Applying Evolutionary Terminology Auditing to the Gene Ontology

Werner CEUSTERS, MD

New York State Center of Excellence in Bioinformatics and Life Sciences

University at Buffalo

Address: 701 Ellicott Street
Suite B2-160
Buffalo, NY - 14203
USA

email: Ceusters@buffalo.edu

Phone: 1-716-881-8971

(no fax)

Abstract:

Evolutionary Terminology Auditing (ETA) is a novel way to assess the quality of terminologies using reality as benchmark. The key idea is that terms added to each new version of a terminology reflect unjustified absences and terms that are deleted unjustified presences in previous versions of the terminology. The method requires that terminology authors not only keep track of changes in successive versions, but also motivate the changes introduced. In this paper, we report on how our method has been applied to the Gene Ontology (GO), a collection of three structured, controlled vocabularies for use in annotating genes, gene products and sequences. We demonstrate that even where the basic requirements for its application are only partially satisfied, the approach can still yield results which are useful for quantifying and forecasting the evolution of a terminology's quality over time.

Keywords:

terminology
evolutionary terminology auditing
realism
ontology
quality assessment
evaluation metrics

1 Introduction

Auditing, in general, is an activity conducted to verify the correctness of some sort of documentation or process. In accounting, for instance, auditing consists of reviewing the financial statements and accounts of a company for their adherence to two criteria: (1) they should contain in a specific format the data elements required by applicable laws and regulations, and (2) the data should be a reliable representation of what is the case in reality. The rationale for the first criterion is the assumption that conformance to it makes it easier to verify adherence to the second criterion. Financial statements that are 'in good shape' allow shareholders and investors to assess reliably whether the company is 'in good shape'. Moreover, several such statements, representing the financial status of the company over the years, can be used to make predictions over the financial growth of the company in the future, hence its expected shareholder value.

Just like financial statements are (or should be) a representation of the financial reality of a company, so are (or should be) biomedical terminologies representations of biomedical reality. Auditing such terminologies must thus also include the assessment of faithfulness to reality in addition to being in a format that allows inconsistencies and mistakes to be identified easily. However, although much research has been devoted to the latter, the former has thus far largely been neglected, which is the specific problem addressed in this paper. We explain how realism-based principles introduced for *ontology evolution* can be used for *terminology auditing* and demonstrate this by applying these principles to the Gene Ontology [1].

1.1 Realism-based approaches to terminology

The realist orientation in biomedical terminology is based on the view that terms in terminologies are to be aligned not on '*concepts'* but rather on entities in reality [2]. Central to this view are three assumptions. The first is that biological reality exists *objectively* in itself, i.e. independent of the perceptions or beliefs of cognitive beings. Thus not only do a wide variety of entities exist in reality (human beings, stomachs, bacteria, disorders, ...), but also how these entities relate to each other (that certain stomachs are parts of human beings, that certain bacteria cause disorders in human beings, and so forth) is not a matter of agreements made by scientists but rather of objective fact. The second assumption is that reality, including its structure, is accessible to us and can be discovered: it is scientific research that allows human beings to find out what entities exist and what relationships obtain between them. The third assumption is that an important aspect of the quality of a terminology is determined by the degree to which the structure according to which the terms of the terminology are organized *mimics* the pre-existing structure of reality, rather than being determined – and usually limited – by, for example, what the representation language is able to express [3], by mixing ontology with epistemology [4], or by incidental features related to the context in which the terminology is built, thus confusing the 'model of meaning' with the 'model of use' [5].

Realism-based terminology *development* was introduced into biomedical informatics some ten years ago as a means of detecting and avoiding the systematic mistakes characteristic of *concept-based* terminologies [3, 6-8], mistakes which are not eliminated through the use of description logics or similar computational devices [9]. The Foundational Model of Anatomy [10] and the Gene Ontology (GO) [1] were among the early adopters of a realist methodology along these lines. The methodology acquired broader acceptance after it was used to develop the Relation Ontology [11] under the auspices of the Open Biomedical Ontologies (OBO) initiative and which adopted it as a quality requirement for inclusion of any such ontology in the OBO Foundry [12].

The first ideas towards realism-based terminology *auditing*, in contrast to *development*, were proposed in 2006 as a means to assess how successive versions of terminologies and ontologies evolve over time [13]. Hence the name '*Evolutionary Terminology Auditing*' (ETA). It was first applied in a small-scale feasibility study to SNOMED CT to determine the adequacy of SNOMED CT's *history mechanism* for the treatment of the distinction between changes occurring on the side of entities in reality and changes in our understanding thereof [14]. Here we report on our experience in applying ETA to the vocabularies of the Gene Ontology.

1.2 The Gene Ontology

The Gene Ontology (GO) [1] is, in contrast to what its name suggest, not an ontology of genes, but rather a '*set of structured, controlled vocabularies for community use in annotating genes, gene* *products and sequences*' [15]. There are three vocabularies which comprise the GO, each currently independent of the others. The '*cellular component ontology*' covers sub-cellular structures and macromolecular complexes, including multi-subunit enzymes and other protein complexes, but not individual proteins or nucleic acids, nor multi-cellular anatomical structures. The '*molecular function ontology*' describes the activities, such as catalytic or binding activities, that may occur at the molecular level. The '*biological process ontology*' is designed to include terms that represent collections of processes as well as terms that represent a specific entire process, both based upon the functions ascribed to cellular components.

GO is extremely popular: to date, more than 2000 papers report on how its vocabularies have been used for a variety of purposes. In addition, GO is rapidly growing in size, and new updates are made available on a daily basis.

However, because the GO authors had '*consciously chosen to begin at the most basic level, by creating and agreeing on shared semantic concepts; that is, by defining the words that are required to describe particular domains of biology*' [16], it is no surprise that in its earlier versions it exhibits the sorts of errors manifested by other concept-based terminologies, including confusing functions with functionings (e.g. the function of an '*ATPase inhibitor*' molecule is always to inhibit ATPase, even when it is in a context where there is nothing to inhibit) [17], mixing use and mention – the term '*use/mention confusion*' denotes a well-known problem in semiotics and semantics, more precisely confusing a name with that what the name stands for – (compare '*physiological process is_a biological process*', with '*biological process part_of Gene Ontology*') [8], and, using relationships and definitions in unprincipled ways, primarily in the context of '*sensu*' terms, as in '*larval fat body development part_of larval development (sensu Insecta)*' [18, 19].

In order to prevent such errors, the GO Consortium adopted a strategy based on best practices in terminology development, thereby paying attention to high quality design principles for terms [20] and definitions [21]. In addition, more advanced computational methods for keeping the terminology internally consistent were introduced [22]. What is still lacking, however, is a quantitative approach to assessing GO's *external* consistency, i.e. how adequately it represents the portion of reality it is intended to represent.

2 Hypothesis

In [13] we argued that each time a new version of a terminology is released, or, better still, each time an individual expression is changed, added or deleted, the authors should document that change by indicating the sort of transition they assume to have been effected. We proposed a calculus based on whether such changes were motivated by (1) a change in reality, (2) a change in the terminology authors' (scientific) understanding of reality, or (3) corrections of earlier encoding mistakes. We further argued that this calculus could be used not so much to demonstrate how good an *individual* version of a terminology is, but rather to measure how much it has been *improved* (or believed to have been improved) as compared to its predecessor. We also speculated on the potential of the calculus for the assessment of the skills of terminology authors through the tracking of the history of their revisions.

The questions for which we sought answers in the work reported on here are: (1) can the approach be used in the context of terminologies that do not exhaustively keep track of the reasons why changes in the terminology are introduced and (2) is it possible to make predictions on the future quality gains of a terminology on the basis of past experience.

3 Materials and methods

3.1 Terminological conventions

Our method depends crucially on the distinction between (1) what is *inside* a terminology in contrast to (2) what is *part of* the first-order reality toward which the terminology is directed, thereby assuming that entities in (1) are *about* entities in (2) [23]. Terms in a terminology are of course as real as cellular components, biological processes and molecular functions. But since the former are *about* the latter, and the latter are *not* about anything, we will use the term '*first-order reality*' to denote the latter and their biological kin. Sometimes, the term '*domain of discourse*' is used instead, but this term does not acknowledge the first assumption of the realist agenda, i.e. that first-order reality is the way it is, independently of whether it is talked about or not.

By ‘*portion of reality*’ (PoR) we mean any *part* of reality, including the entities that exist (such as the universal HUMAN BEING, or ***Werner Ceusters***, a particular that instantiates that universal) and the relationships that obtain between them (for instance that ***Werner Ceusters’ brain*** is part of ***Werner Ceusters***). On the side of a terminology, we are – or at least we should be – dealing primarily with entities that *are about* or *denote* entities or relations in first-order reality. In line with [23], we will use the term ‘*representational unit*’, abbreviated as ‘*RU*’, for any symbolic representation (code, character string, icon, …) which denotes a portion of reality.

While in a well-ordered terminology RUs can be classified on the basis of what they *denote*, it is for some terminologies hard to fathom whether their authors consider the RUs to denote entities in first-order reality, or entities (‘*concepts’* as they would have it) inside the terminology itself [24, 25], or even whether they denote anything at all. RUs can also be classified on the basis of their *form*, for instance as *codes* (e.g. ‘*GO:0048869*’), *terms* (e.g. ‘*cellular developmental process*’), or *expressions* (e.g. ‘*GO:0042995 : cell projection ---[i] GO:0019861 : flagellum*’, which under the realist paradigm denotes the portion of reality consisting of the universal FLAGELLUM, the universal CELL PROJECTION, and the sub_kind relation that holds between them).

By convention, we will use the term ‘*term-RU*’ for representational units in a terminology that have the form of a term. This allows us then to express, for example, that the term ‘*cellular developmental process*’ is a term-RU in GO, or, in line with one of the objectives of terminology as a discipline [26], that the term ‘*developmental process*’ would not be an adequate term-RU in GO because it does not express adequately that exclusively *cellular* developmental processes are denoted by it.

3.2 Evolutionary Terminology Auditing

The third item on the realist agenda in terminology development is the requirement that the structure of a terminology should mimic the structure of the PoR that is covered by the terminology. Granular Partition Theory (GPT) provides a formal account of what it means for a structure to mimic (or not) another structure [27]. GPT allows for instance a terminology that represents whales as fish to be recognized as incorrect, where a terminology that classifies whales as animals but not as mammals, while not incorrect, still to be what GPT calls ‘*locally non-transparent*’. GPT does however not provide a means to quantify such differences, nor does it deal with issues such as whether it matters, for the purposes for which the terminology has been designed, whether whales are mammals, or what the reasons are for given sorts of mismatch. This is especially relevant in domains where our scientific understanding of reality is advancing rapidly and so that terminologies seeking to keep pace with these advances need to be updated frequently.

In [13], we built further upon GPT and developed a metric to quantify the quality of terminologies on the basis of four dimensions: (1) type of structural mismatch as defined by GPT, (2) relevance for the purposes for which the terminology is designed, and whether structural mismatches arise (3) from a wrong or incomplete scientific understanding of the relevant parts of reality, or (4) from editorial mistakes.

3.2.1. Quantification of structural mismatches regarding representational units

As shown in **Table 1**, the current version of ETA is based on 17 possible configurations of match or mismatch – 2 more than in our original proposal [13] – which are divided into two groups, labelled '*P*' and '*A*', denoting respectively the presence or absence of an RU. Each group can further be subdivided into two smaller groups on the basis of whether the presence or absence of an RU in a terminology is justified ('*P+*' and '*A+*') or unjustified ('*P-*' and '*A-*').

The configurations reflect the different kinds of mismatch between what the terminology authors *believe* to exist or to be relevant, on the one hand, and matters of *objective* existence and *objective* relevance-to-purpose on the other. The encoding of a belief can be either correct (R+) or incorrect, either (a) because the encoding does not refer (¬R) or (b) because it does refer, but to a PoR other than the one which was intended (R-). The two configurations not considered in our original proposal [13] both involve an RU that denotes an intended and objectively existing PoR that, however, is already denoted by another RU in the terminology (R++).

As an example, configuration P-1 would hold for an RU stating that 'whales are fish': the putative PoR does not exist – hence the 'N' in column (2) of **Table 1** – and therefore objective relevance does not apply, as indicated by the '-' in column (3). The authors of the terminology do however *believe* that whales are fish and consider it to be relevant; therefore this configuration is marked by the presence of ‘Y’ in both columns (4) and (5). Finally, they use the representational machinery offered by the terminology correctly such that the RU is the intended representation – note the ‘Y’ in column (6) – but this in absence of a corresponding PoR, as indicated by ‘¬R’ in column (7).

Of the 17 configurations, only 3 are desirable: P+1, which consists in the justified presence of an RU that correctly refers to a relevant PoR; and A+1 and A+2, which consist in the justified exclusion of an RU, either because there is no PoR to be referred to, or because this PoR is not relevant to the terminology’s purpose. A-3 and A-4 are borderline cases, in which errors made by terminology authors are without deleterious effect, either because something that is erroneously assumed to exist is deemed irrelevant, or because something that is truly irrelevant is overlooked.

There are eleven different kinds of ‘P’ configurations of which, interestingly, only P+1 and P-6 refer correctly to a corresponding PoR: the former reflects our ideal case for presences; the latter is marred by the incorrect inclusion of an RU which lacks relevance. P-9 and P-10 also denote an existing and intended PoR, but the mistake here is that the terminology authors are not aware of their departure from the principle that for each entity in first-order reality there should be maximally one RU of a specific form.

The last column of **Table 1** shows the magnitude of the error committed when an RU reflecting a given type of configuration is included in or left out of a terminology as measured against its corresponding ideal configuration. Because these ideal configurations are P+1, A+1, and A+2, and because for any other configuration the ‘*corresponding’* ideal configuration is the one which has the same values in columns (2) and (3), the number of mistakes committed in P-4, P-5, P-9, A-1 and A-2 need to be measured against P+1. Similarly A+1 is the ideal configuration for P-1, P-2, P-3 and A-3, and A+2 for all the others. The magnitude of an error is calculated by counting the number of differences that a specific configuration exhibits with respect to its ideal configuration in each of the columns (4) to (7) of **Table 1**, with the additional rule that a non-intended encoding which denotes an existing and thus non-intended PoR – the presence of ‘R-‘ in column (7) – counts double. This is because we judge that users of a terminology will be less likely to use RUs which denote nothing than RUs that denote non-intended PoRs: probably far more users will notice that an RU of the type ‘whales are leprechauns’ is a mistake – and thus never use that RU in some annotation – than there would be users that would notice the mistake in an RU of the type ‘whales are fish’.

3.2.2. Quantification of structural mismatches regarding whole terminologies

Theoretically, it would now be an easy exercise to assess the quality of a terminology as a whole: we would have to (1) inspect each RU in the terminology to determine what match/mismatch configuration it exhibits, and (2) examine its coverage domain to see what relevant RUs are missing. Because the magnitude of a mistake in an undesirable configuration is maximally 5, we would give each best case configuration encountered a score of 5, while each deviation there from would receive the difference between 5 and the corresponding penalty for the corresponding sort of deviant case. The total score would be the ratio of the sum of the scores obtained for each present RU, over the sum of five times the number of RUs present and 4 times the number of RUs missing. The latter is because all missing RUs have an error magnitude of 1, and 5-1=4. The general formula is:

$$\frac{\sum_{i=1}^{n}(5-e_i)}{5n+4m} \qquad (1)$$

in which e_i stands for the magnitude of the error (if any) for a given corresponding *RU*, *n* for the number of *RUs* present in the terminology and *m* for the number of RUs unjustifiably absent. Note that in this study we did not assign a higher or lower error magnitude to unjustified absences that occur at the level of leaf nodes in a terminology as compared to absences at higher levels in the hierarchy.

The score itself can be viewed as a variation to the well-known recall and precision metric, but combined in but one metric and adjusted for the magnitude of the errors committed.

Table 2 gives an example of how this metric should be applied. Imagine three terminologies that provide a vocabulary for describing whales. All three terminologies have RUs for WHALE, FISH, ANIMAL and MAMMAL, but they differ in whether whales are asserted to be (1) fish (Terminology 1 - T1), (2) animals without further specification (Terminology 2 - T2), or (3) mammals (Terminology 3 - T3). In reality, of course, whales are mammals. We further assume, for the sake of the example, that the terminology authors did not make encoding mistakes: if there is a mistake in the terminology, then it is because their scientific understanding of reality is erroneous, not because they encoded a known fact erroneously. We also assume that all PoRs in the domain are relevant to the purposes for which the terminologies are built. When we then compare the three terminologies against the benchmark of reality, the latter being expressed in column (2) of **Table 2**, we see that T1 has one erroneous RU, which is an example of a mistake of type P-1, and one unjustified absence of type A-2; T2 exhibits the same unjustified absence, but in contrast to T1 it does not include an erroneous RU; T3, finally, mimics the structure of reality completely. For each RU in each terminology, the corresponding error magnitudes, if any, are shown in columns (4), (6) and (8). Applying the formula described above, this gives a quality score for T1 of 0.84, for T2 of 0.90 and for T3 of 1.00.

Note that we took the justified absence of type A+1 (whales are fish) into account *only* because there is an RU (in T1) that posits the opposite. It is of course *not* a presupposition of our proposal that one should include all putative RUs which do not denote a corresponding PoR – e.g. that animals are fish, that animals are whales, that fish are mammals, that unicorns are leprechauns, and so forth – in any such assessment. Importantly, not doing so does not affect the magnitude of the overall score. This can be seen in relation to T2 and T3 whose quality scores are not influenced by the fact that they do not contain an erroneous RU to the effect that whales are fish. This is one of the desirable mathematical properties that our metric exhibits, of which the complete characterization, however, falls beyond the scope of this paper.

Note also that this procedure reflects what might initially appear to be an unacceptable idealization, because determining the type of configuration an (included or excluded) RU is involved in depends upon two factors – objective relevance-to-purpose, and relation to objective reality – whose assessment is something which could be correctly carried out only by someone able to adopt the perspective of a god-like observer. Less idealistically, this god-like observer might be replaced by another terminology that is used as gold standard [28], and we adopt here a generalization of this latter approach by using successive versions of a terminology as the gold standard relative to its predecessors. This is motivated, as described further in detail, by the assumption that new versions of a terminology are better than previous ones, despite the possibility that with each version new errors are introduced. But if terminology curators take their work seriously, such errors are likely to be corrected in later versions, for instance on the basis of remarks from the community when the version is used in practice. It seems obvious that using other terminologies as gold standard has at least the same risk. Furthermore, if one is sure about the correctness of another terminology covering the same domain, why should one then bother to develop a new one?

3.2.3. Quality assessment of terminologies over successive versions

The minimal requirement for releasing a terminology as expressed in terms of the realist paradigm (though independent of whether or not authors of a given terminology endorse a realist view) is that its authors should assume in good faith that all its constituent expressions are of the P+1 type (requirement R1). A stronger requirement would be that the authors advance the terminology as complete, i.e. as containing RUs designating *all* PoRs deemed relevant to its purpose (requirement R2). Successive versions of a terminology should approximate ever more closely to this latter ideal. To exploit the paradigm completely, one could even argue that it should be part of the standard terminology authoring process to document any changes made in successive versions by means of the typology described in **Table 1** [13]. This requires terminology authors to register whether or not the changes they introduced in a new version of the terminology are dictated by changes in (1) the underlying reality (requirement R3), (2) objective relevance of an included expression to the purposes of the ontology (requirement R4), (3) the ontology authors' understanding of each of these (requirement R5), and also by (4) the correction of encoding errors (requirement R6).

To see how the heuristic of using a new version of a terminology functions as surrogate for a god-like observer in relation to its predecessors, consider again the whale/fish example of **Table 2**. This time, however, we will consider T1, T2 and T3 to be versions of the same terminology, T3 being newer than T2, and T2 being newer than T1. The results of this interpretation are summarized in **Table 3**; with **Table 4** showing how the individual quality scores are calculated.

When the first version of the terminology (T1) is released, the authors assume in good faith that their work is correct, i.e. that all RUs denote the desired PoRs, and that all and only relevant RUs are present. They might believe that some RUs are missing, but of course, they have no clue which ones, otherwise they would have been included. Therefore, version T1 at time t_1 was assumed to be 'state of the art' and therefore of quality 1.00, the maximal attainable score. At time t_2, however, the authors discover that whales are not fish and they make the corresponding RU 'obsolete'. Note that making an RU obsolete by giving the reason for the change, is preferable to just removing it: if, indeed, the only change introduced between T2 and T1 would be the deletion of the RU that whales are fish, external auditors might wonder whether (1) the deletion is an omission brought about by an encoding error, in which case the RU which was believed to be of type P+1 at t_1 has to be believed to be of type P-2 at t_2, or (2) a deletion based on a conscious decision either (2a) that whales are still to be considered to be fish, but that the RU is not relevant for the purposes for which the terminology is being built, hence consisting in an A-3 type of mistake, or (2b) that the right sort of discovery was made and thus the original RU was of type P-1. Because the latter is the case, the quality score of T1 at t_1 can be recalculated according to the state of the art reached at t_2 using Eq. (1).

A similar analysis can be carried out at t_3, but now applied to both T1 and T2; in general, each new version of a terminology allows us to assess the quality of all previous versions of the terminology in light of the state of the art reached when the new version is released (see **Table 5**).

3.3 Applying Evolutionary Terminology Auditing to the Gene Ontology

3.3.1. Data preparation

In light of the above, we analysed the changes made in the GO from January 2001 until September 2007 by using the monthly reports generated by the Gene Ontology Consortium[1]. We used in our analyses the following information for each of the three GO vocabularies:

(1) the term-RUs added since the previous release including the acronym for the corresponding source (the monthly reports use the label 'database' to indicate the provenance of these term-RUs since most acronyms refer directly to source databases, e.g. 'MGI' for '*Mouse Genome Informatics*', 'FB' for '*FlyBase*', and so forth. Other acronyms, however, denote the curator that was responsible for the addition, such as 'MAH' for Midori Harris. We therefore use in this paper the term 'source' to indicate the provenance of the additions), and the lowest GO-SLIM term that subsumes the new term-RU;

[1] http://www.geneontology.org/MonthlyReports/

(2) the RUs made obsolete and the reasons provided by the GO curators for doing so;

(3) RU merges, and

(4) RU movements with respect to GO-SLIM terms.

We classified the various types of changes introduced in the GO vocabularies according to the typology outlined in **Table 1**. Because the GO authors do not give a reason when adding new RUs, we cannot know what type of unjustified absence such an addition reflects in earlier versions; we assumed them to be either A-1 or A-2, and registered such cases using the label '*A-1/2*'. Not knowing what A-type of error has been made does not matter for the calculation of quality scores since all unjustified absences have an error magnitude of 1.

Fortunately, the GO authors do in most cases give explicit reasons – expressed in free text rather than through a controlled vocabulary – for making RUs obsolete. We analysed each of these reasons manually, and classified them into the applicable match/mismatch configurations of **Table 1**. This was achieved through a step-wise process during which we grouped reasons on the basis of their similarity, including the error configuration type to which they belong. The two top levels of groupings that we developed are shown in **Table 6**, together with the error configuration types assigned to them. We used the label '*nP'* for those cases where no explicit reason was given, and assigned to them an error magnitude of 3, assuming on the basis of inspection of a sample that most reasons would be of type P-1.

The merging of two RUs into one was classified as a P-9 error committed before the merge.

3.3.2. Calculation of quality changes

We calculated several statistics for each of the three vocabularies, thereby keeping track of the provenance of the original terms. We computed by means of Eq.(1) the quality scores for each vocabulary for each monthly version, using the last version for which a monthly report was available (September 2007) as the gold standard. To make this possible, we applied a number of principles to project a change made in this last version onto an error – if any at all – in all previous versions. First, if a newly introduced RU was never made obsolete, there had to be an unjustified absence in each version prior to the addition, and a justified presence starting with the version in which the addition was introduced. Second, if an RU was found to have been made obsolete and this action was never undone, there was a justified absence both prior to the introduction of the corresponding RU and after it was made obsolete (including the version in which the RU was made obsolete), and an unjustified presence in each version that contained the RU. Finally, if a RU that was made obsolete previously was found to be re-introduced, then there must have been an unjustified absence prior to the addition, a justified presence after the addition until the RU was made obsolete, again an unjustified absence after the latter change, and finally a justified presence from the point of re-introduction onwards.

3.3.3. Forecasting

To assess whether the methodology allows making predictions about the evolution of a terminology in the future, we calculated in the manner described in the previous section the quality scores for each monthly version of the GO process vocabulary prior to December 2005, using the December 2005 version as the gold standard. We then forecasted the values for each of the variables in Eq.(1) for each (expected) monthly version between January 2006 and September 2007 by using the evolution of the known values for these variables over the period from January 2001 to December 2005, taking the number of months elapsed since the December 2000 version as independent variable. Each forecasted value was computed using Eq.(2) in which x_p is the number of months elapsed since December 2000, $\overline{x}$ the average of the number of months elapsed, and $\overline{y}$ the average of the values obtained for the corresponding variables during the period from January 2001 to December 2005.

$$y_p = \overline{y} - (\overline{x} - x_p)\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (2)$$

We then used Eq.(1) to calculate the expected quality scores for each monthly version from January 2006 to September 2007 using the values for additions, omissions and deletions forecasted by Eq.(2). As a final step, we again used Eq.(2) to assess the values for additions, omissions and deletions for the 2001-2005 period as viewed from the perspective of the forecast versions. The reason for doing this is that forecast additions of RUs at time t have to be interpreted as unjustified omissions in versions earlier than t.

4 Results

4.1 General statistics

The net size of the GO brought about by additions and deletions changed dramatically over the period studied: the cellular component vocabulary grew by 315%, the molecular function vocabulary by 245%, and the biological process vocabulary by 521%, yielding an overall growth of 362%.

The number of structural changes made (thus excluding name changes), as witnessed by **Table 7**, is even more dramatic, particularly in the process vocabulary, which accounts for 51,829 of the 66,627 changes encountered in total. Although the contribution of term-RU additions to the GO vocabularies by the different sources – labelled '*AI'*, '*CB'*, '*EF'*, and so forth in **Table 7** and **Table 9**, with the exception of '*UNK'* which stands for '*unknown'* – varies widely, the percentage of properties added or removed in the context of these term-RUs relative to all structural changes is quite similar for each source and averages to 71%. Clearly, the individual sources are only responsible for *contributing* a term-RU to the Gene Ontology, but not for structural changes made in the context of that RU. The latter is the sole responsibility of the GO editors.

Each GO-term participates on average in 2 structural changes during the period covered, but there is a large variation: 56% of the GO-terms are involved in maximally one structural change whereas an additional 31% are associated with 2, 3 or 4 changes (see **Table 8**). Note that by '*change'*, in this context, we mean any change that is *not* the addition of an RU to the vocabulary, whereas the counts given in **Table 8** do include term-RU additions. One term, '*GO:0030587: sorocarp development*', underwent 23 changes. **Table 9**, finally, displays how the structural changes were translated into error-types.

4.2 Evolution of the quality over time

Figure 1 shows the evolution of the quality scores of the three GO vocabularies as perceived from the viewpoint of the latest version analysed, i.e. September 2007. As can be expected, the quality scores increase over time and almost with each new version, although there are a few exceptions. Surprisingly, however, the evolution of the quality scores over time is different for each of the three GO vocabularies. Throughout its history, the function vocabulary exhibits the highest quality scores, with a remarkable jump upwards August 2003 which runs parallel with an important increase in its size. Increase in size, although an important *contributor* towards higher quality scores, does *not* however *guarantee* an increase in quality score: the component vocabulary, for instance, exhibits a steady increase in size between March 2003 (0.60) and December 2005 (0.82) where over the same period its quality score grows in a less marked fashion from 0.57 to 0.69.

Figure 2 shows the evolution of the quality scores of the process vocabulary for its successive versions from the perspective of a few contributing sources. Some sources have been left out of this figure to make it better readable. Term-RUs introduced through the MGI, MAH, JIC and EF sources show a quality score evolution which lags behind the evolution of the quality score for the Gene Ontology as a whole, whereas the term-RUs introduced through FlyBase and PSU exhibit since the end of 2001 a quality score which is much above the mean quality. This is because the majority of the terms introduced through the latter are of a much earlier date than the majority of the terms introduced through the former.

4.3 Forecasting

Figure 3 shows the forecasted quality scores for the process vocabulary using the period from January 2001 to December 2005 as a reference and the forecasted September 2007 version as gold standard. The visual goodness of fit with the real quality scores (the graph labelled '*Process 2007-09 view*') is remarkable, which is confirmed through the statistical correlation of 0.99. Similar results could not be obtained by using RUs related to individual contributing sources, for instance concerning FlyBase (correlation 0.90) or MGI (0.86).

5 Discussion

5.1 Related work

5.1.1. Concept-based terminology auditing

Terminology auditing is an endeavour which has been thus far conducted primarily using the concept-based approach, by means of criteria such as those put forward by Cimino [29], or by exploiting the power of description-logics and natural language understanding based algorithms (for recent reviews of the domain and some additional proposals, see for instance [30, 31] and other papers in this special issue). At first, the well-known criteria of non-vagueness (each term in a terminology should have at least one meaning) and non-ambiguity (each term should have no more than one meaning), seem to be very reasonable. When applied literally, however, they do not do justice to the fact that synonyms and homonyms are abundantly used in natural language [32]. Therefore, a common strategy is to replace in the criteria '*term'* by '*concept'*, where a '*concept'* stands for the meaning that all terms attached to it share. But, as argued by Smith [2], this does not eliminate the possibility that terms are included that rest on ontologically false beliefs, rather than denoting entities in first-order reality, which leads him to believe – and we with him – that RUs in terminologies should in every case denote universals (such as HUMAN BEING) or defined classes such as *HUMAN BEINGS older than 21* [23]. Interestingly, Cimino, in defense of his desiderata [33], agrees that '*the notion of terminologies that are limited to well-behaved universals, each one clearly understood because of its extension in reality, is appealing*', and suggests '*a path that acknowledges the importance of representing reality, as best we can know it, but accepts the need for concepts to help us, among other things, reason under uncertainty*'. He considers this a '*realistic path*' – rather than a '*realism-based*' one – and argues that in this path '*terminologies contain terms that refer to universals and to concepts, along with various names and unique identifiers for these. Sometimes, a single term will refer to an entity that has both universal and conceptual characteristics*'. But what then with the original criteria of non-vagueness and non-ambiguity? And is this then not mixing epistemology with ontology in a way that leads to problems of the sort outlined by Bodenreider *et al.* when they concluded '*… that epistemology-loaded terms are pervasive in biomedical vocabularies, that the "classes" they name often do not comply with sound classification principles, and that they are therefore likely to cause problems in the evolution and alignment of terminologies and associated ontologies*' [4] ?

Typical for the concept-based approach is its 'inward'-orientation: the rules or criteria designed to help authors make better terminologies have no other basis than the rules themselves; there is no external benchmark. As a consequence, it is very hard to use these rules in any other way than for the purpose of counting. This is witnessed by the vocabulary criteria defended in [34] and applied to the Gene Ontology as reported in [35]: of the 99 criteria deemed important, GO was found to meet 78 criteria totally, 5 partially, and 2 not at all. Furthermore, 13 criteria were found not to be applicable and 1 was not assessed. But how, we ask, do these findings correlate with quality?

As another example, Hartung and colleagues '*consider the evolution in the relative share of leaf (vs. inner) nodes, the number of relationships, the distribution of is-a, part-of and other relationships, as well as in the concept node degrees and number of paths*' [36], but they also give no further indications as to how these metrics as applied by them to the Gene Ontology and other life science terminologies, relate to quality. They recognize in their conclusion, however, opportunities for future work, more specifically that their '*analysis framework can be extended by additional types of change*' and that '*algorithms to generate annotation and ontology mappings can be extended or refined to improve their stability w.r.t. ontology evolution, e.g., by taking obsolete concepts and versioning explicitly into account*'. This is indeed the strategy that we proposed in [13] and have implemented here.

5.1.2. Concept-based ontology auditing

Closely related to *terminology auditing* is *ontology auditing*, not the least in the biomedical domain in which *formal terminologies* grew out of traditional terminologies by adding the requirement that the relationships between the representational units are to be expressed in some form of logic [37]. In [38], ontology evaluation methods were classified in four categories whether based on (1) comparing the ontology to a gold standard, (2) evaluation by humans who try to assess how well the ontology meets a set of predefined criteria, standards, or requirements, (3) involving comparisons with a data source such as a collection of documents about the domain to be covered by the ontology, and (4) using the ontology in an application and evaluating the results. Our method integrates features of techniques from all but the last category, but was nevertheless recognised as constituting a separate category on its own [39].

With respect to the first category, our method uses the last version of a terminology as the gold standard for all previous versions. It is then the evolution of the quality improvements – if any – over time that predicts the quality of the most recent version. By doing so, our method can also be viewed as using the previous versions of an ontology as the data sources to be compared to (category 3), rather than a set of documents containing texts about the domain covered by the ontology [40].

With respect to the second category, methods differ in what sort of criteria are applied, and to what precisely. Recently, a distinction has been made between *internal* and *external criteria* whereby the former are concerned with the ontologies themselves and the latter with their take-up and use within user communities, their role as standards, and embedding within business practices [41]. Internal criteria are further distinguished as bearing on several *layers*: (1) the lexical and vocabulary layer, (2) the structural and architectural layer, (3) the representational and semantic layer, (4) the data and application layer, and (5) the philosophical layer which is, surprisingly, limited to assessing whether the OntoClean method [42] is used or internal consistency checks have been applied. This distinction is typical for the computer science approach towards ontologies which pays little or no attention to whether the ontology represents reality *faithfully*. The latter, in contrast to prevailing approaches, is what drives the realism-based approach which underpins ETA. Under the framework proposed in [41] our method involves layers (2), (3) and (5) in a very specific way: it uses a *referential* rather than a *model-theoretic* semantics (level 3) which requires assessing whether the structure of the ontology mimics the structure of reality (level 2) and this under the realist agenda rather than under the conceptualist or nominalist view (level 5). The OntoClean method mentioned before exploits some features of the realist agenda, not to mimic the structure of reality as it is perceived in line with the advance of science [43] but according to what the ontology authors want the representational units to mean irrespective of what reality suggests [44].

5.2 Applicability of Evolutionary Terminology Auditing

The results obtained indicate that our method can indeed be applied to existing terminologies even if the latter do not track exhaustively the reasons for which changes are introduced when moving from one version to the next. Of course, terminology auditors could manually inspect changes made in a new version and by doing so try to assess what sort of mistake has been corrected. For very large terminologies such as the GO, however, this is hard to do, so it might be more convenient to start with the assumptions that we applied in our analysis.

A first assumption that can be made – specifically for terminologies for which new versions are created very often and of which the GO is the most conspicuous example – is that first-order reality typically does not change in relevant ways at the level of universals during the time span between two releases. Thus if an RU is introduced in a new version of a terminology, then this is *not* because some new entity came into existence, but rather because (1) an existing entity was discovered and a reference to it deemed relevant, or (2) it was already discovered but not assessed as being relevant. This is, we believe, a safe assumption in the domain of the GO: thus we do not believe that evolution has brought forth new types of cellular components, molecular functions or biological processes that were not already there before 1998, when the GO project was initiated. This assumption, in contrast, would not hold for domains that are heavily influenced by human inventions: new versions of a drug terminology will have to make reference quite often to types of molecules that did not exist before a previous release.

Another assumption that can be made is that encoding errors do not happen frequently. The sort of error we have in mind here is, for example, that in some terminology an author wants to assert the property that A has-part B, but because of inattention selects the wrong term out of the picking list which is offered to him and asserts that A has-part C; or, as another example, makes a misspelling such that what should have been 'EMG' becomes 'ECG'. Although to the best of our knowledge there has thus far not been any study reporting on the number of such encoding mistakes in terminologies, we can assume that these errors are not so frequent and that therefore mistakes of type P-2, P-3, P-4, P-5, P-7 and P-8 would be rather uncommon. This is confirmed in the context of GO by our analysis of the reasons for making RUs obsolete: we found only 25 P-2 errors and 4 P-3 errors, corresponding to, respectively, 2.5% and 0.4% of all deletions (**Table 6**) and we found no examples of the other types of erroneous encoding.

From the list of reasons for RU-deletion as explicitly given by the GO curators, very strong arguments in favour of ETA can be derived: most reasons given do indeed correspond directly with one or other representation/reality mismatch as categorized in **Table 1** and quantified in **Table 6**. Examples are: '*the function it represents does not exist*' (non-existence, P-1), and '*2,4-dichlorophenoxyacetic acid is not synthesized by living organisms and GO does not cover non-biological processes*' (relevancy error, P-6). On the other hand, analyzing this list revealed that our original proposal in [13] was not sufficiently discriminatory, which led to the addition of P-9 and P-10 errors.

5.3 Interpretation of quality scores over successive versions

There is a high correlation (0.95) between the increase in size of the GO as a whole and the quality scores as viewed from the perspective of September 2007. Correlations for the three vocabularies differ slightly: 0.96 for the function vocabulary, 0.95 for the process vocabulary and 0.92 for the component vocabulary. This is, at first sight, surprising since our metric is not directly based on the size of a terminology – by '*size'* we mean here the number of term-RUs – but rather on the amount of changes introduced, both in terms of term-RUs and representational units for properties. Over the period studied, only 8.47% of term-RUs disappeared, either through deletions or mergers, the function vocabulary exhibiting the highest turn-over (13.57%) followed by the component vocabulary (10.94%) and the process vocabulary (5.67%). Representational units for properties change far more often than the addition or deletion of term-RUs, and changes are also more often undone: in the process and component vocabularies 28.94% and 27.88% of property changes were deletions, whereas in the function vocabulary this was 52.11%. Whereas adding or deleting term-RUs can be expected to go hand in hand with corresponding changes in referring to properties, this is not always the case: the component vocabulary for instance exhibits a steady increase in size from March 2003 until December 2005 without a similar increase in quality score. This is because of a major change in the component vocabulary's structure: whereas in December 2005 only 26 term-RUs were added, 825 changes in property-RUs were introduced.

Adding term-RUs to the GO vocabularies, although a continuous process, happens in bursts of which the size, for instance for the process vocabulary, varies between 3 and 952 per month (average 151, mode 118, standard deviation 180). Furthermore, term-RU additions to a specific version often come in groups originating from a specific source: 75% of the FlyBase term-RUs in the process vocabulary were added in September and October 2001 with these term-RUs accounting for 90% of the term-RU additions for these months; 64% of the term-RUs coming from Mouse Genome Informatics were added in September 2006, accounting for 89% of the term-RU additions in that version. These data explain why forecasting the quality scores computed over the RUs related to term-RUs coming exclusively from specific sources does not produce as accurate results as those obtained when forecasting over the process vocabulary as a whole. They explain also why the evolutions of the quality scores computed over the term-RUs related to individual sources such as FlyBase and Mouse Genome Informatics differ considerably from the overall quality score for the GO as a whole.

5.4 Outstanding issues, limitations and future work

The Gene Ontology vocabularies allowed us to test the applicability of Evolutionary Terminology Auditing (ETA) but the results obtained raise some further questions both concerning the auditing methodology itself and the ontology authoring process adhered to by the Gene Ontology editors.

With respect to the former, our findings seem to suggest that the impact on the overall evolutionary quality is more significant for term-RU additions than for property-RU changes. However, this might be the result of calculating the property changes with respect to GO-SLIM terms only – indeed, only changes with respect to GO-SLIM terms are covered in the GO's monthly reports which formed the basis of our analysis – and not with respect to the complete hierarchy of each full vocabulary itself.

Another area for further research is the calculation of the error magnitude associated with each type of mistake with respect to its base line (column 8 in **Table 1**). The current method uses the same error magnitude for each type of unjustified absence in a previous version of the terminology. This turns out to be an advantage in the context of GO because (1) its authors do not supply explicit reasons for adding term-RUs and because the magnitudes for unjustified absences are equal, it does not matter that no reason is given, and (2) we can assume that new additions all relate to the discovery of PoRs which existed already before the development of GO was started. But the current method does not take into account at what level in the classification hierarchy the mistake is made. This applies not only to omissions but also to unjustified additions and property-RU changes. Therefore, another strategy, to be tested in the future, is to base the magnitude of the error also on the difference in the hierarchical position of an RU in an older version as compared to the newer one. We expect a metric based on the information content of an RU [45] rather than on path length differences to be more promising, especially in terminologies that exhibit a large number of references to compositional classes which lead to an artificial increase in path length and which do not mimic the structure of reality. In GO, 6% of the deletions involved the removal of such references.

An additional metric to be considered is the '*life expectancy*' or '*survival*' of an RU over the history of a terminology. Although in GO the ratio of term-RU deletions relative to term-RU additions is rather low (less than 3%), the amount of RUs referring to properties that were made obsolete is high: in the function vocabulary, there are as many deletions as additions and because in this portion of our work thus far we have worked with GO-SLIM terms only, this ratio is probably seriously underestimated. However, again because no explicit reasons for changes in GO, other than deletions, are given, it is thus far not possible to track whether RUs referring to specific properties have just been *deleted*, or rather *replaced*.

6 Conclusion

Evolutionary terminology auditing is based on determining how successive versions of a terminology do a better job in mimicking the structure of reality. It is a novel technique of which the foundations were outlined in [13] by distinguishing on a theoretical basis 15 types of ways in which representational units may or may not correspond to portions of reality. In [14], using SNOMED as an example, it was demonstrated by means of an exploratory analysis that the majority of these types of mismatches actually do occur. The work on the Gene Ontology on which we report here consists of the first systematic application of the theory. Not only did we identify the need for two more types of mismatches, we were also able to demonstrate that the approach is feasible and that it allows for quantifying, even forecasting, the quality of a terminology. To be maximally beneficial, however, it requires not only a metric that is more sensitive to the types of changes introduced in successive versions of a terminology, but also that terminology authors provide greater insight into the underlying reasons for the changes they introduced and that they do this in a way that supports computation. To make that possible, terminology authoring systems should offer facilities to register and quantify such changes in a formal way and to apply one or more metrics of the sort described above.

[1] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature Genetics. 2000;25(1):25-9.

[2] Smith B. From Concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies. Journal of Biomedical Informatics. 2006;39(3):288-98.

[3] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: Where do they come from and how can they be detected? In: Pisanelli DM, ed. *Ontologies in Medicine Studies in Health Technology and Informatics*. Amsterdam, The Netherlands: IOS Press 2004:145-64.

[4] Bodenreider O, Smith B, Burgun A. The ontology-epistemology divide: A case study in medical terminology. *Formal Ontology and Information Systems* 2004:185-95.

[5] Rector AL, Qamar R, Marley T. Binding Ontologies & Coding Systems to Electronic Health Records and Messages. In: Bodenreider O, ed. *Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation: "Biomedical Ontology in Action" (KR-MED 2006)*. Baltimore 2006.

[6] Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of Information in Medicine. 2005;44:498-507.

[7] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT®. In: Fieschi M, Coiera E, Li Y-CJ, eds. *MEDINFO 2004*. Amsterdam, The Netherlands: IOS Press 2004:482-6.

[8] Johansson I. Bioinformatics and Biological Reality. Journal of Biomedical Informatics. 2006;39(3):274-87.

[9] Ceusters W, Smith B. Ontology and Medical Terminology: why Descriptions Logics are not enough. *Towards an Electronic Patient Record (TEPR 2003)*. San Antonio 2003.

[10] Rosse C, Kumar A, Mejino JL Jr, Cook D, Detwiler L, Smith B. A strategy for improving and integrating biomedical ontologies. *AMIA Annu Symp Proc* 2005:639-43.

[11] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biology. 2005;6(5):R46.

[12] Smith B, Ashburner M, Ceusters W, Goldberg L, Mungall C, Shah N, et al. The OBO Foundry: Remolding Biomedical Ontologies to Support Data Integration. Nature Biotechnology. 2007;25:1251-5.

[13] Ceusters W, Smith B. A Realism-Based Approach to the Evolution of Biomedical Ontologies. *Proceedings of AMIA 2006* 2006:121-5.

[14] Ceusters W, Spackman KA, Smith B. Would SNOMED CT benefit from Realism-Based Ontology Evolution? In: Teich J, Suermondt J, Hripcsak C, editors. American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy; 2007 November 10-14; Chicago IL: American Medical Informatics Association; 2007. p. 105-9.

[15] Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Res. 2006 January;34:D322–D6.

[16] Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. Genome Research. 2001;11(8):1425-33.

[17] Smith B, Williams J, Schulze-Kremer S. The Ontology of the Gene Ontology. *AMIA Annual Symposium Proceedings* 2003:609-13.

[18] Smith B, Köhler J, Kumar A. On the application of formal principles to life science data: A case study in the Gene Ontology. *Data Integration in the Life Sciences (DILS) 2004* 2004:79-94.

[19] Smith B, Kumar A. On controlled vocabularies in bioinformatics: A case study in the Gene Ontology. BioSilico: Drug Discovery Today. 2004;2:246-52.

[20] Gene Ontology Consortium. The Gene Ontology: editorial style guide. 2008 March 25, 2008 [cited 2008 May 8]; Available from: http://www.geneontology.org/GO.usage.shtml?all

[21] Gene Ontology Consortium. The Gene Ontology: Logical Definitions. 2008 April 30, 2008 [cited 2008 May 5]; Available from: http://wiki.geneontology.org/index.php/Logical_Definitions

[22] Aranguren ME, Wroe C, Goble C, Stevens R. In situ migration of handcrafted ontologies to reason-able forms. Data & Knowledge Engineering. 2008 (in press).

[23] Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. *KR-MED 2006, Biomedical Ontology in Action*. Baltimore MD, USA 2006.

[24] Smith B, Ceusters W, Temmerman R. Wüsteria. In: Engelbrecht R, Geissbuhler A, Lovis C, Mihalas G, eds. *Connecting Medical Informatics and Bio-Informatics Medical Informatics Europe 2005*. Amsterdam: IOS Press 2005:647-52.

[25] Smith B. Beyond concepts: ontology as reality representation. *Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)*. Amsterdam: IOS Press 2004:73-84.

[26] Sager JC. A Practical Course in Terminology Processing. Amsterdam: John Benjamins Publishing Company. 1990.

[27] Bittner T, Smith B. A Theory of Granular Partitions. In: Duckham M, Goodchild MF, Worboy MF, eds. *Foundations of Geographic Information Science*. London: Taylor & Francis Books 2003:117-51.

[28] Ceusters W. Towards A Realism-Based Metric for Quality Assurance in Ontology Matching. In: Bennett B, Fellbaum C, eds. *Formal Ontology in Information Systems*. Amsterdam: IOS Press 2006:321-32.

[29] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods of Information in Medicine. 1998;37(4-5):394-403.

[30] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as Part of the Terminology Design Life Cycle. Journal of the American Medical Informatics Association. 2006;13:676-90.

[31] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. A Model for Evaluating Interface Terminologies. Journal of the American Medical Informatics Association. 2008;15:65-76.

[32] Baud R, Ceusters W, Ruch P, Rassinoux A-M, Lovis C, Geissbühler A. Reconciliation of Ontology and Terminology to cope with Linguistics. In: Kuhn K, Warren J, Leong T, eds. *Proceedings of MEDINFO 2007, Brisbane, Australia, August 2007*. Amsterdam: Ios Press 2007:796-801.

[33] Cimino JJ. In Defense of the desiderata. Journal of Biomedical Informatics. 2006;39(3):299-306.

[34] Hayamizu T, Ringwald M, De Coronado S, Davis B, Keller M, Komatsoulis G, et al. Evaluation of Vocabulary Review Criteria: Final Report to V/CDE WS; 2006 12/14/2006.

[35] Hayamizu T, Ringwald M, De Coronado S, Davis B, Keller M, Komatsoulis G, et al. Evaluation of the Gene Ontology (GO) using the proposed V/CDE WS Vocabulary Review Criteria; 2007 4/05/2007.

[36] Hartung M, Kirsten T, Rahm E. Analyzing the Evolution of Life Science Ontologies and Mappings. Leipzig: Interdisciplinary Centre for Bioinformatics; 2008.

[37] Ceusters W. Formal terminology management for language-based knowledge systems: resistance is futile. In: Temmerman R, Lutjeharms M, eds. *Trends in Special Language and Language Technology*. Antwerpen: Uitgeverij De Boeck 2001:135-53.

[38] Brank J, Grobelnik M, Mladenić D. A survey of ontology evaluation techniques. *SIKDD 2005*. Ljubljana, Slovenia 2005.

[39] Obrst L, Ceusters W, Mani I, Ray S, Smith B. The Evaluation of Ontologies: toward Improved Semantic Interoperability. In: Baker CJO, Cheung K-H, eds. *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. Heidelberg: Springer 2007:139-58.

[40] Brewster C, Alani H, Dasmahapatra S, Wilks Y. Data Driven Ontology Evaluation. *International Conference on Language Resources and Evaluation*. Lisbon, Portugal: European Language Resources Association 2004.

[41] Kehagias DD, Papadimitriou I, Hois J, Tzovaras D, Bateman J. A Methodological Approach for Ontology Evaluation and Refinement. *ASK-IT Final Conference*. Nuremberg, Germany 2008.

[42] Guarino N, Welty C. Evaluating ontological decisions with OntoClean. Communications of the ACM. 2002;45(2):61-5.

[43] Smith B. Ontology (Science). In: Eschenbach C, Grüninger M, eds. *Formal Ontology in Information Systems - Proceedings of the Fifth International Conference (FOIS 2008)*. Amsterdam: IOS Press 2008:21-35.

[44] Welty C. OntOWLClean: Cleaning OWL ontologies with OWL. In: Bennett B, Fellbaum C, eds. *Formal Ontology in Information Systems : Proceedings of the Fourth International Conference (FOIS 2006)*. Amsterdam: IOS Press 2006:347-59.

[45] Van Buggenhout C, Ceusters W. A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology. International Journal of Medical Informatics. 2005;74(2-4):125-32.

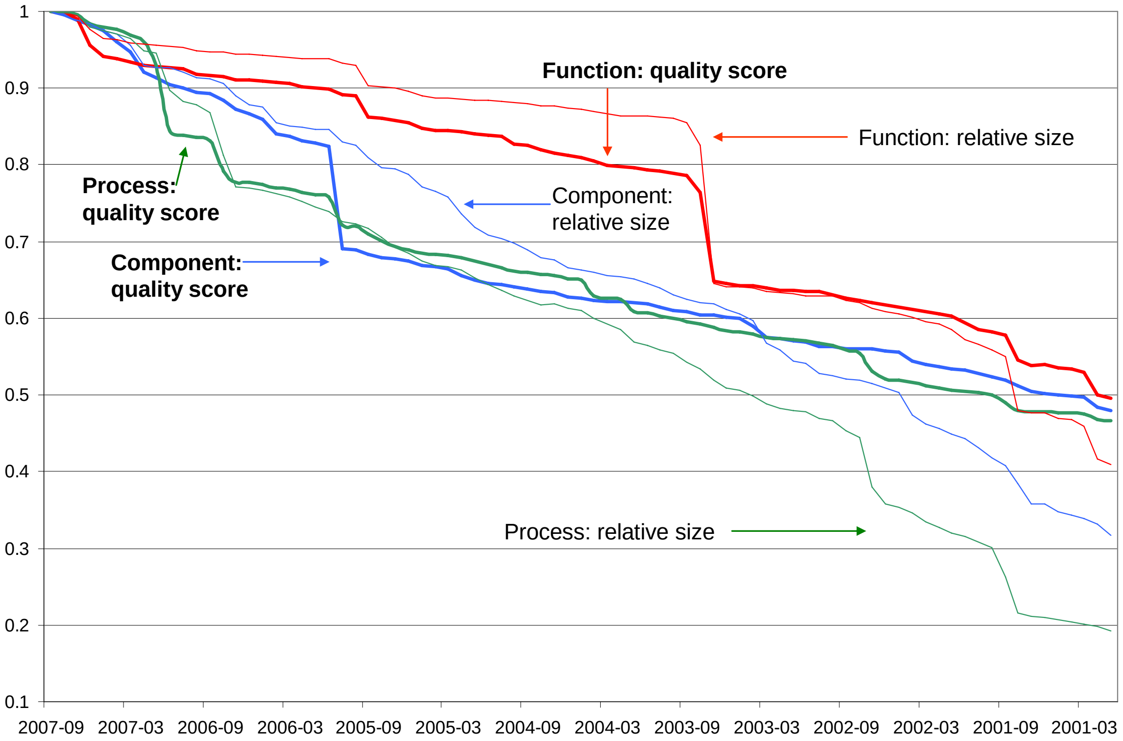

Figure 1: Comparison of the realism-based quality scores and the relative size of the three Gene Ontology Vocabularies measured over time

Figure 2: Quality score changes in the process vocabulary as a whole and in a few contributing sources.

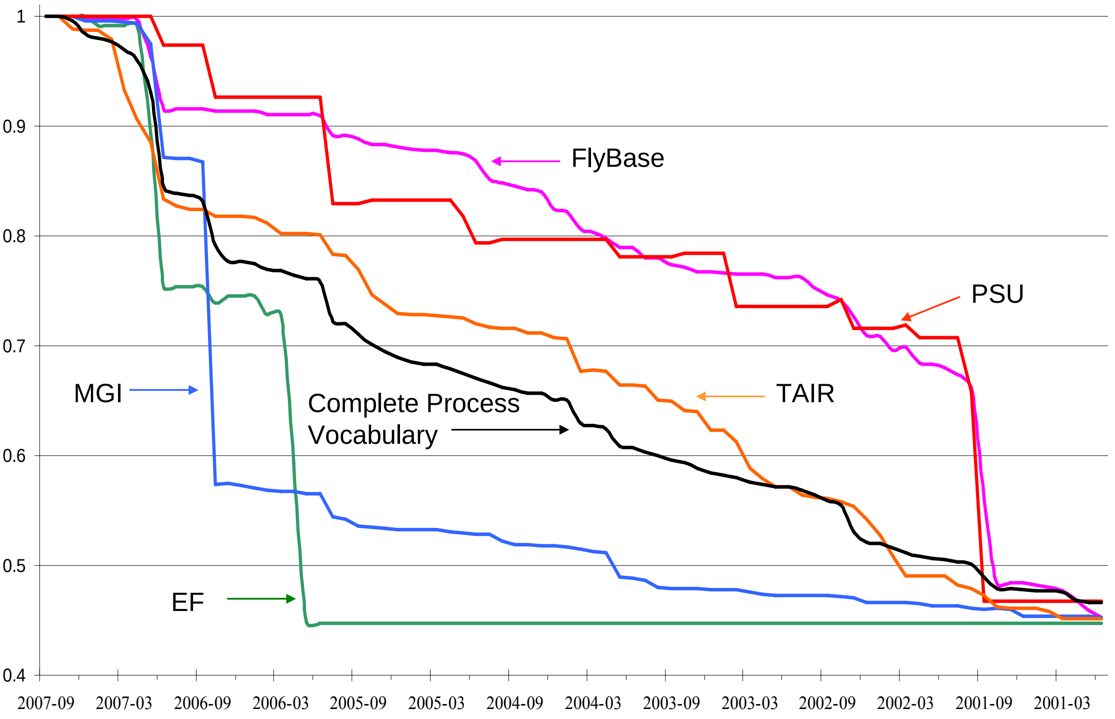

Figure 3: Forecasted quality scores for GO's process vocabulary computed using the versions from January 2001 to December 2005 as reference.

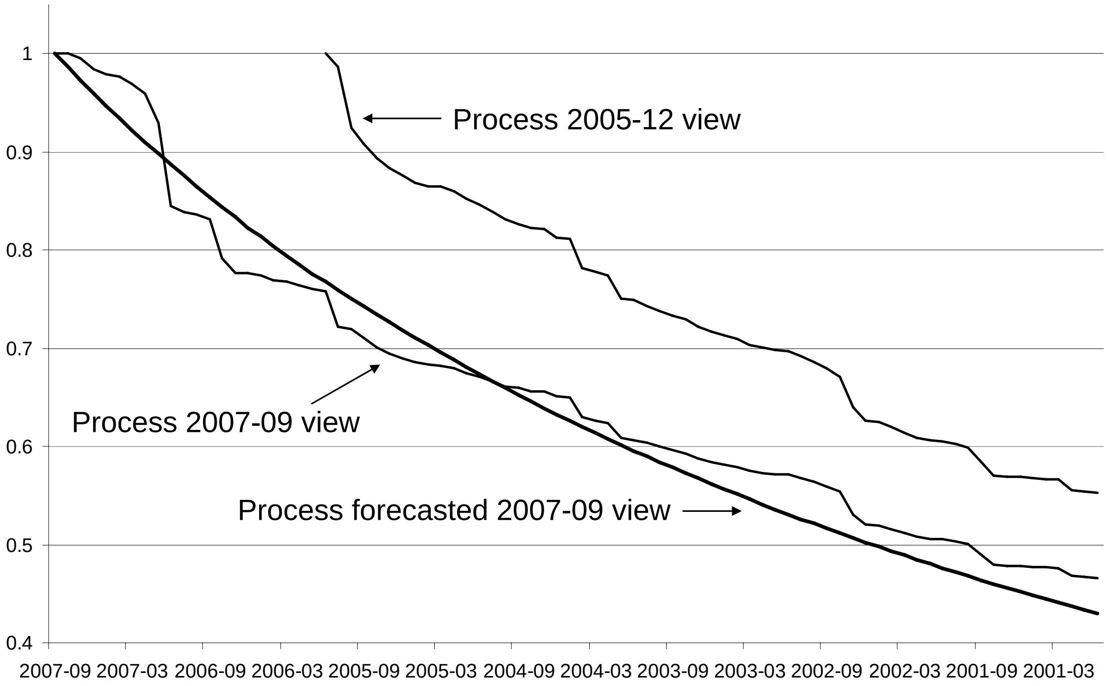

Table 1: Typology of expressions included in and excluded from an ontology in light of relevance and relation to external reality

Configuration	Reality		Representation				Magnitude of error
			Authors' Belief		Encoding		
	Objective Existence	Objective Relevance	In existence	In relevance	Intended encoding	Type of reference	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
P+1	**Y**	Y	**Y**	Y	Y	R+	0
A+1	**N**	-	**N**	-	-	-	0
A+2	**Y**	N	**Y**	N	-	-	0
P-1	**N**	-	**Y**	Y	Y	¬R	3
P-2	**N**	-	**Y**	Y	N	¬R	4
P-3	**N**	-	**Y**	Y	N	R-	5
P-4	**Y**	Y	**Y**	Y	N	¬R	1
P-5	**Y**	Y	**Y**	Y	N	R-	2
P-6	**Y**	N	**Y**	Y	Y	R+	1
P-7	**Y**	N	**Y**	Y	N	¬R	2
P-8	**Y**	N	**Y**	Y	N	R-	3
P-9	**Y**	Y	**Y**	Y	Y	R++	1
P-10	**Y**	N	**Y**	Y	Y	R++	2
A-1	**Y**	Y	**Y**	N	-	-	1
A-2	**Y**	Y	**N**	-	-	-	1
A-3	**N**	-	**Y**	N	-	-	1
A-4	**Y**	N	**N**	-	-	-	1

Table 2: Scoring the quality of terminologies using reality as benchmark

	Reality	**Terminology 1**		**Terminology 2**		**Terminology 3**	
RU(1)	**Config.** **(2)**	**Config.** **(3)**	**Error** **(4)**	**Config.** **(5)**	**Error** **(6)**	**Config.** **(7)**	**Error** **(8)**
animal	P+1	P+1	0	P+1	0	P+1	0
fish	P+1	P+1	0	P+1	0	P+1	0
whale	P+1	P+1	0	P+1	0	P+1	0
mammal	P+1	P+1	0	P+1	0	P+1	0
fish are animals	P+1	P+1	0	P+1	0	P+1	0
mammals are animals	P+1	P+1	0	P+1	0	P+1	0
whales are fish	A+1	P-1	3	A+1	0	A+1	0
whales are animals	P+1	P+1	0	P+1	0	P+1	0
whales are mammals	P+1	A-2	1	A-2	1	P+1	0
SCORE	8*5/ ((8*5)+(0*4)) = 1.00	((7*5)+(1*2))/ ((8*5)+(1*4)) =0.84		7*5/ ((7*5)+(1*4)) =0.90		8*5/ ((8*5)+(0*4)) =1.00	

Table 3: Scoring the quality of terminologies using new versions

	Time t1		Time t2				Time t3					
	T1		T1		T2		T1		T2		T3	
	C.	E.	C.	E.	C.	E.	C.	E.	C.	E.	C.	E.
animal	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0
fish	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0
whale	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0
mammal	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0
fish are animals	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0
mammals are animals	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0
whales are fish	P+1	0	P-1	3	A+1	0	P-1	3	A+1	0	A+1	0
whales are animals	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0	P+1	0
whales are mammals	-	-	-	-	-	-	A-2	1	A-2	1	P+1	0
SCORE	1.00		0.93		1.00		0.84		0.90		1.00	

Table 4: Calculation of quality scores for terminology versions at different times

Terminology	Time of assessment	Formula for quality score	Quality Score
T1	t1	(8*5)/(8*5)	1.00
	t2	((7*5)+(1*2))/(8*5)	0.93
	t3	((7*5)+(1*2))/((8*5)+(1*4))	0.84
T2	t2	(7*5)/(7*5)	1.00
	t3	(7*5)/((7*5)+(1*4))	0.90
T3	t3	(8*5)/((8*5)+(0*4))	1.00

Table 5: Views on the quality of a terminology through successive versions

	Time		
Terminology version	**t1**	**t2**	**t3**
T1	1.00	0.93	0.84
T2	-	1.00	0.90
T3	-	-	1.00

Table 6: Error types for terms made obsolete and classification of motivations

Error type	Level 1 reason	Level 2 reason	N
nP (N=167)	No reason given.		101
	RU's intended referent unclear	RU with ambiguous definition	4
		RU with inaccurate name	20
		RU with non-sensical name	3
		RU without definition	8
		other ambiguities	31
P-1 (N=536)	False belief in existence		4
	RU does not denote anything existing		38
	Wrong property ascription	RU denotes a biological process rather than a function	48
		RU denotes a cellular component rather than a function	14
		RU denotes a gene product <class> rather than a function.	67
		RU denotes a gene product rather than a function.	144
		RU denotes a gene product rather than a process.	30
		RU denotes a molecular function rather than a process.	50
		RU denotes a multifunctional gene product rather than a function.	26
		RU denotes a protein <class> rather than a function.	14
		RU denotes a single gene product and not a complex.	40
		RU denotes more than one molecular function rather than a component.	1
		RU does not denote a biological process.	14
		RU does not denote a cellular component.	2
		RU does not denote a molecular function.	44
P-2 (N=25)	RU with wrong definition	RU with too specific definition for the name	8
		RU with wrong definition (explicitly stated as such)	17
P-3 (N=4)	RU denotes a non-intended entity	RU changed referent	1
		RU denotes a non-intended entity	3
P-6 (N=237)	Irrelevant RU	Defined class irrelevant	23
		entity is not synthesized by living organisms, and GO does not cover non-biological processes.	33
		GO restructuring	1
		Referent outside the scope of GO	2
		RU contains info from more than one ontology.	45
		RU does not denote a single biological process.	4
		RU with multiple referents	3
		Specification in RU's name not needed	1
		RU denotes a gene product <class>.	1
		RU denotes a gene product.	7
		RU denotes a mereological sum	1
		RU denotes a phenotype	29
		RU denotes a substrate-specific process.	5
		RU denotes irrelevant compositional class	2
		RU denotes irrelevant compositional class - and/or	27
		RU denotes irrelevant compositional class - other	32
	RU denotes more than one molecular function.		21
P-9	RU denotes same entity as another RU		15
P-10	RU with erroneous 4D-view on continuant		10

Table 7: Number of changes made to the GO vocabularies from Jan 2001 to September 2007

		Source															
		AI	CB	EF	FB	JIC	JL	MAH	MGI	OS	PSU	SGD	TAIR	TIGR	UNK	WB	**Total**
Component	term-RU Added	143		1	170	33	263	684	45	11	25	124	49	7			**1,555**
	Deletion reversed																
	Merged	5			8	2	7	17				3	3		38		**83**
	Property Added	142			213	36	149	641	28		47	125	41	12	861		**2,295**
	Property Removed	30			103	13	17	140	22		25	15	12	2	508		**887**
	Made obsolete				5	2	2	8	13				2		76		**108**
	Grand Total	320		1	499	86	438	1,490	108	11	97	267	107	21	1,483		**4,928**
	% of non term-RU changes	55%		0%	66%	62%	40%	54%	58%	0%	74%	54%	54%	67%	100%		**68%**
Function	term-RU Added	2,166	4		1,408	124	482	899	156	4	1	54	114				**5,412**
	Deletion reversed				1		2		1								**4**
	Merged	29			71	3	7	10	7			2	3		167		**299**
	Property Added	149	4		245	21	55	86	47			15	10		1,094		**1,726**
	Property Removed	142			264	5	43	53	13			14	5		1,339		**1,878**
	Made obsolete	15			90	2	13	15	6				1		409		**551**
	Grand Total	2,501	8		2,079	155	602	1,063	230	4	1	85	133		3,009		**9,870**
	% of non term-RU changes	13%	50%		32%	20%	20%	15%	32%	0%	0%	36%	14%		100%		**45%**
Process	term-RU Added	3,614	16	73	1,693	1,406	1,260	2,010	1,137	35	10	262	638	4		43	**12,201**
	Deletion reversed	1															**1**
	Merged	43	2	1	45	71	49	34	16	1	2	2	12		118	2	**398**
	Property Added	7,299	16	138	3,151	2,294	2,678	2,960	1,913	25	24	684	1,216	9	5,148	85	**27,640**
	Property Removed	2,349	8	16	1,553	631	1,087	1,102	970	6	21	333	466	3	2,668	41	**11,254**
	Made obsolete	75			36	3	4	28	4			2	13		170		**335**
	Grand Total	13,381	42	228	6,478	4,405	5,078	6,134	4,040	67	57	1,283	2,345	16	8,104	171	**51,829**
	% of non term-RU changes	73%	62%	68%	74%	68%	75%	67%	72%	48%	82%	80%	73%	75%	100%	75%	**76%**
Total	**Grand Total**	16,202	50	229	9,056	4,646	6,118	8,687	4,378	82	155	1,635	2,585	37	12,596	171	**66,627**
	% of non term-RU changes	63%	60%	68%	64%	66%	67%	59%	69%	39%	77%	73%	69%	70%	100%	75%	**71%**

Table 8: Distribution of term-RUs in function of the number of changes

Count of Changes	Changes per Representational Unit																				
Vocabulary	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**24**	**Grand Total**
Component	678	667	343	169	99	59	23	17	2		2	1	1								**2,061**
Function	5,029	755	449	208	100	74	14	9	3		1										**6,642**
Process	3,129	2,554	2,704	1,649	1,396	907	530	404	397	186	117	73	73	27	17	2	5	8	3	1	**14,182**
Grand Total	8,836	3,976	3,496	2,026	1,595	1,040	567	430	402	186	120	74	74	27	17	2	5	8	3	1	**22,885**
No. of changes	**8,836**	**7,952**	**10,488**	**8,104**	**7,975**	**6,240**	**3,969**	**3,440**	**3,618**	**1,860**	**1,320**	**888**	**962**	**378**	**255**	**32**	**85**	**144**	**57**	**24**	**66,627**

Table 9: Distribution of error types over the sources that contributed term-RUs to the Gene Ontology

Count of Error	Source														
Error	AI	CB	EF	FB	JIC	JL	MAH	MGI	OS	PSU	SGD	TAIR	TIGR	UNK	WB	Grand Total
A-1	1	0	0	1	0	2	0	1	0	0	0	0	0	0	0	5
A-1/2	13,513	40	212	6,880	3,914	4,887	7,280	3,326	75	107	1,264	2,068	32	7,103	128	50,829
nP	6	0	0	10	2	3	10	4	0	0	1	3	0	128	0	167
P-1	2,543	8	16	1,984	653	1,156	1,326	1,021	6	46	362	487	5	4,901	41	14,555
P-10	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	10
P-2	1	0	0	2	0	2	3	0	0	0	0	1	0	16	0	25
P-6	58	0	0	52	1	5	6	3	0	0	1	7	0	104	0	237
P-9	80	2	1	126	76	63	62	23	1	2	7	19	0	331	2	795
P-3	0	0	0	1	0	0	0	0	0	0	0	0	0	3	0	4
Grand Total	16,202	50	229	9,056	4,646	6,118	8,687	4,378	82	155	1,635	2,585	37	12,596	171	66,627