Analyzing the Evolution of Life Science Ontologies and Mappings

Michael Hartung[†], Toralf Kirsten[†], Erhard Rahm[†‡]

{hartung, tkirsten}@izbi.uni-leipzig.de, rahm@informatik.uni-leipzig.de
[†]Interdisciplinary Center for Bioinformatics, University of Leipzig
[‡]Dept. of Computer Science, University of Leipzig

Abstract. Ontologies are heavily developed and used in life sciences and undergo continuous changes. However, the evolution of life science ontologies and references to them (e.g., annotations) is not well understood and has received little attention so far. We therefore propose a generic framework for analyzing both the evolution of ontologies and the evolution of ontology-related mappings, in particular annotations referring to ontologies and similarity (match) mappings between ontologies. We use our framework for an extensive comparative evaluation of evolution measures for 16 life science ontologies. Moreover, we analyze the evolution of annotation mappings and ontology mappings for the Gene Ontology.

Keywords: Ontology evolution, ontology matching, mapping evolution

1 Introduction

Ontologies become increasingly important in life sciences. Usually, they provide a harmonized vocabulary describing and structuring a specific domain of interest, e.g., molecular functions of proteins or the anatomy of a species. The vocabulary consists of concepts, which are typically structured within trees or acyclic graphs where the concept nodes are interconnected by "*is-a*" and "*part-of*" relationships. Biological objects, such as genes and proteins, can be semantically and uniformly described or annotated by ontologies by associating them with the respective ontology concepts. For example, proteins are associated to concepts of the Gene Ontology to describe their protein functions and to specify processes they are involved in. The proliferation of ontologies has also generated interest in interrelating different ontologies by so called ontology mappings [1,2,7], e.g., to see which molecular functions are involved in which biological processes or which functions are localized on which cellular component.

Due to the rapid development of life science research we observe that ontologies evolve continuously, i.e., they are frequently changed to incorporate new domain knowledge into them. Typical ontology modifications include the addition of new concepts and new relationships or the deletion of outdated concepts and relationships. To still provide some stability for applications and users of ontologies, the ontology developers typically support a version concept. An ontology version represents the state of the ontology at a specific point in time (release date). While older ontology versions remain stable (unchanged), a new ontology version may reflect an arbitrary number of changes. However, these changes, e.g., deletions, may impair the correctness of previous use cases of the ontology within annotations or ontology mappings. Hence, annotations and ontology mappings affected by ontology changes may have to be identified and corrected. Furthermore, new knowledge represented by added concepts and added relationships should be utilized as quickly as possible.

So far, the evolution of life science ontologies and change impact for annotations and ontology mappings has received almost no attention and is therefore not well understood. As a first step in dealing with ontology evolution in life sciences we therefore propose to analyse how existing ontologies evolve, e.g., to answer immediate questions such as "How volatile (stable) are different ontologies?" "What is the frequency of different types of modifications?" and "Which structural changes occur within ontologies?". Furthermore, we want to analyze the consequences of ontology changes, e.g., to what degree do they imply changes of ontology-based annotation and previously determined ontology mappings.

To that end, we make the following contributions in this paper:

- We propose a generic framework allowing us to systematically study the evolution of ontologies and instance data sources (e.g., representing biological objects such as proteins), as well as the evolution of ontology-related mappings, i.e., annotation mappings and ontology mappings. The framework supports the computation of several general measures to describe individual ontology versions and mappings as well as their evolution.
- In a comprehensive evaluation, we apply the framework to 386 versions of 16 life science ontologies including the sub-ontologies of Gene Ontology and the NCI (National Cancer Institute) thesaurus. In particular, we use the proposed framework measures to analyze the major change types and other evolution characteristics.
- We further evaluate the evolution of annotation mappings and correlate between changes of instances/ontologies and the ontology-based annotations. Furthermore, we analyze the impact of ontology evolution to differently generated ontology mappings.

The analysis results are expected to be helpful for both ontology developers and ontology users to better understand the consequences of ontology changes. Furthermore, the results may help guide the development of algorithms to generate mappings that remain comparatively robust against ontology changes.

The rest of the paper is organized as follows. In Section 2 we introduce a general framework to measure different types of evolutionary changes of ontologies, their associations to biological objects and on interconnecting ontology mappings. In Section 3 we apply the framework and show results for a selected set of life science ontologies whereas Section 4 illustrates the evolution results of protein objects established ontology mappings we observed. Section 5 discusses related work. We finally conclude and outline future work.

2 Evolution and Measurement Framework

Our evolution framework distinguishes between two basic types of evolution as illustrated in Figure 1. On the one side, we investigate the evolution in single *sources*, specifically *ontologies* (1) and *instance sources* (2). For both source types, the evolution is reflected in a series of versions. On the other side, we consider the evolution of *mappings*. Such mappings exist between versions of different instance sources (*instance-instance-mapping* (3)), between versions of instance sources and ontologies (*annotation mapping* (4)) and between versions of different ontologies (*ontology mapping* (5)). In the following we define the models and measures of our framework. A simple example (Figure 2) will illustrate these models and their evolution.

2.1 Framework models

2.1.1 Ontology model

An ontology $ON_v = (C, R, t)$ is defined by its name *ON*, a version number *v*, *concepts* $C = \{c_1, ..., c_m\}$, *relationships* $R = \{r_1, ...,r_n\}$ and a creation timestamp *t*. Concepts represent entities of the domain to be modeled; they are interconnected by the relationships in *R,* e.g., is-a and part-of relationships. Concepts with no relationships to any super concept act as the *roots* $\subseteq C$ of ON_v. Together, *C*, *R* and the *roots* form the ontology's graph structure which is assumed to be a directed acyclic graph (DAG).

A concept can have a varying number of *attributes*. Typical attributes in biomedical ontologies are accession ID, concept name, concept synonyms, concept definition, and obsolete status. In our evolution framework we heavily take into account accession ID and obsolete status information. The accession IDs unambiguously identify concepts and can be used to determine new and deleted concepts when comparing different versions of an ontology. Furthermore, these IDs are used within annotation and ontology mappings. The obsolete status is not generally supported but allows the specification of outdated concepts which may still be in use but should not be used anymore for new applications.

R defines directed binary relationships between concepts. We distinguish between three types of relationships, namely *is-a* (R_{is_a}), *part-of* (R_{part_of}) and *miscellaneous* (R_{mis}). As we will see, is-a and part-of relationships are the most common relationship types in biomedical ontologies. Other ("miscellaneous") relationship types are specific to ontologies of a certain domain, e.g., anatomy, chemistry or molecular biology.

2.1.2 Instance model

An instance source $IS_v = (I, t)$ of version number *v* consists of a set of *instances* $I = \{i_1, ..., i_n\}$, e.g., molecular biological objects such as genes or proteins, and a creation

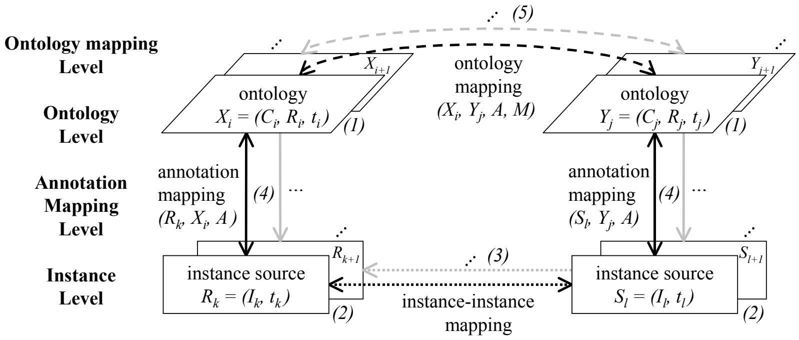


Figure 1: Evolution of sources (1, 2) and mappings (3, 4, 5)

timestamp *t.* Instances are described by a set of attributes including an accession ID attribute and *IS*-specific attributes. The ID attribute is used in mappings between different instance sources (instance-instance mapping) and in annotation mappings.

2.1.3 Annotation mapping model

An annotation mapping $AM = (IS_u, ON_v, A)$ describes a mapping between an instance source *IS* of version *u* and an ontology *ON* of version *v*. The mapping itself, denoted by *A*, is a set of binary associations between instances *I* of IS_u and concepts *C* of ON_v. A single association or correspondence $a_j = (i_j, c_j) \in A$ annotates an instance item $i_j \in I$ with an ontology concept $c_j \in C$. Note that annotation mappings are (implicitly) versioned by the use of versioned instance sources and versioned ontologies. Hence, the combination of the version numbers *u* and *v* can be thought of as the version number of the mapping.

2.1.4 Ontology mapping model

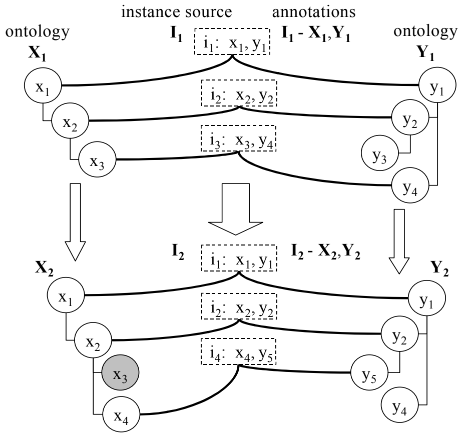
We define an ontology mapping $OM = (X_u, Y_v, A, M)$ between two different ontology versions X_u and Y_v as a set of correspondences *A* based on a match algorithm *M.* A single correspondence $n_k = (x_k, y_k, sim_k) \in A$ comprises two ontology concepts (concept x_k of X_u, concept y_k of Y_v) and a similarity value sim_k. The similarity value indicates the strength of similarity between two ontology concepts and is typically a numerical value from the interval [0,1]. Similarity values are determined by an ontology match algorithm *M.* For example, metadata-based matching algorithms use metadata for matching such as concept names and often apply string similarity measures to estimate the similarity of ontology concepts. On the other hand, instance-based matchers may consider the number of shared instances, i.e., instances associated to both ontology concepts, to compute a similarity value [7].

Similar to annotation mappings, ontology mappings are implicitly versioned by the use of versioned ontologies.

2.1.5 Common evolution model

In order to analyze the evolution of single sources and of mappings, we define a generic evolution model that is applicable to all defined models, in particular ontologies, instances, annotations and ontology mappings. The basis of our evolution model are *object sets* O_{vi} of a version v_i of a source that evolves. Possible objects are ontology concepts or relationships (ontology evolution), instance data (instance evolution), annotation associations (annotation mapping evolution) and ontology correspondences (ontology mapping evolution).

We focus on three change operations that may occur during evolution: *add*, *delete* and *toObs*. Whereas *add* is used to insert new objects in a source or mapping, the *delete* operation directly removes objects which are outdated or no longer required. *ToObs* is a special operation preferentially used in ontologies to mark objects as obsolete. In contrast to *delete*, obsolete objects remain in an evolved source. For simplicity and to preserve the applicability of our evolution model to both ontologies and mappings, we do not consider more complex evolution operations in this study, e.g., moves of concepts within is-a /part-of hierarchies or changes of relationship types.



source	*add*	*del*	*toObs*
ontology X	$\{x_4\}$	{}	$\{x_3\}$
ontology Y	$\{y_5\}$	$\{y_3\}$	{}
instance source I	$\{i_4\}$	$\{i_3\}$	{}
annotation mapping I-X	$\{i_4\text{-}x_4\}$	$\{i_3\text{-}x_3\}$	{}
annotation mapping I-Y	$\{i_4\text{-}y_5\}$	$\{i_3\text{-}y_4\}$	{}
ontology mapping X-Y	$\{x_4\text{-}y_5\}$	$\{x_3\text{-}y_4\}$	{}

Figure 2: Evolution example with ontologies (X,Y), instance sources (I), annotation mappings (I-X,Y) and an ontology mapping (X-Y)

To quantify the evolution behavior, for each change operation we determine the sets of affected objects in the considered source and mapping versions:

- $add_{vi,vj} = O_{vj} / O_{vi}$: *added objects* between version v_i and v_j
- $del_{vi,vj} = O_{vj} / O_{vi}$: *deleted objects* between version v_i and v_j
- $toObs_{vi,vj} = O_{vj,obs} \cap O_{vi,nonObs}$: *objects* that were marked as *obsolete* between version v_i and $v_{j.}$ Here, the subsets $O_{vi,nonObs}$ and $O_{vi,obs}$ are used to distinguish between normal and obsolete objects in a version v_i, together they form the set of all objects O_{vi} in version v_i.

These sets can be quite easily determined for existing ontologies, instance sources, and mappings by analyzing and comparing the accession attributes of objects. For example, if an object ID is present in a newer version of a source and not in the older one, we assign this object to the *add* set, and vice versa for the *delete* set.

A simple yet comprehensive example for both ontology evolution and mapping evolution is shown in Figure 2. The example captures the evolution of two ontologies X (X_1 to X_2) and Y (Y_1 to Y_2), the evolution of one instance source I (I_1 to I_2), the evolution of two annotation mappings I-X (I_1-X_1 to I_2-X_2) and I-Y (I_1-Y_1 to I_2-Y_2), and the evolution of one ontology mapping X-Y (X_1-Y_1 to X_2-Y_2). So in ontology version X_2 there is one new concept, x_4, while concept x_3 has been declared as obsolete. For x_4, there is a new instance annotation (i_4-x_4) as well as a new ontology correspondence (x_4-y_5). For x_3, the previous instance annotation i_3-x_3 and ontology correspondence x_3-y_4 have been deleted in the new mappings.

2.2 Framework measures

Based on the introduced framework, we determine a variety of statistical measures on the investigated sources (ontologies, instance sources) and mappings, as well as on their evolution and growth characteristics. We first present the source- and mapping-specific measures, followed by the evolution and growth measures.

2.2.1 Descriptive statistics for sources and mappings

For all kinds of object sets (instances, concepts, relationships, correspondences), we consider their cardinality in a given version of an instance source, ontology or mapping. For ontologies, we additionally determine structural characteristics such as the used relationship types (is-a, part-of), concept types (obsolete or non-obsolete, leaf or inner concepts), in-degrees and out-degrees, as well as the number of paths and path lengths:

$\lvert O_{vi} \rvert$	number of *objects* in version v_i of a source or mapping $O \in$ {ontology concepts C, ontology relationships R, instance data I, annotation mapping A, ontology mapping A}
$\lvert C_{leaf} \rvert$, $\lvert C_{inner} \rvert$	number of leaf and inner concepts
$\lvert C_{obs} \rvert$, $\lvert C_{nonObs} \rvert$	number of obsolete and non obsolete concepts
$\lvert R_{is\_a} \rvert$, $\lvert R_{part\_of} \rvert$, $\lvert R_{mis} \rvert$	number of is-a, part-of or miscellaneous relationships
$\varnothing d_{in} = \lvert C_{inner} \rvert / (\lvert R_{is\_a} \rvert + \lvert R_{part\_of} \rvert)$	average in-degree of inner concepts
$\varnothing d_{out} = \lvert C \rvert / (\lvert R_{is\_a} \rvert + \lvert R_{part\_of} \rvert)$	average out-degree of concepts
$\varnothing ppc$, $\varnothing ppl$	average number of paths per concept or per leaf concept (path as way to a root concept using is-a or part-of relationships)
$\varnothing pl$, $\varnothing pl_{leaf}$	average path length of all concepts or leaf concepts

For mappings, let $X_{A,u} \subseteq X_u$ and $Y_{A,v} \subseteq Y_v$ be two object sets of version u and v, such that a mapping A interrelates each element of $X_{A,u}$ with at least one element of $Y_{A,v}$ and each element of $Y_{A,v}$ have at least one counterpart in $X_{A,u}$. Then, we can determine the relative coverage of X_u and Y_v for mapping A by $X_{A,u}$ and $Y_{A,v}$, respectively, i.e., the fraction of objects of X_u (Y_v) for which at least one counterpart (and thus correspondence) in mapping A exists.

$cov_{A,Xu} = \lvert X_{A,u} \rvert / \lvert X_u \rvert$ $cov_{A,Yv} = \lvert Y_{A,v} \rvert / \lvert Y_v \rvert$	relative coverage of objects X_u and Y_v by the mapping A

2.2.2 Evolution and growth statistics

Our measures make use of the generic evolution model to compute evolution statistics for all evolution types (ontologies, instance sources, mappings). To determine the number of changes or changed objects we either directly compare two versions v_i and v_j of a source or mapping. Alternatively, we quantify the changes with respect to a certain time interval, e.g., for an entire observation period p or a regular time interval t within p, e.g., per month or per year.

$Add_{vi,vj} = \lvert add_{vi,vj} \rvert$	number of *added objects* between version v_i and v_j
$Del_{vi,vj} = \lvert del_{vi,vj} \rvert$	number of *deleted objects* between version v_i and v_j
$Obs_{vi,vj} = \lvert toObs_{vi,vj} \rvert$	number of objects that changed to obsolete between version v_i and v_j
$Add_{p,t}$ $Del_{p,t}$ $Obs_{p,t}$	average number of *added / deleted / obsolete* objects per time interval t within p

Based on these basic frequencies we determine relative fractions of newly added and deleted objects as well as an *add-delete ratio* (adr) between two versions. Further, we quantify relative fractions relating to a certain time interval t within a period p:

$adr_{vi,vj} = Add_{vi,vj} / (Del_{vi,vj} + Obs_{vi,vj})$	add-delete ratio for changes between version v_i and v_j
$add-frac_{vi,vj} = \frac{Add_{vi,vj}}{\lvert O_{vj} \rvert}$	fraction of objects in version v_j that have been added between version v_i *and* v_j
$del-frac_{vj,vi} = \frac{Del_{vi,vj}}{\lvert O_{vi} \rvert}$	fraction of objects in version v_i that have been deleted between version v_i *and* v_j
$obs-frac_{vj,vi} = \frac{Obs_{vi,vj}}{\lvert O_{vi} \rvert}$	fraction of objects in version v_i that have been marked as obsolete between version v_i and v_j
$add\text{-}frac_{p,t}$ $del\text{-}frac_{p,t}$ $obs\text{-}frac_{p,t}$	average fractions of *added* / *deleted* / *obsolete* objects per time interval t within p based on the version-related *frac* measures

We further define *growth rates*

$$growth_{O,vi,vj} = |O_{vj}| \,/\, |O_{vi}| \in [0,\infty] \subseteq R$$

for most of the measures above as the ratio between the objects O (O∈{ontology concepts C, ontology relationships R, instance data I, annotation mapping A, ontology mapping A}) of version v_j and v_i. The growth rate describes an increase when the rate is greater than 1, a decrease when the rate is less than 1 or no change for $growth_{vi,vj}=1$. Moreover, the growth rate can also be determined for relative measures, such as fractions or coverages, e.g., an increase from 50% to 60% for the ontology coverage between two versions of an ontology mapping corresponds to a growth rate of 1.2.

3 Analysis of Ontology Evolution

We study the evolution of ontologies of different life science domains, ranging from popular Gene Ontology (GO) [3] and NCI Thesaurus [12] to more specific ontologies of the OBO foundry [16], e.g., SequenceOntology or ZebrafishAnatomy. In order to comparatively analyze these ontologies, we set up a central repository with a generic schema suitable for management of heterogeneous ontologies and their versions. Overall, we integrated 386 versions of 16 currently developed life science ontologies.

In the following, we first give an overview of the analyzed ontologies and their versions. We then use the introduced measures to study the evolution behavior of the ontologies including structural ontology changes. Exemplary evolution trend charts for GO Biological Processes and Molecular Functions will be presented in Section 4.2. Detailed information and evolution trend charts for all analyzed ontologies can be found in [5] and online (http://dbs.uni-leipzig.de/ls_ontology_evolution).

3.1 Overview and versioning aspects

Table 1 lists the ontologies and gives details about their size, the number of versions during the observation period, the growth ratio as well as domain and use characteristics. For clarity, we group the analyzed ontologies into 3 groups (*large*, *medium*, *small*) based on their current number of concepts |*C*|. Our evaluation considers ontology versions for an observation period of 45 months, from May 2004 until Feb. 2008. The timestamps t_{start} (t_{last}) of the first (latest) version and the number of versions (*k*) provide information about the versioning rate of an ontology, i.e., how often an ontology releases versions and how long they are actively used. While some ontologies, particularly the Gene Ontology, currently release versions every day we consider at most one version per month (for several versions per month, we pick the first one). We observe that the oldest and most frequently released ontologies are the two largest ontologies, NCI Thesaurus and Gene Ontology. Other ontologies such as FlyBaseCV or CellType have not been updated since a longer period (6-8 months) which may indicate that these ontologies have reached a near-final state. The average number of versions per ontology is 25, i.e., a version is typically current for less than 2 months.

In terms of number of concepts, we observe a considerable growth during the observation period. On average, the number of concepts has increased by 60% during the last 45 months; the maximum (minimum) growth rate is 4.22 (1.02). The largest ontology, NCI Thesaurus has increased its size by 80% to almost 64,000 concepts.

Ontology	*size*	$\lvert C\rvert_{start}$	$\lvert C\rvert_{last}$	$grow_{\lvert C\rvert, start, last}$	t_{start}	t_{last}	*k*	*characteristics, domain and use*
NCI Thesaurus	large	35,814	63,924	1.78	May. 04	Dec. 07	39	broad coverage of cancer domain
GeneOntology		17,368	25,995	1.50	May. 04	Feb. 08	44	aggregation of all GO sub ontologies
-- Biological Process		8,625	15,001	1.74	May. 04	Feb. 08	44	annotation of gene products (biological role)
-- Molecular Function		7,336	8,818	1.20	May. 04	Feb. 08	44	annotation of gene products (molecular function)
-- Cellular Components		1,407	2,176	1.55	May. 04	Feb. 08	44	annotation of gene products (cellular location)
ChemicalEntities		10,236	18,007	1.76	Oct. 04	Jan. 08	28	chemical compounds of biological relevance
FlyAnatomy	medium	6,090	6,222	1.02	Nov. 04	Dec. 07	16	anatomy of Drosophila melanogaster
MammalianPhenotype		4,175	6,077	1.46	Aug. 05	Jan. 08	15	terms for annotating mammalian phenotypic data
AdultMouseAnatomy		2,416	2,745	1.14	Aug. 05	Sep. 07	15	adult anatomy of the mouse (Mus)
ZebrafishAnatomy		1,389	2,172	1.56	Nov. 05	Oct. 07	12	anatomy and development of the Zebrafish
Sequence		981	1,463	1.49	Aug. 05	Feb. 08	26	structured CV for sequence annotation
ProteinModification		1,074	1,128	1.05	Jun. 06	Nov. 07	14	description of protein chemical modifications
CellType	small	687	857	1.25	Jun. 04	Jun. 07	19	cell types from prokaryotes to mammals
PlantStructure		681	835	1.23	Jul. 05	Feb. 08	22	plant morphological and anatomical structures
ProteinProteinInteraction		194	819	4.22	Aug. 05	Feb. 08	19	annotation of protein interaction experiments
FlyBaseCV		658	693	1.05	Nov. 05	Apr. 07	7	used for various aspects of annotation by FlyBase
Pathway		427	593	1.39	Nov. 05	Jan. 08	22	CV for pathways, annotation of gene products
Overall		***82,190***	***131,530***	***1.60***			***386***	

Table 1: Overview and versioning statistics of analyzed ontologies

Size categories - small: $|C| < 1000$, medium: $1000 < |C| < 10000$, large: $|C| > 10000$

The largest and fastest growing GO subontology is Biological Processes (74% increase); on the other hand, the number of Molecular Functions concepts has merely increased by 20% during the observation period.

Table 2 shows more detailed and time-normalized statistics on the evolution behavior of the considered ontologies. In particular, it indicates the average number of newly added, deleted and obsolete concepts *per month* for both the entire observation period and the last year only. In addition, the relative fractions of concepts are specified which are added, deleted or declared obsolete per month.

We observe that the largest ontologies experience the highest numbers in changes. On average, they have approx. 360 (25) additions (deletions) per month compared to approx. 86 (6) additions (deletions) in all analyzed ontologies. Furthermore, the study shows that additions are the dominant change operation for all ontologies. Still, some ontologies experience a significant number of deletions, e.g., ChemicalEntities and Gene Ontology. The add-delete ratio (*adr*) indicates the relative frequency of these two main change types. NCI Thesaurus has the maximal value of 42, indicating that there are 42 times more additions than deletions or new obsolete cases. On the other hand, for ChemicalEntities this ratio is merely 4, i.e., about 20% of the changes are deletes. The relative change fractions reveal that some small and medium ontologies have high evolution rates. In terms of additions, ProteinProteinInteraction has the highest relative change frequency (2.7% new concepts per month).

Another interesting observation is the usage of the obsolete paradigm in different ontologies. Some ontology designers do not mark outdated ontology concepts as obsolete, but strictly delete them, e.g., ChemicalEntities or AdultMouseAnatomy. Most ontologies (13 of 16) follow a hybrid approach, i.e., they use both to-obsolete and delete operations. Some ontologies (NCI Thesaurus, MammalianPhenotypes), perform few deletes but primarily use the obsolete status to mark outdated concepts.

Comparing the evolution rates of the last year with the ones of the overall observation period allows us to see recent evolution trends for the different ontologies. A first group of ontologies exhibits high evolution rates in both periods, e.g., NCI Thesaurus, GO or MammalianPhenotype. This indicates that the knowledge in the domains of these ontologies is continuously evolving and that these ontologies refer to active

	Full period (May. 04 - Feb. 08)							Last year (Feb. 07 - Feb. 08)		
Ontology	*Add*	*Del*	*Obs*	*adr*	*add-frac*	*del-frac*	*obs-frac*	*Add*	*Del*	*Obs*
NCI Thesaurus	627	2	12	42.4	1.3%	0.0%	0.0%	416	0	5
GeneOntology	200	12	4	12.2	0.9%	0.1%	0.0%	222	20	5
-- Biological Process	146	7	2	16.2	1.2%	0.1%	0.0%	133	10	2
-- Molecular Function	36	3	2	6.8	0.4%	0.0%	0.0%	69	7	3
-- Cellular Components	18	2	0	8.9	1.0%	0.1%	0.0%	19	3	0
ChemicalEntities	256	62	0	4.1	1.8%	0.5%	0.0%	384	67	0
FlyAnatomy	5	1	1	3.3	0.1%	0.0%	0.0%	6	0	0
MammalianPhenotype	65	2	9	6.0	1.2%	0.0%	0.2%	74	2	3
AdultMouseAnatomy	11	0	0	30.9	0.4%	0.0%	0.0%	1	0	0
ZebrafishAnatomy	33	5	1	5.5	1.8%	0.3%	0.1%	45	2	1
Sequence	19	3	2	4.1	1.5%	0.3%	0.2%	19	0	0
ProteinModification	5	2	1	1.5	0.4%	0.2%	0.1%	7	0	2
CellType	5	1	0	2.8	0.7%	0.2%	0.1%	1	0	0
PlantStructure	5	0	1	6.1	0.7%	0.0%	0.1%	3	0	0
ProteinProteinInteraction	21	0	0	41.7	2.7%	0.0%	0.2%	4	0	0
FlyBaseCV	1	0	1	2.1	0.2%	0.0%	0.1%	0	0	0
Pathway	7	1	0	7.9	1.3%	0.2%	0.0%	6	2	0

Table 2: Evolution of analyzed life science ontologies (interval *t* = 1 month)

research fields. A second group of ontologies has considerably higher evolution rates in the last year indicating an increased research activity in the respective domains, e.g., for ChemicalEntities or GO Molecular Function. Finally, we discover ontologies with few changes in the recent past, e.g., AdultMouseAnatomy, CellType or FlyBaseCV. Work on these ontologies may have almost been finished so that rather stable ontology versions can be used.

3.2 Influence of evolution on ontology structures

Due to space limitations, we analyze the evolution of structural properties only for the largest ontologies. Table 3 summarizes structural measures for the first and last version of the considered 6 ontologies as well as the resulting growth rates (lower third of the table). We 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.

We observe that for the considered ontologies, the majority of concepts is represented by leaf nodes, i.e., these concepts are not further refined by is-a or part-of relationships. However, the relative share of leaf nodes has reduced during the observation period from about 70% to 67% indicating a corresponding increase of inner nodes and in structured knowledge. For one ontology, GO Biological Process, there are now even fewer leaf concepts (46%) than inner concepts due to a strong decline in the fraction of leaf nodes ("growth" rate 0.89).

The number of relationships increased similarly or faster than the number of concepts (Table 1) during the observation period. The largest increase occurred for ChemicalEntities (growth factor 2.7 for relationships vs. 1.76 for concepts). The considered ontologies are dominated by is-a relationships (ca. 91% of all relationships), while part-of (4%) and miscellaneous (5%) relationships are similarly infrequent[1]. Some ontologies are pure is-a hierarchies, e.g., NCI Thesaurus, GO Molecular Func-

[1] With respect to all 16 ontologies, the relative shares for is-a / part-of / miscellaneous relationships are 86% / 7% / 7%.

	Ontology	$\|C_{leaf}\|$ (%)	$\|R\|$	$\|R_{isa}\|$ (%)	$\|R_{partof}\|$ (%)	$\|R_{mis}\|$ (%)	$\varnothing\ d_{out}$	$\varnothing\ d_{in}$	$\varnothing\ pl_{leaf}$	$\varnothing\ ppl$
First version	***NCI Thesaurus***	79	41,281	100			1.2	5.6	8.2	3.3
	GeneOntology	66	23,589	88	12		1.4	4.0	7.3	3.7
	-- Biological Process	52	13,358	85	15		1.5	3.2	8.0	7.1
	-- Molecular Function	82	8,459	100			1.2	6.4	5.3	1.4
	-- Cellular Components	67	1,772	52	48		1.3	3.8	5.5	1.8
	ChemicalEntities	70	11,593	100			1.1	3.8	8.3	2.3
	MammalianPhenotype	68	4,620	100			1.1	3.4	5.5	1.5
Last version	***NCI Thesaurus***	79	72,466	100			1.1	5.4	8.0	3.0
	GeneOntology	60	41,396	88	12		1.6	3.8	8.6	22.9
	-- Biological Process	46	27,141	84	16		1.8	3.3	8.8	38.7
	-- Molecular Function	81	10,195	100			1.2	5.9	6.2	1.7
	-- Cellular Components	64	4,060	79	21		1.9	5.0	8.3	52.6
	ChemicalEntities	69	31,233	76	1	23	1.4	4.3	12	18.6
	MammalianPhenotype	64	6,875	100			1.2	3.1	7.5	2.5
Growth	***NCI Thesaurus***	1.00	1.8	1.00			1.0	1.0	1.0	0.9
	GeneOntology	0.91	1.8	1.00	1.03		1.2	1.0	1.2	6.2
	-- Biological Process	0.89	2.0	0.99	1.06		1.2	1.0	1.1	5.5
	-- Molecular Function	1.00	1.2	1.00			1.0	0.9	1.2	1.2
	-- Cellular Components	0.95	2.3	1.51	0.44		1.5	1.3	1.5	28.7
	ChemicalEntities	0.99	2.7	0.76	undef.	undef.	1.3	1.1	1.4	8.0
	MammalianPhenotype	0.95	1.5	1.00			1.1	0.9	1.4	1.7

Table 3: Changes in ontology structures

tion or MammalianPhenotype. The biggest changes occurred for ChemicalEntities which started as a pure is-a ontology but introduced part-of and other relationship types in recent versions. We also observe interesting differences between the GO subontologies. While Molecular Function only uses is-a relationships, Biological Process and Cellular Components contain both is-a and part-of relationships. However, the relative share of part-of evolved differently: Biological Process now relies more on part-of than in the beginning (growth: 1.06) while Cellular Components has a sharp relative reduction for part-of (0.44).

With respect to the in-degrees and out-degrees of concept nodes we notice little changes during the observation period, especially for is-a ontologies. The out-degrees of these ontologies is typically lower than 1.2, i.e., their concepts have mostly only one super concept. On the other side, ontologies such as GO Cellular Components or GO Biological Process have about two ancestor concepts per concept since they use is-a and part-of relationships in combination. Lastly, we look at the evolution of path lengths and number of paths in leaf concepts. We notice that except NCI Thesaurus all ontologies increased in their average path length of leaf concepts (up to 50%). The number of paths per leaf ($\varnothing\ ppl$) heavily increased, especially for ontologies which are not limited to is-a relationships (average growth rate: 14). The highest growth rate (28) occurred for the GO Cellular Components which apparently experienced a major restructuring as already observed for the development of is-a vs. part-of relationships.

4 Evolution of Annotation and Ontology Mappings

In further experiments we studied the evolution of the annotation and ontology mappings. We start with a short overview of the scenario we used in the evaluation before we describe the obtained results.

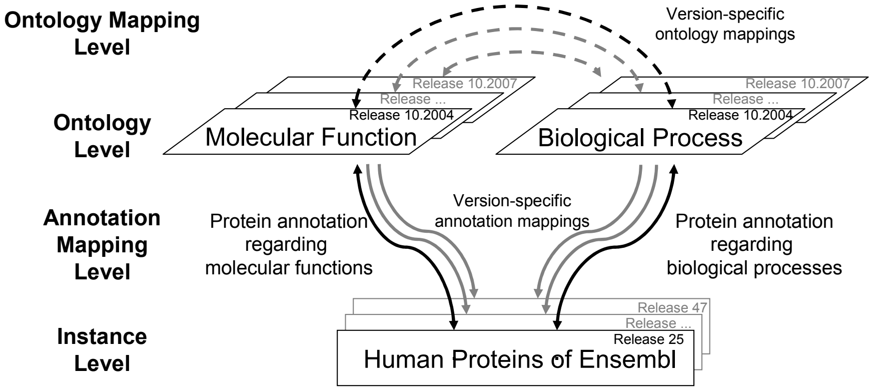


Figure 3: Overview about the evaluation scenario

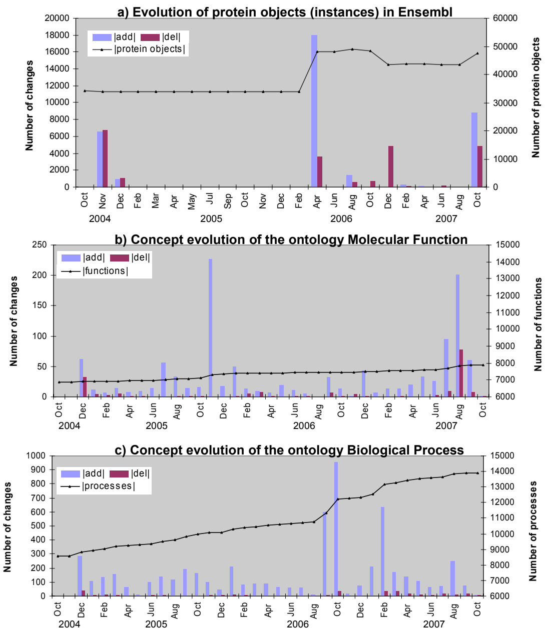
4.1 Evaluation scenario

Figure 3 shows a schematic overview of the evaluation scenario. To reduce the complexity we focus on two ontologies, namely the GO subontologies Molecular Functions and Biological Processes. Both ontologies are usually used to describe properties of proteins, i.e., the function and process concepts of the ontologies are associated with proteins. We therefore evaluate protein instances, namely protein objects of the human species available in the data source Ensembl [6]. Furthermore, we analyze the annotation mappings, as provided by Ensembl, between these proteins and the two ontologies. To interrelate the two ontologies, we determine different ontology mappings using either metadata-based or instance-based match algorithms. We will give some more details below.

4.2 Evolution of instance source vs. ontologies

The Ensembl instances and annotations as well as the two ontologies underlie frequent changes. The evaluation scenario includes 23 versions of Ensembl from Oct. 2004 to Oct. 2007 (36 months). Table 4 shows the Ensembl release numbers together with their release month and year. While in 2004 and 2005 the Ensembl releases appeared irregularly, since 2006 a new Ensembl release is created every two months. The Ensembl information is heavily based on the genome assemblies made public by NCBI; since 2004, three such assemblies (namely 34, 35, and 36) have appeared. Moreover, Table 4 also shows which GO releases have been used for the annotation mappings provided in Ensembl. As one can see, the annotations typically do not refer to the most recent but an older GO version. For example, the annotation mapping in Ensembl release 37 (Feb. 2006) refers to the GO version of March 2005, i.e., there is a time delay of 11 months. The delay has been reduced in recent Ensembl versions.

Figure 4a illustrates the evolution of protein objects (total number, number of added and deleted instances) in Ensembl from Oct. 2004 to Oct. 2007. We observe that a new genome assembly (Nov. 2004, Apr. 2006) led to massive changes of protein objects. The change from version 34 to 35 of the genome assembly caused many



a) Evolution of protein objects (instances) in Ensembl

b) Concept evolution of the ontology Molecular Function

c) Concept evolution of the ontology Biological Process

Figure 4: Evolution of instance data and ontologies

Time		Ensembl Release	NCBI Genome	Used GO Release
2004	Oct.	25	34	02.2004
	Nov.	26	35	
	Dec.	27		09.2004
2005	Feb.	28		
	Mar.	29		
	Apr	30		
	May	31		03.2005
	July	32		
	Sep.	33		
	Oct.	34		
	Nov.	35		
	Dec.	36		
2006	Feb.	37		
	Apr.	38	36	03.2006
	June	39		
	Aug.	40		
	Oct.	41		
	Dec.	42		09.2006
2007	Feb.	43		
	Apr.	44		03.2007
	June	45		05.2007
	Aug.	46		06.2007
	Oct.	47		

Table 4: Release states of protein objects in Ensembl

protein additions and deletions while the total number of proteins remained almost unchanged. However, the change from version 35 to 36 (April 2006), resulted in five times more added than deleted proteins and a corresponding jump in the total number of proteins (about 14,000 more proteins). Recently, there are more changes on protein objects during the utilization of genome assembly 36.

For comparison, Figures 4b and 4c show the evolution of the two considered ontologies during the observation period since Oct. 2004. In contrast to the irregular evolution pattern of Ensembl, we observe that both ontologies experience a continuous evolution with added and deleted/to-obsolete concepts, despite the existence of several peaks in the number of changes. With respect to the growth in the number of objects, the Molecular Function (MF) ontology evolved the least (growth 1.2) and slower than the number of protein instances (growth 1.39 for the entire observation period). The fastest growth is observed for the Biological Processes (BP) ontology (1.74). Furthermore, there are primarily additions and few deletes for the ontologies (add-delete ratios of about 7 and 16 for MF and BP, respectively) while there is significant delete activity for the protein instances (add-delete ratio of 1.6).

4.3 Evolution on annotation mapping level

In this analysis we focus on the evolution of the two Ensembl annotation mappings proteins-MF and proteins-BP. For both cases, we compare two versions namely the annotation mappings of Ensembl release 25 (Oct. 2004, first in this study) with those

Annotation Mappings	Corresp.		Protein obj.		Concepts	
	growth		growth		growth	
	add-frac	del-frac	add-frac	del-frac	add-frac	del-frac
Protein-MF	2.82		1.99		1.39	
	83%	51%	68%	37%	32%	6%
Protein-BP	2.47		1.90		2.25	
	81%	52%	68%	39%	58%	5%

a) Growth rates of annotation mappings

Annotation Mappings	Protein obj.		Concepts	
	cov_{25}	cov_{47}	cov_{25}	cov_{47}
	$growth_{cov}$		$growth_{cov}$	
Protein-MF	47%	67%	28%	35%
	1.43		1.22	
Protein-BP	43%	59%	20%	26%
	1.36		1.39	

b) Coverage statistics

Table 5: Evolution of annotation mappings between Ensembl releases 25 and 47

of release 47 (Oct. 2007, last in this study). Table 5 shows the corresponding evolution measures, introduced in Section 2, in particular growth rates for the number of correspondences, proteins and ontology concepts as well as the add and delete fractions (Table 5a). Table 5b shows coverage measures indicating which shares of the protein source and ontologies participate in the annotation mappings and how these shares changed (growth rates) between the two Ensembl versions.

We observe both annotation mappings show a rather similar evolution behaviour. For both mappings, the growth rates for the total number of correspondences (annotation associations) of 2.82 and 2.47 are very high. These rates are not only higher than the growth for the total number of proteins or ontology concepts (factors between 1.2 and 1.74, see above) but also higher than for the number of annotated proteins (growth factor 1.9 – 1.99) and used ontology concepts (1.39 – 2.25). Similarly, the add and delete activity is much higher for the correspondences than for the individual sources. So the latest annotation mappings of Ensembl release 47 contain 81-83% added (i.e., new) correspondences compared to the initial mapping versions of release 25. Further, more than 50% of the original correspondences have been deleted. These observations reveal that the use of ontologies in annotations grows faster than the ontologies and the number of instances but that there is also a high degree of instability due to many deletions of associations.

This is also confirmed by the coverage ratios shown in Table 5b. The much increased number of correspondences led to an increased annotation coverage for proteins. The coverage values increased during the observation period from 43-47% to 59-67%, i.e., most proteins are now annotated with concepts of the Gene Ontology. Similarly, the coverage of the two ontologies within the annotation mappings improved. Currently, 35% (26%) of the MF (BP) concepts have associated proteins.

4.4 Evolution on ontology mapping level

On the ontology mapping level, we study the evolution of mappings between different versions of the MF and BP ontologies. Such semantic mappings are to specify which molecular functions are involved in which biological processes. The manual creation of such mappings is very time-consuming especially since the ontologies change so frequently. Hence we aim at a (semi-) automatic generation of mappings by using different match algorithms to generate likely correspondences between two ontology versions. For our study we consider four match algorithms of [7]. The first two match approaches are instance-based and assume that two concepts are related if they share a certain number of instances, i.e., associated protein objects in our scenario. The approach termed *Base(5)* assumes that two concepts match if there are at least 5 proteins

Ontology Mappings	Corresp.		Mol. Functions		Biol. Processes	
	|C1|-|C2|, grow		grow		grow	
	add-frac	del-frac	add-frac	del-frac	add-frac	del-frac
Base(5)	2780-8973, 3.2		1.8		2.3	
	78%	29%	52%	16%	62%	12%
Min (1.0)	4795-11564, 2.4		1.4		2.1	
	80%	52%	41%	15%	62%	21%
Name (0.5)	5434-15016, 2.8		2.1		1.4	
	77%	36%	57%	10%	44%	20%
Name (0.7)	389-592, 1.5		1.3		1.3	
	45%	17%	32%	12%	34%	15%

a) Growth rates of ontology mappings

Ontology Mappings	Mol. Functions		Biol. Processes	
	cov_{25}	cov_{47}	cov_{25}	cov_{47}
	grow_{cov}		grow_{cov}	
Base(5)	7%	12%	6%	8%
	1.7		1.3	
Min (1.0)	23%	30%	17%	20%
	1.3		1.2	
Name (0.5)	25%	47%	18%	15%
	1.9		0.8	
Name (0.7)	5%	6%	4%	3%
	1.2		0.7	

b) Coverage statistics

Table 6: Evolution of generated ontology mappings between molecular functions and biological processes of the GeneOntology source

which associate to both concepts. The *Min(1.0)* approach uses the so-called min similarity and threshold 1.0, i.e., two concepts match if all instances associated to the concept with the smaller number of associations are also associated to the other concept. The two other match approaches are metadata-based and utilize the similarity of concept names. We assume a correspondence between concepts when the string (trigram) similarity of their names exceeds a certain threshold, e.g., 0.5 or 0.7; these mappings are named with *Name(0.5)* and *Name(0.7).* With these approaches we generated MF-BP mappings for the ontology versions of Feb. 2004 (associated with Ensembl release 25) and June 2007 (associated with Ensembl release 47).

Table 6a shows the growth rates for the ontology mappings between the two releases as well as the relative fractions for add and delete. In the column "Corresp." we also indicate the absolute number of correspondences in the two versions of the mappings (e.g., for Base(5) the number of correspondences increased from 2780 in the old version to 8973 in the new version of the ontology mapping, growth factor 3.2). Table 6b shows the coverage rates for both ontologies and both mapping versions indicating to what degree the ontologies participate in the mappings. For example, for Base(5) the coverage of MF increased from 7% to 12% between the two versions.

We observe that there are significant differences between the mappings generated by the different match algorithms and their evolution behaviour. For the name matchers, the number of correspondences is heavily dependent on the chosen threshold. A low threshold (0.5) matches many concepts (many correspondences) and leads a relatively high coverage in the ontologies, however with the risk of many false correspondences. A higher threshold (0.7), on the other hand, is very restrictive and matches only few concepts. On the other hand, this restrictive approach leads to the highest evolution stability with the lowest fraction for deleted correspondences (17%). Interestingly, for the name matchers the coverage of the BF ontology decreased, presumably because the BF ontology growed much faster than the MF ontology so that for many new BF names there no MF counterpart is found.

The two instance-based matchers obtain a relatively high number of correspondences (compared to the name matchers) as well as a large increase between the two versions (growth factor 2.4 – 3.2), i.e., the mappings grow faster than the ontologies. The Base(5) matcher is more stable than the Min(1.0) matcher since the delete fraction is merely 29% vs. 52%. On the other hand the Min matcher achieves a much better coverage.

5 Related Work

The evolution and change management of ontologies has so far primarily been addressed in the context of the Semantic Web [18], especially for specific ontology representations such as OWL or Frame Logic. Klein [8,9] investigated the versioning of ontologies, [10] defined change operations to describe the evolution between ontology versions. In [13,14,15], the process of ontology evolution has been formalized and strategies to unambiguously handle critical changes during evolution are proposed. Tools supporting change management of different ontology models are described in [4,11,13].

This line of previous work is complementary to ours and does neither consider life science ontologies nor a quantitative analysis of the evolution behavior. Furthermore, the evolution of ontology-related mappings has not been analyzed before. One recent paper analyzed the evolution of the Gene Ontology [17] using a simple evolution model. We also used some of their measures (e.g., number of paths or path lengths of concept nodes) but propose a more powerful generic evolution model that is applicable to the evolution of ontologies, instance sources, and mappings. Furthermore, we comparatively analyzed not only the Gene Ontology but 16 biomedical ontologies as well as the evolution of annotation and ontology mappings.

6 Conclusions

We proposed a general framework for analyzing the evolution of ontologies and ontology-related mappings. Using the framework we analyzed the recent evolution of 16 life-science ontologies since 2004. We observed that most ontologies are heavily updated and grow significantly. Most changes are additions of new concepts but there is also a surprisingly high number of concepts that are deleted in newer versions or marked as obsolete. The notion of obsolete concepts is supported by most but not all ontologies. This notion is helpful for the stability of ontologies and eases applications the migration to newer ontology versions (without risking invalid references to deleted concepts). The analyzed ontologies are dominated by is-a relationships (>85% of all relationships), although the shares of part-of and domain-specific relationships have slightly increased in recent years. Furthermore, the inner structure of ontologies (share of inner concepts, number of paths, path lengths) increased in the recent past indicating a growth of structured knowledge in life science ontologies.

We further utilized the framework to study the evolution of protein instances, annotation mappings and ontology mappings. Using Ensembl, we observed a large increase in the number of protein annotations to the Gene Ontology (GO). However, the relatively high number of deletes of protein instances caused a rather high instability for the annotation mappings. For the evolution of ontology mappings, we considered several instance- and metadata-based match algorithms to automatically generate correspondences between concepts of two GO subontologies. We observed that the ontology mappings evolved to a larger degree than the ontologies especially for the instance-based methods. Metadata-based methods (e.g., based on concept names) can easily introduce wrong correspondences but may provide improved stability for evolution. This is because they are not dependent on instances and their annotations and thus do not suffer from the higher fluctuation (delete activity) for instances compared to ontologies.

We see several opportunities for future work. First, our analysis framework can be extended by additional types of change (e.g., modification of attribute values) and applied to further ontologies. Second, 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. Third, tools can be developed to help ontology designers to explore the effects of certain ontology changes on existing annotation and ontology mappings, especially for delete operations.

Acknowledgements. This work is supported by BMBF grant 01AK803E "MediGRID - Networked Computing Resources For Biomedical Research".

References

[1] O. Bodenreider, M. Aubry and A. Bugrun: Non-lexical approaches to identifying associative relations in the Gene Ontology. Proc. Pacific Symposium on Biocomputing, 2005.

[2] O. Bodenreider and A. Bugrun: Linking the Gene Ontology to other biological ontologies. Proc. ISMB meeting on Bio-Ontologies, 2005.

[3] The Gene Ontology Consortium: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 32: D258-D261, 2004.

[4] P. Haase, F. van Harmelen, Z. Huang et al.: A framework for handling inconsistency in changing ontologies. Proc. of 4th Intl. Semantic Web Conference, 2005.

[5] M. Hartung, T. Kirsten and E. Rahm: Analyzing the Evolution of Life Science Ontologies and Mappings - Extended Version. Leipzig Bioinformatics Working Paper No. 17, 2008.

[6] T. Hubbard, B. Aken, K. Beal et al.: Ensembl 2007. Nucleic Acids Research 35: D610-D617, 2006.

[7] T. Kirsten, A. Thor and E. Rahm: Instance-based matching of large life science ontologies. Proc. of DILS 2007.

[8] M. Klein and D. Fensel: Ontology versioning on the Semantic Web. Proc. Int. Semantic Web Working Symposium (SWWS), 2001.

[9] M. Klein: Change Management for Distributed Ontologies. PhD thesis, Vrije Universiteit Amsterdam, 2004.

[10] N. Noy and M. Klein: Ontology evolution: Not the same as schema evolution. Knowledge and Information Systems, 6(4):428-440, 2004.

[11] N. Noy, A. Chugh, W. Liu et al.: A Framework for Ontology Evolution in Collaborative Environments. Proc. of the 5th Intl. Semantic Web Conference, 2006.

[12] N. Sioutos, S. de Coronado, M.W. Haber et al.: NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics Vol. 40. 30 –43, 2007.

[13] L. Stojanovic. Methods and Tools for Ontology Evolution. PhD thesis, University of Karlsruhe, 2004.

[14] L. Stojanovic, A. Maedche, B. Motik et al.: User-driven ontology evolution management. Proc. of 13th Intl. Conf. On Knowledge Engineering and Knowledge management, 2002.

[15] L. Stojanovic and B. Motik: Ontology evolution within ontology editors. Proc. of the OntoWeb-SIG3 Workshop, 2002.

[16] B. Smith, M. Ashburner, C. Rosse et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology 25, 1251 - 1255.

[17] Z. Yang, D. Zhang and C. Ye: Ontology Analysis on Complexity and Evolution Based on Conceptual Model. Proc. of DILS 2006.

[18] B. Yildiz: Ontology Evolution and Versioning. Technical Report, TU Vienna, 2006.