Occitan, Arabic and Hebrew, see

[2].), has underlined the need for a flexible and faithful way of encoding the lexical information in such sources as structured data—and since etymological information is very often a core component of many dictionaries, it must, of course, also be given its due consideration. One of the central aims of the work described in this paper has been to try and understand what such a flexible and accurate encoding entails for RDF-based lexicons and, relatedly, which kinds of classes and properties should be regarded as "first class" entities in a conceptual modelling of the etymological domain. In essence then our motivation in this article is dual: both to describe our new RDF vocabulary for etymological data, which is based on the ontolex-lemon model for ontology-lexicons, and to motivate it by arguing for the presence of certain categories and relationships in any such vocabulary, if it is to be a comprehensive one (and note that in this regard much of what we say will apply to any structured format for encoding digital texts and not just to RDF). The rest of the article is structured as follows. The first part of the article begins with a discussion of how we should understand the concept of *etymology* itself, Section 2, before we go on to describe what we feel are some of the main issues which arise out of the attempt to model etymological data as linked data, Section 3. We also briefly describe the ontolex-lemon model, Section 3.2 and go over some related work, Section 3.3. The second part of the article is an attempt to describe and justify the main elements of our model in detail, Section 4.

Please note that this article is an extended version of [3]. The following are the main contributions of this work with respect to the previous version: (a) an extended background discussion of etymologies with the addition of new examples, including a new discussion of etymologies in other kinds of texts than dictionaries or lexicons, and of the generalisation of the notion of etymology to other kinds of lexical phenomena than words; (b) an extended discussion of each of the elements in the vocabulary, and the addition of a new class, *Etymological List*.

## 2. Etymologies—A Little Bit of Background

### 2.1. What Are Etymologies?

It seems apposite to begin this section with an etymology of the word *etymology* itself. The term, which originally came into the English language via Old French in the 14th century, ultimately derives from an ancient Greek word, that is ἐτυμολογία (*étumología*). This latter can be broken up into two separate morphological components, namely, ἔτυμον (*étumon*), meaning "true sense", and the suffix -λογία (*-logía*), denoting "the study of"; hence it originally referred to the study of the "true" senses of words. The term as it is used in modern English today, however, has two primary senses. That is, *etymology* refers both to **a field of study**, one which traditionally occupies itself with the history of linguistic usages and which is typically regarded as a subfield of historical linguistics, as well as to **individual linguistic usage histories themselves**. Etymologies, in this second sense, are usually thought of as descriptions of *word* histories. (Although even here we should always be attentive to the ways in which the notion of word is itself an abstraction, as ten Hacken points out in [4]—even if it is an extremely useful one without which it is hard to see how anyone could get very far in the study of language.) However, in fact, the concept is wide enough to include historic descriptions of any other linguistically interesting phenomena, as Mailhammer writes in *Lexical and Structural Etymology* [5]:

> In principle, anything can be studied etymologically simply by asking the question "where did that come from?". The items investigated may be a phoneme or an entire text, and thus the question might be considered part of different disciplines, not even linguistic ones in the case of texts [. . . ] As a result, it seems unproblematic to extend the notion of linguistic etymology and etymological research beyond words and word histories to anything that can be linguistically described.

In addition, although in this work we will be focusing on the etymologies of words (and, indeed, we will talk about etymologies from a *word*-oriented point of view), it is important not to lose sight of the fact that the concept can, in theory, be applied to any kind of linguistic phenomenon.

Etymologies typically represent the history of a word (or any other kind of linguistic phenomena) by tracing out a sort of linguistic family tree or genealogy. A candidate word is associated with several of its (postulated) etymons and cognates either directly or indirectly (*indirectly* in the sense of taking into explicit consideration other intervening etymons/cognates) via relations which represent historico-linguistic processes, or rather linguistic *mechanisms* of the sort commonly studied by historical linguists. These mechanisms are usually subsumed under one of two headings: namely, either that of **borrowing**, or that of **inheritance**. The former refers to the process by which linguistic elements are transferred from one language into another via language contact: the English language, for instance, can boast no shortage of such examples, such as, as we saw above, the word *etymology* itself; the latter, instead, refers to the inheritance of words (or other linguistic elements) from a parent language, or a prior stage of the same language, such as for instance the Italian word *uomo*, the French word *homme*, the Spanish word *hombre*, the Portuguese word *homem*, and the Romanian word *om*, all of which derive from the Latin *uomo*. In addition we often need to take into consideration the phenomenon of semantic shift in which a word gains or loses a sense over time, as well the creation of new words through regular morphological processes such as affixation.

### *2.2. Example Etymologies*

Let us look, then, at a first example etymology that, notwithstanding its relative simplicity, will allow us to study several of the most typical features of etymologies, and with which we shall be concerned throughout this article; the example in question is adapted from an etymology given in Philip Durkin's excellent *Oxford Guide to Etymology* [6]. The word *friar*, which in Modern English has the meaning of "a male member of a Christian mendicant fraternity", ultimately derives from *frāter*, which was originally the Latin word for "brother". However, how exactly did we get from the one to the other? In fact, as in a good many other cases (including as we saw earlier *etymology* itself) the lexeme first entered the English language as a borrowing from Old French: it derives from the polysemic word *frere* which means both "brother" (as was the case with the Latin original) as well as "member of a religious fraternity". The word in this latter sense was borrowed into Middle English as *frere*, where it maintained its original French pronunciation and had the meaning of "member of a religious fraternity" as well as taking on the more specialised meaning of "member of a mendicant order"—it did not however share the original French word's meaning of "brother" as in "male sibling". Finally, over the years, the second specialised sense came to take on the status of the primary sense of the word *frere*. We can identify several different relationships between the various etymons identified above: Old French ***inherits*** the word *frāter* which, having undergone a sequence of sound changes in the interval, has become *frere*. Next, Middle English ***borrows*** the French word into its vocabulary, indeed, it borrows a single sense of the word; eventually, however, the word having entered the English language, it changes its meaning through a process of ***specialisation***. The following shorthand description for the whole process, which uses the "<" symbol (Please note that "<" is overloaded because it stands both for the development of one word from another or of a word borrowing from another language), is once more taken from [6]:

> Latin *frāter* brother < Old French *frere* brother, also member of a religious order of "brothers"
> < Middle English *frere*, *friar* < modern English *friar*.

In the simplest cases, as in the foregoing, an etymology will describe an ordered sequence of elements, but it is also common for etymologies to include some degree of branching in order to take account of uncertainty as to a word's candidate etymons or cognates—the further back we go, the more uncertainty there will be, the greater the paucity of evidence and the greater the chance of branching. In fact a word's family tree may turn out to be not to be technically speaking a tree at all but more generally a directed acylic graph. Let us illustrate this with an example. (Thanks to Jack Bowers for this example.) *Tempura* is the name of a Japanese dish consisting of seafood fried in a light, wheat-flour based batter (Strictly speaking, however, the concept of tempura has widened, so that it is no longer limited exclusively to fried seafood, as witnessed by the existence, and somewhat modest popularity, of tempura ice cream. We will ignore this complication for the purposes of explication however.) which has achieved a sufficient level of popularity in the anglophone world for the word itself (or at least a close enough approximation of the original Japanese term "天麩羅" (*tenpura*)) to have entered the English lexicon (and especially that part of it having to do with restaurant menus). The original Japanese word has a Portuguese origin (as indeed does the dish itself) and it is in fact, generally recognised to have two main candidate Portuguese etymons. The first, *tenpura*, derives from the Portuguese noun *tempêro* which means "condiment" or "seasoning" and originates from the Latin verb *temperō* and ultimately from the Latin *tempus*. The second is *tempora*, a Latin term which was used by Catholic missionaries in order to refer to feast days in which meat could not be consumed, and which once again finds its origin in the Latin etymon *tempus*. If we were to represent this etymology as a single, connected graph (with the English word *tempura* at its root) it wouldn't be a tree since there are two paths from the root to the word *tempus*. However, we *could* model this etymology as two (or more) separate sequences, although this may involve some measure of redundancy (in a linked data model this redundancy would obviously be minimized because we are able to reuse the same etymons across different sequences). Of course etymologies can, and in more scholarly works likely do, contain additional layers of bibliographic, linguistic, socio-cultural, and historical information, much of which cannot be easily fitted into a simple, pre-defined, schema or mapped (transparently) onto a graph structure. Nonetheless the presence of an underlying graph structure is ubiquitous enough to make etymologies particularly well suited to representation using a graph-based data framework like RDF—something which we argue in further detail below in Section 3.1. To illustrate the modelling challenges that can so easily arise in even moderately complicated etymologies we will present two entries each of which taken from a well-known specialist etymological dictionary and each of which deals with the same word, namely, *girl*. The first etymology is taken from Walter Skeat's influential *Etymological Dictionary of English* originally published in 1886 [7]:

> GIRL, a female child, young woman. (E.) ME. *gerle*, *girle*, *gyrle*, formerly used of either sex, and signifying either a boy or girl. In Chaucer, C.T. 3767 (A 3769) *girl* is a young woman; but in C.T. 666 (A 664), the pl. *girles* means young people of both sexes. In Will. of Palerne, 816, and King Alisander, 2802, it means 'young women'; in P. Plowman, B. i.33, it means 'boys;' cf. B. x. 175. Answering to an AS. form **gyr-el-*, Teut. **gur-wil-*, a dimin. form from Teut. base **gur-*. Cf. NFries. *gör*, a girl; Pomeran. *goer*, a child; O. Low G. *gör*, a child; see Bremen Wörtebuch, ii. 528. Cf. Swiss *gurre*, *gurrli*, a depreciatory term for a girl; Sanders, G. Dict. i. 609, 641; also Norw. *gorre*, a small child (Aasen); Swed. dial. *gårrä*, *guerre* (the same). Root uncertain. Der. *girl-ish*, *girl-ish-ly*, *girl-ish-ness*, *girl-hood*.

The surprising thing about the word *girl* is of course that it was originally used to refer both to young male and to young female persons (The use of this word as an example was inspired by [8].). Skeat traces the word's lineage through Middle English all the way back to Chaucer and *Piers Plowman*—and, indeed, in the latter text he identifies a usage of the word which refers exclusively to "boys"—and ultimately to a reconstructed Anglo-Saxon (and Teutonic) root. Lastly the entry gives some derived forms of the word *girl*. The second etymology is taken from the entry for the word *girl* in Eric Partridge's single volume *Origins: A Short Etymological Dictionary of Modern English* [9]:

> girl, whence **girlish**, derives from ME *girle*, varr *gerle*, *gurle*: o.o.o.: perh of C origin: cf Ga and Ir *caile*, EIr *cale*, a girl; with Anglo-Ir *girleen* (dim -*een*), a (young) girl, cf Ga-Ir *cailin* (dim -in), a girl. However, far more prob, *girl* is of Gmc origin: Whitehall postulates the OE etymon **gyrela* or **gyrele* and adduces Southern E dial *girls*, primrose blossoms, and *grlopp*, a lout, and tentatively LG *goere*, a young p/erson (either sex). Ult, perh, related to L *puer*, *puella*, with basic idea '(young) growing thing'.

Partridge, in a work published seven decades after Skeat's (then) authoritative etymological dictionary, makes three hypotheses regarding the word's origin (the third of which is seemingly compatible with both of the first two). His references to a possible Gaelic root for *girl* as well as to a potential Latin origin of the word—neither of which were mentioned in Skeat's entry—are both accompanied by an explicit note of caution, indicated by the presence of the abbreviation "perh.". This example is useful because it serves to underline the tentative and unavoidably speculative nature of etymologies—unavoidable because etymologists are unable, in most cases, to rely on what was, until quite recently, the modern linguist's favourite source of linguistic data, namely, native speaker intuition—and in addition, and, depending on the language and how far back we want to go, there probably is not very much in the way of corpus data to go on either. On the other hand, however, it is imperative that we are able to properly represent attestations in the cases where an etymology actually **does** cite the evidence of a linguistic usage in a corpus, such as is the case in Skeat's entry for *girl* with its citations of Chaucer's *Canterbury Tales* and Langland's *Piers Plowman*. It also happens to be the case that etymological sources will very often cite the scholarly literature, especially works in etymology and historical linguistics. Fortunately this kind of information, since it involves the description of networks of links and citations, can be represented very naturally in linked data (and this is one of the main advantages of its usage in such contexts). Furthermore, there already exist several comprehensive linked data vocabularies for describing citations and bibliographic data which we can immediately take advantage. (See for instance the SPAR suite of linked data ontologies for the publishing domain (http://www.sparontologies.net/) which include an ontology for bibliographical information (http://www.sparontologies.net/ontologies/biro) and for representing citations (http://www.sparontologies.net/ontologies/cito). In previous work we have developed a vocabulary for attestations in lexical resources too [10].) The take home message, then, is that uncertainty is a hallmark of etymological data—and even non-specialist works that contain etymological information will very often feature more than one candidate etymology for the same entry—and this is something that we should bear in mind when it comes to modelling etymologies in a structured format.

### *2.3. Where Are Etymologies Found?*

Etymologies can frequently be found in general-purpose dictionaries such as the Oxford English Dictionary or *Le Petit Larousse* as well as in more specialist lexical works that focus on disciplines such as etymology or philology. As mentioned in the introduction the representation of etymologies in digital language resources has taken on an increased relevance recently thanks to a burgeoning interest in retrodigitizing legacy resources. However the creation of standards and shared models for the representation of etymological information in structured datasets is valuable and of interest in its own right since it contributes towards making one very important kind of linguistic data more accessible and more amenable to further machine processing—indeed etymologies play a crucial role within the field of historical linguistics; to quote Mailhamer from [5]:

> Etymological research naturally is fundamental to the historical investigations of a language. Etymologies permit generalizations about historical developments, for instance the formulation of sound laws or about a particular synchronically attested phenomenon, such as the occurrence of a particular type of stem formation. Moreover, etymologies are also used as arguments to answer questions about historical relationships among languages."..[E]tymology can be regarded as something like a fundamental auxiliary discipline of historical linguistics

Even though the main focus of this work is on the encoding of etymologies in lexicons or dictionaries, one should bear in mind that etymologies can also be embedded in other kinds of text: indeed entire articles or book chapters are regularly dedicated to the etymology of a single word within the field of historical linguistics. Of course, the presence of etymologies is far from being limited to works that belong to the field of linguistics or to philology, or to popular treatments of the preceding. The use of etymologies in philosophical and theological works can be traced back to antiquity and is a prominent feature of several different cultural and religious traditions. Texts such as Plato's *Cratylus* and Yāska's *Nirukta* describe so-called "semantic" or "speculative" etymologies in which the meaning of a word is elaborated on by comparing it with other words which sound or look similar, without, however, the systematic use of historical evidence that characterises modern "historical" etymologies [11]. More familiar, on the other hand, are the kinds of etymology that are found in early modern philosophical texts such as Thomas Hobbes' *Leviathan* and in the works of the influential Neopolitan thinker Giambattista Vico, and which make up a part of the actual argument of those texts. Take for instance the following passage, from Chapter X of the *Leviathan*, which locates the origins of words for titles of nobility in the history of military practices.

> Titles of *Honour*, such as are Duke, Count, Marquis, and Baron, are honourable; as signifying the value set upon them by the Soveraigne Power of the Common-wealth: Which Titles, were in old time titles of Office, and Command, derived some from the Romans, some from the Germans and French. Dukes, in Latine *Duces*, being generals in war; Counts, *Comites*, such as bare the general company out of friendship; and were left to govern and defend places conquered, and pacified; Marquises, *Marchiones*, were Counts that governed the Marches, or bounds of the Empire. Which titles of Duke, Count, and Marquis came into the Empire about the time of *Constantine* the Great, from the customs of the German *Militia*. However, Baron seems to have been a title of the Gaules, and signifies a Great man; such as were the Kings or Princes men, whom they employed in war about their persons; and seems to be derived from *Vir*, to *Ber*, and *Bar*, that signified the same in the language of the Gaules, that *Vir* in Latin; and thence to *Bero*, and *Baro*: so that such men were called *Berones*, and after *Barones*; and (in Spanish) *Varones*.

Such etymologies (unlike those found in the *Cratylus* and other works originating from antiquity) are comparable with modern etymologies found in dictionaries or other scholarly works. In addition, even if the main use case for our vocabulary is the representation of etymologies in lexical resources there is no reason why it cannot be used for examples of the sort that cited above. In fact our vocabulary will enable such etymologies to be modelled and subsequently published in a standardised way that makes them easier to compare with etymologies from other heterogeneous sources.

## 3. Representing Etymologies on the Semantic Web

### 3.1. Advantages (and Some Drawbacks) of Publishing Etymological Data as Linked Data

The clear presence of an underlying graph structure in most etymologies would clearly tend to favour the use of Linked Data as a means of modelling and publishing such data—as too would the fact that etymologies often make extensive reference to non-linguistic information, and especially to bibliographic, encyclopedic and historical data. Linked data, thanks to its adoption of one single underlying data framework, i.e., RDF, across all fields, topic areas, and genres of resource (from corpora to ontologies to multidimensional statistical data), makes it much more natural to link together resources that are heterogeneous with respect to various different dimensions. In particular the Linked Data approach to conceptual modelling favours the representation of lexemes, word forms, word senses, and numerous other kinds of lexical phenomena, as **individual resources** each of which has its own universal identifier which allows it to be referred to and therefore re-used across different datasets hosted at different locations. We can link these resources together and predicate things of them using **unary** (RDF concepts/classes) and **binary** relations (RDF properties) respectively. Add to this the fact that the Linked Data recommendations promote the re-use of common, standard, vocabularies within and across subject areas and the potential for building (more or less carefully curated) Semantic Web etymological networks that traverse large numbers of individual lexical datasets and take in a myriad other kinds of resource category becomes increasingly plausible. At the same time, however, and despite the graph-like nature of etymological data the representation of even fairly simple etymologies (such as the *friar* and *tempura* examples given above) in RDF turns out not to be as straightforward as it might seem on the basis of the preliminary considerations mentioned above, and such representations can very quickly start to seem overly verbose and full of redundancy. This prolixity is due to two main factors. The first, of course, relates to the complex nature of the data itself (see Section 2); the second to the requirement that we make as much of this data accessible to querying, and more generally to machine processing, as we can under the **expressive limitations** imposed by the RDF data model, the most important of which relates to the fact that the use of RDF restricts us to the use of at most binary predicates. The benefits of a successful modelling, on the other hand, are numerous—in addition, that is, to the possibility, mentioned above, of creating medium to large scale etymological networks. RDF enables us to overcome many of the limitations of both the printed page and of other (relatively) unstructured digital data formats, and to therefore make such data more easily navigable and queryable—one of the major advantages of publishing data in RDF is, of course, that we can subsequently query our datasets using the expressive SPARQL query language (it also gives us access to a whole host of other pre-existing standards and technologies that aren't available with, say, for instance, TEI). Of course just how much more easily queryable, and, generally, *machine actionable* a linked data dataset is depends on *how* it is structured, and how efficiently—and *that* ultimately depends on which classes of "things" we decide to model as RDF individuals and that we predicate attributes of using RDF properties. In this work we have consciously tried to avoid making our vocabulary overly complex by prioritising difficult, limit cases. However, as we hope to have convinced the reader by the end of the article, the classes and properties which we present below are either necessary, or at least very hard to do without, if we want to capture some of the most salient characteristics of etymological data in RDF. In the following sub-section, Section 3.2, we describe the ontolex-lemon model upon which our model is based. Then in Section 3.3 we describe related work, both in terms of previous attempts to model etymological information in RDF as well as in the two data frameworks TEI and the Lexical Markup Framework (LMF).

### *3.2. The Ontolex-Lemon Model*

As has been previously stated our model takes the ontolex-lemon vocabulary as its foundation, and in order therefore to make this article a more self-contained one we will give a brief overview of the latter vocabulary in this section. The use of linked data as a means of publishing language resources is now fairly well established (For instance, the section of the Linked Data cloud dedicated to language resources contains, at the time of writing, more than two hundred different resources (http://linguistic-lod.org/llod-cloud)), and the subject area has its own dedicated bi-annual workshop (Linked Data in Linguistics) as well as numerous projects at a regional, national and international level in which it has been one of the core themes. and it seems to hold out particular promise when it comes to the publication of lexicons (for which task it seems much more suited than say the publication of corpora). The ***lexicon model for ontologies*** [12] otherwise known as the **ontolex-lemon model** (The ontolex-lemon guidelines are available at https://www.w3.org/2016/05/ontolex/), is the most widely used vocabulary for publishing lexicons as linked data. As its name suggests ontolex-lemon was originally proposed as a model for enabling the enrichment of ontological datasets or skos concept hierarchies with linguistic information, rather, that is, than as a general purpose model for creating lexical datasets. Ontolex-lemon was directly inspired by the earlier Lexical Markup Framework, an ISO standard that aimed to provide a common framework for the creation and sharing of computational lexicons [13] (LMF was defined in UML but it is almost exclusively serialised in XML), but is much more simplified than the former, something that is understandable given ontolex-lemon's initial focus on the ontological use case. Nevertheless and despite ontolex-lemon's origins, it has been appropriated as a model for publishing lexicons in linked data, regardless of whether any of the entries describe the linguistic properties associated with concepts in an ontology or not, and it is now regarded as a de facto standard in the area—our primary consideration in using it as the basis of the new vocabulary. At the same time, many of the ontolex-lemon elements which we reference in our vocabulary are shared with, or are semantically close to, the core parts of LMF and the dictionary chapter of the TEI guidlines, leaving open the possibility of introducing our new classes and concepts in extensions of the latter two standards. Figure 1 presents the core ontolex-lemon model.

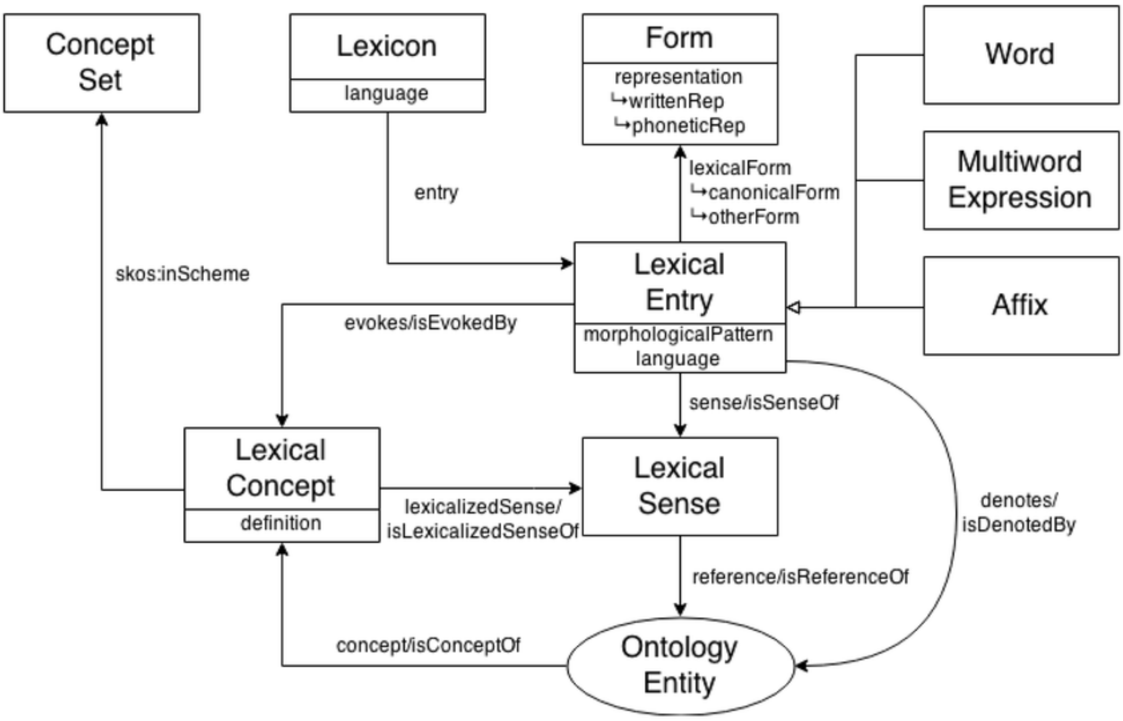

**Figure 1.** The core elements of the ontolex-lemon model (diagram taken from the W3 ontolex-lemon guidelines).

The most relevant parts of the model, for our specific purposes, are the classes `LexicalEntry`, `Form` and `LexicalSense` along with their related properties as shown in the diagram. `LexicalEntry` and `Form` have definitions that are more or less intuitive; less obvious, on the other hand, is the definition of `LexicalSense`: each individual of the class `LexicalSense` is regarded as the reification of a pairing of a `LexicalEntry` with an ontological concept, something that would, strictly speaking, necessitate the existence of an ontological concept for every `LexicalSense`, although we will, sensibly, forgo this latter.

### 3.3. Related Work

Previous work on defining a framework for representing etymological data in digital lexical resources includes Salmon-Alt's proposal for an LMF-based etymology model [14] along with Bowers and Romary's working paper on the *deep* encoding of etymological information in TEI [15] from which we have borrowed the use of the adjective *deep* in reference to the structuring of etymological data in computational lexical resources (as well as taking more general inspiration from that work for the approach presented in the current paper). In [16], Sagot describes the systematic extraction of etymological data from a Wiktionary data dump. The resulting etymological database is then exported with an extension of TEI-XML based on the proposal described in [15], with the addition of dedicated attributes for encoding alternative etymological hypotheses and for encoding etymological chains. Both of these design decisions accord with our own conclusions regarding the importance of representing these etymological phenomena in a wide coverage etymological vocabulary. Additionally, the development of a standard etymological extension to the core LMF model is currently underway as a project (of which the present author is co-project leader) within the ambit of the International Standards Organization working group ISO/TC 37/SC 4/WG 4. With respect to the modeling of etymological information in RDF, previous work includes the *Etymological WordNet* [17] as well as the work of Moran and Bruemmer [18]. In [19] Chiarcos et al. define a minimal extension of the lemon model with two properties for encoding and linking etymological data together: these are the symmetric and transitive `cognate` and the transitive `derivedFrom`. Our intention in this article is to create a vocabulary that is more flexible than the (relatively) minimal models presented in these RDF-based works and which will allow for the representation of a wider range of etymological phenomena.

## 4. A First Proposal for a Linked Data Vocabaulary for Etymology

Our proposed model is called the **Ontolex-lemon Etymological Extension**, or ***lemonETY*** for short. It consists of a small number of what we consider "foundational" classes along with their related properties. In the following sections we will go through each of the main constituents of the model, describing the role that each plays in enabling the representation of etymological information in RDF. In Section 4.1 we define the classes **Etymology** and **Etymon**, respectively, these are the two central components of our model, both of which, taken together, allow us to refer to and predicate over entire etymologies as well as the lexical elements which compose them. In Section 4.2 on the other hand we look at two different ways of constructing ontologies out of their component parts. Given the frequency of etymologies consisting of individual sequences of etymons/lexical entries such as the *friar* example above, we decided that it would be useful to define a means of directly defining such sequences in RDF. We present this construction in Section 4.2.1 as **Etymological List**. On the other hand this construction leaves out several important kinds of case, and so in Section 4.2.2 we present the class **Etymological Link**, which allows for the construction of graph-like etymological structures. The latest version of the *lemonEty* vocabulary can be found at http://lari-datasets.ilc.cnr.it/lemonEty; further documentation and examples can be found at https://github.com/anasfkhan81/lemonEty.

### *4.1. Etymology and Etymon*

The first element of *lemonEty* which we chose to describe here is also arguably the most important, namely, the `Etymology` class. We have chosen to model an etymology as an hypothesis *about* the history of a given lexical element—a hypothesis, moreover, that is usually associated with an arrangement of lexical elements (where we will underspecify what we mean by arrangement for the time being). Making `Etymology` a first class entity in our vocabulary so that it is, in effect, a reification of the association just mentioned (between a lexical element with an arrangement of elements) allows us to give a straightforward treatment of the (common) case in which a single entry in the same resource has more than one etymology associated with it. It also facilitates the attribution to etymologies of numerous other kinds of salient bibliographical and historical information, including, for example, data on those scholarly sources which endorse a certain etymology, or which bring different pieces of evidence to bear on its truth or falsity, as well as making it easier to associate etymologies with differing levels of confidence or with quantitative certainty scores—something that we felt was vital in a field as fundamentally speculative as that of etymology. The next step is to define (object and data) properties which relate individuals of the class `Etymology` to individuals of other (salient) classes. The primary connection here is between an etymology and the lexical element whose history it is intended to describe: this is the role of the object property `etymology`. The range of `etymology` is of course the class `Etymology` itself; we have decided, however, not to impose any axiomatic constraints on the domain of the property in order to avoid delimiting the kinds of things with which we can associate etymologies. We have also introduced two datatype properties `justification` and `likelihood` which take string values and which allow for the description of the justification and likelihood of a single etymology, respectively.

Another class which we felt it necessary to include in our model is `Etymon`. The term *etymon* is fairly standard in the etymological literature (indeed we have made free use of it throughout the article) and designates the role that some lexemes play within a single lexical resource—in particular their status with respect to the "official" lexical entries in a resource: "official" in the sense that it is the main task of the lexical resource to describe them. For us, etymons are lexemes whose sole purpose (in a given lexical resource) is to help describe the etymologies of lexical entries. (Lexemes which play this role in a resource in addition to being standard lexical entries are not classified as etymons.) Aside from this difference of role, and the fact that etymons belong to a different language or to a different language stage from the (main) languages dealt with in a lexical resource—that is, typically, those languages which are listed in the metadata for the resource itself—we should be able to predicate the same kinds of phonetic, morpho-syntactic or semantic property of etymons (insofar as this information plays an explanatory role in the description of another word's history) as we do with "first class" lexical entries—even if it is true that, when it comes to etymons, there is usually a lack of systematicity in the inclusion of any kind of lexical information in a lexicon. Why then, is the distinction between `Etymon` and the ontolex-lemon class `LexicalEntry` a useful one to make at all? Quite simply, it means that in querying a lexical resource we can easily filter out all those "secondary" lexemes which are only present in a dictionary or lexicon because of their historical relationship with the "official" headword-affiliated entries of the resource. This helps to avoid the situation where, for example, in trying to extract a list of entries in a monolingual English lexicon with a SPARQL query we end up with what is effectively a multilingual English-ancient Greek-Latin-French dictionary. Regarding the status of `Etymon` with respect to `Lexical Entry` then, we were faced with two options: the first was to create a new class `Lexeme` of which both `Etymon` and `LexicalEntry` were subtypes; or to make `Etymon` a subclass of `LexicalEntry`. We ended up choosing the latter option because it was less disruptive with respect to the original ontolex-lemon model. We created the class `Cognate`, another subclass of `LexicalEntry`, for similar reasons of diversity of purpose and explanatory role. So that now we have these two new classes under our belt, we can turn to the difficult question of how to actually go about putting etymologies together.

### *4.2. Building up Etymologies*

In this article, we propose two different ways of building up "arrangements" of lexical elements to describe the content of etymologies. The first is described in the next section, Section 4.2.1, and models an etymology as a linear sequence of RDF elements. This will be enough to cover a large number of cases of etymologies found in general purpose lexical resources as well as in more specialist resources. It is still not expressive enough to capture everything, however, and so in Section 4.2.2 we propose a more general, "graph"-based approach centered around the concept of *Etymological Link* which we also introduce in that section.

#### 4.2.1. Etymological List

As is traditionally said to be the case with skinning cats, there is more than one way of modelling ordered sequences in RDF. One option is to make use of the `rdf:seq` or `rdf:list` classes. The former is intended for modeling open lists (that is lists which are allowed to contain more members than those which have already been enumerated) the latter closed lists (those which can contain only the members which have been enumerated in the definition of the list). Unfortunately these two classes suffer from the drawback that they are unavailable in OWL-DL [20] and therefore do not permit the use of OWL-DL reasoning engines, something which seriously limits their practical usefulness. Another possibility would be to use a dedicated ontological vocabulary, or a special ontological pattern in OWL-DL, and this is the approach which we finally decided upon with the lemonEty vocabulary. To make our etymological vocabulary as usable and as easy to work with as possible—and to facilitate both reasoning over linked data ontologies using standard Semantic Web tools—we decided to represent etymological sequences as linked list data structures. We settled on the OWLLIST vocabulary described in [20] for this purpose. (OWLLIST can currently be found at the URL https://users.ugent.be/~pipauwel/ontologies/list_W3ID/20151211/index.html.) With OWLLIST we can model a list of elements $L = [a_0, a_2, a_3, \ldots, a_n]$ by assigning a cell $C_i$ to each sublist (that is, each list which can be formed by removing elements from the original list) $L_i$ of $L$. We can then link each cell $C_i$ (corresponding to the sublist $L_i$) to the cell $C_{i+1}$ (corresponding to the next sublist $L_{i+1}$ in order of length and formed by deleting its first element of $L_i$) using the property hasNext; we link $C_i$ to the first member of the sublist $L_i$, $a_i$ using the object property `hasContents`. See Figure 2. The transitive property, `isFollowedBy`, is a superproperty of `hasNext` and allows us to navigate between a cell representing a list and any of the cells representing its sublists.

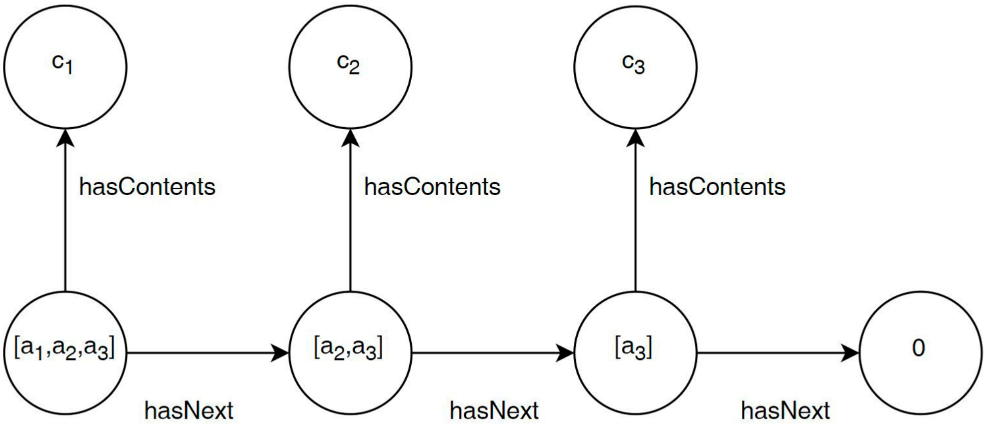

**Figure 2.** How OWLLIST models a three element list $[a_1, a_2, a_3]$. Please note that each of the four bottom nodes are of type **OWLLIST**.

Of course we already have the possibility of using OWLLIST as defined to represent the arrangement of elements in a sequence in an etymology—that is without adding any additional domain specific classes or properties. However, this would leave the types of the relationships between the elements in an etymology underspecified (which would in effect be the equivalent of using the underspecified "<" symbol which we saw earlier). One way to improve on this is to define sub-properties of `hasNext` which elaborate on the type of the etymological relationships between elements in a list representing an etymology. To this end we decided to create a new class representing etymological lists, `EtyList`, as a subclass of `OWLLIST`, along with the following subproperties of the OWLLIST vocabulary object property `hasNext`:

- The object property `hasLink` is a generic etymological relationship between two elements in an `EtyList`; it is a subproperty of `hasNext` and its domain and range are both `EtyList`;
- The object property `borrowingLink` is used to link two cells $C_i$, $C_{i+1}$ together when we want to specify a relationship of borrowing between $a_i$ and $a_{i+1}$; it is a subproperty of `hasLink`;
- The object property `inheritanceLink` is used to link two cells $C_i$, $C_{i+1}$ together when we want to specify an inheritance relationship between $a_i$ and $a_{i+1}$; it is also a subproperty of `hasLink`.

It then becomes a simple matter to continue specialising the types of link between elements in an etymology by creating further sub-properties of `hasLink` depending on the level of detail and the kinds of historical or linguistical information to be represented. For instance we decided to create additional sub-properties corresponding to different kinds of semantic change such as `specialisation`, `amerliorisation`, etc. We use these classes to model the *friar* example given in the turtle notation below and in diagrammatic form in Figure 3.

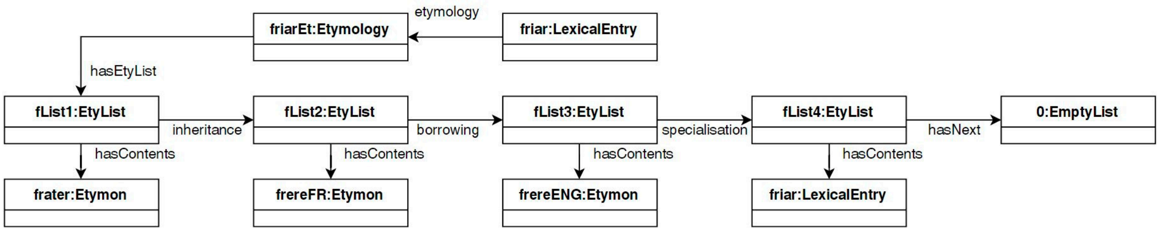

**Figure 3.** Using the `EtyList` class to model the *friar* example.

```
:fList1 a lemonety:EtyList ;
lemonety:inheritance :fList2 ;
<https://w3id.org/list#hasContents> :frater .
:fList2 a lemonety:EtyList ;
lemonety:borrowing :fList3 ;
<https://w3id.org/list#hasContents> :frereFR .
:fList3 a lemonety:EtyList ;
:inheritance :fList4 ;
<https://w3id.org/list#hasContents> :frere .
:fList4 a lemonety:EtyList ;
<https://w3id.org/list#hasContents> :friarENG ;
<https://w3id.org/list#hasNext> :empT .
:empT a  <https://w3id.org/list#EmptyList> .
```

With `EtyList` we can model both single sequences of etymons as well as very many branching etymologies efficiently. In the latter case we can represent each branch as a separate `EtyList`; the fact that the sequences share etymons in common helps to preserve the graph structure of the etymology.

### 4.2.2. Graph-Based Etymologies and the Etymological Link Class

The third of the core elements of our model is *Etymological Link* (`EtyLink`), a reification of an etymologically relevant relationship between two elements in an etymology and something which we represented using the property `hasLink` and its sub properties in the previous section. Such a move might seem to be overkill on first sight (come on, don't we already have object properties that represent etymological links?) or it might seem to be the sort of class that only becomes necessary or really useful in more specialised, highly detailed etymological works. However, this is not entirely true; the need for such a reification (at least when it comes to RDF and its specific technical limitations) can appear even in fairly simple etymologies. Take for instance, the following example (My thanks once again to Jack Bowers for being the source of this example.) which concerns the Portuguese lexeme *nação* "nation", a word which derives from the Latin noun *nātiō* "birth, nation". The relationship between the two words is complicated by the fact that the lemma/singular form of the Portuguese noun actually derives from *nātiōnem*, that is, the accusative singular form of *nātiō*, and not from the nominative lemma form. This would seem to call, in this case at least, for an etymological link relationship between forms rather than between words/etymons. In addition, there is no end of such cases throughout the Romance languages. For instance the Italian word *luce* derives from the accusative of the Latin word *lūx*, that is, *lūcem* rather than from the nominative form. This is something that often varies between Romance cognates as well, so that the Portuguese word for "man", *homem*, derives from the Latin accusative form *hominen*, whereas the Italian for "man", namely, *uomo* is derived from the nominative form *homō*. Where then should we place the etymological links in these kinds of examples: should they be relations between word forms rather than between etymons and etymons/lexical entries? The former approach, although it is, in a certain sense, the more accurate one, does not necessarily accord with how etymologies are often described in such cases, that is, as representing relationships between words, and not just the constituent properties of words. The *friar* example introduced in Section 2 is also salient here: recall that *frere* was borrowed into English from Old French but *only* under a single one of its senses. Again, should we link between senses and senses here or etymons and etymons? To deal with such cases we decided to make *Etymological Link* a separate class, `EtyLink`. Another compelling reason for making such a move relates to the fact that, in specialist etymological resources, individual links within an etymology will have their own detailed justifications (that pertain, for instance, to the existence of sound laws) or include citations to other works. The following list details the object and datatype properties associated with `EtyLink`:

- The object property `hasEtyLink` has `Etymology` as its domain and `EtyLink` as its range, it therefore links an etymology together with its etymological links.
- The object property `etySource` links etymological links with their source etymons/lexical entries, its domain is `EtyLink`and its range is `LexicalEntry` ; similarly for the object property `etyTarget` which links etymological links with their targets.
- The object properties `etySubSource` and `etySubTarget` allows us to deal with cases such as that of *nação* and *frere* where we would like to create additional links between lexical elements such as word senses and forms and etymons and lexical entries in order to add to the explanatory value of an etymology
- The datatype property `type` describes the type of an etymological link as a string value; `type` has the domain `EtyLink`.
- The datatype property `justification` allows us to associate a justification with an etymological link.

In Figure 4 we represent the *nação* example. Note, in particular, the use of the properties `etySource` and `etyTarget` which relate together the lexical entry `nacao` and the etymon `natio` along with the use of `etySubSource` and `etySubTarget` between the relevant form variants of `nacao` and `natio`. (We can model the *frater* example similarly adding extra links between the relevant senses.)

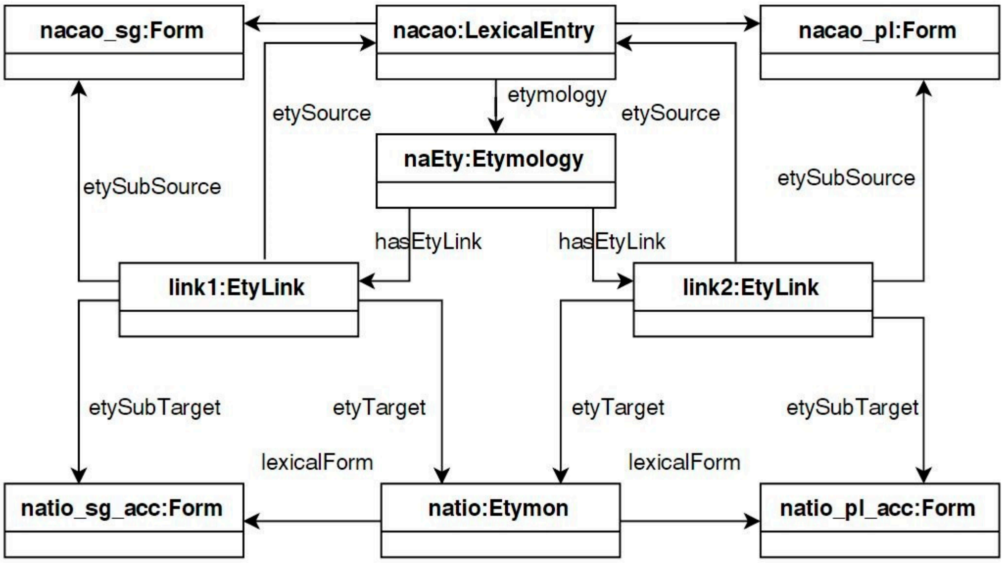

**Figure 4.** Using the `EtyList` class to model the *nação* example.

Although we can arrange `EtyLink` individuals in a list, again using RDF vocabularies like OWLLIST, the class Etymological Link allows us more flexibility and ensures that we need not limit ourselves to linear sequences of lexical elements (or even to trees); indeed it allows us to construct etymological graphs, though these obviously will not be as easy to navigate as etymological lists. We can encode the *tempura* example as described above using `EtyLink` (Please note that although in this case we could use two etymological lists to represent the etymology of *tempura*, the etymology's underlying graph structure would only be implicit in the fact that the two lists share many etymons in common; modelling the example as one single etymology composed of numerous etymological links makes this graph structure more explicit instead of representing it as separate paths.) as in Figure 5.

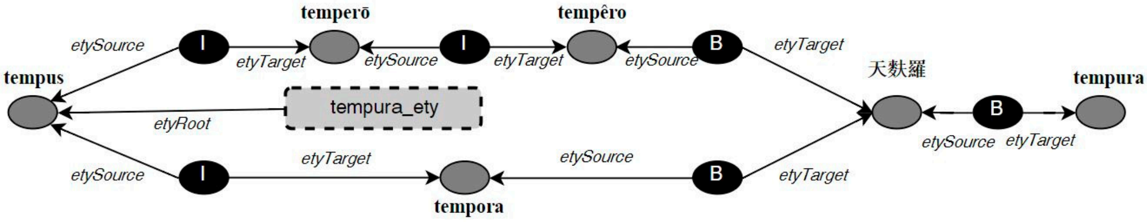

**Figure 5.** Modelling an etymology as a graph structure.

Lexical entries and etymons are represented as grey filled circles in the diagram, the etymological link objects are black filled circles, and the individual representing the whole etymology is a grey rounded rectangle with a dashed outline. Each of the etymological links in the example is typed as either *inheritance* or *borrowing* (represented using a white letter "I" or "B" respectively in the diagram); each of these links is associated with the `tempura_ety` etymological object using the `etymology` and `hasEtyLink` properties mentioned above (these links are not represented in the diagram for reasons of space). The `etySource` and `etyTarget` properties here allow for the representation of the temporal evolution of a word (unlike in the previous example where there was no sequential relationship between the two etymological links). As the diagram demonstrates the extra flexibility offered by `EtyLink` comes at a serious cost both in terms of human readability and ultimately in the navigability and hence queryability of the end result. Nevertheless in many instances, we **are** willing to forfeit efficiency if it means an increase in accuracy and the possibility of capturing the nuances of our data in a natural way; in which case it is hard to see how it would be possible to represent etymological information, and at this level of detail, and in RDF, without incurring these kinds of costs in terms of the prolixity of the final result.

### *4.3. There Is No "One Size Fits All" Approach*

We can encapsulate much of the discussion of the last few sections with the following slogan: when it comes to representing etymologies in RDF there can be no *one size fits all approach*. Notwithstanding this inconvenience—and the unhappy fact that those aspects of etymological data which we have singled out as rendering its efficient representation in RDF a decidedly non-trivial task actually turn out to be in reality very common—we have been able to define a core collection of common concepts and properties which seem to underlie many different kinds of etymologies. Despite their relative prevalence however, users may not always feel that it is necessary to represent them in individual cases, and may therefore be reluctant to use a vocabulary such as that described above due to the ensuing prolixity of the data, opting for a reduced collection of properties. However, this will have a concomitant effect on the kinds of things that it is possible to *systematically* query for; and especially if I want to query across individual resources. Say I want to find all the etymologies of English words that include the Latin etymon *tempus* in an English dictionary, or to write a query to find out how many nouns in Portuguese are directly derived from a non-norminative Latin form in a Portuguese lexicon; this becomes difficult in any reasonably comprehensive etymological source, in which numerous words will have more than one etymology, unless I make the concepts of Etymology and Etymon explicit and in consequence directly queryable. One particularly instructive parallel here is with the publication of digital editions of textual works. When it comes to publishing the electronic version of a text, the fact is that, in most cases, we are simply not interested in encoding information relating to the existence of different copies of the source material. When it comes to creating scholarly editions of texts, however, in the case of texts for which there exist various differing versions, then we find that many of those assumptions which we previously took for granted, including many of those regarding the very notion of a *text* itself, have to be foregrounded, and to be consequently challenged and elaborated upon in a variety of different ways. In the same way, if a user is working on the publication in RDF of an etymological database or indeed any work which contains etymological data, but which does not reference different etymological hypotheses for any single entry, then the existence of a class such as `Etymology` will seem superfluous and by the same token it is highly unlikely that she will want to encode *Etymological Link* as a separate class either. Of course the minute the requirement to model any reasonably complicated etymological entry arises, that same user will find it hard to do without these same concepts or to others that play a conceptually similar role. Even if a user's data only calls for a limited set of object properties to encode the etymological data contained therein, however, she may still want to use 'redundant' classes and properties in order to make her dataset more interoperable with other datasets which do use these elements, and to thereby enable queries that transverse etymological datasets.

Lexical Domain

The final class which we will introduce in this article, `LexicalDomain`, is only indirectly linked with the others which we have looked at thus far. Its importance lies in the fact that it permits us to predicate semantic information of etymons and lexical entries in situations where the use of the ontolex-lemon class **Lexical Sense** may not be appropriate. The intention is that `LexicalDomain` will allow users to associate an overall semantic domain with a lexical element, whether it is an etymon, a lexical entry, or a root, thus bypassing the finer grained description of the meaning of a lexical element offered by **Lexical Sense**. For instance we can associate the English word *lion* both with an ontolex *Lexical Sense* linking it to the extension of the binomial nomenclature *Solanum lycopersicum*, as well as with a *Lexical Domain* corresponding to either of the concepts *Felidae* or *Mammal*. Such a class seems particularly appropriate in the case of reconstructed words, where etymologists have tried to reconstruct a meaning, or a general semantic field, of a root based on the evidence of word senses that are to some degree attested, and where a high degree of semantic vaguenesss is unavoidable. As Watkins (quoted in [6]) points out: "[R]econstructed words are often assigned hazy, vague or unspecific meanings... The apparent haziness in meaning of a given Indo-European root often simply reflects the fact that with the passage of several thousand years the different words derived from this root in divergent languages have undergone semantic changes that are no longer recoverable in detail". We associate the object relations `lexicalDomain` and `domainField` with `LexicalDomain`; the former plays a role corresponding to that played by `sense` with respect to the ontolex-lemon class *Lexical Sense*, and the latter to that played by `reference`, again, with respect to *Lexical Sense*. For instance take the reconstructed proto-Indoeuropean root **ker-tā-*, hypothesised to be the root of the English word *hearth* and which has been given the (reconstructed) meaning "fire". We can represent this state of affairs by associating a lexical domain *d* with the etymon representing **ker-tā-* using the property `lexicalDomain`, and linking *d* to an ontology concept representing "fire", such as e.g., `dbPedia:Fire`, using the property `domainField`.

## 5. Conclusions

The model that we have presented in this article is intended as groundwork for future attempts at defining additional RDF vocabularies that will allow for a fuller description of etymological information. For this reason we have not attempted a more detailed treatment of bibliographic information and citations of secondary scholarly literature in etymological data; nor have we looked in any real detail at the particular challenges associated with modelling semantic and syntactic change or sound laws. Even so these areas will need to be engaged with before we can hope to encode a large number of realistic examples from a variety of sources. Our plan now is to test the robustness and coverage of *lemonEty* by using it to encode entire lexical resources that include a strong etymological component, in particular retrodigitized dictionaries. We are also interested in looking at the possibility in future of extracting etymological graphs/trees/sequences from text, (semi-)automatically and representing them in RDF using our model.

**Funding:** This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731015.

**Acknowledgments:** Thanks to Jack Bowers and Christian Chiarcos for some stimulating discussions on the subject of modelling etymologies.

**Conflicts of Interest:** The author declares no conflict of interest.

## References

1. Bohbot, H.; Frontini, F.; Luxardo, G.; Khemakhem, M.; Romary, L. Presenting the Nénufar Project: A Diachronic Digital Edition of the Petit Larousse Illustré. In Proceedings of the GLOBALEX Workshop at LREC 2018, Miyazaki, Japan, 8 May 2018.
2. Bellandi, A.; Giovannetti, E.; Weingart, A. Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information* **2018**, *9*, 52. [CrossRef]
3. Khan, F. Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web. In Proceedings of the 6th Workshop on Linked Data in Linguistics, Miyazaki, Japan, 12 May 2018.
4. ten Hacken, P. On the Interpretation of Etymologies in Dictionaries. In Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, Slovenia, 17–21 July 2018; p. 144.
5. Mailhammer, R. *Lexical and Structural Etymology: Beyond Word Histories*; Walter de Gruyter: Berlin, Germany, 2013; Volume 11.
6. Durkin, P. *The Oxford Guide to Etymology*; Oxford University Press: Oxford, UK, 2009.
7. Skeat, W.W. *An Etymological Dictionary of the English Language/Rev. Walter W. Skeat*, 4th revised, enlarged and reset. ed.; Oxford University Press: London, UK, 1910; 780p.
8. Hollmann, W.; Semantic Change. In *English Language: Description, Variation and Context*; Palgrave: London, UK, 2009.
9. Partridge, E. *Origins: A Short Etymological Dictionary of Modern English*, 4th ed.; With Numerous Revisions and Some Substantial Additions; Routledge and Kegan Paul: London, UK, 1966; 972p.
10. Khan, A.F.; Boschetti, F. Towards a Representation of Citations in Linked Data Lexical Resources. In Proceedings of the XVIII EURALEX International Congress, Ljubljana, Slovenia, 17–21 July 2018; p. 144.
11. Bronkhorst, J. Etymology and magic: Yāska's Nirukta, Plato's Cratylus, and the riddle of semantic etymologies. *Numen* **2001**, *48*, 147–203. [CrossRef]
12. McCrae, J.; Bosque-Gil, J.; Gracia, J.; Buitelaar, P.; Cimiano, P. The Ontolex-Lemon model: development and applications. In Proceedings of the eLex 2017 Conference, Leiden, The Netherlands, 19–21 September 2017; pp. 19–21.
13. Francopoulo, G.; George, M.; Calzolari, N.; Monachini, M.; Bel, N.; Pet, M.; Soria, C. Lexical markup framework (LMF). In Proceedings of the International Conference on Language Resources and Evaluation-LREC 2006, Genoa, Italy, 22–28 May 2006.
14. Salmon-Alt, S. Data structures for etymology: Towards an etymological lexical network. *Bull. Linguist. Appl. Génér.* **2006**, *31*, 1–12.
15. Bowers, J.; Romary, L. Deep encoding of etymological information in TEI. *arXiv* **2016**, arXiv:1611.10122.
16. Sagot, B. Extracting an Etymological Database from Wiktionary. In Proceedings of the Electronic Lexicography in the 21st century (eLex 2017), Leiden, The Netherlands, 19–21 September 2017; pp. 716–728.
17. De Melo, G. Etymological Wordnet: Tracing The History of Words. In Proceedings of the LREC 2014, Reykjavik, Iceland, 26–31 May 2014; pp. 1148–1154.
18. Moran, S.; Bruemmer, M. Lemon-aid: Using Lemon to aid quantitative historical linguistic analysis. In Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data, Pisa, Italy, 23 September 2013; Association for Computational Linguistics: Stroudsburg, PA, USA, 2013; pp. 28–33.
19. Chiarcos, C.; Abromeit, F.; Fäth, C.; Ionov, M. Etymology Meets Linked Data. A Case Study In Turkic. In Proceedings of the Digital Humanities 2016, Krakow, Poland, 11–16 July 2016.

20. Drummond, N.; Rector, A.L.; Stevens, R.; Moulton, G.; Horridge, M.; Wang, H.; Seidenberg, J. Putting OWL in Order: Patterns for Sequences in OWL. In Proceedings of the OWLED, Athens, GA, USA, 10–11 November 2006.